Germline Variation Affects Tumor Progression and Informs Clinical Therapy Decisions Across Cancers

Ajay Chatrath
Bourbonnais, IL

B.S., Saint Louis University, 2016
B.A., Saint Louis University, 2016
M.S., Johns Hopkins University, 2018

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Biochemistry and Molecular Genetics

University of Virginia
March, 2020

Table of Contents

Abstract

While germline variation has had a rich history of being studied in the context of cancer risk, emerging evidence now suggests that germline variation shapes the landscape of somatic aberrations in cancer and may affect the sensitivity and toxicity of chemotherapy drugs. Given these findings, we hypothesized that germline variation should not only predict the risk of acquiring cancers but also affect the rate at which the tumor progresses. We began our search for germline variants affecting tumor progression by analyzing the genomic sequencing data of approximately 500 patients diagnosed with lower grade gliomas. We identified two germline variants associated with poor outcome in these patients, one in the oncogene *GRB2* and the other in the tumor suppressor gene of *ANKDD1a*. Our results suggested that germline variation is associated with patient outcome and that there is an interaction between common polymorphisms and the somatic landscape in lower grade gliomas.

We then searched for germline variants associated with patient outcome across 33 different types of cancers using sequencing data from over 10,000 cancer patients. In total, we identified 79 prognostic germline variants in individual cancers and 112 prognostic germline variants in groups of cancers. The germline variants identified in individual cancers provide additional predictive power about patient outcomes beyond clinical information currently in use and may therefore augment clinical decisions based on expected tumor aggressiveness. Our results suggested that the idea that germline variation contributes to tumor progression is a general principle of cancer genomics as we

found this to be true across essentially all cancers for which we were sufficiently powered.

Having found that germline variants impact tumor progression, we suspected that the interaction between germline variants and the landscape of somatic events could be exploited therapeutically. To assess this possibility, we developed a pan-cancer approach to identify pathogenic germline variants associated with elevated tumor mutational burden, as high tumor mutational burden is a validated biomarker of immune checkpoint inhibitor efficacy. We identified an association with overall tumor mutational burden in nine genes using a pan-cancer approach, fourteen pathways in individual cancers, and twelve pathways using a pan-cancer approach. Patients with the pathogenic germline variants described in this study may be more likely to respond to treatment with immune checkpoint inhibitors.

Together, our work suggests that germline variation affects tumor progression and is involved with shaping the landscape of somatic events in cancers in a predictable way that can likely be targeted therapeutically. These findings pave the way for future efforts to better individualize patient care.

Advisor and Committee Member Signatures

We hereby approve the thesis of

Ajay Chatrath

For the Degree of Doctor of Philosophy

_____

Anindya Dutta, MD, PhD
Doctoral Advisor
Harrison Family Professor and Dept Chair, Biochemistry and Molecular Genetics
University of Virginia School of Medicine

_____

Aakrosh Ratan, PhD
Thesis Committee Member
Assistant Professor, Public Health Sciences
University of Virginia School of Medicine

_____

Charles Farber, PhD
Thesis Committee Member
Associate Professor, Public Health Sciences
University of Virginia School of Medicine

_____

Todd Stukenberg, PhD
Thesis Committee Member
Professor, Biochemistry and Molecular Genetics
University of Virginia School of Medicine

_____

Alan Bergland, PhD
Thesis Committee Member
Assistant Professor, Department of Biology
University of Virginia School of Medicine

Acknowledgements

Having the opportunity to pursue an MD/PhD has been an amazing blessing and adventure. I am forever indebted to all of the people in my life who have given me the strength and support to pursue my passions.

I thank my PhD mentor, Dr. Anindya Dutta, without whom none of this would have been possible. Working in your lab inspired me to pursue an MD/PhD and has dramatically altered the course of my career. Thank you for always pushing me to do my best work, giving me the freedom to explore my ideas, and believing in my abilities and potential. I will remember the small adages that you have shared with me from time to time, and I will keep them with me as I continue on in my journey.

I thank Dr. Aakrosh Ratan, my computational mentor, for teaching me much of what I know on the practical side of cancer genomics. I feel comfortable with most omic datasets because of all that you taught me. Thank you for the many suggestions for analyses ideas, conversations about genomics, and advice.

I thank my thesis committee members, Dr. Charles Farber, Dr. Alan Bergland, Dr. Aakrosh Ratan, and Dr. Todd Stukenberg. Thank you for your many suggestions for analyses related to my project and career.

I thank Dr. Pankaj Kumar and Dr. Manjari Kiran who were computational researchers in the Dutta lab when I first started in the lab. Thank you for everything you taught me about bioinformatics and performing analyses using publicly available data.

I thank Dr. Tianxi Li, Dr. Xiwei Tang, and Dr. Daniel Keenan for all of the statistical advice related to my projects.

I thank the Dutta lab members, past and present, for their support for this project and friendship. Thank you for your many suggestions during our lab meetings and your help throughout the writing of these papers.

I thank the Medical Scientist Training Program at UVA for all of the support. I thank Dr. Dean Kedes and the MSTP Executive Committee for the opportunity to pursue an MD/PhD at UVA and for their mentorship. I thank Ms. Ashley Woodard for all that you do to keep MSTP functioning. Finally, I thank all of the MSTP students for all of the recommendations and support for my project.

I thank Nancy Rush, Debbie Sites, Helen Norfleet-Shiflett, Katherine Yates, and Jill Clarke for everything that you all do to keep the department, BIMS, and SOM running.

I thank the Research Computing Team at the University of Virginia for all of the help over the years related to performing large-scale computation.

I thank all of my friends from childhood, high school, college, medical school, and graduate school for all of your support as I have pursued my MD/PhD. I thank everyone who has taken the time to mentor me along my journey.

I thank my family members, especially my parents, Dr. Sanjay and Poonam Chatrath, and my siblings, Sheena and Rakesh Chatrath. Thank you for all of your support, molding me into the person I am today, and instilling into me the values that I carry with me today into everything that I do.

I thank Krupa Patel for all of the support throughout the entirety of my adult life, academic and otherwise. I attribute my success to you being such a positive motivator in my life and I am extremely blessed to have you in my life. Thank you for everything.

Finally, I thank all of the patients and their families who donated samples for research to the large-scale sequencing projects described in this work. None of the discoveries described here would have been possible without your donation. The work described here is wholly inspired by the goal to improve outcomes for other patients like you.

Chapter 1: Introduction

Germline Variants Increase the Risk for Cancer

Heritability of Cancer

While cancer has historically been thought of as having both an inherited and environmental basis, exposures and sporadic genetic events occurring as a result of random chance had been believed to be responsible for the tumors observed in most patients. The notion that cancer has a hereditary component had been reinforced through the characterization of a handful of familial cancer syndromes, as individuals with multiple relatives with cancer were found to have a much higher risk of cancer than the general population. However, these familial cancer syndromes were rare and were thought to be driven by the perturbation of genes with dominant effects, leading to the belief that cancer predominantly occurs due to environmental exposures [1-3]. Epidemiologic studies of patients with breast, prostate, ovarian, and uterine cancer suggested that the inherited component of cancer may be much greater than the small number of patients with familial cancer syndromes that had been characterized, though the "inheritance" of cancer risk in these studies of patients with relatives with cancer was confounded by individuals often sharing similar environmental exposures with their relatives [4-8]. The landmark study by Lichtenstein et al. pooled data from over 40,000 pairs of twins to assess the risk of cancer at 28 different anatomical sites and could estimate the genetic and nongenetic components of cancer risk through the comparison of concordant tumor development in monozygotic compared to dizygotic twins. The study estimated that the genetic

component of cancer risk for prostate, colorectal, and breast cancer was 42%, 35%, and 27%, respectively. For most cancers, the familial cancer syndromes had accounted for 1% of cancer cases. This study suggested that there was a major gap in the understanding of the hereditary component of cancer and that focusing solely on the study of DNA repair genes typically perturbed in familial cancer syndromes would not explain the bulk of the genetic contributors of cancer risk [9].

Genetic Understanding of Cancer Risk

Growing clinical evidence suggesting that there existed an inherited component of cancer predisposition fueled interest in studying this risk at the genetic and molecular level [10]. The work by Theodor Boveri first suggested that cancer occurred through somatic events at the genetic level and that inherited perturbations to genetic units, which had not yet been fully characterized, could also be responsible for cancer. Alfred Knudson's "two-hit hypothesis" published in 1971 was in line with this prediction, as Knudson's hypothesis suggested that each allele of a tumor suppressor gene needed to be impaired to allow for tumorigenesis. Further epidemiologic work building off this hypothesis found that two hits, one to each allele of the tumor suppressor gene *RB1*, had typically occurred in patients with retinoblastomas with pathogenic germline variants often being responsible for the first hit [10].

Over 100 cancer predisposition genes perturbed by pathogenic germline variants have been identified through clinical and genomic studies and characterized using experimental approaches. Pathogenic germline variants in

cancer predisposition genes result in loss of function of tumor suppressor genes, such as *ATM*, *BRCA1*, *BRCA2*, *BRIP1*, *PALB2*, *TP53*, *APC*, *NF1*, *PMS2*, and *RB1*. Many of these pathogenic germline variants perturbing tumor suppressor genes disrupt pathways closely tied to tumorigenesis, such as DNA repair, cell proliferation, and cell adhesion [10-12]. Pathogenic germline variants in cancer predisposition genes result in gain of function of proto-oncogenes, such as *MET*, *RET*, *EGFR*, *Ras*, and *Myc* and perturb pathways that predispose to cancer development, such as cell cycle, cell death, and cell growth [10, 11].

Genomic Study of Cancer Predisposition Genes

Large scale sequencing projects have enabled the identification of a substantial number of pathogenic germline variants in cancer predisposition genes [13, 14]. A large study of pathogenic germline variants in around 10,000 cancer patients from The Cancer Genome Atlas project across 33 cancers found that 8% of cancer patients harbor pathogenic germline variants [11]. Other smaller studies have also identified an array of pathogenic germline variants across a wide spectrum of cancers [15-20]. Previous studies have identified common variants associated with differences in risk for cancer, though their effect sizes tend to be smaller in magnitude [21-25].

Despite the explosion in the identification of germline variants associated with differences in cancer risk, the utility of germline variants in clinical oncology has not progressed as rapidly. Although clinical guidelines recommend that patients with pathogenic germline variants that fit under genetic syndromes such as hereditary breast and ovarian cancer syndrome, Lynch syndrome, Li-

Fraumeni syndrome, and Peutz-Jeghers syndrome be considered for earlier screening, these guidelines have not been extended to the full spectrum of germline variants that have been associated with increased risk of cancer [26].

Incorporating germline variants into clinical practice has been challenging for several reasons [11, 19, 21, 26-28]:

(1) The usage of germline variants as part of robust cancer screening regimens requires validation in large cohorts. Germline variants with large effect sizes tend to be rare, with a few exceptions such as the variants in the *APOE4* gene associated with the risk for Alzheimer's disease [29]. Given the rarity of these pathogenic germline variants, they are challenging to validate and require large cohorts of patients to study effectively. On the other hand, common variants are found much more frequently in the population, but their effect sizes tend to be smaller, with a few exceptions [30]. As a result, common variants also typically require fairly large cohorts to validate.

(2) Germline variation should be considered in the context of other clinical factors to maximize clinical utility. As previously described, while there does exist a genetic component to cancer risk, the bulk of cancer risk for many cancers is believed to be due to environmental exposures and is associated with aging. The translation of the discovery of germline variants associated with differences in cancer risk is perhaps best done using datasets with detailed and standardized demographic and

environmental exposure data. These datasets require more foresight and time to generate.

(3) The effect sizes of the germline variants must be large enough to alter clinical decisions. Even if a germline variant is predictive of cancer risk, the magnitude of the effect size must be large enough to warrant the increased cost and inconvenience to the patient to start screening for cancer earlier. Furthermore, clinical trials are necessary to show that the detection of that particular cancer earlier in the disease course can actually be acted upon by oncologists to lengthen overall survival, after adjusting for the lead-time bias.

Further work is necessary to determine when the use of germline variation would be valuable for modifying cancer screening regimens to catch and treat cancer earlier in the disease course.

Germline Variation Affects the Landscape of Somatic Aberrations in Cancer

While most studies of germline variation in cancer have focused on cancer risk, recent studies suggest that germline variation affects the landscape of somatic aberrations in cancer. Alfred Knudson's "two-hit hypothesis" published in 1971 predicted this, as his hypothesis suggested that a germline variant that affected the function of a tumor suppressor gene could be followed by a somatic mutation that affected the other allele of that tumor suppressor gene to cause tumorigenesis [10]. While Alfred Knudson's ideas were initially focused on mutations in *RB1* in children with retinoblastoma, the idea that germline variants in tumor suppressor genes predisposed patients to developing tumors with

somatic mutations in the other allele of the same tumor suppressor gene was extended to other tumor suppressors as well. A recent study of 429 patients with ovarian carcinoma found that the majority of patients with germline truncating mutations in the tumor suppressor genes *BRCA1* and *BRCA2* exhibited loss of heterozygosity [31]. Similarly, genomic studies of myeloproliferative neoplasms identified a germline *JAK2* haplotype associated with increased risk for the development of $JAK2^{V617F}$ somatic mutations, which is one of the most common and well-characterized drivers of myeloproliferative neoplasms [32-35]. In line with these findings, a previous study identified functional germline variants in the *EGFR* tyrosine kinase associated with an increased risk for subsequent somatic mutations in *EGFR* [36].

A large study by Carter et al. analyzed the interaction between inherited polymorphisms and somatic aberrations in almost 6,000 tumors across 22 different cancer types. Carter et al. identified and validated 412 genetic interactions between germline variants and somatic aberrations. While the previous studies discussed here identified somatic associations occurring in the other allele of the same gene perturbed by a germline variant, the study by Carter et al. identified somatic aberrations in genes that were not always the same as the one perturbed by the germline variant. In some cases, the germline variants were associated with increased risk for somatic mutations in genes of the same pathway but not always in the same gene. These findings suggested that the interaction between germline variation and somatic events is much more complex than the field had believed, as germline variants could increase the

susceptibility for somatic aberrations in genes other than the ones that they are found in. Furthermore, this finding suggested that more complex computational methods are necessary to attain an integrated understanding of tumorigenesis and the large number of factors that likely influence which somatic events will occur [37].

A study of the interaction between germline variation and somatic aberrations in cancer using whole genome sequencing data from 2,658 patients across 38 tumor types by the Pan-Cancer Analysis of Whole Genomes Consortium found that germline variation is predictive of somatic mutational processes across cancers. For example, their analysis identified germline variants at the 22q13.1 locus associated with decreased *APOBEC* mutagenesis in cancer. They found rare variants in *BRCA1* and *BRCA2* to be associated with a higher abundance of small somatic structural variant deletions and tandem deletions, consistent with a role of these proteins in error-free homologous recombination directed repair of double-strand breaks. Germline *MBD4* variants were associated with an elevated rate of C>T somatic mutations at CpG dinucleotides. This result is consistent with the role of MBD4 in binding to methylated CpGs and correcting G:T or G:U mismatches in the vicinity. Finally, they identified 114 germline source L1 elements that were capable of active somatic retrotransposition. Overall, their results suggest that germline variation can shape somatic processes at a genome wide scale [38].

Numerous additional studies have explored the link between germline variation and somatic aberrations [39, 40]. Research from our group has also

suggested a link between germline variation and somatic aberrations and our findings are detailed in **Chapter 2**, **Chapter 3**, and **Chapter 4** [41, 42]. Briefly, in **Chapter 2** we find that patients with lower grade gliomas with a germline variant in the *GRB2* oncogene in the *Ras* signaling pathway are at increased risk for somatic mutations in *Capicua transcription repressor* (*CIC*), a driver gene that also regulates the *Ras* signaling pathway. In **Chapter 3**, we describe our discovery of how prognostic germline variants are associated with an increased risk for somatic mutations in driver genes. Finally, in **Chapter 4**, we describe how the tumors found in patients with pathogenic germline variants exhibit predictable perturbations to the transcriptome and somatic mutation profile, further supporting the notion that germline variation shapes the somatic aberration landscape at a genome wide scale.

Understanding the relationship between germline variation and somatic aberrations is particularly promising from a clinical perspective for several reasons [39]:

(1) The existence of a relationship between germline variation and somatic aberration suggests that the aggressiveness of a tumor can be predicted based on the germline status. This could enable clinicians to determine whether or not a tumor will be indolent or aggressive even at the earliest stages of tumorigenesis and could alter the course of treatment.

(2) Germline variation could be used to improve the selection of clinical therapy. The recent large-scale sequencing of tumors and cancer cell lines has helped to identify the genomic determinants of drug sensitivity

[43, 44]. The existence of an interaction between germline variation and somatic aberrations suggests that chemotherapy responsiveness can be predicted using the status of germline variants. In addition, germline variants in the mismatch repair genes predict microsatellite instability, which leads to a greater chance of producing neo-antigens, thus predicting responsiveness to immune checkpoint blockade therapy [45-47].

(3) Understanding the relationship between germline variation and somatic aberration may reinforce the discovery of variants that increase susceptibility to cancer, as it would provide the field with an understanding of the genetic sequence of events by which germline variants contributes to tumorigenesis. This validation could improve the accuracy of polygenic risk scores used to predict an individual patient's risk for cancer.

(4) Understanding the interaction between germline variation and somatic aberrations would inform the creation of complex genomic network-based models integrating germline variation and somatic aberration for predicting cancer risk and progression.

While understanding the interaction between germline variation and somatic aberrations has significant promise for clinical applicability, the investigation of these interactions is riddled with several challenges [31, 37, 39, 40]:

(1) There have been reports of associations between germline variants and somatic aberrations, but associations may be the result of several complex indirect interactions. Gaining a thorough understanding of these

interactions will likely require detailed multi-omic datasets, complex network-based computational approaches, and experimental perturbation.

(2) These interactions may have some context dependence. Some interactions may only be evident in certain contexts, such as in the context of certain environmental exposures or in the presence of germline variants found more commonly in patients of a particularly race.

Germline Variation Affects Tumor Progression

The idea that there is a link between germline variation and somatic events in cancer implies that germline variation may also affect tumor progression and could be used to predict the prognosis of patients with cancer. Several studies in this area identified germline variants predictive of patient outcome in genes with well-characterized driver roles in those cancers, such as *SUFU*, a negative regulator of Hedgehog signaling, in medulloblastoma or *BRCA*1 and *BRCA2* in breast cancer [33, 48, 49].

In **Chapter 2** and **Chapter 3**, we describe studies from our group supporting the idea that germline variation affects tumor progression. In **Chapter 2**, we screen approximately 200,000 germline variants for associations with overall survival in patients with lower grade gliomas and identify two germline variants associated with poor outcome. One germline variant was identified in the *GRB2* oncogene and the other was identified in a tumor suppressor gene, *ANKDD1a* [41]. In **Chapter 3**, we extend our approach to all 33 cancers encompassed in The Cancer Genome Atlas and characterize the landscape of prognostic germline variants using genomic sequencing data from approximately

10,000 patients. Our results suggest that germline variation is associated with patient outcome across cancers and that germline variation seems to affect tumor progression. We found that nearly half of the prognostic germline variants are found in genes with previously reported roles as oncogenes or tumor suppressor genes, a finding which was consistent with our previous study detailed in **Chapter 2**. The other half of the genes with prognostic germline variants were of unknown function and require further study [42].

Understanding the mechanisms by which germline variants are associated with patient outcome and modulate tumor progression is challenging due to genetic linkage between variants, meaning the identified variant may not itself be responsible for the effect on outcome. In addition, the germline variants are present in every tissue in the body, and so could have an effect on outcome through effects on non-tumor cells in the body such as through immune system cells or through changes in the tumor microenvironment. Finally, most of the available datasets are limited to exonic regions, and therefore miss potentially important germline variants in introns and intergenic regions, though this is quickly changing. As a result, determining which nucleotide and which tissue is responsible for the observed phenotype has made understanding the exact molecular mechanisms by which germline variants act quite difficult. Nevertheless, research in this area has suggested possible roles by which the variants may be acting, such as through perturbation of protein structure and function and modulation of gene expression [42, 50, 51].

Germline Variation Affects Drug Responsiveness

While germline variation may be associated with patient outcome directly by modulating aspects of tumor biology, germline variation may also be predictive of patient outcome by modulating responsiveness to therapy [42]. Identifying germline variants that are predictive of differences in therapy responsiveness poses several challenges to the field for two primary reasons [52]:

(1) Patients with the same cancer do not always receive the same treatment. Patients may be treated with different combinations of chemotherapy drugs, radiotherapy, and surgical interventions, making association studies difficult to perform. Furthermore, patients may receive different dosages of chemotherapy drugs or can receive treatment at difference times in their disease course.

(2) Few cohorts have both rich clinical annotation and genomic data availability. Most cohorts typically have either one data type or the other.

Despite these challenges, several studies have begun to address this question. A pan-cancer analysis by Menden et al. of drug sensitivity data from cancer cell lines found that germline variation could be used to predict drug sensitivity across many of the 265 total drugs included in their analysis. In some cases, they found that the germline component of drug sensitivity exceeded the portion of drug sensitivity that could be predicted using somatic mutations. They replicated previous associations, such as germline loss of function mutations in *BRCA1* and *BRCA2* being associated with olaparib and cisplatin sensitivity,

DPYD loss of function germline mutations being associated with 5-flurouracil sensitivity, *WFS1* variants being associated with cisplatin toxicity, and *MGMT* variants being associated with temozolomide toxicity [53-57]. Many of these associations can be explained. For example, the loss of *BRCA1* or *BRCA2*, makes cells susceptible to PARP inhibitors (discussed below) and to cisplatin, which causes cross-linking DNA damage, the repair of which uses HR-dependent steps that are impaired upon mutations in these genes. Similarly, *DPYD* is involved in the catabolism of thymidine and uracil, and so loss of *DPYD* increases the levels of 5-FU in tumor cells, while *MGMT* is a DNA methyltransferase that repairs the alkylating DNA damage caused by temozolomide, so that a decrease in *MGMT* activity increases the toxic effect of temozolomide. Their results suggested that germline variation could affect drug sensitivity through perturbation of protein structure or by being associated with differences in gene expression [53].

Several associations between germline variation and drug sensitivity have been reported in the literature and approved for clinical use by the Food and Drug Associations [58]. A few of the most well-studied associations are highlighted below.

PARP Inhibitors and Germline Mutations in *BRCA1* and *BRCA2*

Olaparib (Lynparza), Rucaparib (Rubraca), Veliparib are chemotherapy drugs that act by inhibiting poly ADP ribose polymerase (PARP). PARPs catalyze poly ADP-ribosylation reactions, which involve the transfer of ADP-ribose groups to target proteins [59]. While PARPs are involved in a variety of biological

processes, they are of particular interest as targets in cancer because of their participation in base excision repair and nucleotide excision repair [60-62]. The failure of base excision repair, in particular, predisposes cells to double-strand breaks. In addition, because PARPs participate in DNA repair, they facilitate the repair of DNA damage caused by alkylating agents, chemotherapy drugs frequently used in the treatment of a variety of cancers. Furthermore, the importance of PARPs for DNA repair increases as the functions of other DNA repair damage pathways are lost [63].

Pathogenic germline variants in *BRCA1* or *BRCA2* impair homologous recombination (HR) directed repair, a DNA repair process used to repair double strand breaks. PARP inhibitors are particularly effective in patients with pathogenic germline variants in *BRCA1* or *BRCA2* because inhibition of PARPs substantially decreases the tumor cells' ability to repair DNA damage and increase double strand breaks that require HR directed repair. This ultimately results in cell death. Olaparib has been approved by the Food and Drug Administration of the United States to treat patients with pathogenic germline variants in *BRCA1* or *BRCA2* with advanced ovarian cancer who have failed three or more previous lines of chemotherapy. Olaparib has also been approved for the treatment of metastatic HER2-negative breast cancer in patients with pathogenic germline variants in *BRCA1* or *BRCA2* [64-68]. Rucaparib is a PARP inhibitor that has been approved for the treatment of adults with deleterious germline mutations in *BRCA1* or *BRCA2* with epithelial ovarian, fallopian tube, or

primary peritoneal cancer who have been previously treated with two or more chemotherapy drugs [64].

Germline Deletions in *BIM* Predispose Patients to Imatinib Resistance

Imatinib (Gleevec) is a tyrosine kinase inhibitor that is used to treat chronic myelogenous leukemia. Imatinib inhibits the constitutively active tyrosine kinase *BCR-ABL*, which is a fusion protein formed by a chromosomal translocation and the driver of chronic myelogenous leukemia. Although imatinib has been massively successful for converting a previously deadly disease into one that can be chronically managed, a small percentage of patients exhibit resistance to imatinib treatment [69, 70].

The study of these resistance patterns resulted in the identification of a 2,903 germline base pair deletion in the *BIM* gene that is associated with decreased responsiveness to imatinib in patients with chronic myelogenous leukemia. The deletion is present in roughly 15% of East Asians and Latin Americans but is not found in Europeans or Africans. BIM functions as an apoptotic activator. Functionally, BIM is regulated through alternative splicing. The three major proapoptotic isoforms of BIM are BIMEL, BIML, and BIMS. BIM has two other isoforms, BIMγ1 and BIMγ2, which are not proapoptotic. The proapoptotic isoforms all contains exon 4, whereas the two isoforms that are not proapoptotic do not contain exon 4. The deletion discovered in these studies occurs over exon 4. As a result, the proapoptotic isoforms are either absent or diminished. Although imatinib is still able to inhibit the activity of *BCR-ABL* in

patients with this deletion, the cells' subsequent apoptotic response is impaired, resulting in resistance to therapy [30, 69, 71-76].

Germline Mutations in Mismatch Repair Genes and 5-Fluorouracil Sensitivity

5-Fluorouracil is a chemotherapy drug commonly used to treat gastrointestinal (esophageal, stomach colon, and pancreatic) cancers, breast cancer, and cervical cancer. 5-Fluorouracil functions by inhibiting thymidylate synthase, which methylates deoxyuridine monophosphate to form thymidine monophosphate. 5-Flurouracil causes cell death because thymidine monosphosphate is essential for DNA synthesis. It is particularly effective for treating cancer because DNA synthesis is necessary for cell division [77, 78].

Lynch syndrome is characterized by pathogenic germline variants in mismatch repair genes such as *MSH2*, *MSH3*, *MSH6*, *MLH1*, *MLH2*, *MLH3*, *PMS1*, and *PMS2*. The mismatch repair pathway functions to repair DNA damage that has occurred as a result of mispairing between nucleotides, including those resulting from small insertions or deletions, most commonly after DNA replication. The small insertions or deletions that occur in areas with one to six nucleotide repeats (microsatellites) due to polymerase slippage are corrected by mismatch repair pathways. Patients with mutations in the mismatch repair pathway thus exhibit a high degree of microsatellite instability and widespread changes in the number of repeating units of microsatellites, [78, 79].

While the widespread microsatellite instability associated with Lynch syndrome predisposes patients to a variety of cancers, patients with Lynch syndrome also exhibit increased resistance to treatment with 5-Fluorouracil. In

wild type cells, treatment with 5-Flurouracil results in the incorporation of 5-Flurodeoxyuridine triphosphate during the generation of new strands of DNA during DNA replication and widespread base pair mismatches. These base pair mismatches are recognized by mismatch repair proteins, which attempt to repair them, but the abundance of the mismatches and resulting repair activities result in cell death. Cells of patients with Lynch syndrome that are mismatch repair deficient are unable to detect these widespread mismatches, resulting in cell survival and therefore resistance to 5-Flurouracil [78, 80-83]. Many clinical studies have confirmed the experimental pre-clinical studies and have shown that patients with evidence of microsatellite instability do not respond as well to treatment with 5-Fluorouracil [45, 79, 84-87].

Patients with Lynch Syndrome are More Likely to Respond to Immune Checkpoint Inhibitors

Immune checkpoint inhibitors are cancer immunotherapy drugs that target immune checkpoints. Biologically, immune checkpoints act to downregulate the immune system and prevent autoimmune diseases. However, cancer cells frequently take advantage of these immune checkpoints as a means of downregulating the immune response, enabling the survival and further proliferation of cancer cells. These cancer cells harbor neoantigens that would otherwise result in the cancer cells being targeted by the immune system. Cytotoxic T-lymphocyte-associated protein 4 (CTLA4) and programmed cell death protein 1 (PD-1) are examples of cell surface proteins that downregulate immune system activity. Several immune checkpoint inhibitors have been approved by the Food and Drug Administration in the United States. Ipilimumab

is an antibody against CTLA-4. Nivolumab, Pembrolizumab (Keytruda), and Spartalizumab are PD-1 inhibitors. Atezolizumab is a PD-L1 inhibitor [88-90].

Clinically, immune checkpoint inhibitors have been shown to be effective in only a subset of cancer patients. Although identifying biomarkers of immune checkpoint inhibitor efficacy is still an ongoing area of research, overall somatic tumor mutation burden has emerged as a biomarker that has been shown to predict immune checkpoint inhibitor response in multiple cancers and in multiple cancer cohorts. Overall somatic mutation burden is believed to correlate with immune checkpoint inhibitor response because tumors with higher overall somatic mutation burden tend to produce a larger number of proteins with mutations that could act as neoantigens that can recognized by the immune system. As a result, following the inhibition of CTLA-4 or PD-1 by immune checkpoint inhibitors, the immune system is better able to target and kill cells harboring neoantigens. Clonal non-synonymous tumor mutation burden has correlated better with immune checkpoint inhibitor response than overall non-synonymous mutation burden or overall somatic mutation burden. Non-synonymous mutations can cause changes in protein structure whereas synonymous mutations do not cause changes in protein structure. Tumors with one (or few) large clone(s) carrying a somatic mutation are believed to be better targeted by the immune system because a single (or few) antibody or T cell is able to target a large number of tumor cells [46, 47, 88, 89, 91, 92].

Patients with Lynch syndrome have been shown to be more likely to respond to treatment with immune checkpoint inhibitors than patients without

Lynch syndrome. As explained above, these patients have pathogenic germline variants in genes necessary for mismatch repair, resulting in widespread microsatellite instability. This genomic instability results in the production of more neoantigens, meaning these tumors are more likely to be targeted by the immune system following inhibition of CTLA-4 or PD-1 [93].

In **Chapter 4**, I describe our approach to identifying pathogenic germline variants that may be associated with immune checkpoint inhibitor responsiveness. Because a small number of tumors from patients treated with immune checkpoint inhibitors have been sequenced, it is challenging to identify germline variants directly associated with responsiveness to immune checkpoint inhibitors. Instead, we use overall somatic mutation burden, non-synonymous mutation burden, and clonal non-synonymous mutation burden as proxies for immune checkpoint inhibitor efficacy. This enabled us to analyze the sequencing data from the approximately 10,000 patients from The Cancer Genome Atlas to identify germline variants associated with increased somatic mutation burden. We hypothesize that because these germline variants are associated with an increase in somatic mutation burden, they are also likely to associate with an increase in immune checkpoint inhibitor efficacy.

Germline Variation Affects Drug Toxicity

Germline variation is often the focus of studies in pharmacogenomics centered on drug toxicity. While chemotherapy drugs are often studied in the context of somatic aberrations when studying the efficacy of drugs against tumor cells, studies of drug toxicity are concerned with the effects of the drug on the

other non-mutated cells in the rest of the body [58, 94, 95]. An example of an association found between germline variation and toxicity for chemotherapeutic drugs is discussed below. This example suggests that studying germline variation in molecular and clinical oncology could enable individualization of cancer therapy selection to minimize the risk of adverse drug reactions based on a patient's genotype.

Germline Variants in *CYP2B6* Affect Cyclophosphamide Toxicity

Cyclophosphamide is a chemotherapy drug used in clinical oncology to treat ovarian, breast, small cell lung cancer, hematologic cancers, and several other solid tumors. Cyclophosphamide is a prodrug and requires activation to exert an effect on cells. Cyclophosphamide is activated by one of several cytochrome P450 enzymes in the liver to form 4-hydroxycyclophosphamide. 4-hydroxycyclophosphamide undergoes several additional reactions to ultimately form a phosphoramide mustard, which is an active cytotoxic agent. Phosphoramide mustard is able to form DNA crosslinks between and within DNA strands. These crosslinks impair DNA replication and transcription and ultimately result in cell death [96].

CYP2B6 is one of the primary cytochrome P450 enzymes that activates cyclophosphamide, resulting in the production of 4-hydroxycyclophosphamide. As a result, CYP2B6 activity and expression level impacts the toxicity of cyclophosphamide. Several pharmacokinetic studies have found that germline variants in *CYP2B6*, such as rs2279343, rs3211371, and rs3745274, alter the function and expression of CYP2B6. These polymorphisms either decrease the

function of CYP2B6 or are associated with decrease in the expression of *CYP2B6*, leading to decreased production of active 4-hydroxycyclophosphamide. Clinically, patients with these polymorphisms are less likely to exhibit complications of cyclophosphamide treatment, such as grade 4 neutropenia [96-98].

Additional examples, such as the effect of germline variants in *DPYD* on 5-FU toxicity and in *MGMT* on temozolomide toxicity have been discussed earlier.

Overview of this Dissertation

In this chapter, I have described the studies of germline variation in the context of oncology, from the initial studies of germline variation and cancer risk to the studies of germline variation in the context of tumor progression and pharmacogenomics. Together, the evidence from the field suggests that germline variation should be studied in the context of cancer risk and tumor progression. Furthermore, unbiased analyses are necessary to identify new means by which germline variation could perturb known oncogenes and tumor suppressor genes and also identify other genes perturbed by germline variation. In **Chapter 2**, I describe our study of germline variants associated with overall survival in lower grade glioma patients. In **Chapter 3**, I extend our study of germline variation to all 33 cancers included within The Cancer Genome Atlas to argue that germline variation contributes to tumor progression across cancers. Finally, in **Chapter 4**, I describe our study of pathogenic germline variants associated with differences in overall somatic mutation burden which suggest that germline variation can be

used to predict immune checkpoint inhibitor efficacy in cancer patients. In **Chapter 5**, I discuss unfinished studies and future directions in which the results in Chapters 2-4 should be advanced and the general implications of our results. I list other papers from the Dutta lab that I am an author on and indicated my contribution to each of those papers in the **Appendix**.

References

1. Li FP: **The 4th American Cancer Society Award for Research Excellence in Cancer Epidemiology and Prevention. Phenotypes, Genotypes, and Interventions for Hereditary Cancers.** *Cancer Epidemiol Biomarkers Prev* 1995, **4:**579-582.

2. Fearon ER: **Human cancer syndromes: clues to the origin and nature of cancer.** *Science* 1997, **278:**1043-1050.

3. Easton DF: **The inherited component of cancer.** *Br Med Bull* 1994, **50:**527-535.

4. Lesko SM, Rosenberg L, Shapiro S: **Family history and prostate cancer risk.** *Am J Epidemiol* 1996, **144:**1041-1047.

5. Gruber SB, Thompson WD: **A population-based study of endometrial cancer and familial risk in younger women. Cancer and Steroid Hormone Study Group.** *Cancer Epidemiol Biomarkers Prev* 1996, **5:**411-417.

6. Easton DF, Matthews FE, Ford D, Swerdlow AJ, Peto J: **Cancer mortality in relatives of women with ovarian cancer: the OPCS Study. Office of Population Censuses and Surveys.** *Int J Cancer* 1996, **65:**284-294.

7. Auranen A, Grenman S, Makinen J, Pukkala E, Sankila R, Salmi T: **Borderline ovarian tumors in Finland: epidemiology and familial occurrence.** *Am J Epidemiol* 1996, **144:**548-553.

8. Auranen A, Pukkala E, Makinen J, Sankila R, Grenman S, Salmi T: **Cancer incidence in the first-degree relatives of ovarian cancer patients.** *Br J Cancer* 1996, **74:**280-284.

9. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland.** *N Engl J Med* 2000, **343:**78-85.

10. Rahman N: **Realizing the promise of cancer predisposition genes.** *Nature* 2014, **505:**302-308.

11. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al: **Pathogenic Germline Variants in 10,389 Adult Cancers.** *Cell* 2018, **173:**355-370.e314.

12. Agarwal D, Nowak C, Zhang NR, Pusztai L, Hatzis C: **Functional germline variants as potential co-oncogenes.** *NPJ Breast Cancer* 2017, **3:**46.

13. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MD, Huang KL, Wyczalkowski MA, Jayasinghe R, Banerjee T, et al: **Patterns and functional implications of rare germline variants across 12 cancer types.** *Nat Commun* 2015, **6:**10086.

14. Southey MC, Goldgar DE, Winqvist R, Pylkas K, Couch F, Tischkowitz M, Foulkes WD, Dennis J, Michailidou K, van Rensburg EJ, et al: **PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS.** *J Med Genet* 2016, **53:**800-811.

15. Cheng DT, Prasad M, Chekaluk Y, Benayed R, Sadowska J, Zehir A, Syed A, Wang YE, Somar J, Li Y, et al: **Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing.** *BMC Med Genomics* 2017, **10:**33.

16. Muskens IS, de Smith AJ, Zhang C, Hansen HM, Morimoto L, Metayer C, Ma X, Walsh KM, Wiemels JL: **Germline cancer predisposition variants and pediatric glioma: a population-based study in California.** *Neuro Oncol* 2020.

17. Nanda N, Roberts NJ: **ATM Serine/Threonine Kinase and its Role in Pancreatic Risk.** *Genes (Basel)* 2020, **11**.

18. Merideth MA, Harney LA, Vyas N, Bachi A, Carr AG, Hill DA, Dehner LP, Schultz KAP, Stewart DR, Stratton P: **Gynecologic and reproductive health in patients with pathogenic germline variants in DICER1.** *Gynecol Oncol* 2020.

19. Chirita-Emandi A, Andreescu N, Zimbru CG, Tutac P, Arghirescu S, Serban M, Puiu M: **Challenges in reporting pathogenic/potentially pathogenic variants in 94 cancer predisposing genes - in pediatric patients screened with NGS panels.** *Sci Rep* 2020, **10:**223.

20. Guo R, DuBoff M, Jayakumaran G, Kris MG, Ladanyi M, Robson ME, Mandelker D, Zauderer MG: **Brief Report: Novel Germline Mutations in DNA Damage Repair in Patients with Malignant Pleural Mesotheliomas.** *J Thorac Oncol* 2019.

21. Sud A, Kinnersley B, Houlston RS: **Genome-wide association studies of cancer: current insights and future perspectives.** *Nat Rev Cancer* 2017, **17:**692-704.

22. Chang CQ, Yesupriya A, Rowell JL, Pimentel CB, Clyne M, Gwinn M, Khoury MJ, Wulf A, Schully SD: **A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes.** *Eur J Hum Genet* 2014, **22:**402-408.

23. Galvan A, Ioannidis JP, Dragani TA: **Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer.** *Trends Genet* 2010, **26:**132-141.

24. Hosking FJ, Dobbins SE, Houlston RS: **Genome-wide association studies for detecting cancer susceptibility.** *Br Med Bull* 2011, **97:**27-46.

25. Varghese JS, Easton DF: **Genome-wide association studies in common cancers--what have we learnt?** *Curr Opin Genet Dev* 2010, **20:**201-209.

26. Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S, Tsimberidou AM, Vnencak-Jones CL, Wolff DJ, Younes A, Nikiforova MN: **Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists.** *J Mol Diagn* 2017, **19:**4-23.

27. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, et al: **Germline Mutations in Predisposition Genes in Pediatric Cancer.** *N Engl J Med* 2015, **373:**2336-2346.

28. Bodmer W, Tomlinson I: **Rare genetic variants and the risk of cancer.** *Curr Opin Genet Dev* 2010, **20:**262-267.

29. Shen L, Jia J: **An Overview of Genome-Wide Association Studies in Alzheimer's Disease.** *Neurosci Bull* 2016, **32:**183-190.

30. Zhao M, Zhang Y, Cai W, Li J, Zhou F, Cheng N, Ren R, Zhao C, Li X, Ren S, et al: **The Bim deletion polymorphism clinical profile and its relation with tyrosine kinase inhibitor resistance in Chinese patients with non-small cell lung cancer.** *Cancer* 2014, **120:**2299-2307.

31. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, Wendl MC, Zhang Q, Koboldt DC, Xie M, Kandoth C, et al: **Integrated analysis of germline and somatic variants in ovarian cancer.** *Nat Commun* 2014, **5:**3156.

32. Jones AV, Chase A, Silver RT, Oscier D, Zoi K, Wang YL, Cario H, Pahl HL, Collins A, Reiter A, et al: **JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms.** *Nat Genet* 2009, **41:**446-449.

33. Kilpivaara O, Mukherjee S, Schram AM, Wadleigh M, Mullally A, Ebert BL, Bass A, Marubayashi S, Heguy A, Garcia-Manero G, et al: **A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms.** *Nat Genet* 2009, **41:**455-459.

34. Campbell PJ: **Somatic and germline genetics at the JAK2 locus.** *Nat Genet* 2009, **41:**385-386.

35. Olcaydu D, Harutyunyan A, Jager R, Berg T, Gisslinger B, Pabinger I, Gisslinger H, Kralovics R: **A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms.** *Nat Genet* 2009, **41:**450-454.

36. Liu W, He L, Ramirez J, Krishnaswamy S, Kanteti R, Wang YC, Salgia R, Ratain MJ: **Functional EGFR germline polymorphisms may confer risk for EGFR somatic mutations in non-small cell lung cancer, with a predominant effect on exon 19 microdeletions.** *Cancer Res* 2011, **71:**2423-2427.

37. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, et al: **Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer.** *Cancer Discov* 2017, **7:**410-423.

38. **Pan-cancer analysis of whole genomes.** *Nature* 2020, **578:**82-93.

39. Mamidi TKK, Wu J, Hicks C: **Mapping the Germline and Somatic Mutation Interaction Landscape in Indolent and Aggressive Prostate Cancers.** *J Oncol* 2019, **2019:**4168784.

40. Yu Y, Hu H, Chen JS, Hu F, Fowler J, Scheet P, Zhao H, Huff CD: **Integrated case-control and somatic-germline interaction analyses of melanoma susceptibility genes.** *Biochim Biophys Acta Mol Basis Dis* 2018, **1864:**2247-2254.

41. Chatrath A, Kiran M, Kumar P, Ratan A, Dutta A: **The Germline Variants rs61757955 and rs34988193 Are Predictive of Survival in Lower Grade Glioma Patients.** *Mol Cancer Res* 2019, **17:**1075-1086.

42. Chatrath A, Przanowska R, Kiran S, Su Z, Saha S, Wilson B, Tsunematsu T, Ahn J-H, Lee KY, Paulsen T, et al: **The Pan-Cancer Landscape of Prognostic Germline Variants in 10,582 Patients.** *medRxiv* 2019**:**19010264.

43. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER, 3rd, Barretina J, Gelfand ET, Bielski CM, Li H, et al: **Next-generation characterization of the Cancer Cell Line Encyclopedia.** *Nature* 2019, **569:**503-508.

44. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al: **COSMIC: the Catalogue Of Somatic Mutations In Cancer.** *Nucleic Acids Res* 2019, **47:**D941-d947.

45. Battaglin F, Naseem M, Lenz HJ, Salem ME: **Microsatellite instability in colorectal cancer: overview of its clinical significance and novel perspectives.** *Clin Adv Hematol Oncol* 2018, **16:**735-745.

46. Duffy MJ, Crown J: **Biomarkers for Predicting Response to Immunotherapy with Immune Checkpoint Inhibitors in Cancer Patients.** *Clin Chem* 2019, **65:**1228-1238.

47. Arora S, Velichinskii R, Lesh RW, Ali U, Kubiak M, Bansal P, Borghaei H, Edelman MJ, Boumber Y: **Existing and Emerging Biomarkers for Immune Checkpoint Immunotherapy in Solid Tumors.** *Adv Ther* 2019, **36:**2638-2678.

48. Guerrini-Rousseau L, Dufour C, Varlet P, Masliah-Planchon J, Bourdeaut F, Guillaud-Bataille M, Abbas R, Bertozzi AI, Fouyssac F, Huybrechts S, et al: **Germline SUFU mutation carriers and medulloblastoma: clinical characteristics, cancer risk, and prognosis.** *Neuro Oncol* 2018, **20:**1122-1132.

49. Baretta Z, Mocellin S, Goldin E, Olopade OI, Huo D: **Effect of BRCA germline mutations on breast cancer prognosis: A systematic review and meta-analysis.** *Medicine (Baltimore)* 2016, **95:**e4975.

50. Le Page C, Rahimi K, Rodrigues M, Heinzelmann-Schwarz V, Recio N, Tommasi S, Bataillon G, Portelance L, Golmard L, Meunier L, et al: **Clinicopathological features of women with epithelial ovarian cancer and double heterozygosity for BRCA1 and BRCA2: A systematic review and case report analysis.** *Gynecol Oncol* 2019.

51. Musa J, Cidre-Aranaz F, Aynaud MM, Orth MF, Knott MML, Mirabeau O, Mazor G, Varon M, Holting TLB, Grossetete S, et al: **Cooperation of cancer drivers with regulatory germline variants shapes clinical outcomes.** *Nat Commun* 2019, **10:**4128.

52. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.** *Cell* 2018, **173:**400-416.e411.

53. Menden MP, Casale FP, Stephan J, Bignell GR, Iorio F, McDermott U, Garnett MJ, Saez-Rodriguez J, Stegle O: **The germline genetic component of drug sensitivity in cancer cell lines.** *Nat Commun* 2018, **9:**3385.

54. Ledermann J, Harter P, Gourley C, Friedlander M, Vergote I, Rustin G, Scott CL, Meier W, Shapira-Frommer R, Safra T, et al: **Olaparib maintenance therapy in patients with platinum-sensitive relapsed serous ovarian cancer: a preplanned**

retrospective analysis of outcomes by BRCA status in a randomised phase 2 trial. *Lancet Oncol* 2014, **15:**852-861.

55. Gorodnova TV, Sokolenko AP, Ivantsov AO, Iyevleva AG, Suspitsin EN, Aleksakhina SN, Yanus GA, Togo AV, Maximov SY, Imyanitov EN: **High response rates to neoadjuvant platinum-based therapy in ovarian cancer patients carrying germ-line BRCA mutation.** *Cancer Lett* 2015, **369:**363-367.

56. Wheeler HE, Gamazon ER, Frisina RD, Perez-Cervantes C, El Charif O, Mapes B, Fossa SD, Feldman DR, Hamilton RJ, Vaughn DJ, et al: **Variants in WFS1 and Other Mendelian Deafness Genes Are Associated with Cisplatin-Associated Ototoxicity.** *Clin Cancer Res* 2017, **23:**3325-3333.

57. Teh LK, Hamzah S, Hashim H, Bannur Z, Zakaria ZA, Hasbullani Z, Shia JK, Fijeraid H, Md Nor A, Zailani M, et al: **Potential of dihydropyrimidine dehydrogenase genotypes in personalizing 5-fluorouracil therapy among colorectal cancer patients.** *Ther Drug Monit* 2013, **35:**624-630.

58. Romero Lagunes ML, Vera Badillo FE: **Design and Implementing Pharmacogenomics Study in Cancer.** *Adv Exp Med Biol* 2019, **1168:**43-77.

59. Morales J, Li L, Fattah FJ, Dong Y, Bey EA, Patel M, Gao J, Boothman DA: **Review of poly (ADP-ribose) polymerase (PARP) mechanisms of action and rationale for targeting in cancer and other diseases.** *Crit Rev Eukaryot Gene Expr* 2014, **24:**15-28.

60. Flohr C, Burkle A, Radicella JP, Epe B: **Poly(ADP-ribosyl)ation accelerates DNA repair in a pathway dependent on Cockayne syndrome B protein.** *Nucleic Acids Res* 2003, **31:**5332-5337.

61. de Murcia JM, Niedergang C, Trucco C, Ricoul M, Dutrillaux B, Mark M, Oliver FJ, Masson M, Dierich A, LeMeur M, et al: **Requirement of poly(ADP-ribose) polymerase in recovery from DNA damage in mice and in cells.** *Proc Natl Acad Sci U S A* 1997, **94:**7303-7307.

62. Satoh MS, Lindahl T: **Role of poly(ADP-ribose) formation in DNA repair.** *Nature* 1992, **356:**356-358.

63. McCabe N, Turner NC, Lord CJ, Kluzek K, Bialkowska A, Swift S, Giavara S, O'Connor MJ, Tutt AN, Zdzienicka MZ, et al: **Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition.** *Cancer Res* 2006, **66:**8109-8115.

64. Tung NM, Garber JE: **BRCA1/2 testing: therapeutic implications for breast cancer management.** *Br J Cancer* 2018, **119:**141-152.

65. Kaufman B, Shapira-Frommer R, Schmutzler RK, Audeh MW, Friedlander M, Balmana J, Mitchell G, Fried G, Stemmer SM, Hubert A, et al: **Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation.** *J Clin Oncol* 2015, **33:**244-250.

66. Robson M, Im SA, Senkus E, Xu B, Domchek SM, Masuda N, Delaloge S, Li W, Tung N, Armstrong A, et al: **Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation.** *N Engl J Med* 2017, **377:**523-533.

67. Robson ME, Tung N, Conte P, Im SA, Senkus E, Xu B, Masuda N, Delaloge S, Li W, Armstrong A, et al: **OlympiAD final overall survival and tolerability results: Olaparib versus chemotherapy treatment of physician's choice in patients with a germline BRCA mutation and HER2-negative metastatic breast cancer.** *Ann Oncol* 2019, **30:**558-566.

68. Exman P, Mallery RM, Lin NU, Parsons HA: **Response to Olaparib in a Patient with Germline BRCA2 Mutation and Breast Cancer Leptomeningeal Carcinomatosis.** *NPJ Breast Cancer* 2019, **5:**46.

69. Liu J, Bhadra M, Sinnakannu JR, Yue WL, Tan CW, Rigo F, Ong ST, Roca X: **Overcoming imatinib resistance conferred by the BIM deletion polymorphism in chronic myeloid leukemia with splice-switching antisense oligonucleotides.** *Oncotarget* 2017, **8:**77567-77585.

70. Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, Deininger MW, Silver RT, Goldman JM, Stone RM, et al: **Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia.** *N Engl J Med* 2006, **355:**2408-2417.

71. Cardona AF, Rojas L, Wills B, Arrieta O, Carranza H, Vargas C, Otero J, Corrales-Rodriguez L, Martin C, Reguart N, et al: **BIM deletion polymorphisms in Hispanic patients with non-small cell lung cancer carriers of EGFR mutations.** *Oncotarget* 2016, **7:**68933-68942.

72. Ng KP, Hillmer AM, Chuah CT, Juan WC, Ko TK, Teo AS, Ariyaratne PN, Takahashi N, Sawada K, Fei Y, et al: **A common BIM deletion polymorphism mediates intrinsic resistance and inferior responses to tyrosine kinase inhibitors in cancer.** *Nat Med* 2012, **18:**521-528.

73. Isobe K, Hata Y, Tochigi N, Kaburaki K, Kobayashi H, Makino T, Otsuka H, Sato F, Ishida F, Kikuchi N, et al: **Clinical significance of BIM deletion polymorphism in**

non-small-cell lung cancer with epidermal growth factor receptor mutation. *J Thorac Oncol* 2014, **9:**483-487.

74. Lee JH, Lin YL, Hsu WH, Chen HY, Chang YC, Yu CJ, Shih JY, Lin CC, Chen KY, Ho CC, et al: **Bcl-2-like protein 11 deletion polymorphism predicts survival in advanced non-small-cell lung cancer.** *J Thorac Oncol* 2014, **9:**1385-1392.

75. Nie W, Tao X, Wei H, Chen WS, Li B: **The BIM deletion polymorphism is a prognostic biomarker of EGFR-TKIs response in NSCLC: A systematic review and meta-analysis.** *Oncotarget* 2015, **6:**25696-25700.

76. Ma JY, Yan HJ, Gu W: **Association between BIM deletion polymorphism and clinical outcome of EGFR-mutated NSCLC patient with EGFR-TKI therapy: A meta-analysis.** *J Cancer Res Ther* 2015, **11:**397-402.

77. Longley DB, Harkin DP, Johnston PG: **5-fluorouracil: mechanisms of action and clinical strategies.** *Nat Rev Cancer* 2003, **3:**330-338.

78. Zhang CM, Lv JF, Gong L, Yu LY, Chen XP, Zhou HH, Fan L: **Role of Deficient Mismatch Repair in the Personalized Management of Colorectal Cancer.** *Int J Environ Res Public Health* 2016, **13**.

79. Cox VL, Saeed Bamashmos AA, Foo WC, Gupta S, Yedururi S, Garg N, Kang HC: **Lynch Syndrome: Genomics Update and Imaging Review.** *Radiographics* 2018, **38:**483-499.

80. Bras-Goncalves RA, Pocard M, Formento JL, Poirson-Bichat F, De Pinieux G, Pandrea I, Arvelo F, Ronco G, Villa P, Coquelle A, et al: **Synergistic efficacy of 3n-butyrate and 5-fluorouracil in human colorectal cancer xenografts via modulation of DNA synthesis.** *Gastroenterology* 2001, **120:**874-888.

81. Carethers JM, Chauhan DP, Fink D, Nebel S, Bresalier RS, Howell SB, Boland CR: **Mismatch repair proficiency and in vitro response to 5-fluorouracil.** *Gastroenterology* 1999, **117:**123-131.

82. Meyers M, Wagner MW, Hwang HS, Kinsella TJ, Boothman DA: **Role of the hMLH1 DNA mismatch repair protein in fluoropyrimidine-mediated cell death and cell cycle responses.** *Cancer Res* 2001, **61:**5193-5201.

83. Tokunaga E, Oda S, Fukushima M, Maehara Y, Sugimachi K: **Differential growth inhibition by 5-fluorouracil in human colorectal carcinoma cell lines.** *Eur J Cancer* 2000, **36:**1998-2006.

84. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, Hamilton SR, Laurent-Puig P, Gryfe R, Shepherd LE, et al: **Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer.** *N Engl J Med* 2003, **349:**247-257.

85. Sargent DJ, Marsoni S, Monges G, Thibodeau SN, Labianca R, Hamilton SR, French AJ, Kabat B, Foster NR, Torri V, et al: **Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer.** *J Clin Oncol* 2010, **28:**3219-3226.

86. Carethers JM, Smith EJ, Behling CA, Nguyen L, Tajima A, Doctolero RT, Cabrera BL, Goel A, Arnold CA, Miyai K, Boland CR: **Use of 5-fluorouracil and survival in patients with microsatellite-unstable colorectal cancer.** *Gastroenterology* 2004, **126:**394-401.

87. Jover R, Zapater P, Castells A, Llor X, Andreu M, Cubiella J, Pinol V, Xicola RM, Bujanda L, Rene JM, et al: **Mismatch repair status in the prediction of benefit from adjuvant fluorouracil chemotherapy in colorectal cancer.** *Gut* 2006, **55:**848-855.

88. Havel JJ, Chowell D, Chan TA: **The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy.** *Nat Rev Cancer* 2019, **19:**133-150.

89. Darvin P, Toor SM, Sasidharan Nair V, Elkord E: **Immune checkpoint inhibitors: recent progress and potential biomarkers.** *Exp Mol Med* 2018, **50:**165.

90. Wei SC, Duffy CR, Allison JP: **Fundamental Mechanisms of Immune Checkpoint Blockade Therapy.** *Cancer Discov* 2018, **8:**1069-1086.

91. Pirker R: **Biomarkers for immune checkpoint inhibitors in advanced nonsmall cell lung cancer.** *Curr Opin Oncol* 2019, **31:**24-28.

92. Bianco A, Perrotta F, Barra G, Malapelle U, Rocco D, De Palma R: **Prognostic Factors and Biomarkers of Responses to Immune Checkpoint Inhibitors in Lung Cancer.** *Int J Mol Sci* 2019, **20**.

93. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, et al: **Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade.** *Science* 2017, **357:**409-413.

94. Weinshilboum RM, Wang L: **Pharmacogenomics: Precision Medicine and Drug Response.** *Mayo Clin Proc* 2017, **92:**1711-1722.

95. Daly AK: **Pharmacogenetics: a general review on progress to date.** *Br Med Bull* 2017, **124:**65-79.

96. Helsby NA, Yong M, van Kan M, de Zoysa JR, Burns KE: **The importance of both CYP2C19 and CYP2B6 germline variations in cyclophosphamide pharmacokinetics and clinical outcomes.** *Br J Clin Pharmacol* 2019, **85:**1925-1934.

97. Tsuji D, Ikeda M, Yamamoto K, Nakamori H, Kim YI, Kawasaki Y, Otake A, Yokoi M, Inoue K, Hirai K, et al: **Drug-related genetic polymorphisms affecting severe chemotherapy-induced neutropenia in breast cancer patients: A hospital-based observational study.** *Medicine (Baltimore)* 2016, **95:**e5151.

98. Hedrich WD, Hassan HE, Wang H: **Insights into CYP2B6-mediated drug-drug interactions.** *Acta Pharm Sin B* 2016, **6:**413-425.

Chapter 2: The Germline Variants rs61757955 and rs34988193 are Predictive of Survival in Lower Grade Glioma Patients

Ajay Chatrath, Manjari Kiran, Pankaj Kumar, Aakrosh Ratan, and Anindya Dutta

Adapted From:

Chatrath A, Kiran M, Kumar P, Ratan A, Dutta A: **The Germline Variants rs61757955 and rs34988193 are Predictive of Survival in Lower Grade Glioma Patients.** *Mol Cancer Res* 2019.

- I conceived of the idea and design for this project, wrote the code for this analysis, analyzed the data, wrote the original draft of the manuscript, and made most of the figures in this manuscript. Some of the computational funding for this project was derived from the Cancer Genomics Cloud proposal for which I was the Project Leader.

Author Contributions

Conceptualization, A.C., A.D.; Methodology, A.C., A.R., A.D., M.K., P.K., Software, Formal Analysis, Investigation, Writing – Original Draft, and Visualization, A.C.; Resources and Funding Acquisition, A.D., A.C.; Writing – Review and Editing, all authors; Supervision and Administration, A.D., A.R., P.K.

The Germline Variants rs61757955 and rs34988193 are Predictive of Survival in Lower Grade Glioma Patients

Ajay Chatrath[1], Manjari Kiran[1], Pankaj Kumar[1], Aakrosh Ratan[2], and Anindya Dutta[1]

[1] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia

[2] Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia

Financial Support:
This work was supported by grants from the NIH R01 CA166054 and CA60499 and a Cancer Genomics Cloud Collaborative Support grant. The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

Potential Conflicts of Interest:
The authors do not have any conflicts of interest to disclose.

Abstract

Lower grade gliomas are invasive brain tumors that are difficult to completely resect neurosurgically. They often recur following resection and progress, resulting in death. Although previous studies have shown that specific germline variants increase the risk of tumor formation, no previous study has screened many germline variants to identify variants predictive of survival in glioma patients. In this study, we present an approach to identify the small fraction of prognostic germline variants from the pool of over four million variants that we variant called in The Cancer Genome Atlas whole exome sequencing and RNA sequencing datasets. We identified two germline variants that are predictive of poor patient outcomes by Cox regression, controlling for eleven covariates. rs61757955 is a germline variant found in the 3' UTR of *GRB2* associated with increased *KRAS* signaling, *CIC* mutations, and 1p/19q co-deletion. rs34988193 is a germline variant found in the tumor suppressor gene *ANKDD1a* that causes an amino acid change from lysine to glutamate. This variant was found to be predictive of poor prognosis in two independent lower grade glioma datasets and is predicted to be within the top 0.06% of deleterious mutations across the human genome. The wild type residue is conserved in all 22 other species with a homologous protein.

Implications: This is the first study presenting an approach to screening many germline variants to identify variants predictive of survival and our application of this methodology revealed the germline variants rs61757955 and rs34988193 as being predictive of survival in lower grade glioma patients.

Introduction

Grade II and grade III (low grade) gliomas are primary brain tumors that are derived from glial cells and include astrocytomas and oligodendrogliomas. They are most commonly found in the cerebral hemispheres. They are highly invasive and therefore difficult to completely resect neurosurgically without significant patient morbidity. Following surgery, patients are typically treated with chemotherapy and radiation, though these tumors typically recur or progress to grade IV gliomas and are fatal.[1] The median survival following lower grade glioma diagnosis is around 7 years.[2]

While the 2007 World Health Organization's (WHO) classification of central nervous system neoplasms differentiated between neoplasms primarily based on histological features, the updated 2016 WHO classification system now utilizes both molecular and histological parameters.[1] Isocitrate dehydrogenase mutation (*IDH*) status, 1p/19q co-deletion status, telomerase reverse transcriptase (*TERT*) promoter mutation status, *MGMT* promoter methylation, *TP53* mutation status, and *ATRX* mutation status may be used to molecularly characterize gliomas.[1,3] The availability of genomic data from patient glioma samples from groups such as The Cancer Genome Atlas (TCGA), the Chinese Glioma Genome Atlas (CGGA), and the Ivy Glioblastoma Atlas Project has substantially contributed to our understanding of these tumors.[4,5]

Many studies have utilized these datasets to identify gene expression signatures, microRNA expression patterns, somatic mutation status, and imaging characteristics that are predictive of survival in lower grade gliomas.[6–8] While

studies have shown that germline mutations can increase an individual's susceptibility for specific cancers,[9–12] including a recent study that identified 853 pathogenic or likely pathogenic germline variants found in 8% of 10,389 cancer patients,[13] no study has comprehensively screened all of the germline variants in a given cancer type to discover the prognostic variants in that cancer type. Although germline mutations have been shown to be prognostic in breast cancer[14] and medulloblastoma[9] in genes that have been well-characterized in the context of these cancers, these variants were not identified using an unbiased approach that screened a large number of germline variants. Identifying prognostic germline variants is challenging due to the limited effect size of germline variants, the large number of germline variants, and confounding clinical factors that may be associated with germline variants. Here we present a novel methodology for identifying prognostic germline variants and report two germline variants that we have found to be associated with survival in lower grade glioma patients.

Methods

Glioma Datasets

491whole exome sequenced normal blood samples (WXS normal), 503 whole exome sequenced tumor samples (WXS tumor), and 501 RNA sequenced tumor samples (RNA tumor) from TCGA lower grade glioma[4] patients available on the Cancer Genomics Cloud (CGC)[15] platform were used as part of this analysis. The clinical information was downloaded directly from the TCGA data portal using the GenomicDataCommons

(https://bioconductor.org/packages/release/bioc/html/GenomicDataCommons.html) R package available through Bioconductor. Additional molecular characteristics about these TCGA patients were acquired by downloading the supplement from Ceccarelli et. al.[16] The raw sequencing data from the Chinese Glioma Genome Atlas patients was downloaded using accession number SRP027383 from the Sequence Read Archive. Clinical information for these patients was downloaded directly from the project's website (http://www.cgga.org.cn/).

Variant Calling

Variant calling was performed on the TCGA lower grade glioma whole exome sequenced normal blood samples (WXS normal), whole-exome sequenced tumor samples (WXS tumor), and RNA sequenced tumor samples (RNA tumor) using VarDict[17] on CGC. The VarDict settings were set at default except for requiring mapping quality greater than 30, base quality greater than 25, a minimum of 3 variant reads, minimum allele frequency of 5%, and the removal of duplicate reads. We compiled a list of all of the unique variants and ran 'samtools[18] depth' on all sequencing files requiring a mapping quality greater than 30. We determined the status of each variant in each patient from the three datasets (WXS normal sample, WXS tumor sample, and RNA tumor sample). The variant status at positions with fewer than ten reads for a given patient was changed to unknown. We used the WXS tumor samples to insert variant calls into the WXS normal samples at positions at which a variant status was listed as unknown in the WXS normal samples. If the variant status was still missing in a

given patient, we then used the RNA tumor sample to insert variant calls into the combined WXS variant call set, allowing us to create the combined set of variant calls.

The same program parameters and approach were used to variant call and process the CGGA RNA sequencing dataset. All computation on the CGGA dataset was performed locally and not on CGC.

Quality Control

We used annovar[19] to determine the allele frequencies of the variants called by VarDict as listed in gnomAD (http://gnomad.broadinstitute.org/). We calculated the allele frequency of the variants in our study using the following formula:

$$\frac{2 * Number\ of\ Minor\ Allele\ Homozygotes + Number\ of\ Heterozygotes}{2 * Total\ Number\ of\ Patients}$$

The R package GGally (https://cran.r-project.org/web/packages/GGally/index.html) was used to calculate the correlation between the four variant call sets and to display their correlations with each other. Only variants with an allele frequency of greater than 5% in gnomAD and found in 15 or more of the TCGA lower grade glioma patients were tested for an association with survival by Cox regression.

Because we used the WXS tumor and RNA tumor samples to fill in missing variant calls, we evaluated whether somatic mutations were affecting the validity of our results. We first determined the percentage of variants called in the WXS tumor sample that were somatic mutations. To do this, we downloaded the

set of somatic mutations generated by the TCGA Research Network.[20] We then calculated the number of somatic mutations called in each patient in this variant call set and divided that number by the total number of variants called in that patient's WXS normal sample. To assess whether somatic mutations were affecting the integrity of our results, we counted the number of times that a somatic mutation called by the TCGA Research Network overlapped with the set of germline variants that we were testing for an association with survival.

Since we used the RNA tumor sample to fill in missing variant calls, we evaluated whether RNA editing was having a significant impact on our analysis. To do this, we downloaded the set of over 2.5 million known RNA editing sites from a rigorously annotated database of RNA editing sites, RADAR.[21] We counted the number of times that the germline variants that we were testing for an association with survival overlapped with any of the known 2.5 million RNA editing sites.

Principal Component Analysis

In order to calculate principal components that could separate patients on the basis of race, we used PLINK[22] to create a pruned set of germline variants to avoid bias from variants in linkage disequilibrium. Pruning was performed using a window size of 50 variants and a variance inflation factor of 2. These variants were used to calculate principal components using base R.

Cox Regression and Receiver Operator Characteristic Curves

Lasso in the R package glmnet[23] was run on 17 covariates (**Table 1**). Information about patient age, gender, tumor location, grade, treatment site, and

TP53 mutation status was acquired from the TCGA data portal, while data for patient somatic mutation count, percent aneuploidy, TERT expression, IDH mutation status, 1p/19q co-deletion status, MGMT promoter methylation status, and chromosome 7 gain with chromosome 10 loss status was acquired from Ceccarelli et. al.[16] The principal components were calculated as described above. 11 of these 17 covariates were selected for inclusion in the final model for survival prediction. The R packages survival[24] and survminer[25] were used to run Cox regression and create Kaplan-Meier curves. For each minor allele, we our model tested whether the minor allele was associated with a difference in survival outcomes with respect to the reference allele. False discovery rate correction was performed through Bonferroni correction.

Receiver operator characteristic (ROC) curves were created and evaluated using the survivalROC (https://cran.r-project.org/web/packages/survivalROC/survivalROC.pdf) and pROC (https://cran.r-project.org/web/packages/pROC/pROC.pdf) R packages. In order to test whether rs61757955 significantly improves the survival model consisting of the eleven covariates selected by Lasso, we compared the two ROC curves using the bootstrap method with 1000 iterations. We also used this bootstrapping approach to determine whether ANKDD1a expression levels, GRB2 expression levels, rs61757955, and rs34988193 together improve the survival model with respect to the eleven covariates selected by Lasso.

RNA-Sequencing Data Processing

We downloaded the HTSeq FPKM quantification files for each patient from the Genomic Data Commons data portal. We only used gene quantification files from primary tumor samples as part of this analysis. Replicate samples from a single patient were averaged.

Variant Correlation to Covariates and Somatic Mutations

In order to test for associations between the germline variants and genomic and histological tumor characteristics, we divided patients based on their germline variant status. We used the Wilcoxon rank-sum test to test for significant differences in each of the continuous variables between patients with and without a given variant. We used Fisher's exact test to test for differences in each of the discrete variables using a similar approach. Somatic mutation calls were downloaded from Ellrott et. al.[20]

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) of mRNA changes associated with rs61757955 and rs34988193 was performed by dividing the patients into two groups for each variant based on whether or not they had the reference allele at the position of the variant. For each germline variant, we calculated the log fold change for all genes expressed greater than one fragment per kilobase per million mapped reads (FPKM) between patients with the variant and without the variant. For each gene, fold change was calculated by dividing the median expression of the gene in patients with the variant by the median expression of

the gene in patients without the variant. We used the log fold change to rank the genes from greatest log fold change to smallest log fold change. This file was used as input for GSEA.[26]

Variant Annotation

In order to identify deleterious mutations, we annotated all variants by combined annotation dependent depletion (CADD) scores and only analyzed the variants predicted to be within the top 0.1% of all deleterious variants (CADD > 30).[27] This led us to identify rs34988193 in *ANKDD1a* as a potentially deleterious variant predictive of survival. Because rs34988193 causes an amino acid change from positively charged lysine to negatively charged glutamate, we ran a BLASTp (httpps://blast.ncbi.nlm.nih.gov/Blast.cgi) search so that we could determine how many species have a protein homologous to ANKDD1a and how consistently the wild type lysine residue was conserved. We identified homologous sequences in 22 other species. These sequences were aligned using ClustalW in MEGA.[28] We also annotated this variant with its PhyloP score.[29] Because the crystal structure for *ANKDD1a* was not available, we downloaded the predicted model for this protein from Modbase (https://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi) and calculated the Gribskov score using prophecy on EMBOSS.[30] We retrieved linked variants from Ensembl using the population of Utah residents with Northern and Western European ancestry which is demographically similar to the TCGA lower grade glioma patient population.

Results

Identification of High Quality Germline Variants

Our variant calling pipeline is shown in **Figure 1**. Briefly, we used the variant caller VarDict on Cancer Genomics Cloud to identify variants from whole exome sequencing (WXS) and RNA sequencing samples in about 500 lower grade glioma patients. In total, we found 4,453,701 unique variants. We used 'samtools depth' to determine the sequencing depth at each of these variants for each patient and changed the variant status to 'unknown' for patients with sequencing coverage less than 10 reads at a given position. We created a set of combined variant calls by using the WXS and RNA tumor samples to fill in unknown values in the whole exome sequenced normal samples that resulted from having a sequencing coverage of less than 10 reads at a given position. This approach increased our sample size and enabled us to include many more variants in our analysis than if we had solely used variant calls from the whole exome sequenced normal blood samples. Ultimately, this left us with four sets of variants – WXS normal, WXS tumor, RNA tumor, and a combined set that resulted from merging the other three variant call sets, giving preference to the WXS normal and then WXS tumor variant calls. We used the combined variant call set when testing variants for an association with survival. We only tested variants found in 15 or more lower grade glioma TCGA patients and listed in gnomAD as having an allele frequency of greater than 5%.

Tumor Variant Calls are not Significantly Affected by Somatic Mutations or RNA Editing After Filtering

Because we used sequencing data from the WXS tumor and RNA tumor samples to fill in missing calls in the WXS normal samples, we evaluated our variant calls for contributions from somatic mutations and RNA editing. We first showed that the majority of variant calls in the tumor sample are germline variant calls. To do this, we counted the number of somatic mutations called by the TCGA Research Network's analysis in each patient and divided that number by the number of variants that we called in the WXS normal sample.[20] The median number of somatic mutations called per patient was 39. The median number of variants called in the WXS normal sample was 95,794. We therefore estimated that over 99.9% of variants called in the WXS tumor sample consisted of germline variants and that the percentage of somatic mutations in the WXS tumor sample across all patients was quite small (**Figure S1**). Because somatic mutations rarely occur at the same position, we suspected that the number of somatic mutations included in our study was extremely small since we limited our analysis to variants found in 15 or more of the lower grade glioma patients and found in gnomAD with an allele frequency of greater than 5%. Indeed, only one of the 196,022 variants that we tested overlapped with a somatic mutation. This somatic mutation occurred in only a single patient (**Table S1**). Ultimately, we did not find any evidence to suggest that somatic mutations were impacting the quality of our analysis.

We next determined whether RNA editing was affecting our analysis by downloading the 2.5 million known RNA editing sites from the rigorously

annotated RNA editing database, RADAR.[21] Only 215 of the 196,022 variants that we tested were located at a position that overlapped with a known RNA editing site. We did not find any of these variants to be prognostic as part of our analysis. We therefore did not find any empirical evidence to suggest that somatic mutations or RNA editing impacted our findings (**Table S1**).

Finally, we established that our four variant call sets (WXS normal, WXS tumor, RNA tumor, and combined) were concordant with each other by calculating the allele frequency of each variant called in the four sets and demonstrating a very strong correlation between all pairs of variants ($r > 0.98$ for all pairs, **Figure S2**). To further evaluate the quality of our variants calls, we calculated the frequency of each allele and compared it to the frequency of these alleles as listed in gnomAD. Our alleles frequencies were well correlated with gnomAD ($r > 0.93$ for all four variant sets, **Table S2**). As expected, the distribution of allele frequencies is negatively skewed as the majority of the identified variants are rare (**Figure S2**). We used the variants from the WXS normal samples to determine the principal components. As expected, these principal components effectively separate patients on the basis of reported race (**Figure S3**).

Identification of 271 Prognostic Germline Variants that are Independent of Clinical Covariates

In order to identify clinically relevant germline variants, we restricted our analysis to variants found in at least 15 patients in the TCGA dataset and found in gnomAD with an allele frequency of greater than five percent. This restricted our analysis to 196,022 testable variants (**Figure 2A**). In order to reduce the risk

of identifying variants that are prognostic because they are confounded by other covariates known to be associated with survival, we used the machine learning algorithm Lasso to determine which of 17 covariates should be controlled for in our Cox regression model. Lasso regression was useful in the screening of these 17 covariates because it penalizes models based on the number of coefficients, allowing for the elimination of less predictive coefficients from the model. The algorithm selected 10 covariates known to be associated with differences in survival in lower grade glioma (age, somatic mutation count, percent aneuploidy, histological subtype of astrocytoma, tumor grade, treatment site, *IDH* mutation status, 1p/19q co-deletion status, *MGMT* promoter methylation status, chromosome 7 gain/chromosome 10 loss status) along with the third principal component that we calculated (**Table 1**). Although the first two principal components are more effective in stratifying patients on the basis of race than the third principal component, the selection of the third principal component over the first two suggests that the third principal component contributes more information to the survival model than the first two principal components. This third principal component primarily separates African Americans from each other, suggesting that a subpopulation of African Americans experienced worse clinical outcomes in this dataset compared to other groups. We ran Cox regression on all 196,022 variants one at a time, controlling for these 11 covariates, to identify germline variants predictive of survival.

We identified 271 germline variants that are predictive of survival ($p < 0.001$) (**Figure 2A**). As is the case with germline variants in general, the majority

of these germline variants are found in protein-coding genes (**Figure 2B**), are located in introns (**Figure 2C**), and are single nucleotide polymorphisms (**Figure 2D**). Most single nucleotide polymorphisms are transitions (**Figure 2E**).

The Germline Variant rs61757955 in *GRB2* is Associated with Poor Prognosis

We identified two germline variants that are highly predictive of survival after false discovery rate correction (FDR < 0.10) (**Figure 3A, Table 2A**). rs61757955 results in a mutation in the 3' UTR of Growth Factor Receptor Bound Protein 2 (*GRB2*) and is associated with a poor prognosis (p=7.08E-10, hazard ratio(HR)=20.4, **Figure 3B, Table 2A**). To determine whether rs61757955 enhances the survival model compared to the eleven clinical covariates alone, we calculated a risk score for each patient using a Cox regression model with rs61757955 and the other 11 covariates and a risk score using the 11 covariates alone. Using these risk scores, we determined the rate at which a patient would be correctly labeled as alive or dead at 7 years with a given false positive rate to create a receiver operator characteristic curve. The increased area under the curve suggests that rs61757955 enhances the survival model compared to the eleven clinical covariates alone (p=0.0489, **Figure 3C**). The allele frequency of rs61757955 is close to 0% according to the 1000 Genomes Project[31] in the Chinese population and, as expected, did not show up in the Chinese Glioma Genome Atlas. We also found rs28672782, a germline variant found in the intron of BRSK2, to be associated with a favorable prognosis, though the testable sample size for this variant was small and the maximum follow up for patients

with this variant was only three years. Therefore, we did not investigate this variant further (**Figure S4, Table 2A**).

In order to test whether rs61757955 in *GRB2* is associated with an increased risk of other genomic abnormalities, we separated patients on the basis of this variant to see if there was a difference in the incidence of the genomic or histological variables (**Table 3**). We found this variant to be associated with an increased incidence of 1p/19q co-deletions (p=0.038). Because 1p/19q co-deletions are frequently seen in Capicua transcriptional repressor (*CIC*) mutated gliomas[32] and *CIC* aberrations are known to be a driver in lower grade glioma tumorigenesis,[33] we tested whether there was a difference in the incidence of *CIC* mutations in patients with this variant. 38% of patients with this variant had *CIC* mutated gliomas, whereas only 16% of patients without the variant had a *CIC* mutation (p=0.0168, **Table 3**). Although the incidence of oligodendrogliomas was elevated in patients with the variant compared to patients without the variant, consistent with reports from the literature that 1p/19q co-deletions and *CIC* mutations are enriched in oligodendrogliomas,[32] this difference was not statistically significant (p=0.475). Since rs61757955 is in a non-coding region, we also tested whether this variant is associated with differences in gene expression. We separated patients based on their variant status and calculated the log fold change of each gene between patients with the variant and patients without the variant. This data was used as the input for gene set enrichment analysis (GSEA). We found rs61757955 to be associated with increased *KRAS* signaling (FDR=0.015) (**Figure 3D**).

Because we only have whole exome sequencing and RNA sequencing data from The Cancer Genome Atlas, we do not know whether the upregulation of genes in the *KRAS* signaling pathway and the increased incidence of *CIC* mutations and 1p/19q deletions are due to this variant or a linked variant in a regulatory region that we would be able to analyze with whole genome sequencing data. Therefore, we identified the four other variants that are genetically linked to rs61757955 in the European population, the population which is most similar to the TCGA lower grade glioma patient population (**Table S3**). These variants did not pass the criteria to be included within the 196,022 testable variants that we had identified at the beginning of this study but could become useful in the future.

rs34988193 is a Deleterious Germline Variant Present in ANKDD1a Associated with Poor Outcomes

In order to identify prognostic variants that are predicted to be deleterious due to effects on the encoded protein, we repeated our analysis but restricted it to only variants with a combined annotation dependent depletion (CADD) score greater than 30 and expression greater than one FPKM on average. 81 variants met this criteria. These variants correspond to the top 0.1% of deleterious mutations as predicted by this scoring system. We found the germline variant rs34988193 in the tumor suppressor gene *ANKDD1a* to be associated with poor prognosis in the TCGA dataset ($p=0.001$, $HR=1.73$, $FDR < 0.10$, **Figure 4A-B, Table 2B**). Because this variant is found in both the European and Asian populations, we were able to test whether this variant is also predictive of survival in the independent Chinese Glioma Genome Atlas (CGGA) dataset. We found

this variant to be predictive of survival in the CGGA dataset and we found the hazard ratio that we calculated in CGGA to be very similar to the hazard ratio calculated in the TCGA dataset (p=0.0743, HR=1.79, **Figure 4C, Table 2B**). rs34988193 is not linked with any other variant in the European population. We did not find any enriched pathways after performing gene set enrichment analysis and this variant was not associated with differences of any of the genomic or histological variables (**Table S4**).

ANKDD1a contains ten ankyrin repeat domains and one death-like domain. This variant causes a non-synonymous mutation in the last codon of the ninth ankyrin repeat domain. The AAG to GAG codon change results in the incorporation of negatively charged glutamate instead of the wild type positively charged lysine residue in the loop between ankyrin repeats nine and ten (**Figure 4D**). This variant has a CADD score of 32 and is therefore predicted to be in the top 0.06% of deleterious mutations across the human genome. We performed a BLASTp search using the *ANKDD1a* protein sequence to identify homologous sequences in 22 other species. We aligned these sequences using ClustalW and found that this lysine residue is conserved in all 22 of these species (**Figure 4E**). The PhyloP score at this position is 8.42, suggesting that evolution is occurring much more slowly than expected at this residue assuming no selection pressure. We determined the position-specific profile Gribskov's score for a lysine to glutamate amino acid change at this position using the multiple sequencing alignment from 23 species to be 15 to 3, suggesting that this variant is highly unfavorable.

Combined Model Predicts Survival Better Than Clinical Covariates Alone

As a result of this analysis, we found the germline variants rs61757955 in the 3' UTR of GRB2 and rs34988193 in the protein-coding region of ANKDD1a to be predictive of survival in lower grade glioma patients. We constructed a survival model consisting of the eleven clinical covariates, rs61757955, rs34988193, GRB2 expression, and ANKDD1a expression and generated a receiver operator characteristic curve by using this model to categorize patients as alive or dead after seven years of follow up. This combined model is significantly better at predicting survival compared to the eleven clinical covariates alone (p=0.0279, **Figure 4F**).

Discussion

Up until this point, the identification of prognostic features in gliomas has been limited to clinical factors, somatic mutations, gene expression changes, and methylation pattern changes.[6–8] Although many studies have commented on how germline variants could enable physicians to better individualize patient care by being able to better predict how a patient might respond to chemotherapeutic treatment,[34–36] most large-scale studies have focused on identifying germline variants that predispose or protect an individual to a disease.[13,37] These studies have not focused on understanding how germline variants can be used to individualize patient care following diagnosis. Identifying prognostic germline variants is difficult due to the large number of germline variants, the limited effect of any single germline variant, and clinical factors that may confound the effect of germline variants. In this study, we have developed a novel method that can be

used to identify prognostic germline variants and we have used that method to identify two variants that are predictive of survival in the TCGA dataset. The germline variant rs61757955 in *GRB2* is not found in the Asian population and so could not be confirmed in an independent dataset. In contrast, the germline variant rs34988193 in *ANKDD1a* is found in both the European and Asian populations, and remarkably, was found to be prognostic with very similar hazard ratios in both the TCGA and CGGA datasets.

Studies of germline variants using TCGA datasets typically solely utilize the WXS normal blood samples.[13,38] One major disadvantage to this approach is that it limits the analysis to genes within the capture regions of the whole exome sequencing kits used by the study.[4] In this study, we combined the information from both the whole exome sequencing and RNA sequencing datasets for a given patient to identify germline variants outside of the whole exome sequencing capture region. Our approach had the added benefit of providing us with more information for a given variant for variants with low sequencing depth in the whole exome sequencing datasets. We do not believe that this approach significantly affected the accuracy of our variant calls because the allele frequencies calculated from the RNA sequencing dataset were well correlated with the allele frequencies from gnomAD and with the allele frequencies calculated from the whole exome sequencing datasets. We showed that somatic mutations and RNA editing did not affect the integrity of our finding. Only one somatic mutation in a single patient overlapped with the 196,022 variants that we tested in our analysis and only 215 of the 196,022 variants that we tested

overlapped with the 2.5 million known RNA editing sites. We did not find any of these variants to be predictive of survival. Instead, we feel that the increased sample size resulting from the additional sequencing coverage greatly outweighs any effect that somatic mutations or RNA editing had on our results.

We next needed to devise an approach to using these germline variants in a Cox regression model. We first had to decide how to deal with the absence of a variant in the variant call file. The variant could be absent because the patient was wild type for that allele or because the sequencing depth at that position was too low to make the variant call. We therefore determined the sequencing depth of each variant at each position so that we could exclude patients with low sequencing depths for the testing of specific variants. Testing a large number of variants increased the probability of a variant being significant solely because it was confounded with another significant variable. To avoid this issue, we tested each variant while controlling for 11 other covariates that we found to be predictive of survival. In this study, we found rs61757955 to be associated with differences in 1p/19q co-deletion status. By including the 1p/19q co-deletion as a covariate in our model, we were able to estimate the effect of rs61757955 independent from the 1p/19q co-deletion status and the other ten covariates.

GRB2 is a signal transduction adaptor protein that plays an oncogenic role in a variety of cancers.[39–42] *GRB2* plays an important role in the *RAS/RAF/ERK* pathway. Its SH2 domain binds the phosphotyrosine of activated growth factor receptor, while its two SH3 domains bind the guanine nucleotide exchange factor son of sevenless (*SOS*) protein, resulting in *SOS* recruitment to the plasma

membrane and subsequent *RAS* activation. *RAS* binds and activates the kinase *RAF*, which phosphorylates the kinase *MEK*. *MEK* phosphorylates and activates extracellular signal-regulated kinase (*ERK*) which transmits the signal to transcription factors in the nucleus. This results in cell proliferation.[43] We found the variant rs61757955 located in the 3' UTR of *GRB2* to be associated with poor prognosis in glioma patients. Separating patients on the basis of this variant revealed that the *KRAS* signaling pathway is upregulated in patients with this variant. As described above, *GRB2* plays a well-characterized role in this pathway.[43] We also found this variant to be associated with an increased incidence of *CIC* mutations and 1p/19q co-deletions. *CIC* is a known driver of lower grade glioma pathogenesis.[33] Mutations in *CIC* are common in oligodendrogliomas and are associated with poor prognosis.[4,32] Although patients with rs61757955 variant exhibited an elevation in the incidence of oligodendrogliomas which we expected given the increased incidence of *CIC* mutations and 1p/19q co-deletions,[32] this difference was not statistically significant. It is possible that this germline variant or the four other germline variants that it is linked with increase a patient's risk for oligodendrogliomas with the *CIC* mutation and 1p/19q co-deletion.

In this study, we were only able to study variants in the whole exome or RNA sequencing data. Although it is possible that the 3' UTR of *GRB2* has regulatory activity or affects *GRB2* protein translation efficiency, it is also possible that one of the variants that rs61757955 is linked to regulates the *KRAS* signaling pathway. None of the four linked variants are in the protein coding sequence of

GRB2 so that if they upregulate RAS activity, like the rs61757955, they likely do so by regulating the expression of *GRB2*. While recent large-scale sequencing studies have published patient whole genome sequences,[44] this data is not yet available for gliomas. We will be able to apply our approach to variants in regulatory regions in the future to specifically identify these prognostic variants when whole genome sequencing data for gliomas is available. Our inability to further study this variant in the CGGA dataset due to this variant being rare in Asian populations is a limitation of this study which could be addressed in the future with the availability of additional glioma sequencing datasets. This result also suggests that the clinical usefulness of specific germline variants is dependent on the frequency of that germline variant in the population.

ANKDD1a is a tumor suppressor gene that has been shown to inhibit cell autophagy and induce apoptosis in glioblastoma multiforme (GBM). It directly interacts with and upregulates *FIH1*, resulting in inhibition of *HIF1α* activity and decreased *HIF1α* half-life. This induces apoptosis in GBM cell lines in hypoxic microenvironments. Hypermethylation of this gene is common in GBM and leads to decreased *ANKDD1a* expression and increased cell proliferation.[45] We found the germline variant rs34988193, located at the end of the ninth of ten ankyrin repeat domains in this protein, to be associated with a poor prognosis in lower grade glioma patients in both the TCGA and CGGA datasets. The hazard ratio independently calculated using the two datasets is remarkably similar. The wild type lysine residue is conserved in all 22 species with a homologue to *ANKDD1a* and this position has a high PhyloP score. This variant is predicted to be within

the top 0.06% of deleterious mutations in the human genome by CADD score[27] because it causes a change from a positively charged lysine residue to a negatively charged glutamic acid residue in the loop of this ankyrin repeat. Ankyrin repeats are common domains known for their involvement with protein-protein interactions.[46,47] Previous studies have suggested that mutations in the loops of ankyrin repeats may disrupt protein-protein interactions.[48–50] The change from a positively to negatively charged amino acid resulting from the germline variant rs34988193 in the loop of *ANKDD1a* may disrupt *ANKDD1a*'s protein interaction partners and could explain the poor prognosis associated with this variant seen in two independent datasets. Given the amino acid change, further studies involving rs34988193 in *ANKDD1a* could be directed towards experimentally determining whether or not this variant alters *ANKDD1a*'s protein-protein interactions.

rs61757955 in *GRB2* and rs34988193 in *ANKDD1a* could also be used to enhance predictions made by survival models clinically, as we found that these variants are significant predictors of prognosis even after controlling for eleven covariates. The prognostic effect of rs34988193 in *ANKDD1a* seems to be fairly reliable, as we found that this variant had a similar hazard ratio in both the TCGA and CGGA datasets. Our approach could be used in the future to identify sets of germline variants that together enhance the predictions made by survival models, though the current number of lower grade glioma sequencing samples is small relative to the large number of possible combinations of germline variants. Focused studies on particular sets of genes or pathways could potentially get

around this low sample size problem by drastically limiting the number of variants studied. We believe that this study provides researchers with an effective approach to identifying biologically significant germline variants and provides clinicians with germline variants that could enhance currently existing survival models.

Acknowledgements

We thank Dr. Wei Min Chen at the University of Virginia for his statistical input on our Cox regression models. We thank Dr. Ana Damljanovic and Dr. Liz Williams for their assistance with computation on the Cancer Genomics Cloud Platform. We thank dbGAP for providing us with access to The Cancer Genome Atlas data. We thank all of the patients and their families that participated in The Cancer Genome Atlas and Chinese Glioma Genome Atlas studies. We thank the Dutta lab members for the valuable feedback during the drafting of this manuscript.

References

1. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol*. 2016. doi:10.1007/s00401-016-1545-1

2. Bauman G, Fisher B, Watling C, Cairncross JG, Macdonald D. Adult Supratentorial Low-Grade Glioma: Long-Term Experience at a Single Institution. *Int J Radiat Oncol Biol Phys*. 2009. doi:10.1016/j.ijrobp.2009.01.010

3. Jiao Y, Killela PJ, Reitman ZJ, et al. Frequent ATRX, CIC, FUBP1 and IDH1 mutations refine the classification of malignant gliomas. *Oncotarget*. 2012;3(7):709-722. doi:10.18632/oncotarget.588

4. Network CGAR, Brat DJ, Verhaak RGW, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med*. 2015. doi:10.1056/NEJMoa1402121

5. Yan W, Zhang W, You G, et al. Molecular classification of gliomas based on whole genome gene expression: A systematic report of 225 samples from the Chinese Glioma Cooperative Group. *Neuro Oncol*. 2012. doi:10.1093/neuonc/nos263

6. Qian Z, Li Y, Fan X, et al. Prognostic value of a microRNA signature as a novel biomarker in patients with lower-grade gliomas. *J Neurooncol*. 2018. doi:10.1007/s11060-017-2704-5

7. Hu X, Martinez-Ledesma E, Zheng S, et al. Multigene signature for predicting prognosis of patients with 1p19q co-deletion diffuse glioma. *Neuro Oncol*. 2017. doi:10.1093/neuonc/now285

8. Zhou H, Vallières M, Bai HX, et al. MRI features predict survival and molecular markers in diffuse lower-grade gliomas. *Neuro Oncol*. 2017. doi:10.1093/neuonc/now256

9. Guerrini-Rousseau L, Dufour C, Varlet P, et al. Germline SUFU mutation carriers and medulloblastoma: Clinical characteristics, cancer risk, and prognosis. *Neuro Oncol*. 2018. doi:10.1093/neuonc/nox228

10. Dudley B, Karloski E, Monzon FA, et al. Germline mutation prevalence in individuals with pancreatic cancer and a history of previous malignancy. *Cancer*. 2018. doi:10.1002/cncr.31242

11. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med*. 2015.

doi:10.1056/NEJMoa1508054

12. Carter H, Marty R, Hofree M, et al. Interaction landscape of inherited polymorphisms with somatic events in cancer. *Cancer Discov*. 2017. doi:10.1158/2159-8290.CD-16-1045

13. Huang K lin, Mashl RJ, Wu Y, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*. 2018. doi:10.1016/j.cell.2018.03.039

14. Baretta Z, Mocellin S, Goldin E, Olopade OI, Huo D. Effect of BRCA germline mutations on breast cancer prognosis: A systematic review and meta-analysis. *Medicine (Baltimore)*. 2016. doi:10.1097/MD.0000000000004975

15. Lau JW, Lehnert E, Sethi A, et al. The cancer genomics cloud: Collaborative, reproducible, and democratized - A new paradigm in large-scale computational research. *Cancer Res*. 2017. doi:10.1158/0008-5472.CAN-17-0387

16. Ceccarelli M, Barthel FP, Malta TM, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*. 2016. doi:10.1016/j.cell.2015.12.028

17. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016. doi:10.1093/nar/gkw227

18. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009. doi:10.1093/bioinformatics/btp352

19. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010. doi:10.1093/nar/gkq603

20. Ellrott K, Bailey MH, Saksena G, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst*. 2018. doi:10.1016/j.cels.2018.03.002

21. Ramaswami G, Li JB. RADAR: A rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res*. 2014. doi:10.1093/nar/gkt996

22. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007. doi:10.1086/519795

23. Friedman AJ, Hastie T, Simon N, Tibshirani R, Hastie MT. Lasso and

Elastic-Net Regularized Generalized Linear Models. *Available online https//cran.r-project.org/web/packages/glmnet/glmnet.pdf (Verified 29 July 2015)*. 2015.

24. Therneau TM, T. Lumley. Package ‘ survival .’ *R Top Doc*. 2015. doi:10.1016/j.jhydrol.2011.07.022.

25. *Package “survminer” Type Package Title Drawing Survival Curves Using “Ggplot2.”*; 2018. https://cran.r-project.org/web/packages/survminer/survminer.pdf. Accessed August 27, 2018.

26. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. 2005. doi:10.1073/pnas.0506580102

27. Kircher M, Witten DM, Jain P, O’roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014. doi:10.1038/ng.2892

28. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016. doi:10.1093/molbev/msw054

29. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010. doi:10.1101/gr.097857.109

30. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*. 2000. doi:10.1016/S0168-9525(00)02024-2

31. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015. doi:10.1038/nature15393

32. Gleize V, Alentorn A, Connen De Kérillis L, et al. CIC inactivating mutations identify aggressive subset of 1p19q codeleted gliomas. *Ann Neurol*. 2015. doi:10.1002/ana.24443

33. Bailey MH, Tokheim C, Porta-Pardo E, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 2018;173(2):371-385.e18. doi:10.1016/j.cell.2018.02.060

34. Pesenti C, Gusella M, Sirchia SM, Miozzo M. Germline oncopharmacogenetics, a promising field in cancer therapy. *Cell Oncol*.

2015. doi:10.1007/s13402-014-0214-4

35. Pinto N, Cohn SL, Dolan ME. Using germline genomics to individualize pediatric cancer treatments. *Clin Cancer Res*. 2012. doi:10.1158/1078-0432.CCR-11-1938

36. Schärfe CPI, Tremmel R, Schwab M, Kohlbacher O, Marks DS. Genetic variation in human drug-related genes. *Genome Med*. 2017. doi:10.1186/s13073-017-0502-5

37. Lu C, Xie M, Wendl MC, et al. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun*. 2015. doi:10.1038/ncomms10086

38. Koire A, Katsonis P, Lichtarge O. REPURPOSING GERMLINE EXOMES OF THE CANCER GENOME ATLAS DEMANDS A CAUTIOUS APPROACH AND SAMPLE-SPECIFIC VARIANT FILTERING. *Pac Symp Biocomput*. 2016.

39. Gay B, Suarez S, Caravatti G, Furet P, Meyer T, Schoepfer J. Selective Grb2 SH2 inhibitors as anti-Ras therapy. *Int J Cancer*. 1999. doi:10.1002/(SICI)1097-0215(19991008)83:2<235::AID-IJC15>3.0.CO;2-B

40. Giubellino A, Burke TR, Bottaro DP. Grb2 signaling in cell motility and cancer. *Expert Opin Ther Targets*. 2008. doi:10.1517/14728222.12.8.1021

41. Haines E, Saucier C, Claing A. The adaptor proteins p66Shc and Grb2 regulate the activation of the GTPases ARF1 and ARF6 in invasive breast cancer cells. *J Biol Chem*. 2014. doi:10.1074/jbc.M113.516047

42. Yu GZ, Chen Y, Wang JJ. Overexpression of Grb2/HER2 signaling in Chinese gastric cancer: Their relationship with clinicopathological parameters and prognostic significance. *J Cancer Res Clin Oncol*. 2009. doi:10.1007/s00432-009-0574-8

43. Lowenstein EJ, Daly RJ, Batzer AG, et al. The SH2 and SH3 domain-containing protein GRB2 links receptor tyrosine kinases to ras signaling. *Cell*. 1992. doi:10.1016/0092-8674(92)90167-B

44. Bolouri H, Farrar JE, Triche T, et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat Med*. 2018. doi:10.1038/nm.4439

45. Feng J, Zhang Y, She X, et al. Hypermethylated gene ANKDD1A is a candidate tumor suppressor that interacts with FIH1 and decreases HIF1α stability to inhibit cell autophagy in the glioblastoma multiforme hypoxia

microenvironment. *Oncogene*. August 2018:1. doi:10.1038/s41388-018-0423-9

46. Sedgwick SG, Smerdon SJ. The ankyrin repeat: A diversity of interactions on a common structural framework. *Trends Biochem Sci*. 1999. doi:10.1016/S0968-0004(99)01426-7

47. Shi DJ, Ye S, Cao X, Zhang R, Wang KW. Crystal structure of the N-terminal ankyrin repeat domain of TRPV3 reveals unique conformation of finger 3 loop critical for channel function. *Protein Cell*. 2013. doi:10.1007/s13238-013-3091-0

48. Sullivan JM, Zimanyi CM, Aisenberg W, et al. Novel mutations highlight the key role of the ankyrin repeat domain in TRPV4-mediated neuropathy. *Neurol Genet*. 2015;1(4):e29. doi:10.1212/NXG.0000000000000029

49. Lamandé SR, Yuan Y, Gresshoff IL, et al. Mutations in TRPV4 cause an inherited arthropathy of hands and feet. *Nat Genet*. 2011. doi:10.1038/ng.945

50. Li J, Mahajan A, Tsai MD. Ankyrin repeat: A unique motif mediating protein-protein interactions. *Biochemistry*. 2006. doi:10.1021/bi062188q

Figures

Figure 1. A flowchart describing the steps involved in identifying prognostic germline variants.

Figure 1

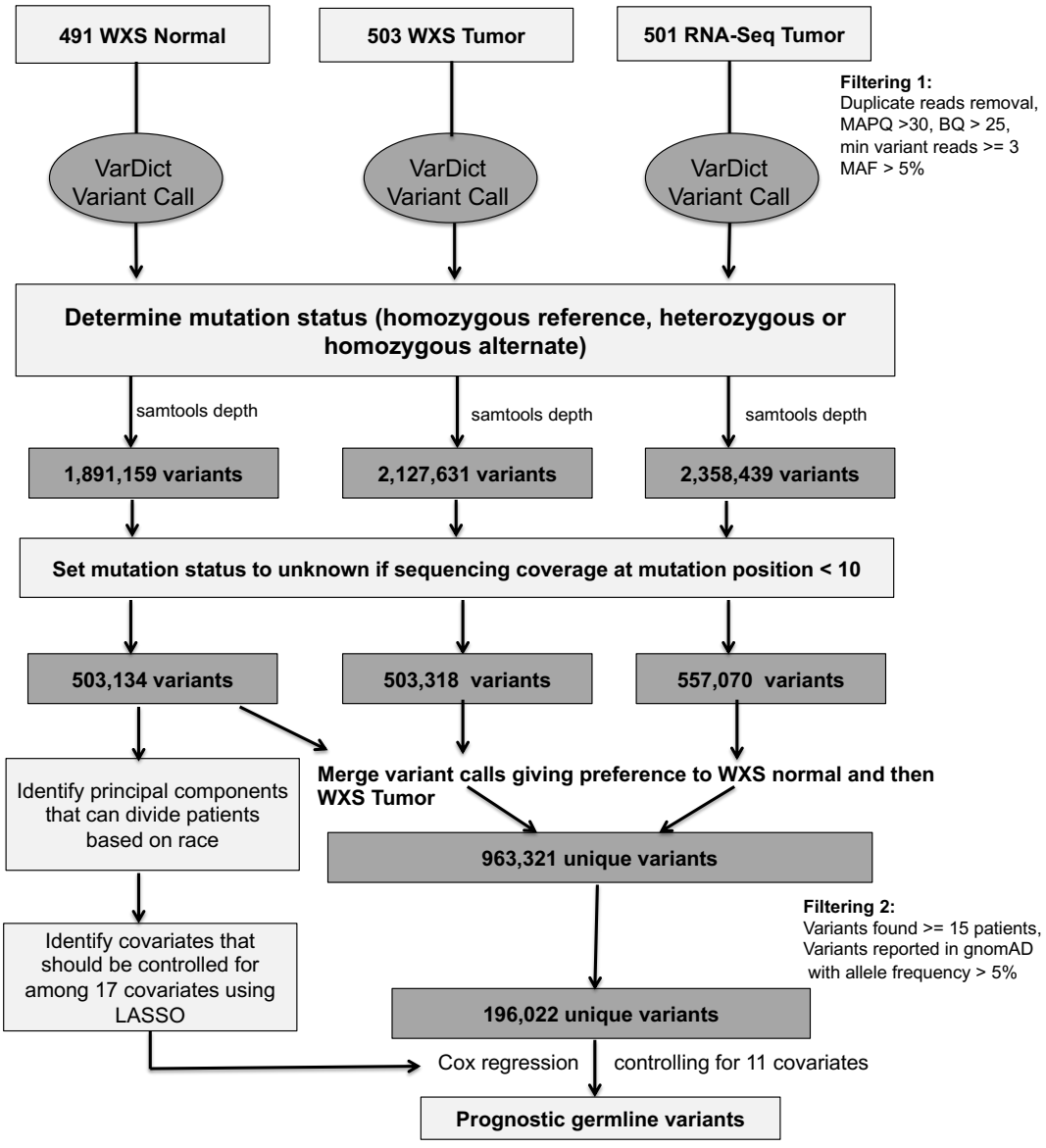


Figure 2. Prognostic germline variants in the TCGA dataset.
(A) Of the 4.4 million unique variants called in the TCGA dataset, we ran Cox regression on the 196,022 germline variants found in gnomAD with an allele frequency greater than 5% and found in 15 or more of the TCGA lower grade glioma patients.
(B-E) Similar to the 196,022 germline variants, the 271 prognostic variants are mostly found in **(B)** protein-coding genes, **(C)** are located in introns, and are **(D)** single nucleotide polymorphisms (SNP). **(E)** Most single nucleotide polymorphisms cause transitions.

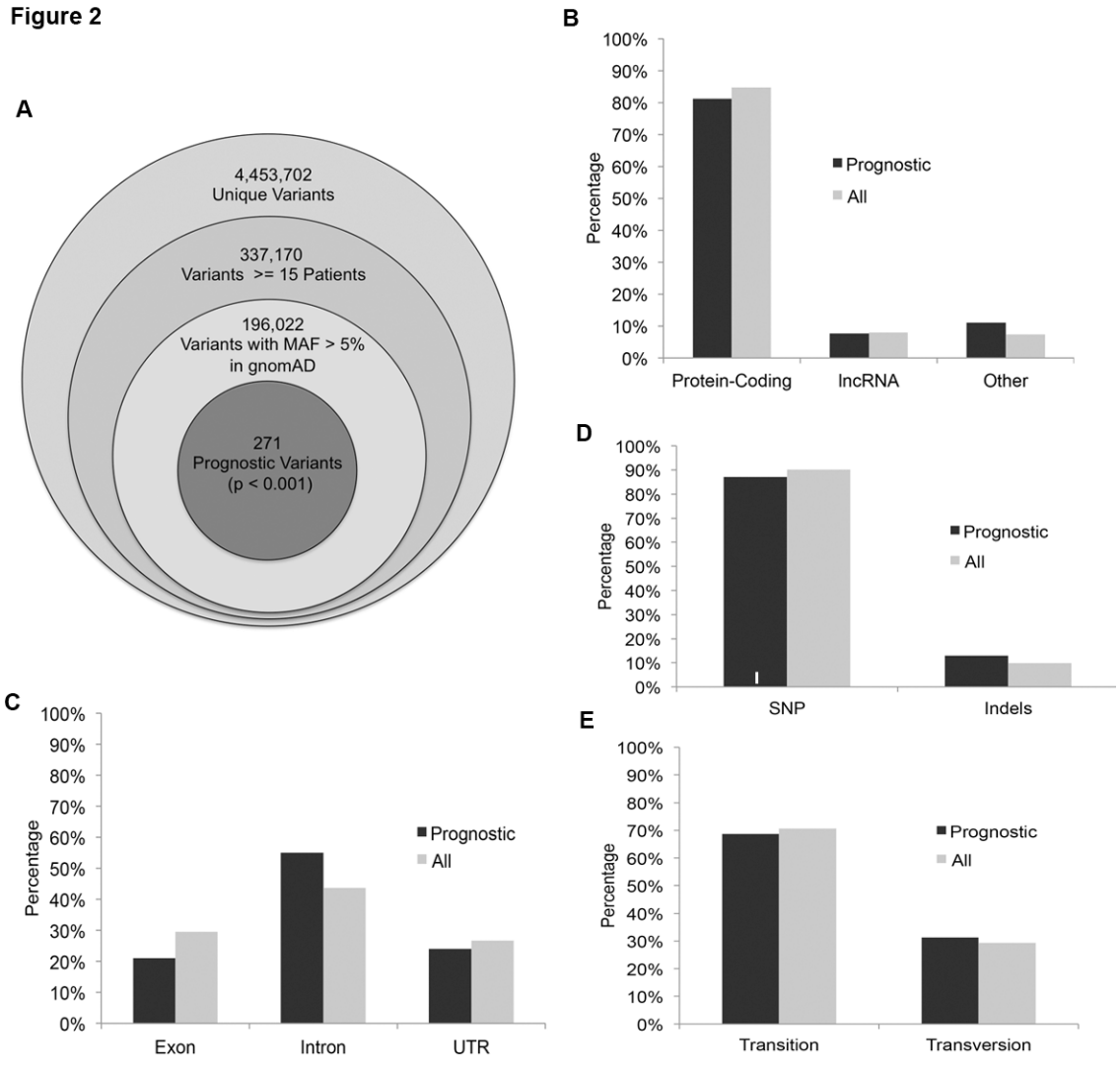


Figure 3. rs61757955 is a highly prognostic germline variant identified in the TCGA dataset.
(A) Manhattan plot showing the p-values resulting from testing each germline variant by Cox regression, controlling for the 11 variables in bolded in Table 1.Two variants passed the FDR threshold in the TCGA dataset.
(B) A Kaplan-Meier plot depicting the deleterious outcome associated with rs61757955, adjusting for the eleven covariates.
(C) Receiver operator characteristic curve at 7 years. rs61757955 increases the area under the curve compared to the 11 covariates alone, suggesting that it improves the clinical model.
(D) Separation of patients on the basis of whether or not they have this germline variant to determine which genes are induced or repressed in patients with rs61757955. Subsequent gene set enrichment analysis reveals that patients with this germline variant exhibit upregulation of the genes involved with *KRAS* signaling.

Figure 3

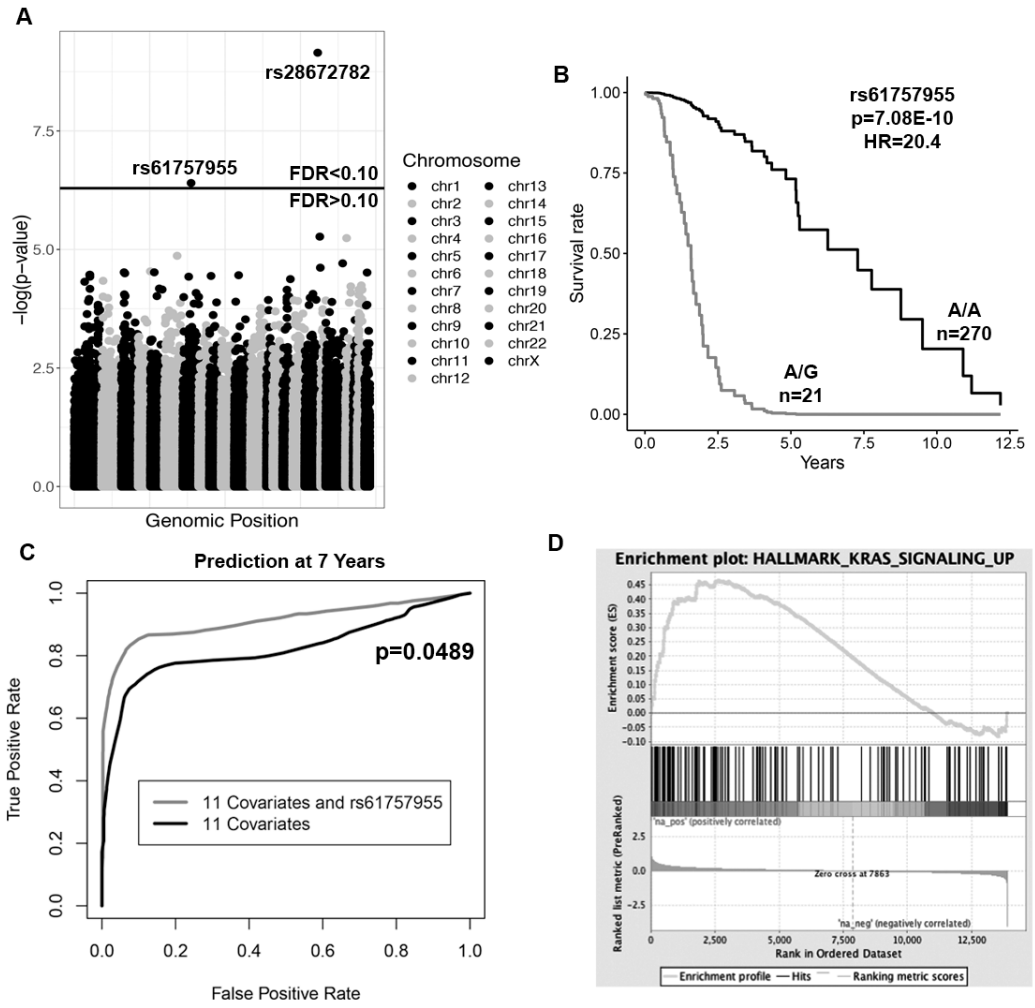
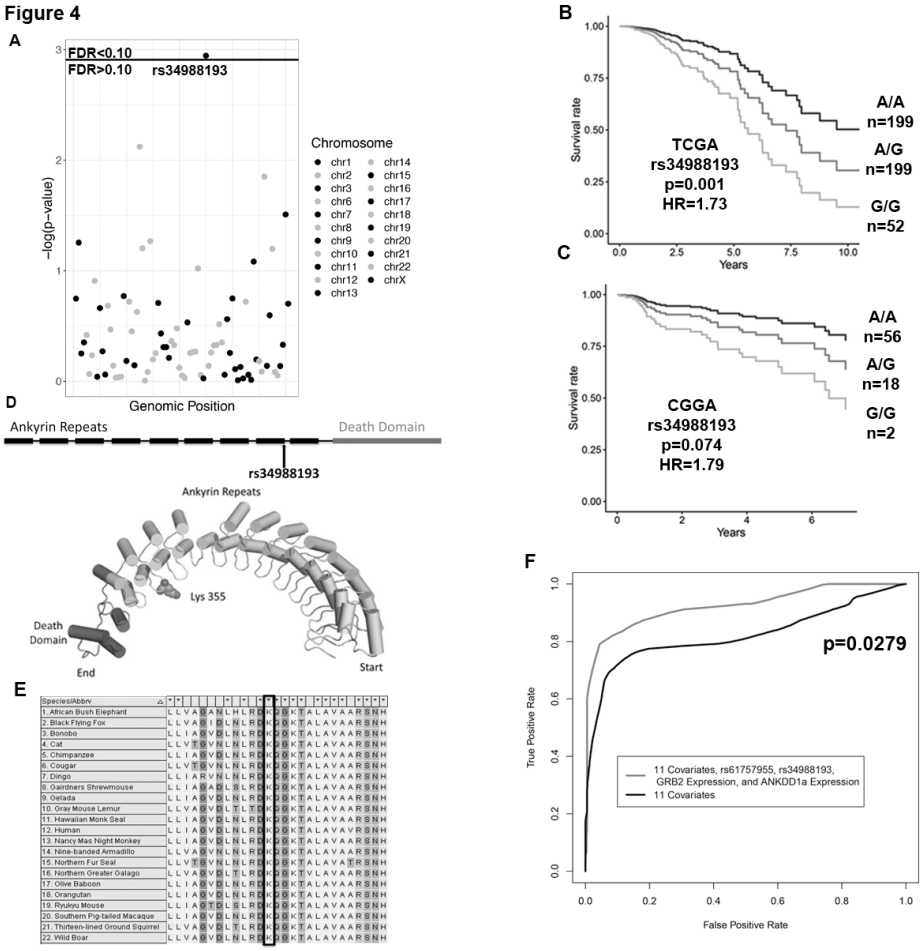


Figure 4. rs34988193 is a prognostic variant predicted to be highly deleterious.
(A) A Manhattan plot with the p-values resulting from testing each germline variant by Cox regression, controlling for the eleven covariates in bolded in Table 1. rs34988193 is prognostic (FDR<0.10) in the TCGA when restricting the analysis to the top 0.1% most deleterious variants by combined annotation dependent depletion (CADD).
(B-C) Kaplan-Meier plots depicting the deleterious outcome associated with rs34988193 in the **(C)** TCGA and **(D)** CGGA datasets, adjusted for the eleven covariates.
(D) A schematic showing that this variant is located in the ninth ankyrin repeat of *ANKDD1a*. The predicted protein structure of *ANKDD1a* reveals that this variant leads to an amino acid change from lysine to glutamate on the loop of an ankyrin repeat.
(E) Multiple sequence alignment of *ANKDD1a* in 22 species showing that lysine is conserved at this position in all of the species with this protein.
(F) Receiver operator characteristic curves comparing the ability of two survival models to label patients as alive or dead after seven years of follow up. The inclusion of rs61757955 variant status, rs34988193 variant status, GRB2 expression and ANKDD1a expression significantly improves the survival prediction compared to the eleven covariates bolded in table 1 alone (p=0.0279).



Tables

Table 1. List of variables that are known to be associated with differences in survival in lower grade glioma patients. 11 variables (bolded) were selected by Lasso for inclusion in the survival model. We used these 11 variables as covariates in our Cox regression model when testing each germline variant.

Covariate		**Median (Min-Max) or Number of Patients**
Age		41 (14 - 87)
Gender	Female	250
	Male	200
Somatic Mutation Count		50 (0 - 12255)
Percent Aneuploidy		11% (5.2E-4% - 95%)
log(TERT Expression)		1.0 (0.0 - 9.1) FPKM
Principle Component 1 (PC1)		0.043 (-0.091 - 0.064)
Principle Component 2 (PC2)		-0.017 (-0.23 - 0.17)
Principle Component 3 (PC3)		-0.53 (1.34E-4 - 0.33)
Histological Type	**Astroctyoma**	172
	Oligoastrocytoma	113
	Oligodendroglioma	165
Tumor Location	Frontal Lobe	265
	Temporal Lobe	125
	Parietal Lobe	42
	Other	18
Grade	G2	212
	G3	237
	Cannot Be Assessed	1
Treatment Site	**Henry Fords Hospital**	82
	Case Western St. Joes	90
	Other	278

IDH Mutant	Wild Type	83
	Mutant	367
1p/19q Co-deletion	Absent	303
	Present	147
MGMT Promoter Methylation	Unmethylated	80
	Methylated	370
Chr 7 gain/Chr 10 loss	Absent	399
	Present	51
TP53 Mutant	Absent	232
	Present	218

Table 2. Description of the prognostic germline variants identified in this study.
(A) A description of the two prognostic germline variants (FDR<0.10) in the TCGA dataset identified when testing all 196,022 germline variants.
(B) A description of the prognostic germline variant (FDR<0.10) rs34988193 in *ANKDD1a* identified when the analysis was restricted to only germline variants with a combined annotation dependent depletion (CADD) score greater than 30 in the TCGA and CGGA datasets.

Table 2A.

Variant	Chrom	Pos	Ref	Alt	Population Frequency	Sample Size	Number of Heterozygotes	Number of Homozygotes	Gene Name	Median Expression (FPKM)	p-value	Hazard Ratio
rs61757955	17	75318086	A	G	5.01%	291	21	0	GRB2	42.2	7.08E-10	20.4
rs28672782	11	1446622	C	T	16.27%	50	15	5	BRSK2	7.29	<1E-16	1.15E-10

Table 2B.

Variant	Chrom	Pos	Ref	Alt	Population Frequency	CADD Score	PhyloP Score	Gene Name	Dataset	Sample Size	Number of Heterozygotes	Number of Homozygotes	p-value	Hazard Ratio
rs34988193	15	64943580	A	G	30.90%	32	8.42	ANKDD1a	TCGA	450	199	52	0.00113	1.73
									CGGA	76	18	2	0.0743	1.79

Table 3. The association between the germline variant rs61757955 and genomic and histological variables. Patients were divided based on whether or not they had the germline variant rs61757955. Patients with the germline variant rs61757955 were more likely to have *CIC* mutated gliomas and the 1p/19q co-deletion.

Variable	Mean or Percentage (Wild Type)	Mean or Percentage (Mutant)	p-value
CIC Mutated	**15.9%**	**38.1%**	**0.017**
1p/19q Co-deletion	**25.2%**	**47.6%**	**0.038**
Oligodendroglioma	33.7%	42.9%	0.475
Total Somatic Mutation Count	30.9	30.0	0.766
Percent Aneuploidy	15.1%	11.7%	0.524
Astrocytoma	38.1%	42.9%	0.651
Grade 3	53.0%	42.9%	0.497
IDH Mutated	78.1%	85.7%	0.583
1p/19q Co-deletion	25.2%	47.6%	0.038
MGMT Promoter Methylation	77.8%	81.0%	1.000
Chr 7 Gain/Chr 10 Loss	13.0%	9.5%	1.000
Expression of GRB2 (FPKM)	45.7	44.4	0.636

Supplementary Figures

Figure S1. A boxplot representing the percentage of variants called in the whole exome sequenced (WXS) tumor sample that is likely somatic mutations. This value was calculated by counting the number of somatic mutations called in each patient by The Cancer Genome Atlas (TCGA) Research Network and dividing that number by the number of variants called in the WXS normal sample. Even before filtering, most variants called in the WXS tumor sample are germline variants.

Figure S1

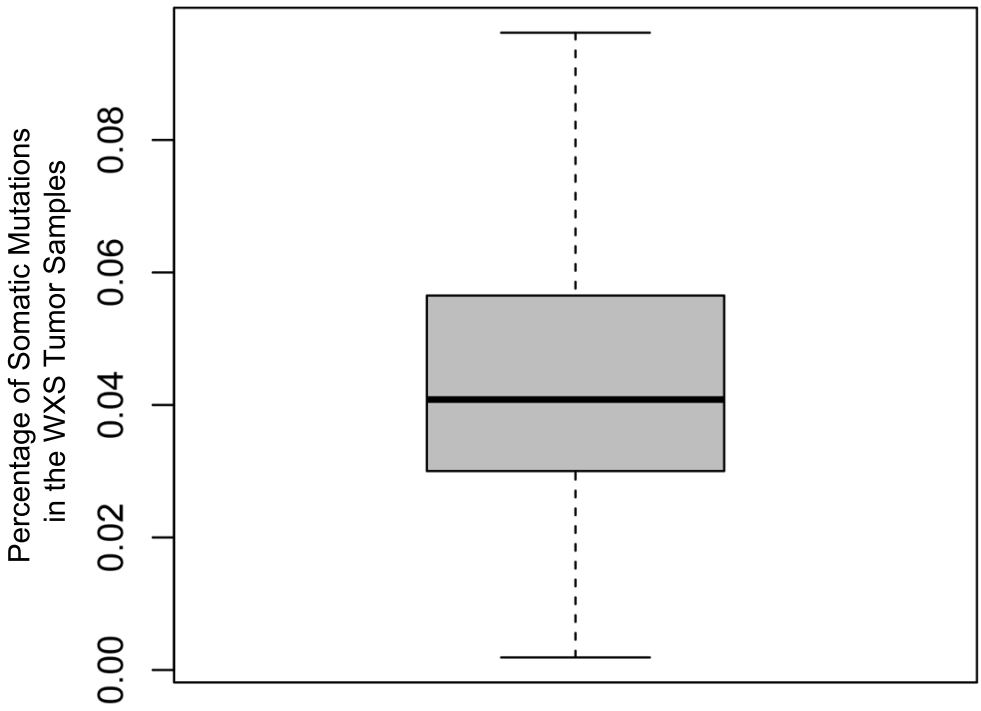


Figure S2. Correlation between the variant allele frequencies calculated from the four variant sets and the distribution of allele frequencies. The Pearson correlation panels on the top right (red) indicate that the calculated allele frequencies of the four variant sets are well-correlated with each other. This is depicted graphically in the bottom left panels with scatterplots. The distribution of allele frequencies is plotted along the diagonal. As has been shown in other studies, the distribution is negatively skewed because most minor alleles in the human population are rare.

Figure S2

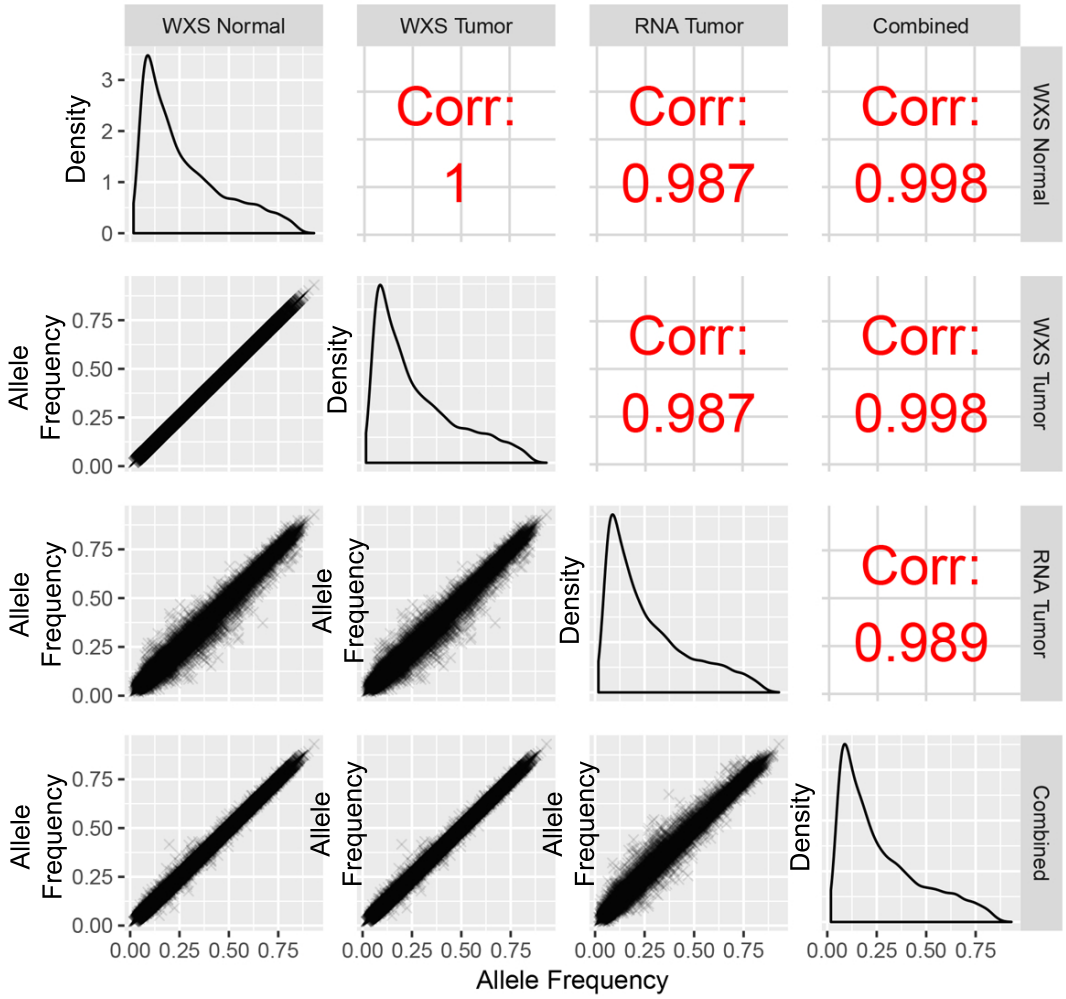


Figure S3. Principal components calculated from germline variants from the whole exome sequencing data from the non-tumor samples. Our principal components effectively stratify patients on the basis of patient-reported race.

Figure S3

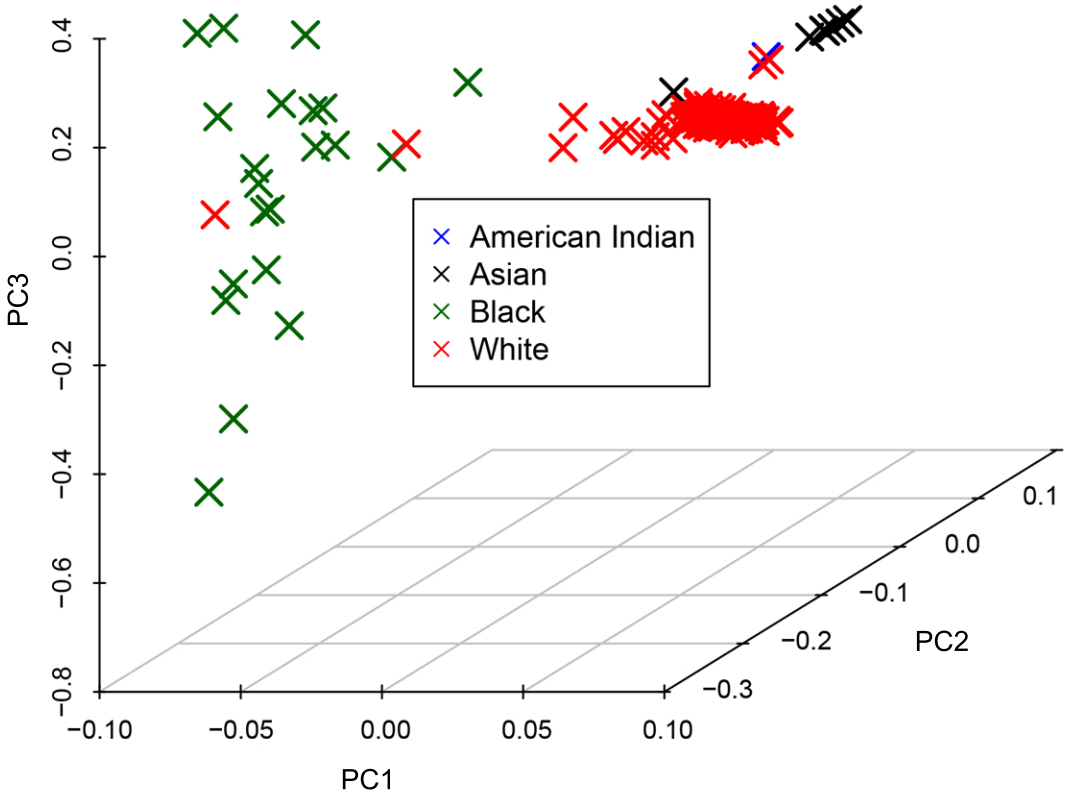
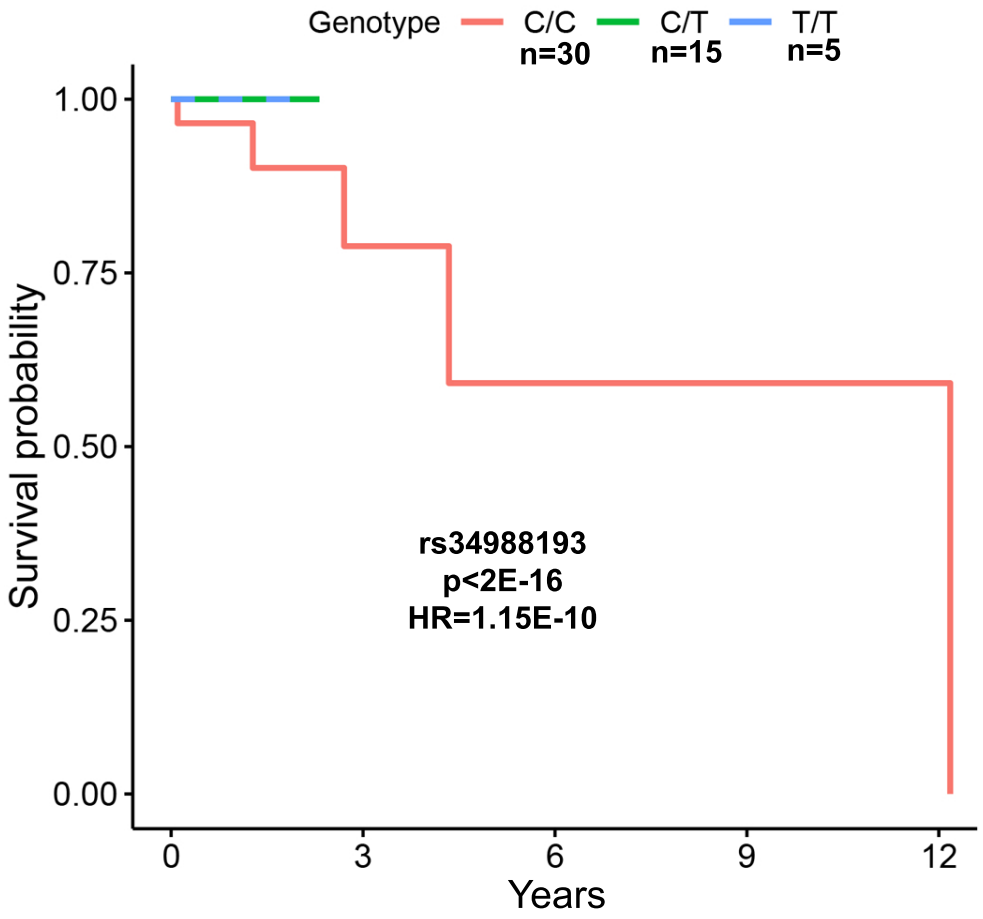


Figure S4. Kaplan-Meier plot for the germline variant rs28672782 in BRSK2. Although we found this germline variant to be prognostic (FDR<0.10), we decided not to further investigate this germline variant due to only 50 patients having sufficient sequencing depth at this position and due to the maximum follow up of patients with this germline variant being only three years.

Figure S4



Supplementary Tables

Table S1. Quality control checks reveal that somatic mutations and RNA editing did not affect the results of our analysis. Less than 0.1% of variants called in the whole exome sequenced (WXS) tumor sample are somatic mutations in the TCGA lower grade glioma patients. After our filtering steps, only one somatic mutation of the 196,022 variants tested in a single patient persists as part of our analysis. Of the 196,022 variants that we tested, only 0.1% of these variants overlap with a known RNA editing site. We did not find any of these variants as prognostic in our analysis, implying that somatic mutations and RNA editing did not compromise the quality of our analysis.

Quality Check	Percentage
Percentage of Somatic Mutations Called in the WXS Tumor Sample	< 0.1%
Percentage of Somatic Mutations Included in this Analysis After Filtering	< 0.001%
Percentage of Germline Variants Included in this Analysis that Overlap with a Known RNA Editing Site	0.10%

Table S2. Correlation between the allele frequencies calculated in our four variant sets and the allele frequencies reported by gnomAD. Our calculated allele frequencies are well-correlated with the allele frequencies reported by gnomAD, suggesting that our variant calls are high quality.

Variant Set	Pearson Correlation Coefficient with gnomAD
WXS Normal	0.963
WXS Tumor	0.964
RNA Tumor	0.937
Combined	0.947

Table S3. Variants genetically linked to rs61757955 in the European population, the population that is most similar to the TCGA lower grade glioma patient population. The upregulation of KRAS signaling may be due to rs61757955 or to one of these four linked variants.

Variant	Chromosome	Position	Distance from rs61757955 (bp)	Correlation (r^2)	Location
rs56298430	17	75341627	23541	1	Intron of GRB2
rs41282071	17	75320799	2713	0.936	Intron of GRB2
rs55771008	17	75342476	24390	0.936	Intron of GRB2
rs72850335	17	75298213	19873	0.879	18,863 Base Pairs Upstream of GRB2

Table S4. Results from testing for an association between the germline variant rs34988193 and genomic and histological variables. We did not find the germline variant rs34988193 to be associated with any changes in genomic or histological variables by separating patients on the basis of whether or not they had the germline variant rs34988193.

Variable	Mean or Percentage (Wild Type)	Mean or Percentage (Mutant)	p-value
Total Somatic Mutation Count	31.5	76.9	0.436
Percent Aneuploidy	13.4%	15.1%	0.466
Astrocytoma	41.7%	35.5%	0.204
Grade 3	53.8%	51.8%	0.704
IDH Mutated	82.9%	80.5%	0.542
1p/19q Co-deletion	28.6%	35.9%	0.107
MGMT Promoter Methylation	82.4%	82.1%	1.000
Chr 7 Gain/Chr 10 Loss	10.6%	12.0%	0.657
Expression of ANKDD1a	2.53	2.6	0.473

Chapter 3: The Pan-Cancer Landscape of Prognostic Germline Variants in 10,582 Patients

Ajay Chatrath, Roza Przanowska, Shashi Kiran, Zhangli Su, Shekhar Saha, Briana Wilson, Takaaki Tsunematsu, Ji-Hye Ahn, Kyung Yong Lee, Teressa Paulsen, Ewelina Sobierajska, Manjari Kiran, Xiwei Tang, Tianxi Li, Pankaj Kumar, Aakrosh Ratan, and Anindya Dutta

Adapted From:

Chatrath A, Przanowska R, Kiran S, Su Z, Saha S, Wilson B, Tsunematsu T, Ahn J, Lee K, Paulsen T, Sobierajska E, Kiran M, Tang X, Li T, Kumar P, Ratan A, Dutta A: **The Pan-Cancer Landscape of Prognostic Germline Variants**. *Genome Medicine* 2020.

- I conceived of the idea and design for this project, wrote the code for this analysis, analyzed the data, wrote the original draft of the manuscript, and made most of the figures in this manuscript. Some of the computational funding for this project was derived from the Cancer Genomics Cloud proposal for which I was the Project Leader.

Author Contributions

Conceptualization, A.C., A.D.; Methodology, A.C., A.R., A.D., X.T., T.L., P.K., M.K.; Software, Formal Analysis, Investigation, Writing – Original Draft, and Visualization, A.C.; Resources and Funding Acquisition, A.D., A.C.; Data Curation, A.C., R.P., Z.S., S.S., S.K., B.W., T.T., K.L., J.H.A., T.P., E.S., M.K.; Writing – Review & Editing, all authors; Supervision and Administration, A.D., A.R., P.K.

The Pan-Cancer Landscape of Prognostic Germline Variants in 10,582 Patients

Ajay Chatrath[1], Roza Przanowska[1], Shashi Kiran[1], Zhangli Su[1], Shekhar Saha[1], Briana Wilson[1], Takaaki Tsunematsu[1], Ji-Hye Ahn[1], Kyung Yong Lee[1], Teressa Paulsen[1], Ewelina Sobierajska[1], Manjari Kiran[2], Xiwei Tang[3], Tianxi Li[3], Pankaj Kumar[1], Aakrosh Ratan[4], Anindya Dutta[1]

[1] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia

[2] Department of Systems and Computational Biology, School of Life Sciences, University of Hyderabad, Telangana, India

[3] Department of Statistics, University of Virginia, Charlottesville Virginia

[4] Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia

Financial Support:
This work was supported by grants from the NIH R01 CA166054 and CA60499 and a Cancer Genomics Cloud Collaborative Support grant. The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

Potential Conflicts of Interest:
The authors do not have any conflicts of interest to disclose.

Abstract

Background: While clinical factors such as age, grade, stage, and histological subtype provide physicians with information about patient prognosis, genomic data can further improve these predictions. Previous studies have shown that germline variants in known cancer driver genes are predictive of patient outcome but no study has systematically analyzed multiple cancers in an unbiased way to identify genetic loci that can improve patient outcome predictions made using clinical factors.

Methods: We analyzed sequencing data from the over 10,000 cancer patients available through The Cancer Genome Atlas to identify germline variants associated with patient outcome using multivariate Cox regression models.

Results: We identified 79 prognostic germline variants in individual cancers and 112 prognostic germline variants in groups of cancers. The germline variants identified in individual cancers provide additional predictive power about patient outcomes beyond clinical information currently in use and may therefore augment clinical decisions based on expected tumor aggressiveness. Molecularly, at least twelve of the germline variants are likely associated with patient outcome through perturbation of protein structure and at least five through association with gene expression differences. Almost half of these germline variants are in previously reported tumor suppressors, oncogenes or cancer-driver genes with the other half pointing to genomic loci that should be further investigated for their roles in cancers.

Conclusions: Germline variants are predictive of outcome in cancer patients and specific germline variants can improve patient outcome predictions beyond predictions made using clinical factors alone. The germline variants also implicate new means by which known oncogenes, tumor suppressor genes, and driver genes are perturbed in cancer and suggest roles in cancer for other genes that have not been extensively studied in oncology. Further studies in other cancer cohorts are necessary to confirm that germline variation is associated with outcome in cancer patients as this is a proof-of-principle study.

Background

Large-scale sequencing projects increased our molecular understanding of cancers to the point where using sequencing data to augment clinical decisions seems promising [1, 2]. Somatic mutations in cancers have received substantial attention in oncology as they can be used to individualize drug selection [2, 3]. While much effort has been directed towards characterizing somatic mutations in cancer, recent studies suggest that germline variants also have significant clinical utility.

In line with the heritability of some cancers, several germline variants predict a patient's risk for developing cancer and are useful for individualizing cancer screening guidelines [4-13]. Germline variation can affect drug sensitivity, predict drug toxicity, and could help select therapy to minimize side-effects [14-26]. Some germline variants increase patient risk for specific somatic aberrations, suggesting that germline variation may impact disease course [27].

We hypothesized that the effects of germline variants on cancer progression may be strong enough to identify associations with patient outcome. Previous studies tested for an association between patient outcome and a small number of germline variants in genes well-characterized in a given cancer [28, 29]. We published an unbiased method of testing for an association between a large number of germline variants and patient outcome in patients with lower grade gliomas [30]. In this study, we identify prognostic germline variants using sequencing data from 10,582 patients from The Cancer Genome Atlas (TCGA). These germline variants significantly improve predictions of patient outcome

compared to clinical variables alone, identify biological mechanisms by which germline variants affect patient outcomes, and identify genes and pathways that impact cancer biology and therapy.

Methods

Data Sources, Variant Calling, and Quality Control

The results in this manuscript are based upon data generated by The Cancer Genome Atlas (TCGA) Research Network: https://www.cancer.gov/tcga. We determined the germline variant statuses of 10,582 cancer patients by variant calling the patients' whole exome sequenced normal samples (WXS normal), whole exome sequenced tumor samples (WXS tumor), and RNA sequenced tumor samples (RNA tumor) available on Cancer Genomics Cloud using VarDict (mapping quality > 30, base quality > 25, variant reads > 2, minimum allele frequency > 5%, no duplicate reads) and determined the sequencing depth at each position using samtools (mapping quality > 30) [31-33]. We set variant calls to unknown if the position at which the variant was called was covered by fewer than 10 reads. We then merged these three variant call sets, giving preference to WXS normal then WXS tumor then RNA tumor. We only included variants with an allele frequency of greater than 5% in the non-Finnish European population of gnomAD, variants found in more than 14 patients in a given cancer, and variants whose calls were greater than 90% concordant with each other in a given cancer in our final analysis [34]. These thresholds had been selected in our previous study in order to better tune the allele frequencies of the European patients in our study to previously reported population frequencies [30]. Our quality control tests

for setting these thresholds yielded similar results across the other cancers outside of the lower grade gliomas. We labeled variant calls as concordant for a given variant if they gave the exact same variant call (homozygous for the reference allele, heterozygous, or homozygous for the alternate allele) in the WXS normal, WXS tumor, and RNA tumor samples. Variant calls were therefore discordant if the variant call differed in any of the three samples. The percentage concordance was calculated for each germline variant by dividing the total number of concordant variant calls by the total number of patients and multiplying the result by 100%.

We retrieved clinical outcomes data for each patient using the TCGA Pan-Cancer clinical data resource [35]. We used TCGAbiolinks to obtain patient clinical information and we downloaded patient race composition from The Cancer Genome Ancestry Atlas (TCGAA) [36, 37]. Additional clinical information for the lower grade glioma and glioblastoma patients was downloaded from a previous analysis [38]. We used Lasso-regularization to determine which clinical covariates should be controlled for in our models, while using patient race composition from TCGAA in place of patient-reported race [39, 40]. The patient race composition reported in the TCGAA more accurately captured the genetic ancestry of the TCGA patients compared to patient reported race as patient race composition is quantitative and multidimensional. Where we did not control for patient race composition in cancers where patient race composition was not identified as a significant predictor of patient outcome by Lasso-regularized Cox regression, we later retested the set of prognostic germline variants by adding

back patient race composition as a covariate into our Cox regression models. As expected, because patient race composition was not a significant predictor of patient outcome in these cancers, we still found all of our originally identified prognostic germline variants to be statistically significant predictors of patient outcome. We also found that the hazard ratios estimated in the original models (without race) with the retested models (with race) were highly correlated (Spearman rho=0.983, p=7.63E-47).

We were not able to control for treatment. As discussed in greater detail by Liu et al., it is very difficult to control for treatment in the TCGA dataset [35]. Detailed treatment information was not submitted in a consistent manner for many of the patients in TCGA and absence of submitted treatment information does not necessarily mean that the patient did not receive treatment. Furthermore, treatment regimens are quite complex and depend on chemotherapy drug selection and dosage, extent of surgical excision, and radiation therapy, among other factors. The broad spectrum of treatment options makes treatment challenging to control for. As discussed by Liu et al., the TCGA treatment information will likely need to be evaluated by panels of cancer specialists before it can be used for modeling in pan-cancer studies [35]. Nevertheless, it is unlikely that differences in treatment accounted for the bulk of the associations observed in this study. The most natural way for treatment differences to account for the observation that germline variation is associated with patient outcome is due to socioeconomic differences associated with patient race or unconscious or conscious biases in treatment selection based on patient

race. However, we accounted for calculated genetic ancestry as part of our pipeline, making these possibilities unlikely.

We determined the number of somatic mutations in the cancer samples and evaluated the overlap between germline variants and somatic mutations and RNA editing sites as previously described [30]. To ensure that our variant calls from the four variant call sets (WXS normal, WXS tumor, RNA tumor, and Combined) were concordant with each other, we calculated the allele frequency of each variant as in our previous analysis and calculated the Spearman correlation coefficient of these allele frequencies with each other.

Power Analysis

We performed a power analysis in individual cancers to evaluate our ability to detect associations between germline variants and patient outcome using Cox regression. The power to detect an association between a germline variant and patient outcome is dependent on the sample size, effect size, correlation with other covariates in the model, the number of individuals with the germline variants, and the number of individuals without a germline variant, among other factors. As a result, the power to detect an association differs between germline variants, even assuming the same hazard ratio. To estimate our power, we therefore randomly sampled 10,000 germline for each cancer from the pool of germline variants to be tested in that cancer. We calculated statistical power using the powerSurvEpi R package (https://cran.r-project.org/web/packages/powerSurvEpi/index.html). We calculated our power to detect a significant association at a significance level (α) of:

$$\frac{0.10}{\text{Total Number of Germline Variants Tested in That Cancer}}$$

This threshold would be as stringent or slightly more stringent than false discovery correction using the Benjamini-Hochberg procedure which we ultimately used in our analysis. We then calculated the percentage of germline variants for which we had greater than 80% statistical power to detect a significant association at hazard ratios of 2, 3, 4, 5, 10, 15, and 20.

Identification of Prognostic Germline Variants

We utilized six total approaches for identifying prognostic germline variants. In all analyses, we tested variants for an association with outcome using a Cox regression model, controlling for the covariates that we identified previously for each cancer using Lasso-regularization. We used the R packages survminer (https://cran.r-project.org/web/packages/survminer/index.html) and survival (https://cran.r-project.org/web/packages/survival/index.html) to perform Cox regression and generate Kaplan-Meier plots. p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure. The circos plots were generated using the R package circlize [41].

In analysis 1, we tested variants for an association with patient outcome in individual cancers, setting an adjusted p-value threshold (FDR) less than 0.10. We reported all statistically significant results and did not filter our results based on a hazard ratio threshold, as it is difficult to know what hazard ratio threshold would be clinically and biologically relevant. In the second analysis, we filtered our results from analysis one to identify germline variants that were recurrently associated ($p<0.05$) with favorable (hazard ratio (HR)<1) or poor (HR>1)

outcome relative to the reference allele in seven or more cancers, such that the most recurrent prognostic variants would be reported. Given that molecular similarities between some of the TCGA cancers may have made it more likely that certain germline variants would be picked up in this second analysis than others, we did not think that it would be statistically valid to estimate the probability of variants being pulled out by this analysis by chance. In the third analysis, we grouped the cancers based on clinical understanding about the cancers and clustering patterns observed previously by the TCGA research network [42]. We tested germline variants for associations with patient outcome (FDR<0.10) in these larger groups to detect germline variants with smaller effect sizes. In pooling cancers, we implicitly assumed that the germline variant had similar effects in the grouped cancers. If this assumption was not true for a particular germline variant, then that germline variant would actually be less likely to be associated with patient outcome. Only variants found in 15 or more patients across all grouped cancers were tested, resulting in fewer variants being tested in this analysis.

Analyses 4-6 were quite similar to analyses one through three, except that we restricted our analysis to only germline variants that caused significant amino acid changes with a combined annotation dependent depletion (CADD) score greater than 25 [43]. This enabled us to identify associations that we did not capture in analyses one through three due to the relatively higher stringency in that analysis resulting from multiple hypothesis correction. In analysis four, we tested variants with CADD score > 25 in individual cancers for an association

with patient outcome (FDR<0.10). In analysis five, we filtered the results from analysis four to identify germline variants with CADD score > 25 that were recurrently associated (p<0.05) with favorable (HR<1) or poor (HR>1) prognosis in 5 or more patients. In analysis six, we tested germline variants with CADD > 25 for a significant association (FDR<0.10) with patient outcome in the previously described patient groups.

The Cox regression models that we fit for individual cancers controlled for the covariates that we found to be prognostic in those cancers (**Table S1**). The Cox regression models that we fit for patient groups controlled for the covariates that we found to be prognostic in individual cancers with each term containing an interaction term associating that variable with the cancer that it was associated with patient outcome in. We also controlled for cancer type in these combined groups. As an example, suppose that variable A is associated with patient outcome in cancer X and variable B is associated with patient outcome in cancer Y. Then we would fit two Cox regression models to identify prognostic germline variants in individual cancers and a third Cox regression model to identify germline variants prognostic in the pooled cohort, as illustrated below.

(1) Identifying germline variants associated with patient outcome in cancer X

$$\text{Patient Outcome} \sim \beta_0 + \beta_1(\text{Variable A}) + \beta_2(\text{Germline Variant Status})$$

(2) Identifying germline variants associated with patient outcome in cancer Y

$$\text{Patient Outcome} \sim \beta_0 + \beta_1(\text{Variable B}) + \beta_2(\text{Germline Variant Status})$$

(3) Identifying germline variants associated with patient outcome when the patients with cancer X and the patients with cancer Y are pooled together

$$\text{Patient Outcome} \sim \beta_0 + \beta_1(\text{Cancer X Status}) + \beta_2(\text{Cancer X Status})(\text{Variable A}) + \beta_3(\text{Cancer Y Status})(\text{Variable B}) + \beta_4(\text{Germline Variant Status})$$

In model (3) above, cancer X status is a dummy variable that can be 0 or 1. The value of this variable is 0 for patients with cancer Y and 1 for patients with cancer X. The opposite is true for the cancer Y status variable. This allowed us to group patients to test for an association with patient outcome, while controlling for differences between different cancers and relevant clinical differences between patients with the same cancer.

Concordance and Correlation of Hazard Ratios for the Prognostic Germline Variants

We tested whether germline variants associated with patient outcome ($p<0.05$) in three of more cancers were typically recurrently associated with increased risk of poor outcome or recurrently associated with decreased risk of poor outcome more often than would be expected by random chance and if the hazard ratios estimated for these prognostic germline variants in different cancers were correlated with each other.

To test for concordance, we first counted the number of times that germline variant was found to be associated ($p<0.05$) with poor patient outcome ($HR<1$) or favorable patient outcome ($HR>1$). We then calculated the following value for each prognostic germline variant:

$$\frac{\max(Poor\ Outcome, Favorable\ Outcome)}{Poor\ Outcome + Favorable\ Outcome}$$

where poor outcome is the number of times that the germline variant was associated with poor outcome (HR<1) and favorable outcome is the number of times that the germline variant was associated with favorable outcome (HR>1). If a germline variant was perfectly concordant, then the calculated value would be 1. While theoretically the expected value would be 0.5 for a random germline variant, we empirically estimated the expected value by the following calculation:

$$\frac{\max(Total\ Number\ of\ Poor\ Outcome, Total\ Number\ of\ Favorable\ Outcome)}{Total\ Number\ of\ Poor\ Outcome + Total\ Number\ of\ Favorable\ Outcome}$$

In this set of prognostic variants, there were more variants associated with poor patient outcome (HR<1) than favorable patient outcome (HR>1), resulting in the expected index being 0.589. We then used a Wilcoxon rank sum test to determine whether the concordance values that we calculated from the set of prognostic germline variants differed from what we would expect by random chance.

We next tested whether the hazard ratios estimated for a given prognostic germline variant in different cancers were correlated with each other. Because we had previously found the hazard ratios to be concordant, we performed this analysis separately for instances in which a germline variant was found to be associated with increased risk of poor outcome and decreased risk of poor outcome. We identified the set of variants associated with favorable (HR<1) outcome and poor (HR>1) outcome in three or more cancers. The set of variants that were associated with favorable and poor outcome were analyzed separately.

For each analysis, we generated all possible pairs of hazard ratios for a given germline variant. We then ran a Spearman's correlation test to determine whether or not the hazard ratios were correlated to each other. Because the hazard ratio is also correlated to the allele frequency, we repeated the prior analysis with a Spearman partial correlation test to control for germline variant allele frequency. Partial correlation was calculated used the ppcor R package [44].

Characteristics of Prognostic Germline Variants

Having identified the prognostic germline variants, we then aimed to compare the characteristics of prognostic germline variants to the characteristics of germline variants identified in previous genome wide association studies [45]. We decided to use the variants from analysis one and analysis three to understand the characteristics of prognostic germline variants because the other approaches each identified a very small number of prognostic germline variants. We decided not to pool all of the germline variants together due to possible differences in characteristics between these sets of variants. We therefore analyzed the characteristics of the prognostic germline variants from analysis one and from analysis three separately. To avoid considering the same information multiple times, we removed variants that were linked with each other from the analyses in this section and only retained the first variant by genomic position. The actual variant retained did not have a significant effect on our results because the hazard ratios and sample sizes for the linked variants were very similar.

We first tested whether or not the minor allele was typically associated with poor patient outcomes. We sorted the variants into two categories: minor alleles that were associated with poor outcome in the Cox regression model (HR>1) and minor alleles that were associated with favorable outcomes (HR<1). Although the reference allele was often the major allele, this was not always the case. We performed a one-sided Fisher's exact test in R to determine whether or not the minor allele was more likely to be associated with poor outcome. The R package scatterpie (https://cran.r-project.org/web/packages/scatterpie/index.html) was used to display the proportion of homozygous reference, heterozygous, and homozygous alternate individuals. For variants in analysis three that were pulled out in multiple groups, we displayed the proportion of individuals only for the group that contained the largest number of individuals. The largest group always contained all individuals because the smaller groups were made up of smaller number of cancers and was always contained in the larger group. For example, suppose a variant was found to be prognostic in both group 20 (KICH, KIRP) and group 19 (KICH, KIRC, KIRP). In this case, we would perform all calculations using the information from group 19.

We next tested whether or not there was an inverse correlation between effect size and allele frequency. To do this, we calculated the Spearman correlation coefficient between effect size, calculated as $|\ln(\mathrm{HR}) - 0|$, and allele frequency. Finally, we identified the genomic regions (upstream of a gene, 5' UTR, exonic, intronic, 3' UTR, downstream of a gene, or intergenic) in which

each variant was located in using annovar [46]. Some variants were found in multiple different transcripts and therefore mapped to several different genic regions. For the purposes of creating the figures, we allowed a single variant to count once for multiple different regions. Excluding these variants from the figures did not change our interpretation of the results.

Testing Whether the Effects of the Prognostic Germline Variants are at Least Partially Independent

If the effects of the prognostic germline variants are at least partially independent of each other, we would expect that if two prognostic germline variants are found in the same patient that the outcome observed in those patients would be even more extreme than the outcome in patients with only a single germline variant. In other words, a patient with two prognostic germline variants associated increased risk for poor outcome should have a worse outcome than a patient with only one prognostic germline variant associated with poor outcome.

To test this hypothesis, we analyzed the set of prognostic variants identified in individual cancers. We set a few boundaries on our analysis to reduce bias.

(1) We identified prognostic germline variants highly linked to each other and only kept the first prognostic germline variant by chromosomal position in this set. The determination of which germline variant was selected did not substantially alter our results.

(2) We analyzed pairs of variants in individual cancers. Although we could evaluate multiple prognostic variants in each of the cancers, this would make

the analysis more complex, given the differing effect sizes of the prognostic germline variants.

(3) Because most of the prognostic variants in individual cancers were associated with increased risk for poor outcome, we limited this analysis to only variants associated with increased risk for poor outcome and excluded variants associated with favorable outcome.

(4) In the testing of each pair of prognostic germline variants, we excluded individuals who were homozygous for one of the prognostic germline variants. Our Kaplan-Meier plots suggest that for some of the prognostic germline variants, having two copies of the variant has a stronger effect than having a single copy so including homozygotes for the prognostic germline variants could confound our results. The homozygotes for the prognostic germline variant were relatively rare and so we could not test them separately. Since they were relatively rare, the exclusion of homozygotes for the prognostic germline variant did not dramatically reduce our sample size.

Having setup the conditions for this test, we created three groups for each pair of prognostic germline variants associated with poor patient outcome:

(1) Patients homozygous for the reference allele of both prognostic germline variants

(2) Patients heterozygous for one of the two prognostic germline variants and homozygous for the reference allele of the other prognostic germline variant

(3) Patients heterozygous for both of the prognostic germline variants

We then tested for differences in patient outcome between groups (2) and (1) and groups (3) and (1). If the effects of the prognostic germline variants are at least partially independent, we would expect the hazard ratio from the comparison of groups (3) and (1) to be greater than the hazard ratio from the comparison of groups (2) and (1). We calculated these hazard ratios for each pair of prognostic germline variants and ran a paired one-sided Wilcoxon signed-rank test to evaluate whether the hazard ratio from the comparison of groups (3) and (1) was greater than the hazard ratio from the comparison of groups (2).

Association of Prognostic Germline Variants with Somatic Driver Mutations

We tested whether the prognostic germline variants were more likely to be associated with somatic mutations in driver genes than would be expected by random chance. We retrieved the set of driver genes for each cancer and consensus somatic mutation calls for each cancer from TCGA Network analyses [2, 47]. For each cancer, we only considered driver genes with five or more patients with a somatic mutation in that driver gene in that cancer. For each prognostic germline variant, we tested whether the variant associated with increased risk of poor outcome was associated with an increased incidence of somatic mutations in each of the driver genes being considered for that cancer in patients with the allele associated with increased risk of poor outcome compared to patients with the protective allele using a one-sided Fisher's exact test. p-values were adjusted using the Benjamini-Hochberg procedure.

We were then able to determine the number of germline variants that were associated with a somatic mutation in a driver gene. We repeated this approach for all germline variants included in this analysis and performed a one-sided Fisher's exact test to determine whether or not more prognostic germline variants than expected were associated with a somatic mutation in a driver gene.

Area Under the Curve

To assess the clinical relevance of our findings, we tested whether the germline variants enhanced patient outcome predictions made using clinical information alone. While we had identified germline variants associated with outcome controlling for clinical covariates, we aimed to determine whether these variants significantly improved patient outcome predictions beyond predictions made using the clinical model alone, particularly in cancers in which the prediction by the clinical model was already quite accurate. We generated receiver operator characteristic (ROC) curves from the tenth percentile of patient death or patient progression to the ninetieth percentile of patient death or patient progression for each variant in R (https://cran.r-project.org/web/packages/survivalROC/survivalROC.pdf, https://cran.r-project.org/web/packages/timeROC/timeROC.pdf). We generated two ROC curves per variant: (1) the first was made using only patient clinical information (C) and (2) the second was generated using both patient clinical information and germline variant status (C+GV). We ran a one-sided Wilcoxon-rank sum test in R to determine whether the model supplemented with germline variant status consistently yielded better predictions across time for each variant. While our

Cox regression analysis identified variants that were significantly associated with patient outcome, these variants may not necessarily substantially improve clinical outcome predictions in cancers in which the clinical variables are already very good at predicting outcome. Running the one-sided Wilcoxon-rank sum test allowed us to test whether the improvement to the prediction was significant.

Gene Annotation and Literature Review

We annotated the variants resulting from our analysis using biomaRt [48, 49]. We reviewed the literature for the functions of these genes to understand their functions. Many of the authors (RP, SK, ZS, SS, BW, TT, JA, KL, TP, ES, MK) initially reviewed the literature for information about each gene. The literature review was then verified by three of the authors (RP, SK, ZS) to ensure consistency and validity.

Having generated a list of genes that the germline variants are associated with from biomaRt, we first specifically searched the literature to see if these genes had a function in cancer that had been characterized and that fit a category described by Weinberg and Hanahan [50]. This part of the literature review had the largest number of unknowns due to the large amount of specificity required by the studies. We then relaxed our stringency and checked to see whether or not the gene was associated with findings in the literature consistent with oncogenic or tumor suppressor activity in the context of cancer. The classification of the genes as oncogenes or tumor suppressors was based on published biochemical or molecular studies of the genes in the context of cancer. Multiple studies supported the classification as either an oncogene or tumor

suppressor gene for a substantial number of the genes. Finally, to understand in general whether or not these genes are being actively studied by the field, we categorized these genes based on whether or not the literature suggested that the genes are being studied in a cancer in which the germline variant was found to be prognostic, studied in any cancer, or studied in any human disease. We also overlapped our gene list with the list of driver genes generated by the TCGA research network [2].

Variant Mechanisms and Literature Review

We next aimed to understand the mechanisms by which the prognostic germline variants may be exerting their effects. We started with the germline variants that were predicted to cause significant amino acid changes (CADD>25). We determined the position and amino acid change caused by these germline variants using Ensembl [51]. We determined the domain in which these germline variants cause their amino acid changes using the National Center for Biotechnology Information databases (https://www.ncbi.nlm.nih.gov/) and the Ensembl and Uniprot databases [52]. We next identified germline variants that are likely acting as expression quantitative trait loci in *cis* (*cis* eQTLs). For each germline variant, we separated patients based on whether or not they had at least one non-reference allele and then determined whether or not there was a statistically significant difference between the mean expression of the gene associated with the variant between the two groups using a Wilcoxon rank sum test. We then combined our prediction as to whether the germline variant was protective or associated with increased risk of poor outcome with the expression

difference between the two groups to determine whether increased expression of the gene would be expected to be protective or associated with increased risk of poor outcome. We fit Cox regression models using the expression of each of the genes, controlling for clinical covariates, and compared the result to our prediction. We reported variants that are concordant with our predictions. Because the differential expression and Cox regression results had to both be concordant with each other, we used a more relaxed cut-off of $p < 0.10$ for hypothesis generation. Further studies with larger cohorts and more statistically power are necessary to further interrogate these associations. Finally, we checked to see whether the eQTL was also reported in GTEx in the tissue from which the tumor was derived by downloading the list of tissue-specific and pan-tissue eQTLs and comparing the eQTLs identified in our analysis to those reported in GTEx.

We reviewed the literature for previous associations tied to these variants reported in the literature. As was the case with gene annotation, the literature review was first done by multiple authors (RP, SK, ZS, SS, BW, TT, JA, KL, TP, ES, MK) with the final round of quality control and verification being done by a single author (BW).

Correlation with Drug Sensitivity

We found the germline variant rs1800932 in *MSH6* to be associated with favorable patient outcome and increased *MSH6* expression. Because a previous analysis found that *MSH6* knockdown resulted in increased temozolomide resistance, we tested whether *MSH6* expression was correlated with

temozolomide sensitivity in cancer cell lines [53]. To do this, we downloaded *MSH6* expression levels and temozolomide sensitivity for 915 cell lines using data from the Genomics of Drug Sensitivity in Cancer database through CellMinerCDB [54, 55]. We tested for an association using Spearman's correlation test.

Pathway Dysregulation

For selected prognostic germline variants described in the text, we tested whether or not these prognostic germline variants were associated with upregulation or downregulation of genes in specific pathways. For each prognostic germline variant, we separated patients into two groups based on whether or not the variant allele was called in those patients. We calculated the log fold change of each gene expressed greater than a median of 1 fragment per kilobase per million mapped reads and used these values as an input for gene set enrichment analysis [56].

Results

Identification of High Quality Germline Variants

Germline variants were called and filtered as shown **Figure S1** using sequencing data from 10,582 TCGA patients with 33 different types of cancers. In total, 77.6 million unique variants were called. After filtering, we limited our analysis to 519,319 unique variants (**Figure S2**). Because the final variant call set was created by merging variant calls from whole exome sequenced (WXS) normal tissue samples, WXS tumor samples, and RNA sequenced tumor samples, we evaluated our variant calls for contamination by somatic mutations

or RNA editing. Our final germline variant call set did not substantially overlap with somatic mutations or RNA editing sites (**Figures S3-4, Text S1**).

Determination of Prognostic Clinical Models for Each Cancer

To identify prognostic germline variants that provide additional outcome information not already captured by clinical variables, we created clinical models predictive of patient outcome for each cancer using the clinical information previously collected by the TCGA research network along with the components of calculated race from The Cancer Genome Ancestry Atlas. The variables selected for each cancer are summarized in **Table S1**. The study was powered to capture prognostic germline variants with moderate to high effect sizes (beginning at hazard ratios > 2) (**Figure S5, Text S2**).

Identification of Prognostic Germline Variants

The 191 prognostic germline variants from the six analyses are described in **Table S2A-F.**

The first three analyses identified germline variants associated with prognosis in (1) individual cancers, (2) multiple cancers giving roughly equal weight to each cancer, and (3) cancers grouped by organ system, histological, or molecular classifications (**Figure 1A**). Analysis 1 tested 519,139 variants for associations with patient outcome in individual cancers and identified 70 unique prognostic variants (**Figure 1B, Table S2A,** Kaplan Meier plots of selected examples in **Figure 2**).

While analysis 2 identified hundreds of variants recurrently predictive of outcome in >4 cancers, we will only discuss the 5 variants that were predictive in

seven or more cancers (**Figure 1C, Table S2B**). Both the direction of the hazard ratios (increased or decreased risk of poor outcome) and the magnitude of the effect on patient outcome for germline variants across different cancers were highly correlated (**Text S3**).

Analysis 3 increased our statistical power by grouping similar cancer types to increase the number of patients with the minor allele that could be included in the study. 29 different patient groups were created based on organ system, histological, or molecular classification (**Figure 1D,** group justification in **Table S3**). 258,466 unique germline variants were tested and 103 prognostic variants were identified (**Figure 1E, Table S2C,** Kaplan Meier plots of selected examples in **Figure S6**).

Prognostic Germline Variants Causing Significant Amino Acid Changes

Analyses 4-6 repeated analyses 1-3 but limited these analyses to variants within the top 0.3% of deleterious mutants across the human genome with CADD>25 (**Figure 3A**). Analysis 4 tested a total of 981 unique variants and identified nine unique prognostic variants (**Figure 3B, Table S2D**). Of the 16 variants that were recurrently predictive of patient outcomes in 4 or more cancers (analysis 5), we will discuss the one variant that was predictive in five cancers (**Figure 3C, Table S2E**). Analysis 6 tested 903 unique variants for an association with outcome in the patient groups used in analysis 3 and described in **Figure 1D** and identified 3 additional prognostic variants (**Figure 3D, Table S2F**).

The Pan-Cancer Landscape of Prognostic Germline Variants

The large number of prognostic variants identified in analysis 1 and 3 allowed us to compare the characteristics of these germline variants with previously reported characteristics of variants identified by genome wide association studies (GWAS). Three characteristics have been noted in variants identified through GWAS: (1) the minor allele tends to be associated with increased risk for poor outcome when considering the set of variants with large effect sizes, (2) there is a negative correlation between effect size and allele frequency, and (3) most germline variants identified by GWAS do not cause amino acid changes [45].

To test whether the allele associated with increased risk for poor outcome is usually the minor allele, the predictive alternate alleles from analysis 1 were classified as associated with increased risk for poor outcome (HR>1) or decreased risk for poor outcome (HR<1) based on the Cox regression results. Of the prognostic germline variants from analysis 1, the allele associated with increased risk is clearly often the minor allele (p=7.077E-8) (**Figure 4A**). A similar analysis with the predictive variants from analysis 3 (**Figure 4B**) did not show a significant statistical depletion of alternate alleles associated with increased risk for poor outcome from the population (p=0.115). The predictive variants from analysis 3 were detectable only with larger sample sizes and have smaller effect sizes than those identified by analysis 1. Thus the result in **Figure 4B** is still consistent with the first premise that an allele associated with increased risk for

poor outcome with a large effect size (as in analysis 1, but not analysis 3) is usually the minor allele [45].

A negative correlation is seen between effect size and allele frequency with both variants from analysis 1 (Spearman's rho = -0.282, p=0.0184) and analysis 3 (Spearman's rho = -0.667, p<2.2E-16), satisfying the second premise. Finally, the vast majority of predictive variants identified by this study do not cause amino acid changes (**Figure 4C-D**), satisfying the third premise.

If the effects of the prognostic germline variants are at least partially independent of each other, we would expect that patients with two prognostic germline variants that increase the risk for poor outcome should do worse than patients with only one of these prognostic germline variant that increases the risk for poor outcome. Indeed, when tested, we found this to be true (p=8.45E-17, analysis approach detailed in **Methods**).

A previous study had identified germline variants associated with an increased incidence of somatic mutations in cancer related genes [27]. We also found that some of the prognostic germline variants were associated with an increased risk of somatic mutations in cancer driver genes. While more prognostic germline variants were associated with an increased risk of somatic mutations in driver genes than was expected by random chance (OR=1.89, p=0.0001, **Text S4**), not all of the prognostic germline variants were associated with an increased risk of such somatic mutations. A more detailed study of somatic mutations in driver genes is necessary that will take into account differences in genes and cancer types.

Germline Variants Significantly Improve Outcome Prediction Models

The effect sizes of prognostic germline variants from analysis 1 were large enough to hypothesize that germline variants identified in individual cancers could improve clinical outcome models in current use.

The clinical variables predictive of outcome (**Table S1**) were used to generate the first outcome model (Clinical: C). The second outcome model was based on clinical information plus the status of a particular predictive germline variant (Germline Variant: GV) (C+GV). An example receiver operator characteristic (ROC) curve for predicting LAML patient vital status at 366 days of follow-up is shown using C and C+GV for predictive variant rs3003628 (ROC in **Figure 4E**). The area under the ROC curves (∆AUC) for the C model is 0.807 and for the C+GV model is 0.928. The change in AUC (∆AUC) for the C+GV model relative to the C model in this example is 0.12 (12%). To ensure that the change in AUC is consistent at different times of follow-up, ∆AUC was calculated from the 10^{th} to the 90^{th} percentile of patient outcome time. The mean and standard error of ∆AUC was plotted against the p-value of the one-sided test evaluating whether the AUC for C+GV is significantly larger than the AUC for C (**Figure 4F**).

This analysis was repeated for all predictive variants. There is a consistent, statistically significant ($p<0.05$) increase in AUC when the clinical model is enhanced by germline variant information (C+GV) compared to the clinical model alone (C) for 63 of the predictive germline variants out of 70 tested

(**Table S4**). These results demonstrate that adding predictive germline variants to existing clinical criteria will improve the prediction of outcome of many cancers.

Prognostic Variants in Driver Genes, Oncogenes, and Tumor Suppressor Genes

90 of the 193 genes in the proximity of one of the prognostic germline variants have been functionally implicated in nine of the twelve hallmarks of cancer (**Figure 5A, Table S5**) [50].

Roughly 50% of the predictive variants are found in or near genes that possibly have tumor suppressor or oncogenic activity (**Figure 5B**, **Table S5**). About 25% of the predictive genes were previously studied in the cancer in which the germline variant was found to be prognostic, about half were previously studied in at least one cancer, and roughly two-thirds were studied in at least one human disease (**Figure 5C**, **Table S5**). Prognostic variants were identified in or near *MSH6*, *POLQ*, *ARID5B*, and *IDH2*, which are previously reported cancer-driver genes (**Figure 5D**).

Prognostic Germline Variants Can Cause Significant Amino Acid Changes or Act as eQTLs

The 12 prognostic variants identified in analyses 4-6 caused significant amino acid changes (CADD>25), with many of these amino acid changes occurring in protein-coding domains with annotated or known functions (**Figure 5E**).

39 variants could act as *cis* eQTLs, as they were associated with expression differences of the proximate genes. We highlight 5 of these variants because the expression levels of the proximate genes are also predictive of

survival, with the direction of the effect (HR >1 or <1) being concordant with the effect of the variant (**Figure 5F**). Of these 5 variants, 3 were also *cis* eQTLs in the corresponding tissue in GTEx [57].

Prognostic Variants Implicated in Other Diseases

Some of the prognostic variants are linked with diseases that occur in the tissue giving rise to the tumor, suggesting the variant has an important function in that tissue (**Figure 5G**, **Table S6A**). **Table S6B** lists prognostic genes that are linked in the literature to traits in tissues outside the ones bearing the tumors.

Individual Prognostic Variant Characterization

In this section, we characterize three germline variants to illustrate how individual germline variants may be associated with patient outcome. These hypotheses are supported by bioinformatic analyses and require future molecular insight to confirm and fully understand the mechanistic underpinnings of these associations.

rs1800932 in *MSH6* May Be Associated with Favorable Outcome by Increasing Temozolomide Sensitivity

rs1800932 predicts favorable patient outcome in gliomas (LGG and GBM). This variant is an eQTL for increased expression of *MSH6* in many tissues, including nerve, is associated with increased expression of *MSH6* in patients with LGG (p=0.00732), and has previously been reported to be associated with a decreased risk of prostate cancer [57, 58]. We found *MSH6* expression to be correlated with elevated temozolomide sensitivity in cancer cell lines (Spearman's rho=0.165, p=5.01E-7) [54]. Temozolomide is a DNA alkylating agent used in the treatment of most glioma patients and is likely to have been

used in the therapy of most patients with gliomas in TCGA. *MSH6* knockdown increases temozolomide resistance and somatic mutations in *MSH6* are associated with temozolomide resistance in gliomas [53, 59]. Taken together, this suggests that rs1800932 is an eQTL for increased expression of *MSH6* in gliomas, which may increase sensitivity to temozolomide, the primary chemotherapeutic agent for gliomas.

rs55796947 in *MAP2K3* May Result in Cell Cycle Arrest and Apoptosis

rs55796947 in *MAP2K3/MKK3* predicts favorable prognosis in KIRC. This germline variant introduces a stop codon in *MAP2K3* that truncates the kinase domain. *MAP2K3* inhibition results in cell cycle arrest, autophagy-mediated cell death, the unfolded protein response (UPR), and sensitization to chemotherapy drugs [60]. Indeed, tumors in patients with this variant upregulate genes involved with apoptosis ($p<0.001$, **Figure 6A-B**) and downregulate *E2F* targets involved in cell-cycle progression ($p=0.047$, **Figure 6C**). This germline variant likely truncates the kinase domain of *MAP2K3*, resulting in cell cycle arrest, apoptosis, and favorable patient outcome.

rs77903511 is an eQTL for *BIRC5* which Inhibits Apoptosis

rs77903511 predicts poor patient outcome in UVM (**Figure 6D)**. *BIRC5* inhibits apoptosis through interaction with and inhibition of caspase 9 and effector caspases. The alternate allele is associated with increased *BIRC5* expression in the tumors ($p=0.02$, **Figure 6E**). Consistent with a role of *BIRC5* in apoptosis inhibition, *BIRC5* expression is associated with poor patient outcome (**Figure**

6F). This variant, therefore, may be associated with poor outcome because of an increase of the apoptosis inhibitor *BIRC5*.

Discussion

This study shows, as a general principle, that germline variants are associated with cancer patient outcome. The prognostic germline variants enhanced patient outcome predictions compared to models based on currently collected clinical data. We envision germline variants providing clinicians with information about a patient as a supplement to reported history, physical exam findings, and imaging and laboratory tests. These predictions will improve over time with the use of more information available in electronic medical records.

The results of this study are most easily applied at the population level to identify groups of patients at increased risk for poor outcome (for example for clinical trials) and for follow-up mechanistic studies on how the variants affect outcome. This study will serve as the basis for future work to apply these findings at the level of individual patients, as a given variant will need to be considered in conjunction with other variants and with clinical factors to calculate expected survival time or time to progression. While we identified a large number of prognostic germline variants in analysis 1, our sample size for this study was relatively modest. The power calculations and the identification of additional prognostic germline variants by grouping similar cancers suggest that more prognostic germline variants will likely emerge as more tumors are sequenced and will further support the notion that germline variation is associated with patient outcome across cancers. Our study of prognostic germline variants was

limited to common germline variants (allele frequency > 5% in the population) due to statistical limitations derived from sample size in our ability to study pathogenic and low frequency germline variants. However, our results imply that these rarer germline variants may have large effect sizes that may make them particularly valuable for improving clinical outcome model predictions. These variants will likely be studied in the future through more complex approaches or in studies of larger cohorts.

Further study is necessary to validate the associations that we identified, as setting the discovery threshold at FDR<0.10 suggests that some of the associations may have occurred by random chance. The variants identified in analyses 2 and 5 require deeper interrogation, as we were unable to develop an unbiased test to assess the probability of those associations occurring by random chance. While we identified germline variants associated with significant improvements in clinical outcome predictions, further work is necessary to identify situations in which the additional prognostic information would be valuable for treatment decisions or end of life planning.

Given the paucity of studies testing for associations between germline variants and patient outcome in cohorts of cancer patients, we were unsure of the effect sizes that could be expected in this study across the 33 cancers. This uncertainty was further exacerbated by reports of effect sizes being negatively correlated with allele frequency for some traits [45]. The results of this study will provide researchers with a sense for the magnitude of effect sizes that can be expected from germline variants associated with patient outcome along with the

relationship between effect size and allele frequency. These results will help better optimize future studies for detecting significant associations.

It is reassuring that a significant fraction of prognostic germline variants are found in or near possible tumor suppressor genes, oncogenes, or known cancer driver genes. The variants in cancer driver genes, *MSH6*, *POLQ*, *ARID5B*, and *IDH2*, warrant further study to determine the mechanism by which these variant affect cancer progression [61]. The twelve germline variants in **Figure 5E** that cause substantial amino acid changes are prime candidates for experimental follow-up and are discussed in detail in **Text S5.** A handful of the prognostic germline variants have been associated with human disease, some in the same tissue and others in unrelated tissues, suggesting that these pathologies may stem from shared molecular phenomena (**Table S6**).

The mechanisms of action of many of the prognostic variants are currently unknown. There are many possibilities by which the variants that do not cause amino acid changes could affect cancer biology [62]. Many variants are likely acting as *trans* eQTLs, which are difficult to study in datasets with relatively small sample sizes. Some of the variants may also be acting as eQTLs in non-tumor cells, such as immune system cells or cells of the vasculature. The already high involvement of tumor suppressor genes, oncogenes, and driver genes among the prognostic germline variants is promising for future study. This report provides basic science researchers with genes and variants that should be studied to better understand the etiology and progression of cancers, while providing

clinicians with the potential for better clinical predictions that could be made if germline variants are considered in the context of patient care.

Conclusions

While the prediction of outcome for patients with cancer is currently based on clinical factors, the analysis of next-generation sequencing data in clinical oncology has suggested that genomic information can further improve these predictions. Previous studies analyzing the usage of genomic information in clinical oncology have focused primarily on somatic aberrations. In this proof-of-principle study, we systematically analyzed sequencing data from thirty-three different cancers to test whether germline variation could also be used to provide clinicians with information about patient outcome. We identified prognostic germline variants across individual cancers and group of cancers and find that these germline variants provide additional predictive power about patient outcomes beyond the information that can be gathered from clinical factors alone. Mechanistically, twelve of the germline variants seem to be associated with patient outcome through perturbation of protein structure and at least five through association with gene expression differences, though the molecular functions of most of the germline variants are currently unknown. About half of the germline variants are in previously reported tumor suppressor genes, oncogenes, or driver genes with the other half implicating loci that deserve further investigation in oncology. As this is a proof-of-principle study, further studies of germline variation in other cancer cohorts are necessary confirm that germline variation is associated with patient outcome across cancers.

List of Abbreviations

ACC - Adrenocortical Carcinoma

AUC - Area Under the Curve

BLCA - Bladder Urothelial Carcinoma

BRCA - Breast Invasive Carcinoma

C - Clinical

CADD - Combined Annotation Dependent Depletion Score

CESC - Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma

CHOL - Cholangiocarcinoma

COAD - Colon Adenocarcinoma

DLBC - Lymphoid Neoplasm Diffuse Large B-Cell Lymphoma

ESCA - Esophageal Carcinoma

eQTL - Expression Quantitative Trait Loci

FDR - False Discovery Rate

GBM - Glioblastoma Multiforme

GTEx - Genotype Tissue Expression Project

GWAS - Genome Wide Association Study

GV - Germline Variation

HNSC - Head and Neck Squamous Cell Carcinoma

HR - Hazard Ratio

KICH - Kidney Chromophobe

KIRC - Kidney Renal Clear Cell Carcinoma

KIRP - Kidney Renal Papillary Cell Carcinoma

LAML - Acute Myeloid Leukemia

LGG - Brain Lower Grade Glioma

LIHC - Liver Hepatocellular Carcinoma

LUAD - Lung Adenocarcinoma

LUSC - Lung Squamous Cell Carcinoma

MESO - Mesothelioma

OV - Ovarian Serious Cystadenocarcinoma

PAAD - Pancreatic Adenocarcinoma

PCPG - Pheochromocytoma and Paraganglioma

PRAD - Prostate Adenocarcinoma

READ - Rectum Adenocarcinoma

ROC - Receiver Operator Characteristic Curve

SARC - Sarcoma

SKCM - Skin Cutaneous Melanoma

TCGA - The Cancer Genome Atlas

TCGAA - The Cancer Genome Ancestry Atlas

TGCT - Testicular Germ Cell Tumors

THCA - Thyroid Carcinoma

THYM - Thymoma

UCEC - Uterine Corpus Endometrial Carcinoma

UCS - Uterine Carcinosarcoma

UVM - Uveam Melanoma

UPR - Unfolded Protein Response

WXS - Whole Exome Sequencing

WXS Normal - Whole Exome Sequenced Normal Sample

WXS Tumor - Whole Exome Sequenced Tumor Sample

RNA Tumor - RNA Sequenced Tumor Sample

Declarations

Ethics Approval and Consent to Participate: The need for Institutional Review Board Approval at our institution (University of Virginia) was waived for this study as all data used from this project had previously been generated as part of The Cancer Genome Atlas Project and none of the results reported in this manuscript can be used to identify individual patients.

Availability of Data and Materials: All data used for this study is publicly available through The Cancer Genome Atlas project and can be downloaded from the genomic data commons (https://portal.gdc.cancer.gov/). The results in this manuscript are based upon data generated by The Cancer Genome Atlas (TCGA) Research Network: https://www.cancer.gov/tcga.

Competing Interests: The authors declare that they have no competing interests.

Acknowledgements

We thank Drs. Ana Damljanovic, Liz Williams, and Manisha Ray for their assistance with computation on the Cancer Genomics Cloud Platform, the High Performance Computing Team at the University of Virginia for assistance with computation on our university cluster and dbGAP for providing us with access to The Cancer Genome Atlas data. Most importantly, we are indebted to the

patients and all of their families for their participation in The Cancer Genome Atlas project and the opportunity to study these cancers in a clinical context.

References

1. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, et al: **International network of cancer genome projects.** *Nature* 2010, **464:**993-998.

2. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al: **Comprehensive Characterization of Cancer Driver Genes and Mutations.** *Cell* 2018, **173:**371-385.e318.

3. Lee B, Tran B, Hsu AL, Taylor GR, Fox SB, Fellowes A, Marquis R, Mooi J, Desai J, Doig K, et al: **Exploring the feasibility and utility of exome-scale tumour sequencing in a clinical setting.** *Intern Med J* 2018, **48:**786-794.

4. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al: **Pathogenic Germline Variants in 10,389 Adult Cancers.** *Cell* 2018, **173:**355-370.e314.

5. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, et al: **Germline Mutations in Predisposition Genes in Pediatric Cancer.** *N Engl J Med* 2015, **373:**2336-2346.

6. Pearlman R, Frankel WL, Swanson B, Zhao W, Yilmaz A, Miller K, Bacher J, Bigley C, Nelsen L, Goodfellow PJ, et al: **Prevalence and Spectrum of Germline Cancer Susceptibility Gene Mutations Among Patients With Early-Onset Colorectal Cancer.** *JAMA Oncol* 2017, **3:**464-471.

7. Mandelker D, Zhang L, Kemel Y, Stadler ZK, Joseph V, Zehir A, Pradhan N, Arnold A, Walsh MF, Li Y, et al: **Mutation Detection in Patients With Advanced Cancer by Universal Sequencing of Cancer-Related Genes in Tumor and Normal DNA vs Guideline-Based Germline Testing.** *Jama* 2017, **318:**825-835.

8. Cheng DT, Prasad M, Chekaluk Y, Benayed R, Sadowska J, Zehir A, Syed A, Wang YE, Somar J, Li Y, et al: **Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing.** *BMC Med Genomics* 2017, **10:**33.

9. Lee SE, Lee HS, Kim KY, Park JH, Roh H, Park HY, Kim WS: **High prevalence of the MLH1 V384D germline mutation in patients with HER2-positive luminal B breast cancer.** *Sci Rep* 2019, **9:**10966.

10. 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.

11. Gori S, Barberis M, Bella MA, Buttitta F, Capoluongo E, Carrera P, Colombo N, Cortesi L, Genuardi M, Gion M, et al: **Recommendations for the implementation of BRCA testing in ovarian cancer patients and their relatives.** *Crit Rev Oncol Hematol* 2019, **140:**67-72.

12. Tian W, Bi R, Ren Y, He H, Shi S, Shan B, Yang W, Wang Q, Wang H: **Screening for hereditary cancers in patients with endometrial cancer reveals a high frequency of germline mutations in cancer predisposition genes.** *Int J Cancer* 2019, **145:**1290-1298.

13. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K: **Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland.** *N Engl J Med* 2000, **343:**78-85.

14. Menden MP, Casale FP, Stephan J, Bignell GR, Iorio F, McDermott U, Garnett MJ, Saez-Rodriguez J, Stegle O: **The germline genetic component of drug sensitivity in cancer cell lines.** *Nat Commun* 2018, **9:**3385.

15. Pomerantz MM, Spisak S, Jia L, Cronin AM, Csabai I, Ledet E, Sartor AO, Rainville I, O'Connor EP, Herbert ZT, et al: **The association between germline BRCA2 variants and sensitivity to platinum-based chemotherapy among men with metastatic prostate cancer.** *Cancer* 2017, **123:**3532-3539.

16. Low SK, Zembutsu H, Nakamura Y: **Breast cancer: The translation of big genomic data to cancer precision medicine.** *Cancer Sci* 2018, **109:**497-506.

17. Hahnen E, Lederer B, Hauke J, Loibl S, Krober S, Schneeweiss A, Denkert C, Fasching PA, Blohmer JU, Jackisch C, et al: **Germline Mutation Status, Pathological Complete Response, and Disease-Free Survival in Triple-Negative Breast Cancer: Secondary Analysis of the GeparSixto Randomized Clinical Trial.** *JAMA Oncol* 2017, **3:**1378-1385.

18. Li X, Wu N, Li B: **A high mutation rate of immunoglobulin heavy chain variable region gene associates with a poor survival and chemotherapy response of mantle cell lymphoma patients.** *Medicine (Baltimore)* 2019, **98:**e15811.

19. Horak P, Weischenfeldt J, von Amsberg G, Beyer B, Schutte A, Uhrig S, Gieldon L, Klink B, Feuerbach L, Hubschmann D, et al: **Response to olaparib in a PALB2 germline mutated prostate cancer and genetic events associated with resistance.** *Cold Spring Harb Mol Case Stud* 2019, **5**.

20. Crona DJ, Skol AD, Leppanen VM, Glubb DM, Etheridge AS, Hilliard E, Pena CE, Peterson YK, Klauber-DeMore N, Alitalo KK, Innocenti F: **Genetic Variants of VEGFA and FLT4 Are Determinants of Survival in Renal Cell Carcinoma Patients Treated with Sorafenib.** *Cancer Res* 2019, **79:**231-241.

21. de Velasco G, Gray KP, Hamieh L, Urun Y, Carol HA, Fay AP, Signoretti S, Kwiatkowski DJ, McDermott DF, Freedman M, et al: **Pharmacogenomic Markers of Targeted Therapy Toxicity in Patients with Metastatic Renal Cell Carcinoma.** *Eur Urol Focus* 2016, **2:**633-639.

22. Hertz DL, Henry NL, Rae JM: **Germline genetic predictors of aromatase inhibitor concentrations, estrogen suppression and drug efficacy and toxicity in breast cancer patients.** *Pharmacogenomics* 2017, **18:**481-499.

23. Lee SHR, Yang JJ: **Pharmacogenomics in acute lymphoblastic leukemia.** *Best Pract Res Clin Haematol* 2017, **30:**229-236.

24. Singh M, Bhatia P, Khera S, Trehan A: **Emerging role of NUDT15 polymorphisms in 6-mercaptopurine metabolism and dose related toxicity in acute lymphoblastic leukaemia.** *Leuk Res* 2017, **62:**17-22.

25. Guan J, Fransson S, Siaw JT, Treis D, Van den Eynden J, Chand D, Umapathy G, Ruuth K, Svenberg P, Wessman S, et al: **Clinical response of the novel activating ALK-I1171T mutation in neuroblastoma to the ALK inhibitor ceritinib.** *Cold Spring Harb Mol Case Stud* 2018, **4**.

26. Udagawa C, Nakamura H, Ohnishi H, Tamura K, Shimoi T, Yoshida M, Yoshida T, Totoki Y, Shibata T, Zembutsu H: **Whole exome sequencing to identify genetic markers for trastuzumab-induced cardiotoxicity.** *Cancer Sci* 2018, **109:**446-452.

27. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, et al: **Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer.** *Cancer Discov* 2017, **7:**410-423.

28. Guerrini-Rousseau L, Dufour C, Varlet P, Masliah-Planchon J, Bourdeaut F, Guillaud-Bataille M, Abbas R, Bertozzi AI, Fouyssac F, Huybrechts S,

et al: **Germline SUFU mutation carriers and medulloblastoma: clinical characteristics, cancer risk, and prognosis.** *Neuro Oncol* 2018, **20:**1122-1132.

29. Baretta Z, Mocellin S, Goldin E, Olopade OI, Huo D: **Effect of BRCA germline mutations on breast cancer prognosis: A systematic review and meta-analysis.** *Medicine (Baltimore)* 2016, **95:**e4975.

30. Chatrath A, Kiran M, Kumar P, Ratan A, Dutta A: **The Germline Variants rs61757955 and rs34988193 are Predictive of Survival in Lower Grade Glioma Patients.** *Mol Cancer Res* 2019.

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

32. Lau JW, Lehnert E, Sethi A, Malhotra R, Kaushik G, Onder Z, Groves-Kirkby N, Mihajlovic A, DiGiovanna J, Srdic M, et al: **The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research.** *Cancer Res* 2017, **77:**e3-e6.

33. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR: **VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.** *Nucleic Acids Res* 2016, **44:**e108.

34. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al: **Analysis of protein-coding genetic variation in 60,706 humans.** *Nature* 2016, **536:**285-291.

35. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.** *Cell* 2018, **173:**400-416.e411.

36. Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, Hu X, Zhang Y, Wang Y, Jiang J, et al: **Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers.** *Cancer Cell* 2018, **34:**549-560.e549.

37. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, et al: **TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data.** *Nucleic Acids Res* 2016, **44:**e71.

38. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, Morozova O, Newton Y, Radenbaugh A, Pagnotta SM, et al: **Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma.** *Cell* 2016, **164:**550-563.

39. Tibshirani R: **The lasso method for variable selection in the Cox model.** *Stat Med* 1997, **16:**385-395.

40. Friedman J, Hastie T, Tibshirani R: **Regularization Paths for Generalized Linear Models via Coordinate Descent.** *J Stat Softw* 2010, **33:**1-22.

41. Gu Z, Gu L, Eils R, Schlesner M, Brors B: **circlize Implements and enhances circular visualization in R.** *Bioinformatics* 2014, **30:**2811-2812.

42. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al: **Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.** *Cell* 2018, **173:**291-304.e296.

43. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M: **CADD: predicting the deleteriousness of variants throughout the human genome.** *Nucleic Acids Res* 2019, **47:**D886-d894.

44. Kim S: **ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients.** *Commun Stat Appl Methods* 2015, **22:**665-674.

45. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Jr., Chatterjee N: **Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants.** *Proc Natl Acad Sci U S A* 2011, **108:**18026-18031.

46. Wang K, Li M, Hakonarson H: **ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res* 2010, **38:**e164.

47. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, et al: **Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines.** *Cell Syst* 2018, **6:**271-281.e277.

48. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological**

databases and microarray data analysis. *Bioinformatics* 2005, **21:**3439-3440.

49. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nat Protoc* 2009, **4:**1184-1191.

50. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144:**646-674.

51. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al: **Ensembl 2019.** *Nucleic Acids Res* 2019, **47:**D745-d751.

52. Consortium U: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res* 2019, **47:**D506-d515.

53. McFaline-Figueroa JL, Braun CJ, Stanciu M, Nagel ZD, Mazzucato P, Sangaraju D, Cerniauskas E, Barford K, Vargas A, Chen Y, et al: **Minor Changes in Expression of the Mismatch Repair Protein MSH2 Exert a Major Impact on Glioblastoma Response to Temozolomide.** *Cancer Res* 2015, **75:**3127-3138.

54. Rajapakse VN, Luna A, Yamade M, Loman L, Varma S, Sunshine M, Iorio F, Sousa FG, Elloumi F, Aladjem MI, et al: **CellMinerCDB for Integrative Cross-Database Genomics and Pharmacogenomics Analyses of Cancer Cell Lines.** *iScience* 2018, **10:**247-264.

55. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, et al: **Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.** *Nucleic Acids Res* 2013, **41:**D955-961.

56. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.

57. GTEx-Consortium: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45:**580-585.

58. Basu S, Majumder S, Bhowal A, Ghosh A, Naskar S, Nandy S, Mukherjee S, Sinha RK, Basu K, Karmakar D, et al: **A study of molecular signals deregulating mismatch repair genes in prostate cancer compared to benign prostatic hyperplasia.** *PLoS One* 2015, **10:**e0125560.

59. Xie C, Sheng H, Zhang N, Li S, Wei X, Zheng X: **Association of MSH6 mutation with glioma susceptibility, drug resistance and progression.** *Mol Clin Oncol* 2016, **5:**236-240.

60. Baldari S, Ubertini V, Garufi A, D'Orazi G, Bossi G: **Targeting MKK3 as a novel anticancer strategy: molecular mechanisms and therapeutical implications.** *Cell Death Dis* 2015, **6:**e1621.

61. Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, et al: **Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics.** *Cell* 2018, **173:**305-320.e310.

62. Zhang F, Lupski JR: **Non-coding genetic variants in human disease.** *Hum Mol Genet* 2015, **24:**R102-110.

Figures

Figure 1. Prognostic germline variants identified in analyses one through three.

A. A description of the three analyses used to identify prognostic germline variants in this figure.

B. Analysis 1. Germline variants found to be predictive of patient outcome in each cancer. Each dot represents a germline variant that was tested for an association with patient outcome. Variants closer to the outside of the plot are more closely associated with patient outcome. Variants in red are significantly (FDR<0.10) associated with patient outcome. The alternating black and grey colors reflect alternating chromosomes for the germline variants that were not significant predictors of patient outcome.

C. Analysis 2. Germline variants found to be recurrently predictive of patient outcome in multiple different cancers. We identified 5 total germline variants that were recurrently predictive ($p<0.05$) of favorable (HR<1) or poor (HR>1) patient outcomes in 7 or more different cancers.

D. Analysis 3. 29 groups of cancers created to identify germline variants with weaker effect sizes in larger patient cohorts. Justification for these groups is provided in **Table S3**.

E. Analysis 3. Germline variants found to be predictive of patient outcome in the groups described in **Figure 1D**. The format of the figure is the same as in **Figure 1B**.

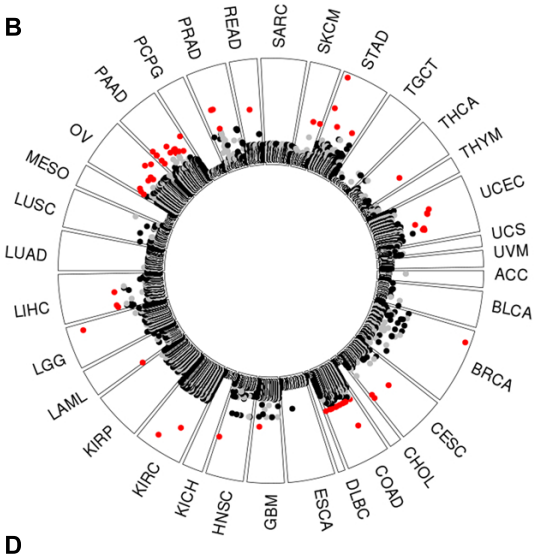
Figures

Figure 1

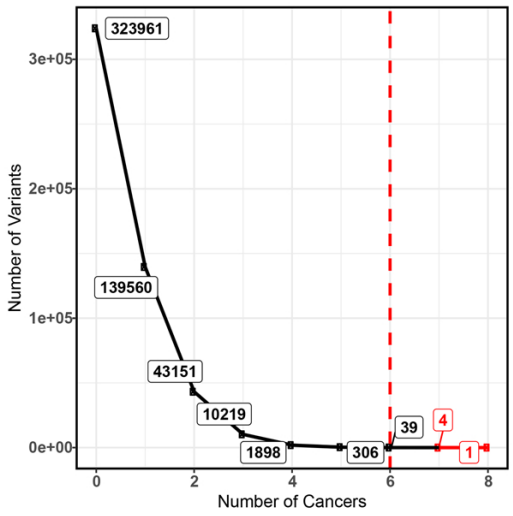
A

Analysis Number	Description	Significance Criteria	Number of Cancers or Groups	Unique Variants Tested	Total Statistical Tests Performed	Unique Prognostic Variants
1	GV Predictive of Patient Outcome in Individual Cancers	FDR < 0.10	33 Cancers	519,319	5,217,214	70
2	GV Consistently Predictive of Patient Outcome in 7 or more cancers	p < 0.05 with HR consistently >1 or <1	33 Cancers	519,319	519,319	5
3	GV Predictive of Patient Outcome in Patient Groups	FDR < 0.10	29 Groups	258,466	2,352,228	103

B



C



D

Group Number	Group
1	ACC, KICH
2	ACC, PCPG
3	BLCA, CESC, HNSC, LUSC
4	BLCA, KICH, KIRC, KIRP
5	BLCA, KIRC, KIRP
6	BRCA, OV, UCEC, UCS
7	CESC, HNSC, LUSC
8	CHOL, COAD, ESCA, LIHC, PAAD, READ, STAD
9	CHOL, LIHC
10	COAD, ESCA, PAAD, READ, STAD
11	COAD, ESCA, READ, STAD
12	COAD, READ
13	COAD, READ, STAD
14	DLBC, LAML
15	DLBC, LAML, THYM
16	DLBC, PCPG, SARC, THYM, UCS
17	GBM, LGG
18	GBM, LGG, PCPG
19	KICH, KIRC, KIRP
20	KIRC, KIRP
21	LAML, PRAD, THCA, THYM
22	LAML, THCA
23	LUAD, LUSC
24	LUAD, LUSC, MESO
25	OV, UCEC
26	PAAD, STAD
27	PRAD, TGCT
28	SKCM, UVM
29	UCEC, UCS

E

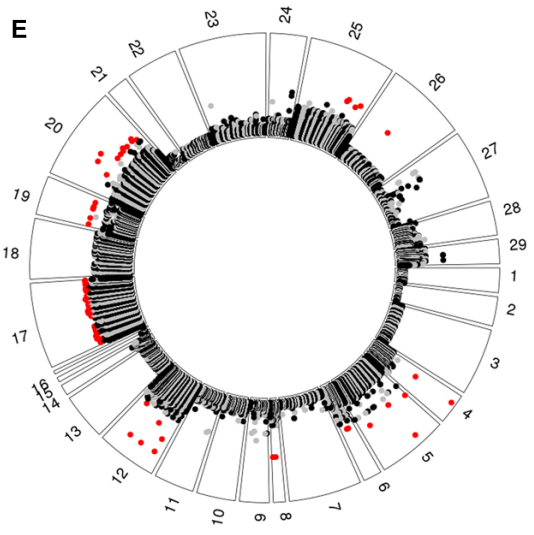


Figure 2. Selected Kaplan-Meier plots of the prognostic germline variants from Analysis 1. The number of patients in each group is indicated next to each line and the patient outcome measure of each disease is given in **Table S1**. The reported p-values and hazard ratios were calculated using univariate regression and are different from the p-values and hazard ratios reported elsewhere which are based on multivariate regression

Figure 2

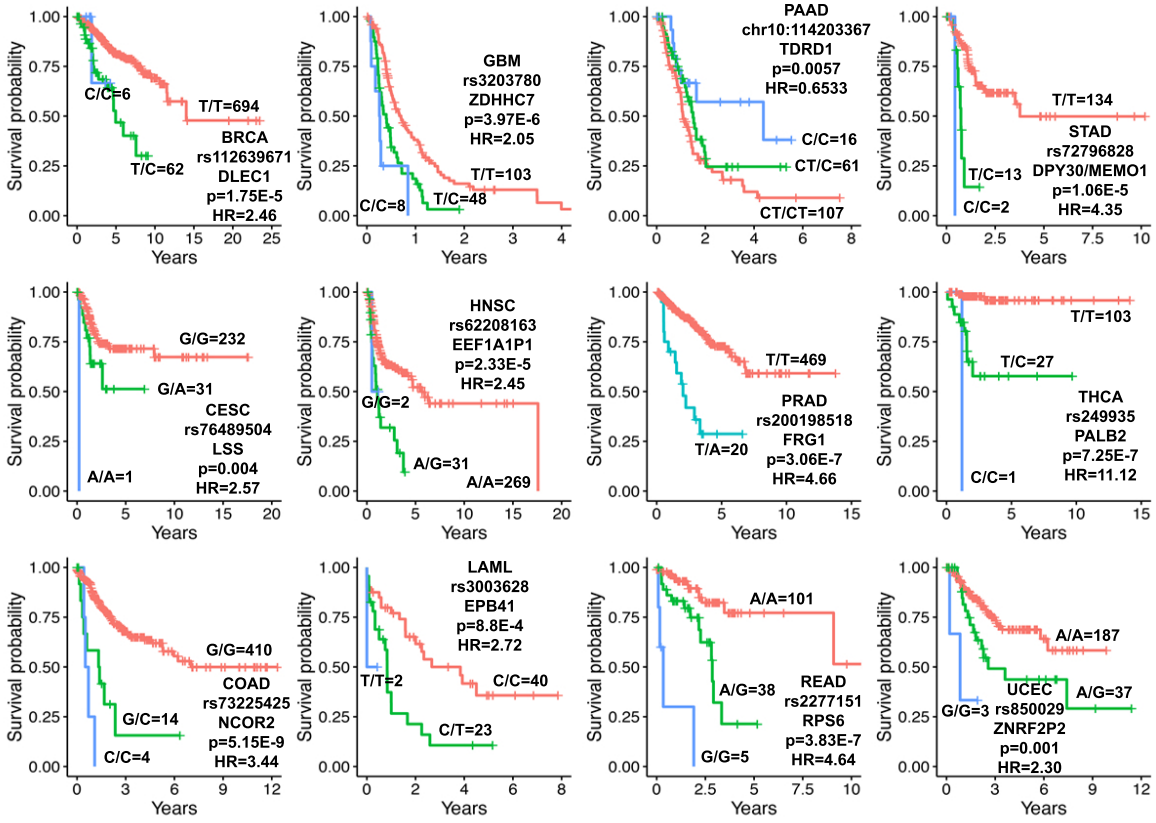


Figure 3. Prognostic germline variants that cause significant amino acid changes (CADD>25) identified in analyses four through six.

A. A description of the three analyses used to identify prognostic germline variants in this figure.

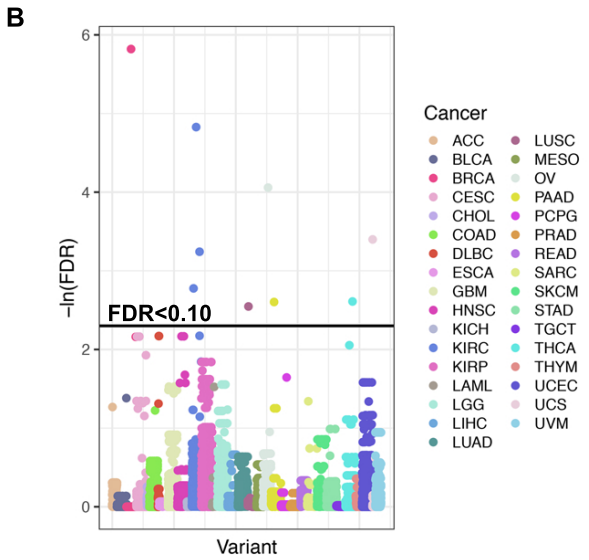
B. Analysis 4. Germline variants causing significant amino acid changes found to be predictive (FDR<0.10) of patient outcome in each cancer.
C. Analysis 5. Germline variants causing significant amino acid changes found to be recurrently predictive ($p<0.05$) of favorable (HR<1) or poor (HR>1) patient outcomes in 5 or more different cancers.

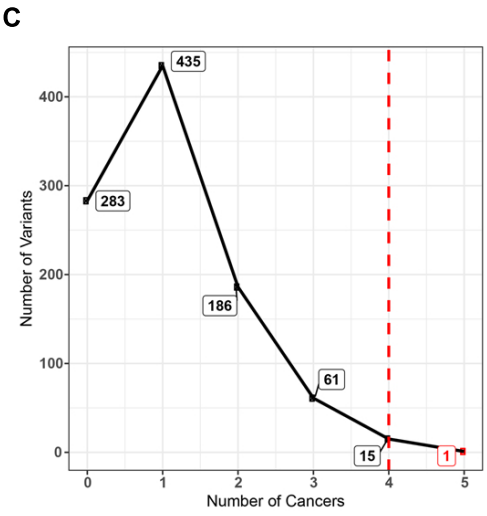
D. Analysis 6. Germline variants causing significant amino acid changes found to be predictive of patient outcome in patient groups defined in **Figure 1D**.

Figure 3

A

Analysis Number	Description	Significance Criteria	Number of Cancers or Groups	Unique Variants Tested	Total Statistical Tests Performed	Unique Prognostic Variants
4	GV with CADD > 25 Predictive of Patient Outcome in Individual Cancers	FDR < 0.10	33 Cancers	981	21,753	9
5	GV Consistently Predictive of Patient Outcome in 5 or more cancers	p < 0.05 with HR consistently >1 or <1	33 Cancers	981	981	1
6	GV Predictive of Patient Outcome in Patient Groups	FDR < 0.10	29 Groups	903	12,216	3





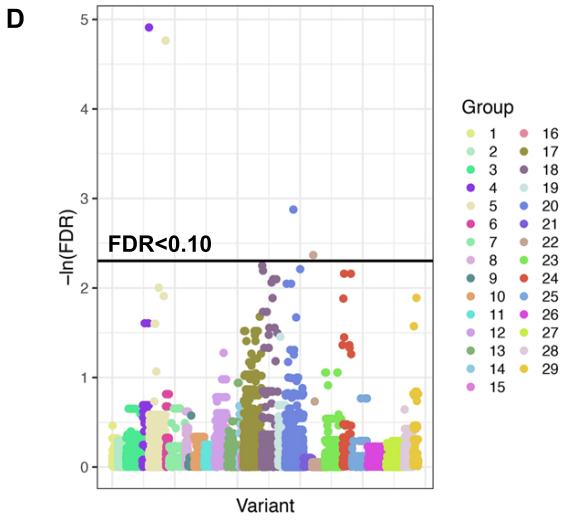


Figure 4. Characteristics of prognostic germline variants and improvement of patient outcome models by the prognostic germline variants.

A-B. Scatterplots of the prognostic germline variants identified in individual cancers in Analysis 1 (**A**) and in groups of cancers in Analysis 3 (**B**). Each pie chart reflects the distribution of patients that are homozygous for the reference allele, heterozygous, and homozygous for the alternate allele for one prognostic variant. The minor allele was much more likely to be associated with increased risk for poor outcome rather than decreased risk for poor outcome (p=7.077E-8) in Analysis 1 though this trend was not significant in Analysis 3 (p=0.115).

C-D. Pie charts displaying the genomic locations of the germline variants in Analysis 1 (**C**) and Analysis 3 (**D**).

E. An example of a receiver operator characteristic (ROC) curve calculated using data from LAML at 366 days of follow-up. The blue line represents the patient outcome predictions made using clinical information alone (C model). The red line represents patient outcome predictions made using clinical information in addition to rs3003628 germline variant status (C+GV model), which we found to be predictive of patient outcomes in LAML. The Area Under the Curve (AUC) was 0.81 for the C model and 0.93 for the C+GV model giving a □AUC of 0.12 (12%).

F. Many of the prognostic germline variants improve clinical outcome model predictions. For each prognostic variant, we created a ROC curve based on the clinical (C) model and the clinical + germline variant (C+GV model), as in **Figure 4E,** at each point in time from the 10^{th}- 90^{th} percentile of patient progression or death for each cancer. The ∆AUC of the C+GV model versus the C model at each time point was calculated (**Table S4**). X-axis: Mean and standard error of ∆AUC. Y-axis: The p-values from testing whether or not the AUC of the C+GV model is significantly greater than that of the C model using a Wilcoxon rank sum test. Four examples of prognostic germline variants that significantly increase the AUC are labeled and highlighted in **Table S4**.

Figure 4

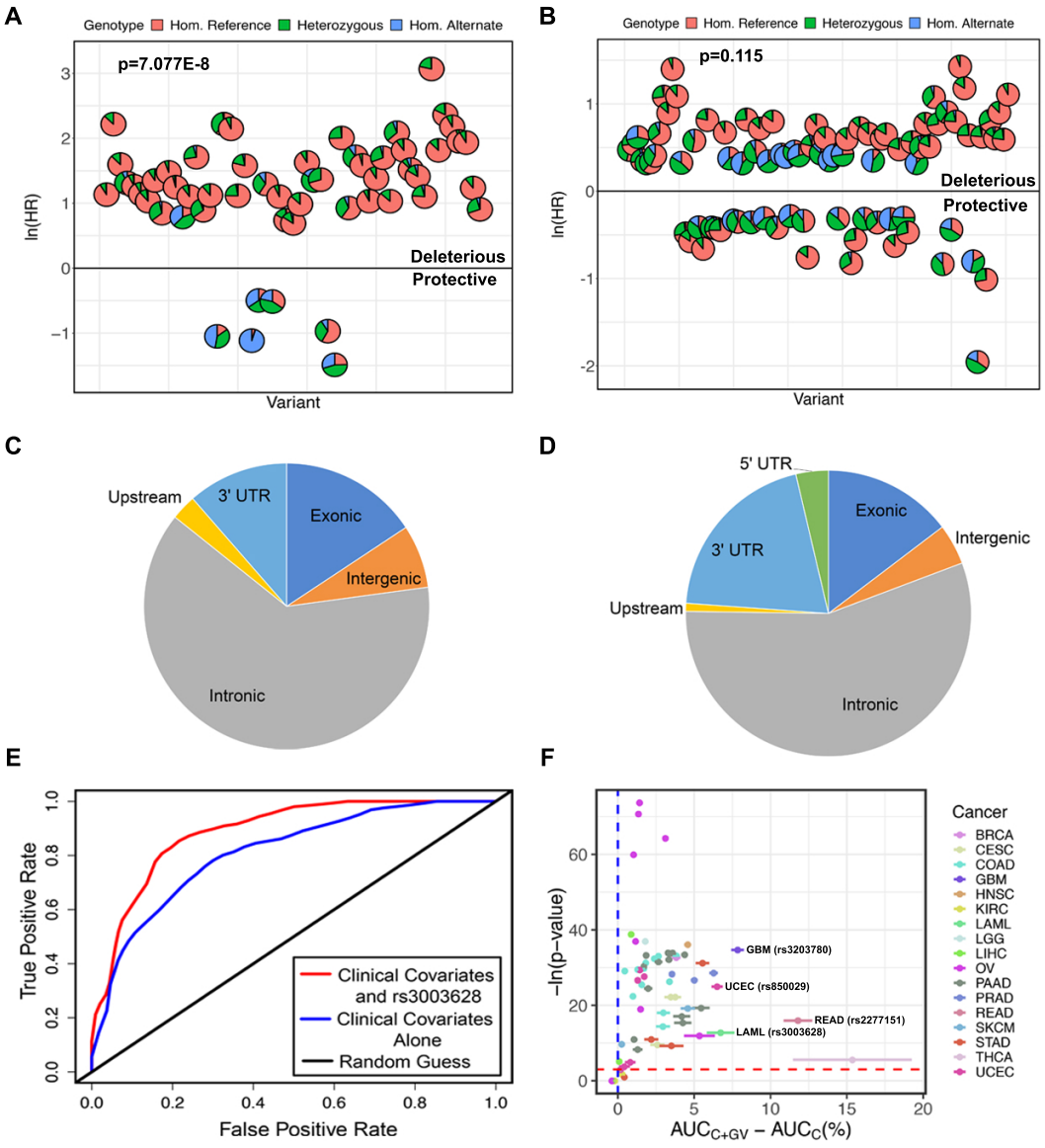

A
Genotype Hom. Reference Heterozygous Hom. Alternate
p=7.077E-8
ln(HR)
Deleterious
Protective
Variant
B
Genotype Hom. Reference Heterozygous Hom. Alternate
p=0.115
ln(HR)
Deleterious
Protective
Variant
C
Upstream
3' UTR
Exonic
Intergenic
Intronic
D
5' UTR
Exonic
Intergenic
3' UTR
Upstream
Intronic
E
True Positive Rate
False Positive Rate
Clinical Covariates and rs3003628
Clinical Covariates Alone
Random Guess
F
−ln(p−value)
AUC C+GV − AUC C(%)
GBM (rs3203780)
UCEC (rs850029)
READ (rs2277151)
LAML (rs3003628)
Cancer
BRCA
CESC
COAD
GBM
HNSC
KIRC
LAML
LGG
LIHC
OV
PAAD
PRAD
READ
SKCM
STAD
THCA
UCEC

Figure 5. Literature review of genes associated with the prognostic germline variants and mechanisms by which prognostic germline variants may exert their effects.

A. The cancer-related functions of genes associated with the prognostic germline variants are quite diverse.

B. Many of the genes associated with the variants have previously been reported to be tumor suppressor genes or oncogenes. We categorized genes as tumor suppressor genes or oncogenes based on phenotypes reported in the literature, even if the exact mechanism through which the genes act have not yet been determined.

C. Although many of the variants have been studied in the field, there are many genes that have not yet been studied in the context of human disease and therefore may warrant investigation by the field.

D. Four of the genes associated with prognostic germline variants are in previously reported cancer driver genes.

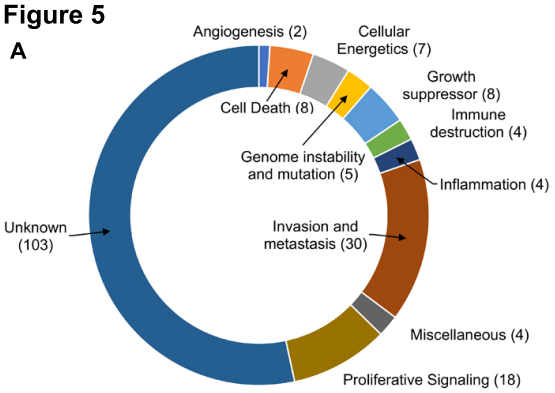
E. Some of the prognostic germline variants cause dramatic amino acid changes and may disrupt well-characterized protein domains.

F. Some of the prognostic germline variants likely act as expression quantitative trait loci in *cis* (*cis* eQTLs) and the expression of these genes are predictive of patient outcome. We found three of these germline variants to also be eQTLs in the genotype tissue expression (GTEx) database in the same tissue that the tumor was derived from.

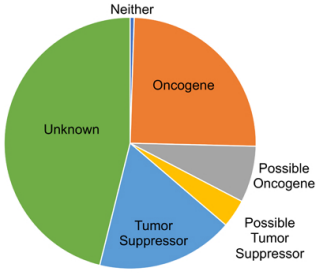
G. Some of the prognostic germline variants have been reported to be associated with other diseases related to the tissue from which the tumor was derived.

Figure 5

A



B



C

D

Gene Name	Function	Cancers	TCGA Driver Gene Classification
ARID5B	Transcriptional coactivator	GBM, LGG	UCEC
IDH2	Krebs cycle	GBM, LGG	Pan-Cancer
MSH6	Mismatch repair	GBM, LGG	UCEC
POLQ	Microhomology-mediated end-joining	GBM, LGG	Pan-Cancer

E

Cancer(s)	Gene	Function	dbSNP ID	Domain	Amino Acid Change
OV	A2ML1	Proteinase inhibitor	rs1558526	Thiol ester domain	C970Y
CESC, ESCA, KIRC, LIHC, PAAD	BORCS5	Lysosomal positioning	rs3751262	Unknown	D191N
PAAD	CRYBG1	Unknown	rs61741114	PFAM 00030	L1235P
KIRC, KIRP	ECD	Inhibits p53 degradation	rs2271904	Conserved protein domain family SGT1	D667G
THCA	EIF2AK4	Protein synthesis regulator	rs35602605	Class II tRNA amino-acyl synthetase-like catalytic core domain	G1306S
UCS	EPHA10	Cell-cell communication	rs6671088	Catalytic domain of the protein tyrosine kinase	G749E
LAML, THCA	FCRL6	MHC Class II Receptor	rs61823162	49 amino acids downstream of ITIM motif	Q423*
BLCA, KICH, KIRC, KIRP	KDELR3	Retention of endoplasmic reticulum proteins	rs12004	Transmembrane region	V199G
KIRC	MAP2K3	MAP kinase signaling	rs55796947	Catalytic domain of the dual-specificity protein kinases	Q102*
BRCA	MYOF	Membrane fusion	rs11594445	Between C2E Ferlin and C2F Ferlin region	R1783Q
LUSC	OR10X1	Olfactory receptor	rs863362	Transmembrane region	W66*
KIRC	SAG	Terminates Rho signaling	rs7565275	Arrestin N region	I76V

F

Variant and Gene Information					Median Expression		Differentially Expressed	Cox Regression Results for Gene		GTEx
dbSNP ID	Cancer	Gene	Gene Function	Variant Deleterious or Protective	Wild Type (FPKM)	Mutant (FPKM)	p-value	p-value	Hazard Ratio	eQTL in Tumor-Derived Tissue
rs77903511	UVM	BIRC5	Apoptosis inhibitor	Deleterious	1.76	3.46	0.02	0.08	1.35	No
rs12418990	PAAD	AC103974.1	Unknown	Deleterious	1.29	0.92	0.06	0.08	0.61	No
rs2303949	LGG	WLS	Wnt protein sorting and secretion regulator	Protective	113.83	21.99	0.06	0.01	1.01	Yes
rs6679382	LGG	WLS	Wnt protein sorting and secretion regulator	Protective	113.83	21.99	0.03	0.01	1.01	Yes
rs62286656	LGG	GFM1	Mitochondrial translation	Protective	8.02	5.69	0	0.05	1.13	Yes

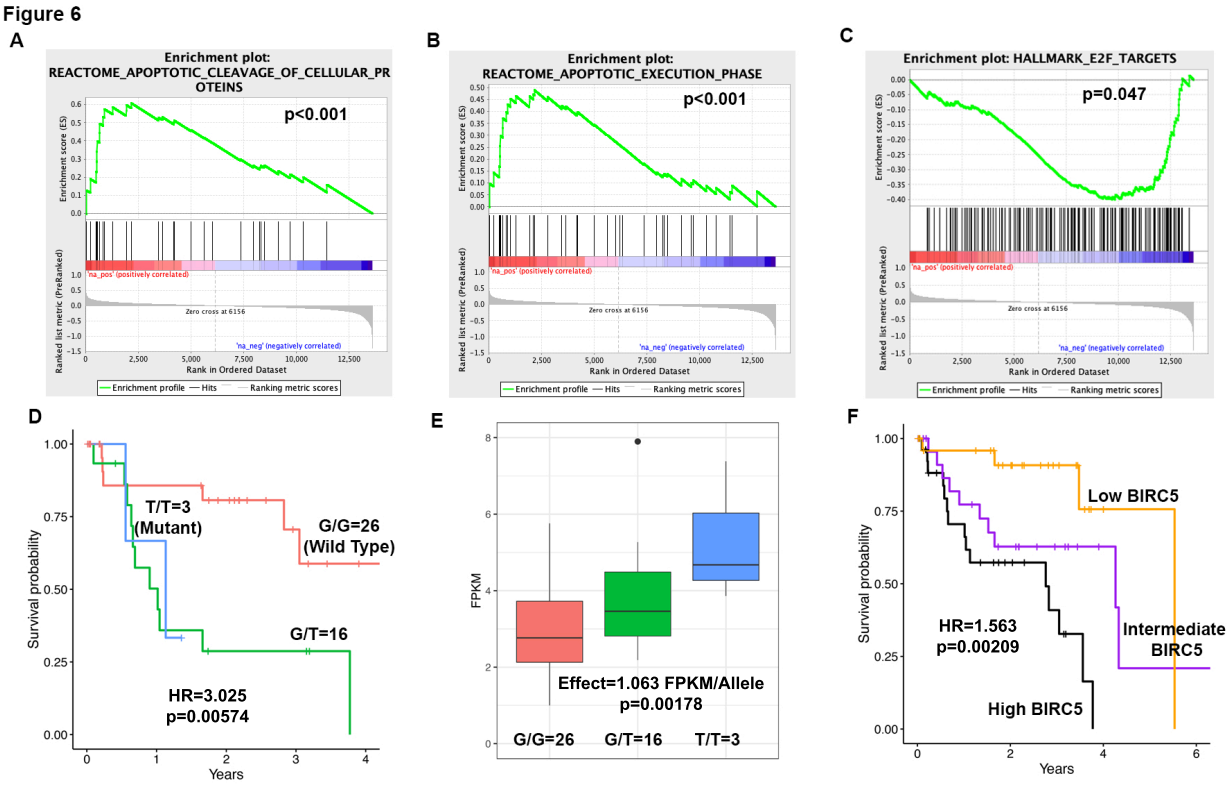
G

Variant	Gene	Gene Function	Cancer(s)	Association
rs3087404	SMUG1	Base excision repair	GBM, LGG	Depression
rs10759	RGS4	GTPase activating protein	GBM, LGG	Depression, Schizophrenia
rs9972327	IDH2	Krebs cycle	GBM, LGG	Alzheimer's Disease
rs12980648	FXYD5	Ion transport regulator	OV, UCEC	Endometriosis
rs2069398	CDK2	Cell cycle regulator	SKCM	Malignant Melanoma

Figure 6. Examples by which two of the prognostic germline variants may be associated with patient outcome.

A-C. rs55796947 in *MAP2K3/MKK3* is associated with favorable patient outcome in KIRC and results in complete loss of *MAP2K3*'s protein kinase domain due to a Q73* amino acid change. *MAP2K3* inhibition has previously been reported to result in cell cycle arrest and response to chemotherapy drugs. Tumors with the variant show upregulation of genes involved with apoptotic cleavage (**A**), genes in the apoptotic execution phase (**B**), and downregulation of E2F targets (**C**) in a Gene Set Enrichment Analysis (GSEA) of RNAseq data.

D-F. rs77903511 in the apoptosis inhibitor *BIRC5* is predictive of poor patient outcome in UVM (**D**). This variant is associated with increased *BIRC5* expression (**E**). Elevated *BIRC5* expression is associated with poor patient outcome (**F**).



Supplementary Figures

Figure S1. An overview of our approach to identifying prognostic germline variants. Whole exome sequenced normal (WXS Normal), whole exome sequenced tumor (WXS Tumor), and RNA sequenced tumor (RNA Tumor) samples from 10,582 cancer patients from The Cancer Genome Atlas (TCGA) were variant called. The three variant call sets were merged to create a single Combined variant call set that was used in the rest of the analysis. The variants were filtered to include only common variants that were concordant between the three sequencing datasets. We tested variants for an association with patient outcomes while controlling for clinical covariates using Cox regression models.

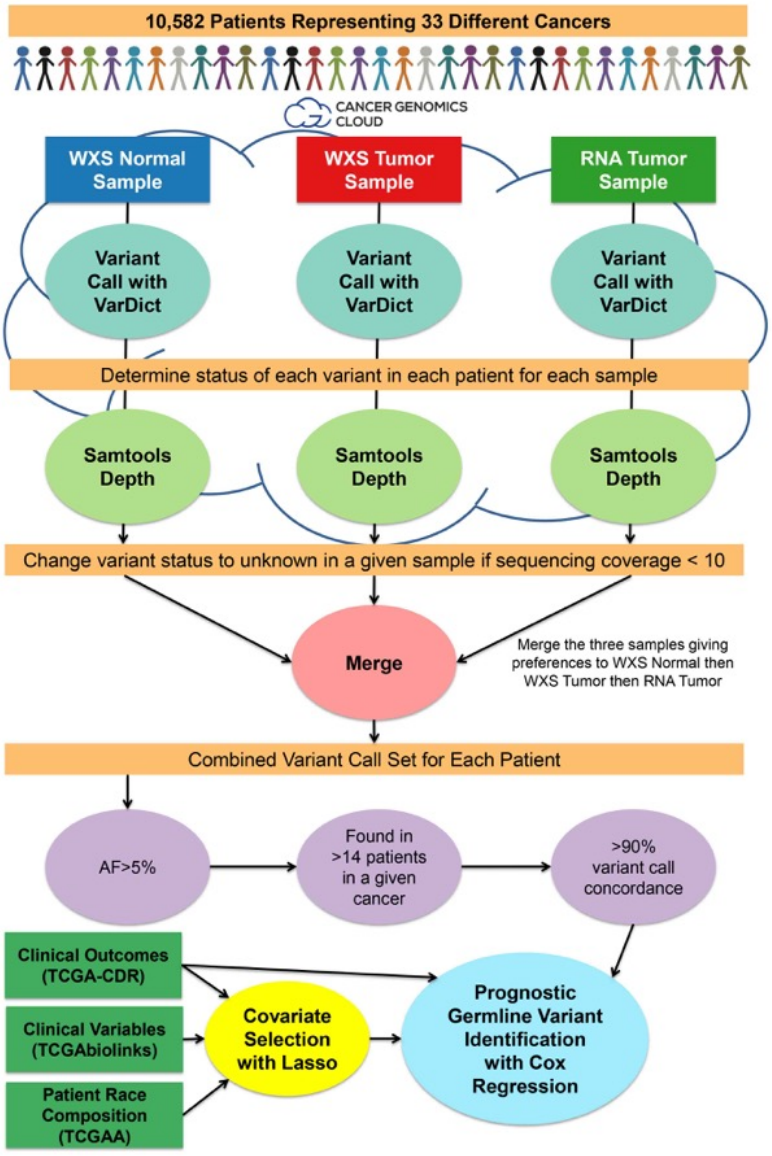


Figure S2. An overview of the total number of germline variants called and removed by the various filters included in this analysis. 519,319 germline variants were analyzed in this study.

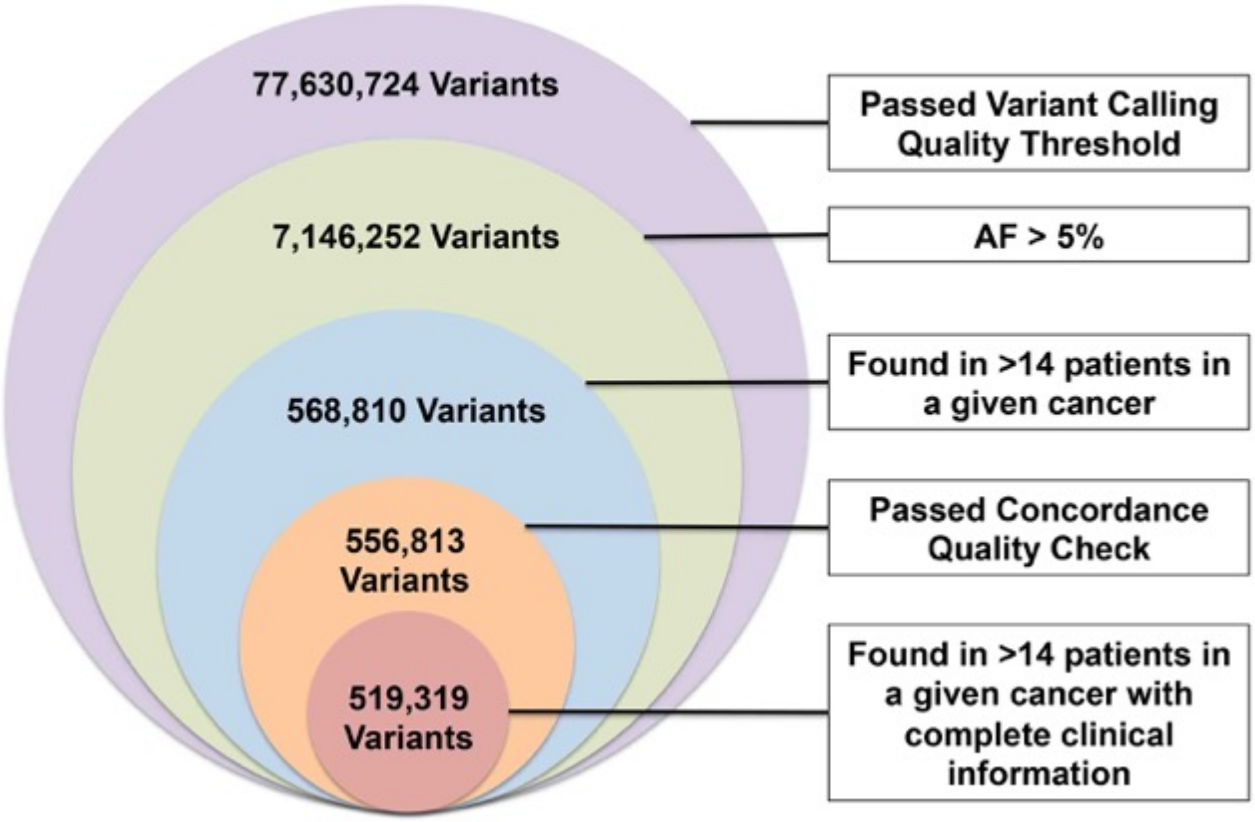


Figure S3. Somatic mutations did not compromise the integrity of this study. **A.** Most variants called from the tumor samples were germline variants. We plotted the percentage of variants called in the whole exome sequenced tumor (WXST) sample that were somatic mutations (SM) across all cancers. **B.** Few germline variants (GV) cause the same base change as a somatic mutation (SM) across all the cancers after filtering. **C.** Few germline variants (GV) included in this analysis overlap in genomic position with a somatic mutation (SM).

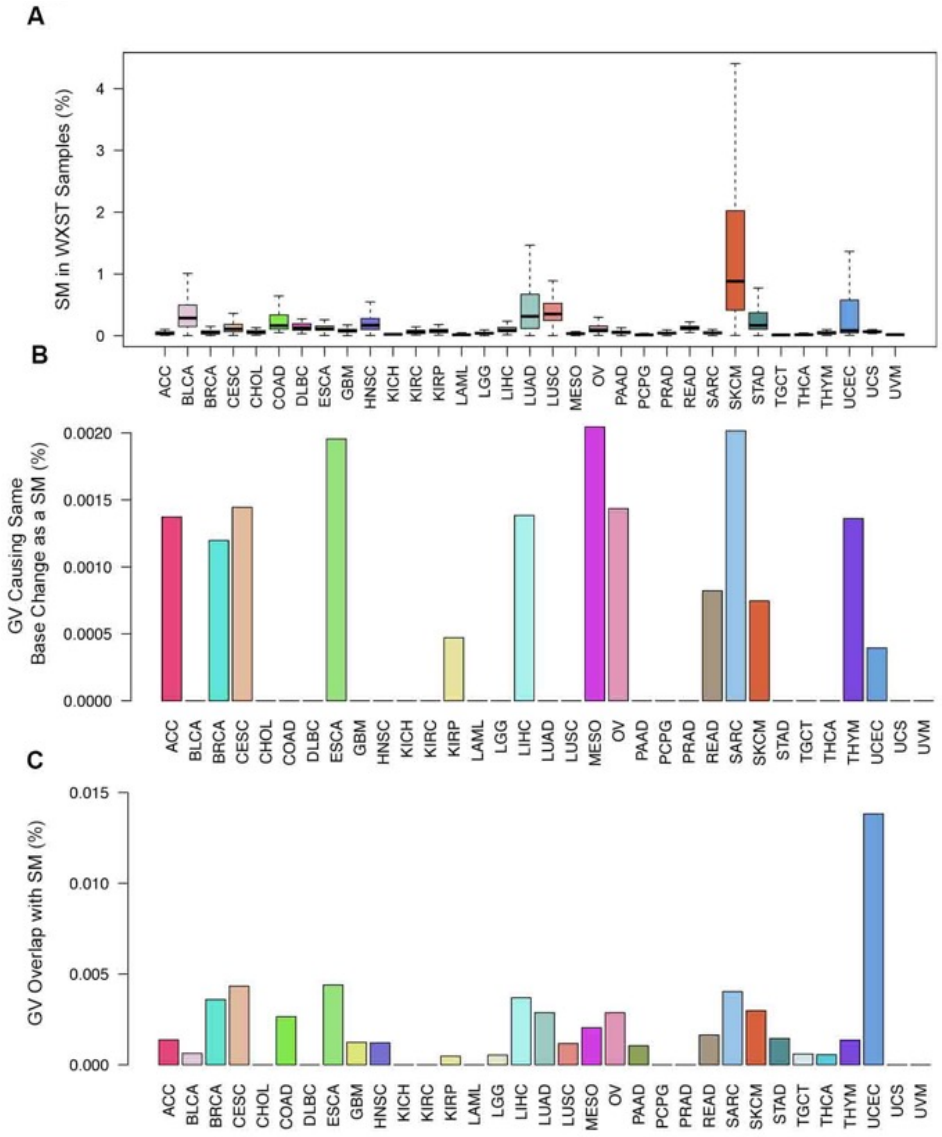


Figure S4. RNA editing did not affect the integrity of this analysis.
A. Few germline variants (GV) included in this study overlap with a known RNA editing site in genomic position.
B. Most germline variants are called in the whole exome sequenced samples (WXS). A relatively small number of germline variants were called solely from the RNA sequenced tumor (RNAT) sample.
C. The variant calls from the whole exome sequenced normal (WXSN), whole exome sequenced tumor (WXST), RNA sequenced tumor (RNAT), and Combined (the three variant call sets merged together) are highly concordant with each other. We calculated the allele frequency of each variant in each variant call set and calculated the Spearman correlation coefficient between all pairs.

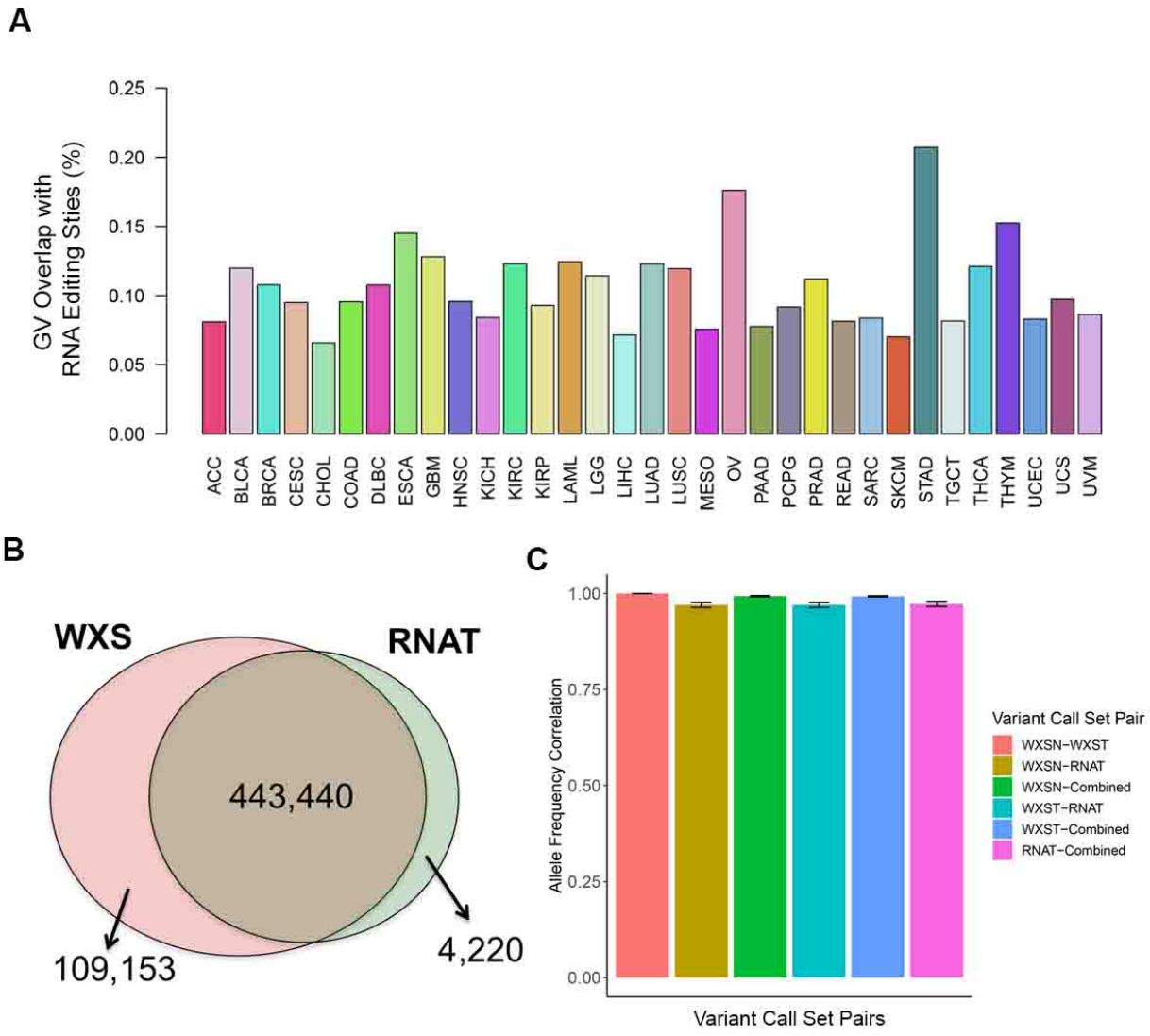


Figure S5. Power analysis results depicting the percentage of germline variants with >80% power to detect an association between variant status and patient outcome in individual cancers assuming varying effect sizes. To estimate our statistical power, we randomly sampled 10,000 germline variants in each cancer in each iteration and calculated our statistical power to detect an association between each germline variant and patient outcome. The results of this analysis separated the cancers out into three groups:

(1) Associations detectable at hazard ratios of moderate magnitudes of 2-3 (BLCA, BRCA, GBM, HNSC, KIRC, LGG, LUAD, LUSC, OV, SKCM, STAD, CESC, COAD, ESCA, LAML, LIHC, MESO, PAAD, PRAD, SARC, THCA, and UCEC)
(2) Associations detectable at hazard ratios of moderately high magnitudes of 4-5 (ACC, KIRP, READ, TGCT, UCS, PCPG, THYM, and UVM)
(3) Associations detectable at hazard ratios of high magnitudes (CHOL, DLBC, and KICH)

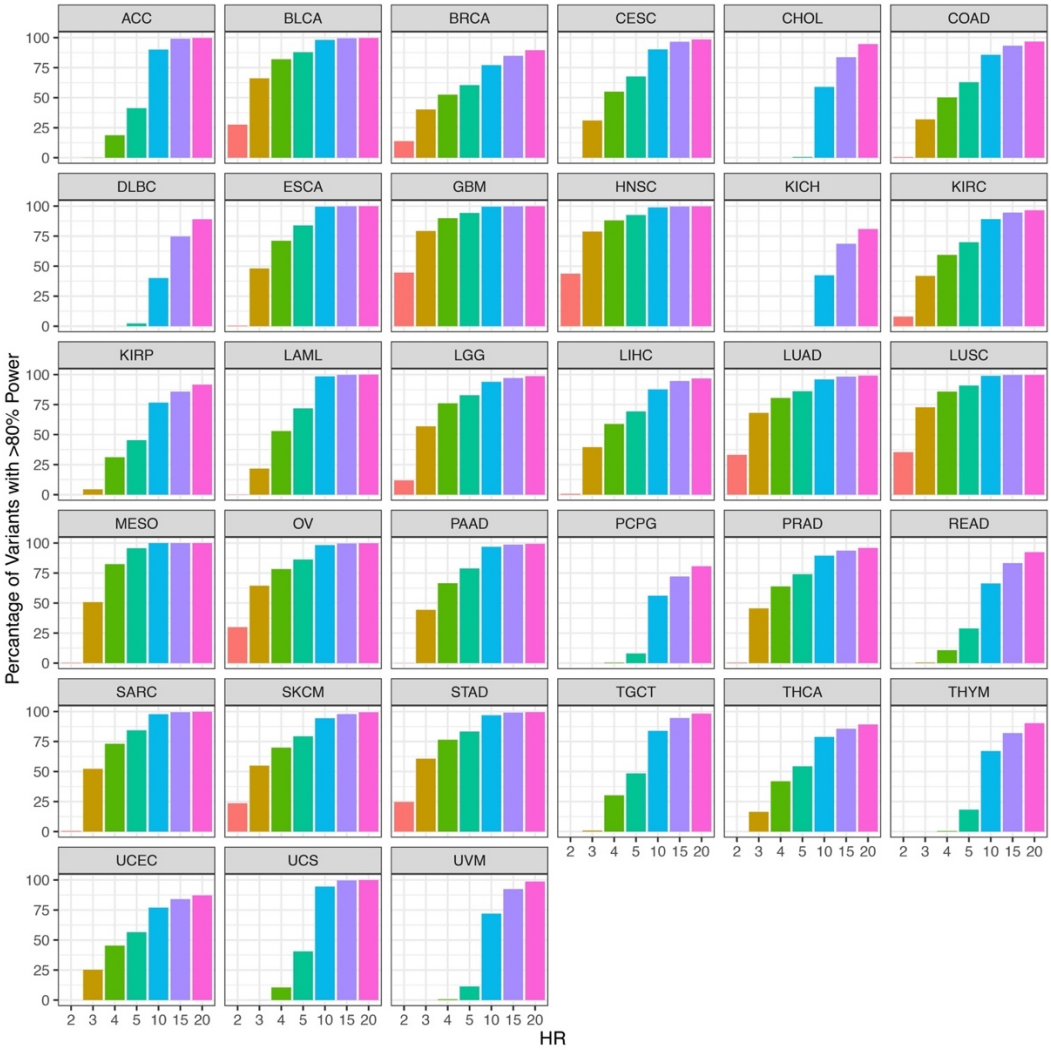


Figure S6. Selected Kaplan-Meier curves from the variants identified in Analysis 3 in which related cancers were grouped together prior to testing for association with survival.

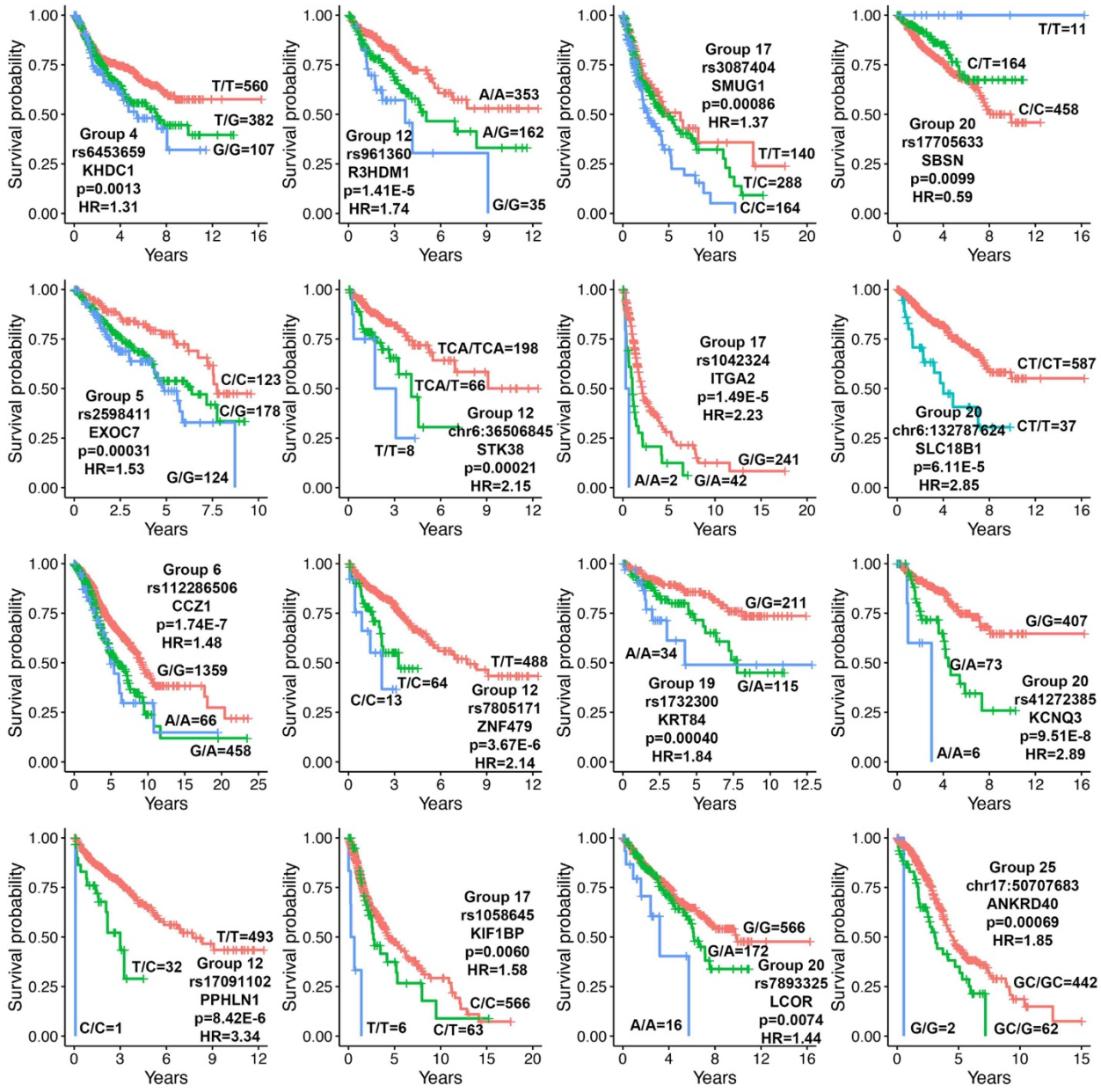
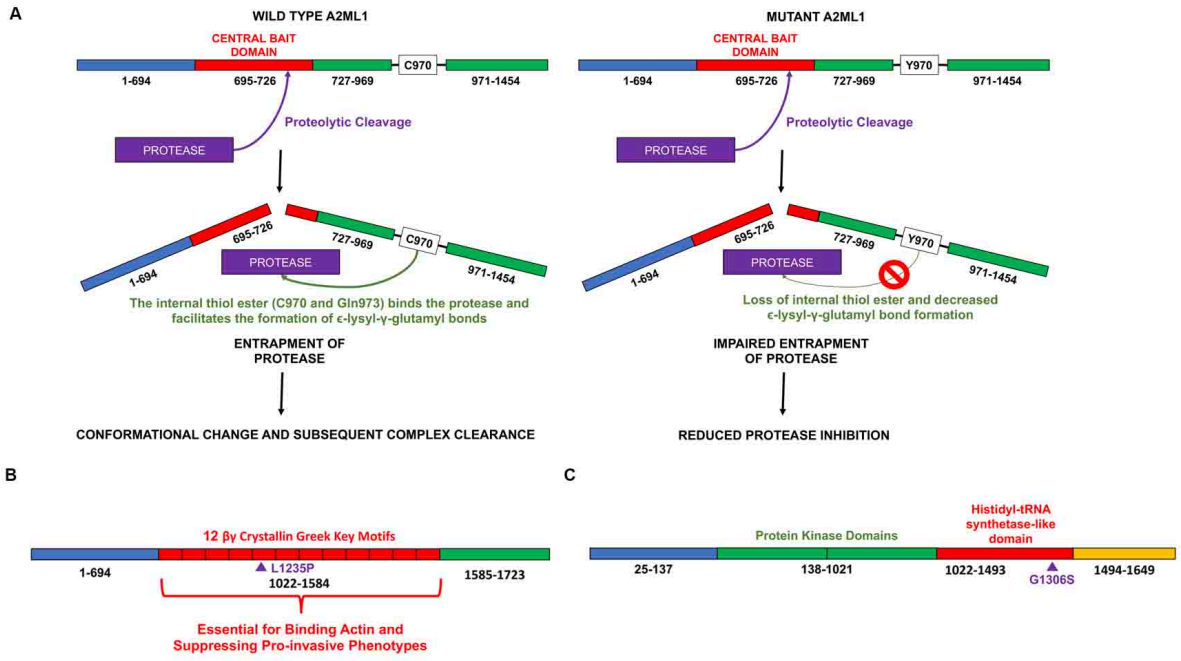


Figure S7. Schematic representations of how rs1558526, rs6174114, and rs35602605 may perturb well characterize protein domains.

A. rs1558526 is associated with favorable patient outcome in OV in the secreted protease inhibitor *A2ML1*. Wild type *A2ML1* inhibits proteases by forming a covalent bond following cleavage of its central bait domain (left). C970 facilitates the formation of this covalent bond. rs1558526 causes a C970Y amino acid change that likely disrupts *A2ML1*'s ability to inhibit proteases (right).

B. rs6174114 in *CRYBG1/AIM1* is associated with poor patient outcome in PAAD. The binding of *CRYBG1* to actin requires its 12 □□ crystallin motifs and results in suppression of pro-invasion phenotypes. rs6174114 causes a L1235P amino acid change in the fifth □□ crystallin motifs that may disrupt the packing of the beta sheets and perturb *CRYBG1*'s function, likely leading to increased tumor invasiveness and poor patient outcome.

C. rs35602605 in *EIF2AK4/GCN2* is associated with poor prognosis in THCA. *EIF2AK4* decreases translation of some proteins and increases translation of others (such as *CDKN1A*) under conditions of stress by binding uncharged tRNAs through its histidyl-tRNA-synthetase domain. rs35602605 results in a G1306S amino acid change in the histidyl-tRNA synthetase-like domain. This variant may disrupt the function of *EIF2AK4* resulting in poor patient outcome.



Supplementary Tables

Table S1. Clinical information about the patients included in this study and the covariates that we controlled for in our Cox regression models that were selected using Lasso-regularization.

Abbreviation	Cancer	Sample Size	Endpoint	Covariates
ACC	Adrenocortical carcinoma	91	OS	Age, Gender, Calculated Race, Stage
BLCA	Bladder Urothelial Carcinoma	410	OS	Age, Height, Stage
BRCA	Breast invasive carcinoma	1079	OS	Age, Estrogen Receptor Status
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	294	OS	Age, Histological Type, Calculated Race, Stage
CHOL	Cholangiocarcinoma	45	OS	Albumin Level, Calculated Race
COAD	Colon adenocarcinoma	441	OS	Age, Anatomic Position, Calculated Race, Stage
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	47	PFI	None
ESCA	Esophageal carcinoma	184	PFI	Histological Type, Anatomic Location, Weight
GBM	Glioblastoma multiforme	390	OS	Age, Chr 19/20 co-gain, Gender, IDH Mutation Status
HNSC	Head and Neck squamous cell carcinoma	523	OS	Age, Anatomic Location, Grade, Calculated Race, Stage
KICH	Kidney Chromophobe	65	PFI	Age, Stage
KIRC	Kidney renal clear cell carcinoma	530	OS	Age, Gender, Grade, Hemoglobin Level, Platelet Count, Calculated Race, Stage, White Blood Cell Count
KIRP	Kidney renal papillary cell carcinoma	286	OS	Stage
LAML	Acute Myeloid Leukemia	131	OS	Age, Cytogenetics Risk, Morphology
LGG	Brain Lower Grade Glioma	510	OS	1p/19q co-deletion status, Age, Chr 7 gain/Chr 10 Loss Status, Grade, IDH Mutation Status
LIHC	Liver hepatocellular carcinoma	369	OS	Age, Alcohol Consumption History, Fetoprotein Value, Grade, Platelet Count, Calculated Race, Stage
LUAD	Lung adenocarcinoma	506	OS	Stage
LUSC	Lung squamous cell carcinoma	497	OS	Age, Anatomic Location, Calculated Race
MESO	Mesothelioma	85	OS	Age, Histological Type
OV	Ovarian serous cystadenocarcinoma	523	OS	Age, Anatomic Location, Grade, Calculated Race, Stage

PAAD	Pancreatic adenocarcinoma	184	OS	Age, Anatomic Location, Gender, Grade, Calculated Race, Smoking History, Stage
PCPG	Pheochromocytoma and Paraganglioma	177	PFI	None
PRAD	Prostate adenocarcinoma	498	PFI	Anatomic Location, Gleason Grade, Calculated Race
READ	Rectum adenocarcinoma	163	PFI	Age, Gender, Calculated Race, Stage
SARC	Sarcoma	260	OS	Age, Pathology Margin Status, Postoperative Treatment, Residual Tumor
SKCM	Skin Cutaneous Melanoma	437	OS	Age, Breslow Depth Value, Calculated Race, Stage
STAD	Stomach adenocarcinoma	416	OS	Age, Anatomic Location, Grade, Stage, Calculated Race
TGCT	Testicular Germ Cell Tumors	134	PFI	Anatomic Location, History of Undescended Testis, Calculated Race, Stage
THCA	Thyroid carcinoma	505	PFI	Histological Type, Stage
THYM	Thymoma	122	PFI	None
UCEC	Uterine Corpus Endometrial Carcinoma	544	OS	Age, Grade, Height, Histological Type, Menopausal Status, Calculated Race, Stage, Total Pelvic Lymph Node Ratio, Total Pelvic Lymph Nodes Positive, Weight
UCS	Uterine Carcinosarcoma	56	OS	Hypertension, Residual Tumor, Total Pelvic Lymph Node Ratio, Tumor Invasion on Primary Pathology
UVM	Uveal Melanoma	80	OS	Age, Morphology, Tumor Diameter, Year of Diagnosis

Table S3. Justification for the groups presented in **Figure 1D**.

Group Number	Group	Group Description
1	ACC, KICH	Clustered by TCGA
2	ACC, PCPG	Adrenal Tumors
3	BLCA, CESC, HNSC, LUSC	Clustered by TCGA
4	BLCA, KICH, KIRC, KIRP	Urinary System
5	BLCA, KIRC, KIRP	Urinary System Without KICH
6	BRCA, OV, UCEC, UCS	Female Reproductive
7	CESC, HNSC, LUSC	Clustered by TCGA
8	CHOL, COAD, ESCA, LIHC, PAAD, READ, STAD	Gastro-intestinal
9	CHOL, LIHC	Bile Production and Storage
10	COAD, ESCA, PAAD, READ, STAD	Digestive System
11	COAD, ESCA, READ, STAD	Gastro-intestinal Tract
12	COAD, READ	Colon
13	COAD, READ, STAD	Lower Gastro-intestinal Tract
14	DLBC, LAML	Blood
15	DLBC, LAML, THYM	Immune System
16	DLBC, PCPG, SARC, THYM, UCS	Clustered by TCGA
17	GBM, LGG	Gliomas
18	GBM, LGG, PCPG	Neuro-endocrine and Gliomas
19	KICH, KIRC, KIRP	Kidney
20	KIRC, KIRP	Kidney without KICH
21	LAML, PRAD, THCA, THYM	Clustered by TCGA
22	LAML, THCA	Clustered by TCGA
23	LUAD, LUSC	Pulmonary without MESO
24	LUAD, LUSC, MESO	Pulmonary
25	OV, UCEC	Pelvic Female Reproductive
26	PAAD, STAD	GI Enzyme Production
27	PRAD, TGCT	Male Reproductive
28	SKCM, UVM	Melanoma
29	UCEC, UCS	Uterus

Supplementary Text

Text S1. The final set of germline variants included in this analysis are not substantially contaminated by somatic mutations or RNA editing.

Because the final variant call set was created by merging variant calls from WXS Normal, WXS Tumor, and RNA Tumor data, we evaluated our variant calls to ensure that they were not significantly contaminated by somatic mutations or RNA editing.

The total number of somatic mutations in each patient were obtained from the TCGA Research Network [47]. <2% of the total number of variants in a patient prior to any filtering or quality control were somatic mutations (**Figure S3A**). After filtering, <0.002% of germline variants in a given cancer included in this analysis caused the same base pair change as a somatic mutation (**Figure S3B**). In fact, <0.02% of germline variants included in this analysis in a given cancer even overlapped in position with a somatic mutation (**Figure S3C**). Therefore our final variant call set after filtering was not significantly contaminated by somatic mutations.

We next checked whether our variant call set was significantly affected by RNA editing. A set of over 2.5 million known RNA editing sites was identified from the rigorously annotated RNA editing database RADAR and overlapped with the germline variants included in this analysis [63]. <0.25% of germline variants in a given cancer included in this analysis overlapped in position with an RNA editing site (**Figure S4A**).

79.6% of germline variants were called in both the WXS and RNA samples, 19.6% were called only in the WXS samples, and 0.8% were called

only in the RNA samples (**Figure S4B).** Because a large number of germline variants were called in both the WXS and RNA samples, we were able to evaluate the concordance between the variant calls between the WXS Normal, WXS Tumor, and RNA Tumor samples. The allele frequency of each variant in each cancer in all four variant call sets (WXS Normal, WXS Tumor, RNA Tumor, and the three variant call sets combined) was calculated and correlated with each other. The allele frequencies in the four variant call sets were very well correlated with each other (**Figure S4C**), implying that the variant calls between the different samples were highly concordant. Taken together, these results suggest that somatic mutations, RNA editing, and pooling of the variant call sets did not lead to spurious germline variant calls.

Germline variant calling of all of the patients included in TCGA had previously been performed by Huang et al. [4]. We found that 93.0% of the variants called by Huang et al. were also found to have the same exact germline variant call in our analysis. For 1.5% of the variant calls there was disagreement between the two tools about whether an individual was heterozygous or homozygous for the alternate allele. 5.53% of the variants were called by GenomeVIP (Huang et al.'s tool) but not VarDict (our tool). <0.07% of the variants were called in VarDict but not GenomeVIP.

The concordance between the two germline variant call sets is quite strong, given the differences between the two studies. Huang et al. had performed variant calling on the WXS Normal samples aligned to hg19 and had performed variant calling using GenomeVIP, which integrates variant calls from

Varscan, GATK, and pindel, whereas our germline variant calls were generated using VarDict from the WXS Normal, WXS Tumor, and RNA sequenced tumor samples aligned to hg38 [33, 64-66]. Huang et al. implemented a variety of filtering criteria, including requiring an unfiltered allelic depth greater than 5 reads. We required a filtered (we excluded reads with a mapping quality less than 30 and base quality less than 25) read depth of 3 reads per sample and allele fraction of 5% The level of discordance that we found was expected, given the differences that could result from the usage of different reference genomes during alignment, filtering criteria, and variant calling tools [33].

Text S2. The results of our power analysis suggest that we can detect associations between germline variants with moderate to high effect sizes and patient outcome.

We evaluated our ability to detect significant associations between germline variants and patient outcome across the thirty-three cancers by calculating statistical power. The power to detect a significant association between a variant and patient outcome is dependent on multiple factors, including sample size, effect size, correlation with other covariates in the survival model, the number of patients with the germline variant, and the number of patients without the germline variant. To get a sense of our likelihood to detect associations across the thirty-three cancers at various effect sizes, we randomly sampled 10,000 germline variants from the pool of testable germline variants and calculated power for each germline variant at hazard ratios of 2, 3, 4, 5, 10, 15, and 20. The results are depicted in **Figure S5**.

The results suggest that our study design would enable us to detect associations beginning around a hazard ratio of 2. With that said, our power study suggests that for every germline variant that we are able to associate with patient outcome at lower hazard ratios, we will likely fail to detect several others due to having limited statistical power for variants with lower effect sizes, even in the cancers with the largest sample sizes. Future studies with larger sample sizes will be able to detect these associations that our current study will likely miss. Furthermore, it should be noted that even if germline variants fail to be associated with patient outcome, our study is not sufficiently powered to claim that those variants are not in reality associated with outcome. Finally, the results

suggest that we are extremely unlike to detect an association with germline variants with low to moderate effect sizes in ACC, CHOL, DLBC, KICH, PCPG, TGCT, THYM, UCS, and UVM.

Text S3. The direction (indicating whether a germline variant is associated with increased or decreased risk of poor outcome) and magnitude of the hazard ratio is correlated across cancers in which the germline variant is prognostic.

When looking at the set of variants associated with patient outcome in three or more cancers, we found that the direction of the hazard ratio for a given variant in different cancers in which it was prognostic (HR>1 implying that the variant is associated with increased risk of poor outcome or HR<1 implying that the variant is associated with decreased risk of poor outcome) was much more concordant ($p<2.2E-16$) than we expected based on random chance. Surprisingly, we even found the magnitude of the hazard ratio to be correlated across cancers. We identified the set of variants associated with favorable (HR<1) outcome and poor (HR>1) outcome in three or more cancers and found the hazard ratios estimated for a variant in different cancers to be correlated for both the variants associated with poor outcome (HR>1) (Spearman rho=0.146, $p=5.36E-157$) and variants associated with favorable outcome (HR<1) (Spearman's rho=0.185, $p=2.71E-101$). Because previous studies have reported a correlation between effect size of variants identified in GWAS and allele frequency, we considered whether this correlation may be confounded by the allele frequency of these variants [45]. After controlling for allele frequency, we still find a significant partial correlation after analyzing both the variants associated with increased risk of poor outcome (Spearman rho=0.0667, $p=4.024E-34$) and decreased risk of poor outcome (Spearman rho=0.0584, $p=2.274E-11$) variants. These findings reinforce the notion that the prognostic germline variants' effects tend to show some consistency across cancers.

Text S4. The alleles associated with increased risk of poor outcome of prognostic germline variants are more likely to be associated with somatic mutations in known cancer driver genes than the alleles of non-prognostic germline variants.

A previous study had identified germline variants that were associated with a significant increased incidence of somatic mutations in cancer related genes.[27] We therefore hypothesized that the prognostic variants were associated with an increased incidence of somatic mutations in driver genes in the cancer in which that variant was prognostic. To test this hypothesis, we created 353 germline variant-cancer pairs and determined the number of prognostic variants for which the allele associated with increased risk of poor outcome was associated with an increased incidence of somatic mutations relative to the protective allele. We repeated this analysis for all of the germline variants included in this analysis. We found that 47 of the 353 (13.3%) germline variant-cancer pairs were associated with an increased incidence of mutations in cancer driver genes which is more than expected by random chance (OR=1.89, p=0.0001).

Text S5. A detailed discussion of the twelve germline variants that cause significant amino acid changes.

To demonstrate that the prognostic germline variants identify genes that could be directly or indirectly linked to cancer progression, below we turn to the twelve germline variants in **Figure 5E** that caused substantial amino acid changes. Of these *MAP2K3* has been discussed in the main text.

A2ML1 is a secreted protease inhibitor that inhibits all classes of proteases. When proteases cleave the central bait domain of *A2ML1*, conformational changes cause an internal thiol ester, formed by C970 and Gln973, to become highly reactive. This thiol ester bond binds the protease and facilitates the formation of covalent bonds between *A2ML1* and the protease, resulting in protease entrapment and inhibition [67]. In our analysis, the germline variant rs1558526 was associated with favorable patient outcome in ovarian cancer patients and resulted in a C970Y change in *A2ML1*. Because the very cysteine residue that forms the internal thiol ester is lost, this amino acid change likely disrupts *A2ML1*'s protease inhibition function (**Figure S7A**). This result suggests that certain extracellular proteases which *A2ML1* may normally inhibit may have anti-tumor effects, for example by degrading angiogenic factors or anti-immune factors.

CRYBG1/AIM1 (absent in melanoma) is a protein that localizes to the cytoskeleton. Loss of *CRYBG1* in prostate cancer cells leads to increased G-actin (relative to F-actin), cell migration, invasion and soft agar colony formation. Binding of *AIM1* to actin requires the six C terminal domains made of 12 βγ crystallin motifs [68]. We found rs6174114 in *CRYBG1* to be associated with poor

patient outcome in pancreatic cancer. This variant changes L1235 to P in the fifth domain of *CRYBG1*. Substitution of proline at this position could disrupt the packing of the beta sheets that make a β or γ motif (**Figure S7B**), resulting in loss of *CRYBG1* function and therefore increase cell migration, invasion, and soft agar colony formation. This would explain the poor patient outcome associated with this germline variant. Somatic mutation or epigenetic suppression of *CRYBG1* has been seen in melanomas, lymphomas, and prostate carcinoma. Decreased expression of the protein associated with metastasis [68].

EIF2AK4/GCN2 is a protein kinase that is activated under stress by binding to uncharged tRNAs through its histidyl-tRNA-synthetase domain. This kinase is important for decreasing protein translation and for activating specific translation of genes like *ATF4* and *p21/CDKN1A* under conditions of stress often seen inside tumors like amino acid starvation and glucose starvation. We found the germline variant rs35602605 in *EIF2AK4* to be associated with poor prognosis. This variant causes a G1306S amino acid change in the histidyl-tRNA synthetase-like domain (**Figure S7C**). This variant may disrupt the ability of the histidyl-tRNA synthetase-like domain to bind uncharged tRNAs and thereby protect the cancer cells from translation of stress-induced genes like *CDKN1A* that restrain tumor proliferation. If true, this would explain the association of this germline variant with poor patient outcome.

The other gene-products identified by prognostic variants in **Figure 5E** also warrant a detailed examination. Two of them could be important for immune response to a tumor. *FCRL6* binds to MHC class II proteins and acts as an

immune checkpoint protein that is often upregulated in Tumor infiltrating lymphocytes [69]. It is particularly interesting that *FCRL6* expression of T lymphocytes is decreased five-fold in acute and chronic myeloid leukemias because the rs61823162 variant (which truncates the protein) is associated with outcome in LAML [70]. *EPHA10* is a non-functional tyrosine kinase receptor for ephrins. The G749E mutation is located in the tyrosine kinase domain, which upregulates PD-L1 protein expression [71]. Three genes are involved in intracellular vesicle transport, membrane fusion and cell migration: *BORCS5* recruits the *ARL8B* GTPase to lysosomes for lysosomal movement and function, *KDELR3* is involved in retaining proteins in the endoplasmic reticulum, and *MYOF* facilitates vesicle fusion. Two are involved in GPCR pathways: *OR10X1 is* an olfactory receptor and *SAG/arrestin1* binds to GPCRs (such as rhodopsin) to terminate signaling. Many olfactory receptors are ectopically expressed in several cancer and their activation decreases cancer cell proliferation and migration and increases apoptosis [72, 73]. The I-76 of *SAG* that is altered by the variation is located in the highly conserved finger loop of motif 2, (E/D)x(I/L)xxxGL, which is extended and buried in the rhodopsin (GPCR)-SAG interface [74]. Finally *ECD/SGT1* associates with many cellular proteins relevant for cancer, *MDM2*, *Rb*, *HSP90*, *SKP1*, and *RUVBL1*, the last in particular using the C-terminal region of *ECD* that is mutated in the prognostic variant.

Supplementary References

1. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, et al: **Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines.** *Cell Syst* 2018, **6:**271-281.e277.

2. Ramaswami G, Li JB: **RADAR: a rigorously annotated database of A-to-I RNA editing.** *Nucleic Acids Res* 2014, **42:**D109-113.

3. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al: **Pathogenic Germline Variants in 10,389 Adult Cancers.** *Cell* 2018, **173:**355-370.e314.

4. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res* 2010, **20:**1297-1303.

5. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet* 2011, **43:**491-498.

6. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al: **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.** *Curr Protoc Bioinformatics* 2013, **43:**11.10.11-33.

7. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR: **VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research.** *Nucleic Acids Res* 2016, **44:**e108.

8. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Jr., Chatterjee N: **Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants.** *Proc Natl Acad Sci U S A* 2011, **108:**18026-18031.

9. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, et al: **Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer.** *Cancer Discov* 2017, **7:**410-423.

10. Galliano MF, Toulza E, Gallinaro H, Jonca N, Ishida-Yamamoto A, Serre G, Guerrin M: **A novel protease inhibitor of the alpha2-macroglobulin family expressed in the human epidermis.** *J Biol Chem* 2006, **281:**5780-5789.

11. Haffner MC, Esopi DM, Chaux A, Gurel M, Ghosh S, Vaghasia AM, Tsai H, Kim K, Castagna N, Lam H, et al: **AIM1 is an actin-binding protein that suppresses cell migration and micrometastatic dissemination.** *Nat Commun* 2017, **8:**142.

12. Johnson DB, Nixon MJ, Wang Y, Wang DY, Castellanos E, Estrada MV, Ericsson-Gonzalez PI, Cote CH, Salgado R, Sanchez V, et al: **Tumor-specific MHC-II expression drives a unique pattern of resistance to immunotherapy via LAG-3/FCRL6 engagement.** *JCI Insight* 2018, **3**.

13. Kulemzin SV, Zamoshnikova AY, Yurchenko MY, Vitak NY, Najakshin AM, Fayngerts SA, Chikaev NA, Reshetnikova ES, Kashirina NM, Peclo MM, et al: **FCRL6 receptor: expression and associated proteins.** *Immunol Lett* 2011, **134:**174-182.

14. Yang WH, Cha JH, Xia W, Lee HH, Chan LC, Wang YN, Hsu JL, Ren G, Hung MC: **Juxtacrine Signaling Inhibits Antitumor Immunity by Upregulating PD-L1 Expression.** *Cancer Res* 2018, **78:**3761-3768.

15. Weber L, Al-Refae K, Ebbert J, Jagers P, Altmuller J, Becker C, Hahn S, Gisselmann G, Hatt H: **Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis.** *PLoS One* 2017, **12:**e0172491.

16. Ranzani M, Iyer V, Ibarra-Soria X, Del Castillo Velasco-Herrera M, Garnett M, Logan D, Adams DJ: **Revisiting olfactory receptors as putative drivers of cancer.** *Wellcome Open Res* 2017, **2:**9.

17. Peterson YK, Luttrell LM: **The Diverse Roles of Arrestin Scaffolds in G Protein-Coupled Receptor Signaling.** *Pharmacol Rev* 2017, **69:**256-297.

Chapter 4: A Pan-Cancer Analysis of Germline Variants Associated with Increased Tumor Mutational Burden

Ajay Chatrath, Aakrosh Ratan, and Anindya Dutta

- I conceived of the idea and design for this project, wrote the code for this analysis, analyzed the data, wrote the original draft of the manuscript, and made all of the figures in this manuscript.

Author Contributions

Conceptualization, A.C., A.D.; Methodology, A.C., A.R., A.D.; Software, Formal Analysis, Investigation, Writing – Original Draft, Visualization, and Data Curation, A.C.; Resources and Funding Acquisition, A.D.; Writing – Review & Editing, all authors; Supervision and Administration, A.D., A.R.

A Pan-Cancer Analysis of Germline Variants Associated with Increased Tumor Mutational Burden

Ajay Chatrath[1], Aakrosh Ratan[2], Anindya Dutta[1]

[1] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia

[2] Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia

Financial Support

This work was supported by grants from the NIH R01 CA166054, R01 CA60499, and T32 GM007267 (AC).

Potential Conflicts of Interest:
The authors do not have any conflicts of interest to disclose.

Abstract

Although rare genetic syndromes have historically been challenging to study and treat, the explosion of next generation sequencing data has made the study of these syndromes much more feasible. While patients with certain rare genetic syndromes are at higher risk for acquiring cancer and are therefore screened for cancer more aggressively, clinical management guidelines for patients with pathogenic germline variants after acquiring cancer are only now beginning to change. In this study, we identify pathogenic germline variants associated with tumor hypermutation by grouping them by gene or by pathway, as a proxy for identifying germline markers of immune checkpoint inhibitor efficacy, as immune checkpoint inhibitor responsiveness has been strongly correlated to overall tumor mutational burden. We identified an association with overall tumor mutational burden in nine genes (*APC*, *FANCL*, *SLC25A13*, *ERCC3*, *MSH6*, *PMS2*, *TP53*, *MSH2*, and *BRIP1*) using a pan-cancer approach, fourteen pathways in individual cancers, and twelve pathways using a pan-cancer approach. We also report evidence of the effects of the pathogenic germline variants on the cells, suggesting that these germline variants affect how the tumor progresses and not just tumor risk. Patients with pathogenic germline variants in *APC* or genes related to beta-catenin degradation exhibit upregulation of genes in the tumors involved with *Wnt* signaling and patients with pathogenic germline variants in genes regulating cell cycle checkpoint exhibit upregulation of *E2F* targets and mitotic spindle genes in the tumors. We found tumor mutational signatures concordant with the expected effects of pathogenic germline variants

in pathways related to mismatch repair, nucleotide excision repair, and homologous recombination. Our findings suggest that tumors of patients with the identified pathogenic germline variants have increased tumor mutational burden compared to tumors of patients without these germline variants. Patients with the pathogenic germline variants described in this study may be more likely to respond to immune checkpoint blockade because of the expected increase in tumor neoantigens.

Introduction

Rare genetic syndromes have historically been challenging to study and treat due to challenges associated with finding enough patients to sufficiently power cohort studies, discouraging companies to invest in drug development for rare diseases due to predicted lack of profitability. While individually these diseases are rare, collectively these diseases affect millions of individuals worldwide, leaving many patients undiagnosed or without treatment [1, 2].

The explosion of next generation sequencing data has helped to identify rare germline variants that cause or contribute to these rare genetic syndromes [3, 4]. In oncology, it is well-established that patients with germline variants in genes mutated in certain genetic syndromes, such as Lynch syndrome, Li-Fraumeni syndrome, Von Hippel-Lindau syndrome, and Fanconi anemia, are at much higher risk of acquiring cancer [5, 6]. While individuals with these pathogenic germline variants are generally screened more aggressively, clinical management for patients with these pathogenic germline variants is occasionally, but not always, differentiated from that of patients without pathogenic germline

variants [7-9]. Patients with Lynch syndrome have pathogenic germline variants in mismatch repair genes, such as *MSH2*, *MSH6*, *PMS2*, and *MLH1*. Patients with pathogenic germline variants in mismatch repair genes exhibit higher levels of microsatellite instability and it has been well documented that patients with high tumor mutational burden have been shown to be more likely to respond to immunotherapy drugs such as pembrolizumab [10, 11]. As expected, patients with Lynch syndrome are more likely to respond to treatment with immune checkpoint blockade [12].

We have previously suggested that germline variants affect tumor progression across a large spectrum of cancers through the analysis of common germline variants with an allele frequency greater than 5% in the population [13, 14]. In this study, we analyze rare, pathogenic germline variants to identify germline variants associated with increased tumor mutational burden, as these germline variants may increase the likelihood of a patient responding to immune checkpoint blockade.

Methods

Patient Data Availability

We downloaded the set of rare, pathogenic germline variants found in the patients in The Cancer Genome Atlas (TCGA) previously published by Huang et al. and the set of somatic mutations in these patients generated by Ellrott et al. [5, 15]. Overall tumor mutational burden for each patient was determined by counting the total number of somatic mutations found in the primary tumor

sample of each patient. Clinical data for the TCGA patients was accessed from the TCGA pan-cancer clinical data resource [16].

Identification of Individual Genes Associated with Tumor Hypermutation

Across all of the TCGA patients, 132 unique genes contained at least one pathogenic germline variant. We limited our analysis only to genes with pathogenic germline variants in at least five different patients. As a result, we decided not to test individual genes in individual cancers since this criteria would only be met for 13 genes. Instead, we pooled all of the TCGA patients together and tested whether individual genes perturbed by pathogenic germline variants (presence or absence of a pathogenic germline variant) were associated with overall tumor mutation burden using linear regression, controlling for tumor type. We tested a total of 73 unique genes in this analysis. P-values were adjusted using the Benjamini-Hochberg procedure throughout this study.

Identification of Pathways Associated with Tumor Hypermutation

To study the association between pathogenic germline variants and tumor mutational burden in individual cancers, we grouped genes by pathways. We tested pathways perturbed by pathogenic germline variants in five or more patients. We downloaded pathway annotation information from Reactome [17]. We tested whether having a pathogenic germline variant in the pathway (presence or absence) was associated with overall somatic mutation burden using linear regression in individual cancers. We tested a total of 117 unique pathways. Finally, we performed a pan-cancer analysis of pathway association

with tumor hypermutation using the same approach, while also controlling for tumor type. We tested a total of 454 unique pathways in this analysis.

While each gene set itself was unique, some of the gene sets entirely overlapped with each other in this analysis based on the genes that the pathogenic germline variants were found in. For example, suppose gene set 1 contains genes A, B, and C and gene set 2 contains genes B, C, and D. While these two gene sets are unique, if pathogenic germline variants are only found in genes B and C, then the resulting statistical test results of gene sets 1 and 2 will be exactly the same meaning our approach lacks the resolution to distinguish between these two gene sets. To address this issue, we have reported all gene sets in **Table 1** that entirely overlap for each statistical test that yielded a significant p-value (adjusted p-value < 0.05).

Gene Set Enrichment Analysis

As part of our analysis, we found that pathogenic germline variants in *APC*, genes involved with the degradation of beta-catenin, and cell cycle checkpoint genes were associated with elevated tumor mutational burden. *APC* forms a complex with beta-catenin, an intracellular signaling transducer of *Wnt* signaling, resulting in the degradation of beta-catenin. We hypothesized that the perturbation of *APC* and genes involved with the degradation of beta-catenin would result in upregulation of genes involved with *Wnt* signaling [18]. Similarly, we hypothesized that perturbation of genes involved with cell cycle checkpoint would result in upregulation of E2F targets and genes involved with mitotic spindle activity.

We tested these hypotheses by performing gene set enrichment analyses. We downloaded the previously released RNA-sequencing quantification files for each patient generated by the TCGA research network (https://portal.gdc.cancer.gov/). We then excluded genes with a median expression level less than 1 FPKM across the patient cohort being tested. The expression values of the remaining genes were then normalized by mean and standard deviation. We ranked the genes from induced to repressed by testing for an association between the expression of each gene and the presence of a pathogenic germline variant in the gene or pathway using logistic regression, controlling for tumor type. We used these ranked gene lists to perform Gene Set Enrichment Analysis [19].

Mutational Signature Analysis

Having identified pathogenic germline variants perturbing well-known DNA repair genes associated with overall somatic burden, we hypothesized that the mutational signatures corresponding to these DNA repair genes would be enriched in patients with these pathogenic germline variants compared to patients without these germline variants. To test this possibility, we downloaded all single base substitution signatures from COSMIC [20]. We determined the optimal contribution of COSMIC signatures to reconstruct the mutational profile observed in each of the patients in TCGA using the R package "MutationalPatterns" [21]. We converted the contribution values to fraction of the total set of somatic mutations explained for that patient, such that the sum of the

fraction contributions of all the COSMIC signatures for each patient was equal to 1.

Because the etiology of many of the COSMIC signatures is known, we were able to generate hypotheses for the mutational signatures that we expected to be enriched in each group of patients. For example, we had found that patients with pathogenic germline variants in *MSH6* had tumors with higher somatic mutation burden. *MSH6* is a well-characterized gene involved with DNA mismatch repair, so we hypothesized that COSMIC signatures related to DNA mismatch repair deficiency would be enriched, such as COSMIC signatures 6, 15, 20, 26, and 44. We evaluated this hypothesis by testing for an association between the fractional contribution of a signature and the presence or absence of pathogenic germline variants perturbing a gene or pathway, controlling for tumor type. We also controlled for the presence or absence of deleterious somatic mutations in the gene or pathway being tested to partially isolate the effect of the germline variant itself apart from any somatic mutations that it may predispose a patient to. We defined deleterious somatic mutations as somatic mutations marked as “probably damaging” by the TCGA research network [5].

Increased Susceptibility to Mutations in Cancer Driver Genes and in Genes in the Same Pathway as the Original Pathogenic Germline Variant

Having identified pathogenic germline variants that predispose patients to increased overall tumor mutational burden, we asked whether or not the pathogenic germline variants were associated with the somatically mutated genes that the were ultimately seen in the tumor. We downloaded the list of driver genes in each cancer released by Bailey et al. [22]. We calculated the

number of “probably damaging” somatic mutations in driver genes in each patient. The set of driver genes reported by Bailey et al. differs across cancers [22]. We tested for an association between the log-transformed number of “probably damaging” somatic mutations in cancer driver genes and the presence or absence of a pathogenic germline variant in a gene or pathway, controlling for tumor type and total tumor mutational burden in that patient. We controlled for total tumor mutational burden because otherwise we would find an enrichment of deleterious somatic mutations across many classes of genes, not just driver genes. This enabled us to test whether the number of deleterious somatic mutations in the cancer driver genes was enriched more than we would expect in patients with pathogenic germline variants given the patients’ overall tumor mutational burden.

We repeated this approach for testing for enrichment of “probably damaging” somatic mutations in the same pathway as the gene or pathway affected by the pathogenic germline variants. When performing this test with our individual gene association, we tested all pathways in which that gene was found. When performing this test with our pathway associations, we only tested for an association with somatic mutations in that particular pathway.

Software

Computation was performed using R version 3.5.2. The R packages “ggplot2” and “scatterpie” were used to generate the figures in this manuscript.

Results

Huang et al. had previously described the set of rare, pathogenic germline variants found in the patients in The Cancer Genome Atlas [15]. The majority of these pathogenic germline variants were predicted to functionally perturb known tumor suppressor genes or oncogenes. Prior to identifying which pathogenic germline variants contribute to elevated tumor mutational burden, we had to address the problem raised by the low frequency of the variants in the study population. While 132 unique genes contained pathogenic germline variants, we recognized that only 13 of them could be analyzed in individual cancers with a modest threshold requiring at least five patients in the cancer cohort carrying the given variant. As a result, this approach could only be applied to a handful of cancers (**Figure 1A**). We therefore felt that we were unable to study most genes containing a pathogenic germline variant using this approach.

We therefore increased the number of patients with related germline variants by three approaches. (1) We pooled all of the patients in The Cancer Genome Atlas (TCGA) together, and, by doing so, were now able to test 73 total genes (**Figure 1B**). (2) We grouped the pathogenic germline variants by pathway in individual cancers (**Figure 1C**). (3) We grouped the pathogenic germline variants by pathway and then studied all the cancers grouped together (**Figure 1D**). Our overall methodology is summarized in **Figure 2**.

Identification of Individual Genes Associated with Tumor Hypermutation

In the first analysis we grouped the pathogenic germline variants based on the gene they were found in and tested each gene for association with overall

tumor mutational burden using all patients in TCGA, but controlling for tumor type. This identified nine genes that when perturbed by a pathogenic germline variant were associated with elevated tumor mutational burden (**Figure 3A, Table 1A**). Three of these genes (*APC*, *FANCL*, and *SLC25A13*) were significant after correcting for multiple hypothesis testing. We further characterized the other significant associations later in this study ($p<0.05$) despite them not reaching the multiple hypothesis corrected p-value cut-off because all of these genes (*ERCC3*, *MSH2*, *MSH6*, *PMS2*, *BRIP1*, and *TP53*) have well-known roles in DNA repair.

Identification of Pathways in Individual Cancers Associated with Tumor Hypermutation

We next grouped pathogenic germline variants in individual cancers by pathway. We tested each germline variants in each pathway for association with overall tumor mutational burden in each of the individual cancers. This identified significant increases in tumor mutational burden in COAD, ESCA, and KIRC due to germline variants in specific pathways (**Figure 3B**, **Table 1B**). While each of the annotated pathways consisted of different and unique gene sets, the genes that empirically contributed to these gene sets sometimes overlapped in this analysis. We have therefore grouped pathways for which the contributing genes entirely overlapped in this particular analysis. In total, we identified 14 associations (1 in COAD, 6 in ESCA, and 7 in KIRC). The significantly associated pathways were primarily related to DNA damage repair and cell cycle control.

Pan-Cancer Identification of Pathways Associated with Tumor Hypermutation

Lastly, we identified pathogenic germline variants in pathways which were associated with elevated tumor mutational burden using a pan-cancer approach, controlling for tumor type as in Analysis 1 (**Figure 3C**, **Table 1C**). In total, we identified twelve significant associations. Four of the gene sets were related to *Wnt* signaling. The pathogenic germline variants in *APC* greatly contributed to these associations, as described in our analysis of individual genes. One association was driven entirely by *SLC25A13* and had also been described in the first analysis with individual genes. Two associations were with pathways related to apoptosis and two other associations were in pathways indicating deficiencies in mismatch repair.

Pathogenic Germline Variants that Predict Increased Tumor Mutational Burden Predict Changes in the Transcriptome in the Corresponding Tumors

Our results suggest that the pathogenic variants not only increase the risk for cancer, as has been previously shown [15], but may also contribute to a patient's tumor having a higher tumor mutational burden than that of a patient without a pathogenic germline variant. In order to support this hypothesis, we searched for other evidence that the pathogenic germline variants affected tumor phenotype, beginning with changes in the transcriptome.

Patients with pathogenic germline variants in *APC* went on to develop tumors with higher tumor mutational burden. *APC* is a well-known negative regulator of beta-catenin, a signal transducer in the *Wnt* signaling pathway [18,

23]. Indeed, we found widespread upregulation of the genes involved with *Wnt* signaling in patients with pathogenic germline variants in *APC* compared to patients without pathogenic germline variants in *APC* ($p<0.001$) in a gene set enrichment analysis. When germline variants are pooled by pathways, we also find that patients with pathogenic germline variants in the *Wnt* signaling pathway exhibit higher tumor mutational load than patients without such variants. We again find that genes involved in *Wnt* signaling are upregulated in tumors of these patients with germline variants in the *Wnt* signaling pathway (**Table 2**).

In Analysis 2, we found that patients with pathogenic germline variants in genes related to cell cycle checkpoint control exhibit high tumor mutational burden. In a gene set enrichment analysis, *E2F* targets and genes related to the mitotic spindle function were upregulated in these patients. This suggests a de-regulation of the cell cycle transcriptional program in tumors of these patients with pathogenic germline variants in cell cycle checkpoint genes (**Table 2**).

Collectively, these results support the hypothesis that specific germline variants that affect the tumor mutational burden can also affect other tumor phenotypes like the gene expression profile. The changes in the tumor gene expression profile could hint at phenotypes that explain the increased tumor mutational burden.

Pathogenic Germline Variants that Predict Increased Tumor Mutation Burden Predict an Enrichment of Expected Mutation Signatures

Previous work has shown that patients with certain germline variants are at higher risk for specific somatic mutations in the tumor [24]. Given that certain germline variants predict increased tumor somatic mutation, we hypothesized

that those variants in genes or pathways associated with DNA repair will also predispose the tumors to certain specific mutational signatures.

To get at this question, we first estimated what fraction of each patient's somatic mutation profile is explained by each of the previously reported COSMIC mutational signatures. We next tested whether there was an enrichment of specific mutational signatures corresponding to the expected effect of the pathogenic germline variants.

We first did this analysis with individual genes predicted to increase the tumor mutational burden. When comparing patients with pathogenic germline variants to those without, we found enrichment of mismatch repair signatures in patients with pathogenic germline variants in the mismatch repair genes *PMS2* (COSMIC signature 44: p=0.032), *MSH2* (COSMIC signature 6: p=0.017, COSMIC signature 15: 1.22E-5), and *MSH6* (COSMIC signature 6: p=0.024, COSMIC signature 15: p=0.0091). We also found enrichment in a transcription-coupled nucleotide excision repair signature in patients with pathogenic germline variants in the nucleotide excision repair gene *ERCC3* (COSMIC signature 29, p=0.034) (**Table 3**).

Upon repeating this analysis using our results from pathogenic germline variants grouped by pathways in individual cancers, we found an enrichment of homology directed repair deficiencies in patients with pathogenic germline variants in pathways related to homology directed repair (COSMIC signature 3: p=3.40E-4) and DNA double strand break repair (COSMIC signature 3: p=0.00267). Finally, in the pan-cancer analysis at the level of pathways, we

found enrichment in mismatch repair signatures in patients with pathogenic germline variants in pathways related to mismatch repair (COSMIC signature 6: p=0.00285, COSMIC signature 15: p=0.000127) and diseases of mismatch repair (COSMIC signature 6: p=0.00207, COSMIC signature 15: p=0.000296) (**Table 3**).

Thus, as hypothesized, germline variants in genes or pathways that predict increased tumor mutational burden, often also predict the enrichment of mutation signatures that are expected from our knowledge of the DNA repair functions of these genes and pathways.

Increased Risk for Somatic Mutations in Driver Genes

Having identified pathogenic germline variants associated with tumor hypermutation and obtained evidence at the level of the transcriptome and somatic mutation profile that the pathogenic germline variants influenced the molecular features of tumors, we next wondered whether the presence of pathogenic germline variants affected the genes perturbed by somatic mutations.

Not surprisingly, because the overall tumor mutational burden was higher in patients with the pathogenic germline variants that we identified, these patients were at higher risk for somatic mutations in driver genes and genes in the same pathway as the pathogenic germline variant (data not shown). To account for this, we reperformed these analyses controlling for each patients' overall tumor mutational burden.

We first tested the nine individual genes predisposing to increased tumor mutational burden identified using the pan-cancer approach. Patients with

pathogenic germline variants in *MSH2* (effect size = 0.806 additional somatic mutations in driver genes, adjusted p-value = 0.0216) and *MSH6* (effect size = 0.483 additional somatic mutations in driver genes, adjusted p-value = 0.00776) had more deleterious somatic mutations in driver genes, controlling for tumor type and total tumor mutational burden (**Table 4A**). We did not find an enrichment of deleterious somatic mutations in driver genes from pathways predisposing to increased tumor mutational burden, after controlling for tumor mutational burden in each sample.

Increased Risk for Somatic Mutations in the Same Pathway

Alfred Knudson's classic two-hit hypothesis stated that many genes, particularly tumor suppressor genes, require two hits to result in a phenotypic change. We hypothesized that if a germline variant served as the first hit to a pathway, that a somatic mutation in the same pathway would be more likely to result in cancer than it would in a patient without a pathogenic germline variant in the pathway. If true, this would suggest that the pathogenic germline variants may contribute not only to the increased tumor mutational burden of a tumor but may also influence the somatic mutations that are selected for during the development and progression of cancer.

We first tested whether patients with pathogenic germline variants in the nine individual genes that we identified to be associated with increased tumor mutational burden were at higher risk of acquiring a deleterious somatic mutation in the same gene. We did not find any genes for which this was the case. We then asked whether these patients were at a higher risk for a deleterious somatic

mutation in the same pathway, controlling for tumor mutational burden. Indeed, we found patients with pathogenic germline variants in *APC*, *MSH2*, and *MSH6* to be at increased risk of deleterious somatic mutations in the same pathway (**Table 4B**). Multiple pathways were tested for each gene because the genes were annotated in several different pathways.

We did not find any examples of this phenomenon when testing the pathways that were significantly associated with overall tumor mutational burden in individual cancers. When analyzing our pan-cancer pathway results, we found that patients with pathogenic germline variants in genes related to mitochondrial protein import (consisting of *SLC25A13* pathogenic germline variants in this study) and beta catenin phosphorylation (consisting mainly of *APC* pathogenic germline variants) pathways were at higher risk for deleterious somatic mutations in the same pathway, controlling for tumor type and overall tumor mutational burden (**Table 4B**).

Discussion

The widespread collection of next generation sequencing data has enabled detailed study of rare genetic syndromes [6, 25]. While patients with pathogenic germline variants are often screened more aggressively for cancer, clinical guidelines for these patients has only changed in a few circumstances [7, 12]. We previously identified common germline variants associated with differences in patient outcome across a multitude of cancers, suggesting that germline variation contributes not only to cancer risk but also to tumor progression [13, 14]. In this study, we identified pathogenic germline variants

associated with tumor hypermutation and have identified molecular fingerprints of their effects by analyzing RNA-sequencing data and somatic mutation profiles. Our findings suggest that these pathogenic germline variants remain relevant after a patient has been diagnosed with cancer and may contribute to the molecular differences in tumors collected from patients with and without pathogenic germline variants. Our results suggest that patients with pathogenic germline variants should be managed differently than patients without pathogenic germline variants in some cases.

We found that tumors from patients with pathogenic germline variants in the mismatch repair genes *MSH2*, *MSH6*, and *PMS2*, and in the mismatch repair pathway exhibit elevated somatic mutation burden. We found enrichment in the of COSMIC mutational signatures related to mismatch repair in these patients' somatic mutation profiles. Germline mismatch repair deficiency has previously been associated with microsatellite instability and increased responsiveness to immunotherapy and so these findings served as an important positive control in our study [12].

Tumors with pathogenic germline variants in the nucleotide excision repair gene *ERCC3* were associated with elevated tumor mutational burden and we observed enrichment for the mutational signature for nucleotide excision repair deficiency in these patients. While a previous study showed that somatic mutations in the nucleotide base excision repair gene *ERCC2* likely contributes to increased overall somatic mutation burden, no previous study has demonstrated an association between nucleotide excision repair gene

perturbation and immune checkpoint inhibitor efficacy [26]. We did not find a significant association between nucleotide excision repair pathway perturbation by pathogenic germline variants and tumor mutational burden at the pathway level, suggesting that the contribution to overall somatic mutation burden may be limited to select genes in the pathway.

We found patients with pathogenic germline variants in *APC*, which binds to beta-catenin and leads to its degradation, and genes involved with beta-catenin degradation to be associated with elevated tumor mutational burden. We observed upregulation of genes involved with *Wnt* signaling in these patients. Aberrations to the *Wnt* signaling pathway are linked to the formation of many cancers [23]. Spranger et al. showed that non-T cell inflamed tumors exhibited high beta-catenin signaling activity and reduced response to immune checkpoint blockade [27]. Further work is necessary to predict whether pathogenic germline variants in *APC* and genes involved with beta-catenin degradation will be associated with increased or decreased response to immune checkpoint blockade, as the elevated tumor mutational burden would be expected to increase efficacy whereas the elevated beta-catenin signaling would be expected to decrease efficacy.

Patients with pathogenic germline variants in *BRIP1* and other genes involved with homology directed repair exhibited high tumor mutational burden and we observed the molecular signature for homology directed repair in our pathway analysis. Mutations in the homology directed repair genes *BRCA1* and *BRCA2* have previously been shown to be associated with increased tumor

mutational burden and increased response to immune checkpoint blockade [28, 29]. Our results suggest that this finding may be extended to other genes involved with homology directed repair as well.

Tumors from patients with pathogenic germline variants in *SLC25A13* exhibited elevated tumor mutational burden. This gene codes for a mitochondrial aspartate/glutamate transporter. Pathogenic germline variants in this gene are associated with the urea cycle disorder type II citrullinemia and neonatal intrahepatic cholestasis [30]. Lee et al. has previously shown that tumors exhibiting urea cycle dysfunction generate nitrogen metabolites, resulting in DNA damage and ultimately better response to immune checkpoint blockade [31]. While Lee et al.'s analysis focused on somatic urea cycle dysfunction, our work suggests that germline urea cycle dysfunction may also be a marker for improved immune checkpoint blockade response.

Overall, the results of our analysis suggest that understanding the germline contribution to tumor mutational burden could identify sets of patients that could benefit from immune checkpoint blockade therapy. More broadly, our work suggests that germline variation informs the landscape of somatic aberrations and that the contribution from germline variation may ultimately contribute to important differences in clinical management, such as the selection of chemotherapy drugs. This implication is consistent with prior work done in cancer genomics [24, 32] Furthermore, our work supports the findings of other studies discussing the association between somatic biomarkers and efficacy of immune checkpoint blockade. Nevertheless, there are several limitations to our

study. While we are predicting that overall tumor mutational burden will predict better efficacy of immune checkpoint blockade, the strength of this association may differ across patients with different genetic syndromes. Although we did observe downstream evidence of the pathogenic germline variants' effects, we were unable to validate our associations in an independent dataset due to the rarity of these pathogenic germline variants. Both of these concerns will be addressed in the future, as the amount of sequencing data available from patients treated with immune checkpoint inhibitors continues to grow.

Figures

Figure 1. An overview of the number of genes or pathways that could be tested requiring pathogenic germline variants in five or more patients. (A) We did not test for associations in individual genes in individual cancers due to small sample size. We were able to test for associations in (B) individual genes using a pan-cancer approach, (C) pathways in individual cancers, and (D) pathways using a pan-cancer approach. The distribution of patients in the pan-cancer approaches is displayed graphically in (B) and (D).

Figure 1

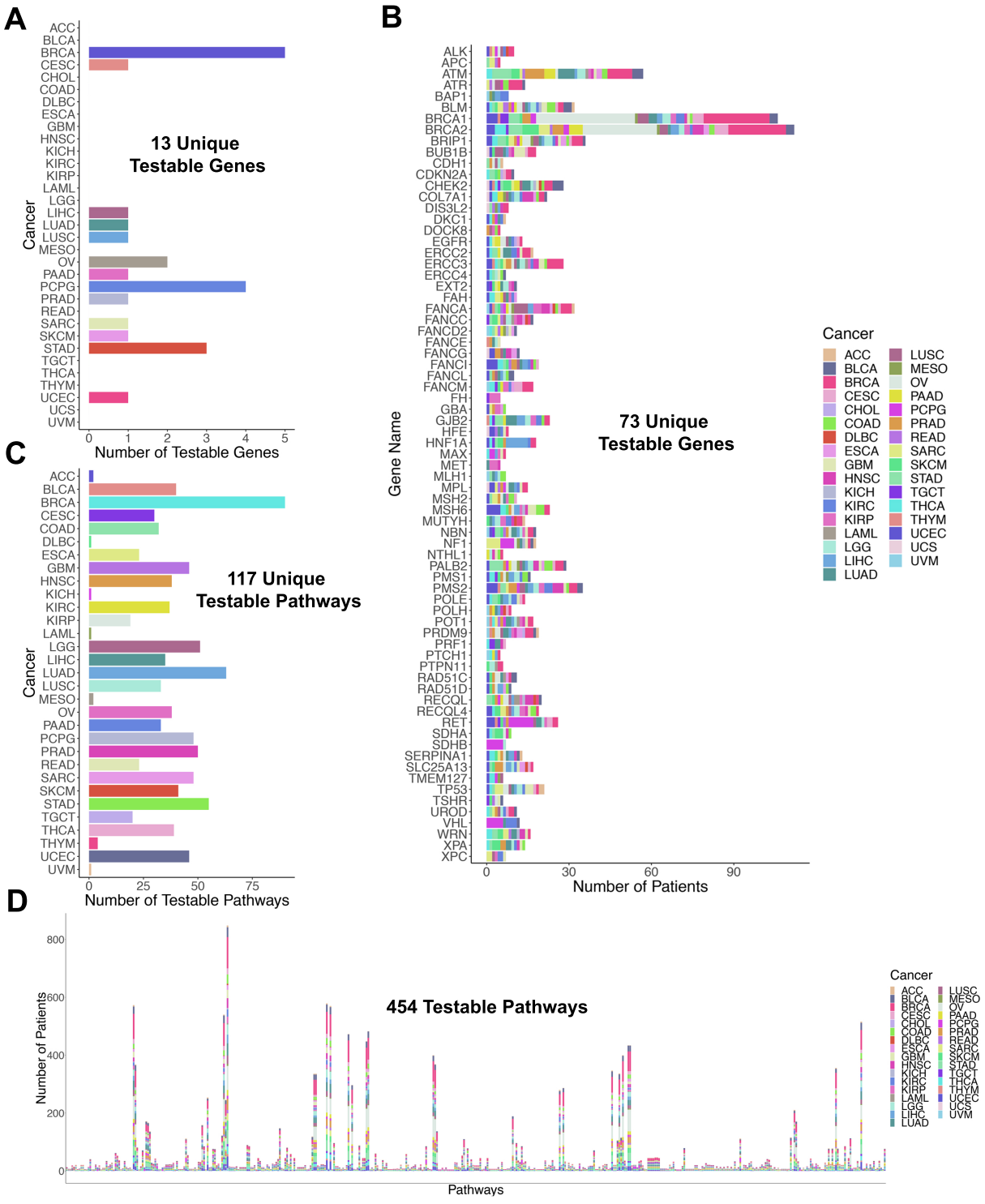

A
13 Unique Testable Genes
Cancer
Number of Testable Genes
B
73 Unique Testable Genes
Gene Name
Number of Patients
Cancer
C
117 Unique Testable Pathways
Cancer
Number of Testable Pathways
D
454 Testable Pathways
Number of Patients
Pathways
Cancer

Figure 2. A summary of the overall approach employed in this study.

Figure 2

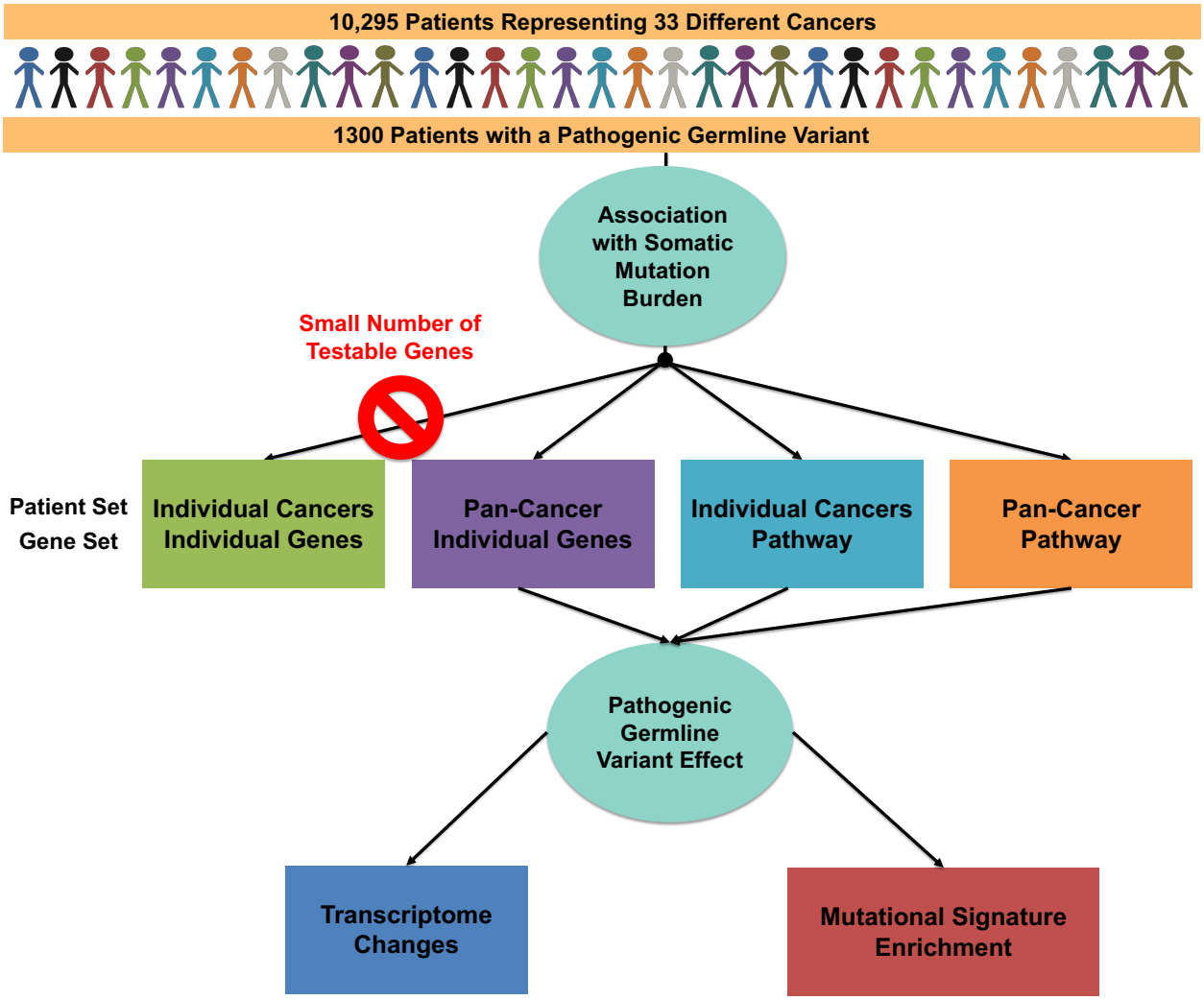
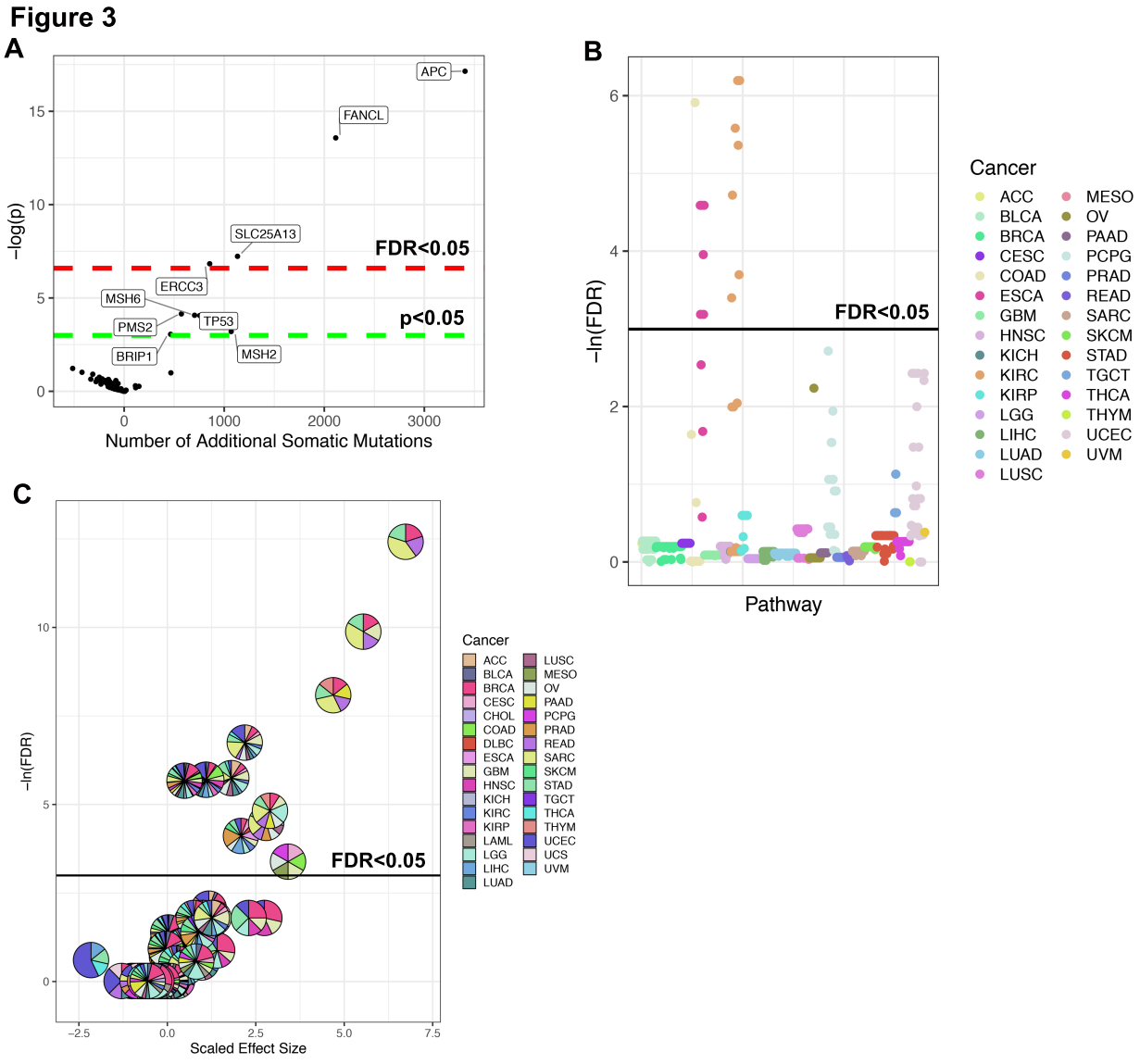


Figure 3. Manhattan plots summarizing the associations with overall somatic mutation burden using our three analyses. We identified associations with elevated somatic mutation burden in (A) genes perturbed by pathogenic germline variants using a pan-cancer approach, (B) pathways perturbed by pathogenic germline variants in individual cancers, and (C) pathways perturbed by pathogenic germline variants using a pan-cancer approach.



Tables

Table 1. A summary of the associations we found with elevated tumor mutational burden in (A) individual genes using a pan-cancer approach, (B) pathways in individual cancers (B), and (C) pathways using a pan-cancer approach.

Table 1A.

Gene	Number of Patients with Mutation	Number of Additional Somatic Mutations	p-value	Adjusted p-value
APC	5	3406.9	3.57E-08	2.61E-06
FANCL	10	2115.6	1.26E-06	4.62E-05
SLC25A13	17	1134.1	7.24E-04	1.76E-02
ERCC3	28	854.2	1.07E-03	1.96E-02
MSH6	22	705.0	1.68E-02	1.80E-01
PMS2	34	572.9	1.57E-02	1.80E-01
TP53	19	755.6	1.73E-02	1.80E-01
MSH2	7	1070.3	4.05E-02	3.70E-01
BRIP1	35	464.2	4.71E-02	3.82E-01

Table 1B.

Cancer	Pathway	Number of Additional Somatic Mutations	p-value	Adjusted p-value	Mutated Genes
KIRC	HDR THROUGH SINGLE STRAND ANNEALING SSA, PROCESSING OF DNA DOUBLE STRAND BREAK ENDS	142.6	1.54E-04	2.04E-03	BRCA1 (3), BLM (1), BRIP1 (1)
KIRC	RESOLUTION OF D LOOP STRUCTURES, RESOLUTION OF D LOOP STRUCTURES THROUGH SYNTHESIS DEPENDENT STRAND ANNEALING SDSA, HOMOLOGOUS DNA PAIRING AND STRAND EXCHANGE	124.7	3.00E-04	2.04E-03	BRCA1 (3), BLM (1), BRCA2 (1), BRIP1 (1)
COAD	DISEASE	1724.2	1.00E-04	2.71E-03	MSH6 (3), MLH1 (2), MSH2 (2), CDKN2A (1), ERCC3 (1), KIT (1), NTHL1 (1), PMS2 (1)
KIRC	G2 M CHECKPOINTS, G2 M DNA DAMAGE CHECKPOINT, REGULATION OF TP53 ACTIVITY, REGULATION OF TP53 ACTIVITY THROUGH PHOSPHORYLATION	113.7	9.97E-04	3.77E-03	BRCA1 (3), BLM (1), BRIP1 (1), TP53 (1)
KIRC	HDR THROUGH HOMOLOGOUS RECOMBINATION HRR, HOMOLOGY DIRECTED REPAIR	95.2	1.51E-03	4.68E-03	BRCA1 (3), POLE (2), BLM (1), BRCA2 (1), BRIP1 (1)

KIRC	CELL CYCLE CHECKPOINTS	94.7	3.14E-03	8.89E-03	BRCA1 (3), BLM (1), BRIP1 (1), CDKN1B (1), TP53 (1)
ESCA	MEIOSIS, MEIOTIC RECOMBINATION, REPRODUCTION	370.3	3.68E-03	1.01E-02	ATM (2), BRCA2 (2), PRDM9 (2), RAD51C (1)
ESCA	CELL CYCLE CHECKPOINTS, G2 M CHECKPOINTS, G2 M DNA DAMAGE CHECKPOINT, HDR THROUGH SINGLE STRAND ANNEALING SSA, PROCESSING OF DNA DOUBLE STRAND BREAK ENDS	456.1	2.34E-03	1.01E-02	ATM (2), BRIP1 (2), BARD1 (1)
ESCA	TP53 REGULATES TRANSCRIPTION OF DNA REPAIR GENES	432.3	3.97E-03	1.01E-02	ATM (2), FANCC (1), FANCD2 (1), RAD51D (1)
ESCA	REGULATION OF TP53 ACTIVITY, REGULATION OF TP53 ACTIVITY THROUGH PHOSPHORYLATION	358.7	9.16E-03	1.92E-02	ATM (2), BRIP1 (2), BARD1 (1), STK11 (1)
KIRC	DNA DOUBLE STRAND BREAK REPAIR	64.1	9.46E-03	2.47E-02	BAP1 (3), BRCA1 (3), POLE (2), BLM (1), BRCA2 (1), BRIP1 (1), TP53 (1)

KIRC	CELL CYCLE	63.5	1.38E-02	3.34E-02	BRCA1 (3), POLE (2), BLM (1), BRCA2 (1), BRIP1 (1), CDKN1B (1), DKC1 (1), TP53 (1)
ESCA	CELL CYCLE	241.2	2.56E-02	4.13E-02	ATM (2), BRCA2 (2), BRIP1 (2), PRDM9 (2), BARD1 (1), RAD51C (1)
ESCA	HDR THROUGH HOMOLOGOUS RECOMBINATION HRR, DNA DOUBLE STRAND BREAK REPAIR, RESOLUTION OF D LOOP STRUCTURES, HOMOLOGY DIRECTED REPAIR, RESOLUTION OF D LOOP STRUCTURES THROUGH SYNTHESIS DEPENDENT STRAND ANNEALING SDSA, HOMOLOGOUS DNA PAIRING AND STRAND EXCHANGE	231.5	3.23E-02	4.13E-02	ATM (2), BRCA2 (2), BRIP1 (2), BARD1 (1), PALB2 (1), RAD51C (1), RAD51D (1)

Table 1C.

Pathway	Number of Additional Somatic Mutations	p-value	Adjusted p-value	Mutated Genes
BETA CATENIN PHOSPHORYLATION CASCADE, DISASSEMBLY OF THE DESTRUCTION COMPLEX AND RECRUITMENT OF AXIN TO THE MEMBRANE, SIGNALING BY WNT IN CANCER, PHOSPHORYLATION SITE MUTANTS OF CTNNB1 ARE NOT TARGETED TO THE PROTEASOME BY THE DESTRUCTION COMPLEX	3406.9	3.57E-08	4.04E-06	APC (5)
DEGRADATION OF BETA CATENIN BY THE DESTRUCTION COMPLEX	2823.0	5.65E-07	5.12E-05	APC (5), AXIN2 (1)
DEACTIVATION OF THE BETA CATENIN TRANSACTIVATING COMPLEX	2408.2	4.08E-06	3.08E-04	APC (5), MEN1 (2)
OVARIAN TUMOR DOMAIN PROTEASES	1186.7	1.82E-05	1.17E-03	TP53 (21), APC (5), PTEN (3)
PROGRAMMED CELL DEATH	1001.9	5.65E-05	3.20E-03	TP53 (21), CDH1 (6), APC (5), STAT3 (1)
MISMATCH REPAIR	647.5	7.36E-05	3.35E-03	PMS2 (35), MSH6 (23), MSH2 (11), MLH1 (7), POLD1 (2)
DISEASES OF MISMATCH REPAIR MMR	656.5	7.40E-05	3.35E-03	PMS2 (35), MSH6 (23), MSH2 (11), MLH1 (7)

DISEASE	355.2	8.36E-05	3.44E-03	PMS2 (35), ERCC3 (28), MSH6 (23), NF1 (18), ERCC2 (17), MUTYH (14), EGFR (13), EXT2 (11), MSH2 (11), CDKN2A (10), MLH1 (7), CDH1 (6), NTHL1 (6), PTPN11 (6), APC (5), MET (5), ABCB11 (4), MAP2K2 (4), CDK4 (3), CDKN1B (3), HRAS (3), KRAS (3), PTEN (3), RAF1 (3), GALNT3 (2), KIT (2), PDGFRA (2), TSC2 (2), CBL (1), EXT1 (1), SMAD4 (1), SOS1 (1), STAT3 (1)
SIGNALING BY WNT, TCF DEPENDENT SIGNALING IN RESPONSE TO WNT	1534.9	2.31E-04	8.06E-03	APC (5), MEN1 (2), TERT (2), AXIN2 (1), SMARCA4 (1)
APOPTOTIC CLEAVAGE OF CELLULAR PROTEINS, APOPTOTIC EXECUTION PHASE	1482.7	3.75E-04	1.13E-02	CDH1 (6), APC (5)

MITOCHONDRIAL PROTEIN IMPORT, GLUCONEOGENESIS, GLUCOSE METABOLISM, ASPARTATE AND ASPARAGINE METABOLISM, PROTEIN LOCALIZATION	1134.1	7.24E-04	1.64E-02	SLC25A13 (17)
REGULATION OF KIT SIGNALING	1784.4	1.57E-03	3.39E-02	KIT (2), SH2B3 (2), CBL (1), SOS1 (1)

Table 2. Gene set enrichment results concordant with the expected effects of the pathogenic germline variants. We observed upregulation of Wnt signaling in patients with pathogenic germline variants in *APC* and genes involved with beta-catenin degradation. We observed upregulation of *E2F* target genes and genes involved with mitotic spindle formation in patients with pathogenic germline variants in genes related to cell cycle checkpoint.

Patient Set	Gene Set	Gene or Pathway	Associated GSEA Result	p-value
Pan-Cancer	Individual Genes	APC	Upregulation of Wnt Signaling Pathway	<0.001
ESCA	Pathway	Cell Cycle Checkpoint	Upregulation of Genes Involved with Mitotic Spindle Formation	0.00875
ESCA	Pathway	Cell Cycle Checkpoint	Upregulation of E2F Target Genes	0.0459
KIRC	Pathway	Cell Cycle Checkpoint	Upregulation of Genes Involved with Mitotic Spindle Formation	<0.001
KIRC	Pathway	Cell Cycle Checkpoint	Upregulation of E2F Target Genes	<0.001
Pan-Cancer	Pathway	Degradation of Beta Catenin	Upregulation of Wnt Signaling Pathway	<0.001

Table 3. Mutational signature results concordant with the expected effects of the pathogenic germline variants.

Patient Set	Gene or Pathway	Signature	Reported Cause of Signature	p-value
Pan-Cancer	MSH6	6	Mismatch Repair Deficiency	2.360E-02
Pan-Cancer	MSH6	15	Mismatch Repair Deficiency	9.100E-03
Pan-Cancer	ERCC3	29	Transcription-Coupled Nucleotide Excision Repair Deficiency	3.350E-02
Pan-Cancer	PMS2	44	Mismatch Repair Deficiency	3.199E-02
Pan-Cancer	MSH2	6	Mismatch Repair Deficiency	1.730E-02
Pan-Cancer	MSH2	15	Mismatch Repair Deficiency	1.220E-05
KIRC	HDR THROUGH HOMOLOGOUS RECOMBINATION HRR, HOMOLOGY DIRECTED REPAIR	3	Homologous Recombination Deficiency	3.400E-04
KIRC	DNA DOUBLE STRAND BREAK REPAIR	3	Homologous Recombination Deficiency	2.670E-03
Pan-Cancer	MISMATCH REPAIR	6	Mismatch Repair Deficiency	2.850E-03
Pan-Cancer	MISMATCH REPAIR	15	Mismatch Repair Deficiency	1.270E-04
Pan-Cancer	DISEASES OF MISMATCH REPAIR MMR	6	Mismatch Repair Deficiency	2.070E-03
Pan-Cancer	DISEASES OF MISMATCH REPAIR MMR	15	Mismatch Repair Deficiency	2.960E-04

Table 4. Patients with certain pathogenic germline variants are more likely to accrue deleterious somatic mutations in (A) cancer-specific driver genes and (B) genes in the same pathway, even after controlling for tumor type and overall somatic mutation burden.

Table 4A.

Gene	Number of Additional Somatic Mutations in Driver Genes	p-value	Adjusted p-value
MSH6	0.483	8.626E-04	7.764E-03
MSH2	0.807	4.797E-03	2.159E-02

Table 4B.

Patient Set	Gene Set	Gene or Pathway	Pathway	Number of Additional Somatic Mutations in Pathway	p-value	Adjusted p-value
Pan-Cancer	Individual Gene	MSH2	DISEASE	3.181	8.097E-08	1.287E-05
Pan-Cancer	Individual Gene	APC	BETA CATENIN PHOSPHORYLATION CASCADE	0.545	1.013E-06	4.027E-05
Pan-Cancer	Individual Gene	APC	PHOSPHORYLATION SITE MUTANTS OF CTNNB1 ARE NOT TARGETED TO THE PROTEASOME BY THE DESTRUCTION COMPLEX	0.547	8.818E-07	4.027E-05
Pan-Cancer	Individual Gene	MSH2	GENE EXPRESSION TRANSCRIPTION	2.867	8.510E-07	4.027E-05
Pan-Cancer	Individual Gene	MSH6	DISEASE	1.041	2.142E-06	6.813E-05
Pan-Cancer	Individual Gene	MSH2	GENERIC TRANSCRIPTION PATHWAY	2.394	3.799E-06	1.007E-04
Pan-Cancer	Individual Gene	MSH2	TRANSCRIPTIONAL REGULATION BY TP53	1.524	6.784E-06	1.541E-04
Pan-Cancer	Individual Gene	MSH2	DNA REPAIR	1.141	1.633E-04	3.245E-03
Pan-Cancer	Individual Gene	MSH6	DNA REPAIR	0.519	2.429E-04	4.291E-03
Pan-Cancer	Individual Gene	APC	SIGNALING BY WNT IN CANCER	0.473	1.099E-03	1.748E-02
Pan-Cancer	Individual Gene	APC	DEGRADATION OF BETA CATENIN BY THE DESTRUCTION COMPLEX	0.522	2.440E-03	3.527E-02
Pan-Cancer	Pathway	BETA CATENIN PHOSPHORYLATION CASCADE	BETA CATENIN PHOSPHORYLATION CASCADE	0.545	1.013E-06	1.064E-05

Pan-Cancer	Pathway	PHOSPHORYLATION SITE MUTANTS OF CTNNB1 ARE NOT TARGETED TO THE PROTEASOME BY THE DESTRUCTION COMPLEX	PHOSPHORYLATION SITE MUTANTS OF CTNNB1 ARE NOT TARGETED TO THE PROTEASOME BY THE DESTRUCTION COMPLEX	0.547	8.818E-07	1.064E-05
Pan-Cancer	Pathway	SIGNALING BY WNT IN CANCER	SIGNALING BY WNT IN CANCER	0.473	1.099E-03	7.694E-03
Pan-Cancer	Pathway	DISASSEMBLY OF THE DESTRUCTION COMPLEX AND RECRUITMENT OF AXIN TO THE MEMBRANE	DISASSEMBLY OF THE DESTRUCTION COMPLEX AND RECRUITMENT OF AXIN TO THE MEMBRANE	0.364	4.512E-03	2.369E-02
Pan-Cancer	Pathway	DEGRADATION OF BETA CATENIN BY THE DESTRUCTION COMPLEX	DEGRADATION OF BETA CATENIN BY THE DESTRUCTION COMPLEX	0.402	7.494E-03	3.147E-02
Pan-Cancer	Pathway	ASPARTATE AND ASPARAGINE METABOLISM	ASPARTATE AND ASPARAGINE METABOLISM	0.094	1.267E-02	4.433E-02

References

1. Hartman AL, Hechtelt Jonker A, Parisi MA, Julkowska D, Lockhart N, Isasi R: **Ethical, legal, and social issues (ELSI) in rare diseases: a landscape analysis from funders.** *Eur J Hum Genet* 2019.

2. Gainotti S, Mascalzoni D, Bros-Facer V, Petrini C, Floridia G, Roos M, Salvatore M, Taruscio D: **Meeting Patients' Right to the Correct Diagnosis: Ongoing International Initiatives on Undiagnosed Rare Diseases and Ethical and Social Issues.** *Int J Environ Res Public Health* 2018, **15**.

3. Vaske OM, Bjork I, Salama SR, Beale H, Tayi Shah A, Sanders L, Pfeil J, Lam DL, Learned K, Durbin A, et al: **Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer.** *JAMA Netw Open* 2019, **2:**e1913968.

4. Sanderson SC, Hill M, Patch C, Searle B, Lewis C, Chitty LS: **Delivering genome sequencing in clinical practice: an interview study with healthcare professionals involved in the 100 000 Genomes Project.** *BMJ Open* 2019, **9:**e029699.

5. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, et al: **Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines.** *Cell Syst* 2018, **6:**271-281.e277.

6. Kamps R, Brandao RD, Bosch BJ, Paulussen AD, Xanthoulea S, Blok MJ, Romano A: **Next-Generation Sequencing in Oncology: Genetic Diagnosis, Risk Prediction and Cancer Classification.** *Int J Mol Sci* 2017, **18**.

7. Lindor NM, Petersen GM, Hadley DW, Kinney AY, Miesfeldt S, Lu KH, Lynch P, Burke W, Press N: **Recommendations for the care of individuals with an inherited predisposition to Lynch syndrome: a systematic review.** *Jama* 2006, **296:**1507-1517.

8. Ballinger ML, Best A, Mai PL, Khincha PP, Loud JT, Peters JA, Achatz MI, Chojniak R, Balieiro da Costa A, Santiago KM, et al: **Baseline Surveillance in Li-Fraumeni Syndrome Using Whole-Body Magnetic Resonance Imaging: A Meta-analysis.** *JAMA Oncol* 2017, **3:**1634-1639.

9. Maher ER, Yates JR, Harries R, Benjamin C, Harris R, Moore AT, Ferguson-Smith MA: **Clinical features and natural history of von Hippel-Lindau disease.** *Q J Med* 1990, **77:**1151-1163.

10. Snyder A, Makarov V, Merghoub T, Yuan J, Zaretsky JM, Desrichard A, Walsh LA, Postow MA, Wong P, Ho TS, et al: **Genetic basis for clinical response to CTLA-4 blockade in melanoma.** *N Engl J Med* 2014, **371:**2189-2199.

11. Van Allen EM, Miao D, Schilling B, Shukla SA, Blank C, Zimmer L, Sucker A, Hillen U, Foppen MHG, Goldinger SM, et al: **Genomic correlates of response to CTLA-4 blockade in metastatic melanoma.** *Science* 2015, **350:**207-211.

12. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, et al: **Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade.** *Science* 2017, **357:**409-413.

13. Chatrath A, Kiran M, Kumar P, Ratan A, Dutta A: **The Germline Variants rs61757955 and rs34988193 Are Predictive of Survival in Lower Grade Glioma Patients.** *Mol Cancer Res* 2019, **17:**1075-1086.

14. Chatrath A, Przanowska R, Kiran S, Su Z, Saha S, Wilson B, Tsunematsu T, Ahn J-H, Lee KY, Paulsen T, et al: **The Pan-Cancer Landscape of Prognostic Germline Variants in 10,582 Patients.** *medRxiv* 2019**:**19010264.

15. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al: **Pathogenic Germline Variants in 10,389 Adult Cancers.** *Cell* 2018, **173:**355-370.e314.

16. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.** *Cell* 2018, **173:**400-416.e411.

17. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, et al: **The Reactome Pathway Knowledgebase.** *Nucleic Acids Res* 2018, **46:**D649-d655.

18. Krishnamurthy N, Kurzrock R: **Targeting the Wnt/beta-catenin pathway in cancer: Update on effectors and inhibitors.** *Cancer Treat Rev* 2018, **62:**50-60.

19. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.

20. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al: **COSMIC: the Catalogue Of Somatic Mutations In Cancer.** *Nucleic Acids Res* 2019, **47:**D941-d947.

21. Blokzijl F, Janssen R, van Boxtel R, Cuppen E: **MutationalPatterns: comprehensive genome-wide analysis of mutational processes.** *Genome Med* 2018, **10:**33.

22. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al: **Comprehensive Characterization of Cancer Driver Genes and Mutations.** *Cell* 2018, **173:**371-385.e318.

23. Anastas JN, Moon RT: **WNT signalling pathways as therapeutic targets in cancer.** *Nat Rev Cancer* 2013, **13:**11-26.

24. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, et al: **Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer.** *Cancer Discov* 2017, **7:**410-423.

25. Sylvester DE, Chen Y, Jamieson RV, Dalla-Pozza L, Byrne JA: **Investigation of clinically relevant germline variants detected by next-generation sequencing in patients with childhood cancer: a review of the literature.** *J Med Genet* 2018, **55:**785-793.

26. Van Allen EM, Mouw KW, Kim P, Iyer G, Wagle N, Al-Ahmadie H, Zhu C, Ostrovnaya I, Kryukov GV, O'Connor KW, et al: **Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma.** *Cancer Discov* 2014, **4:**1140-1153.

27. Spranger S, Bao R, Gajewski TF: **Melanoma-intrinsic beta-catenin signalling prevents anti-tumour immunity.** *Nature* 2015, **523:**231-235.

28. Hugo W, Zaretsky JM, Sun L, Song C, Moreno BH, Hu-Lieskovan S, Berent-Maoz B, Pang J, Chmielowski B, Cherry G, et al: **Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma.** *Cell* 2016, **165:**35-44.

29. Nolan E, Savas P, Policheni AN, Darcy PK, Vaillant F, Mintoff CP, Dushyanthen S, Mansour M, Pang JB, Fox SB, et al: **Combined immune checkpoint blockade as a therapeutic strategy for BRCA1-mutated breast cancer.** *Sci Transl Med* 2017, **9**.

30. Song YZ, Zhang ZH, Lin WX, Zhao XJ, Deng M, Ma YL, Guo L, Chen FP, Long XL, He XL, et al: **SLC25A13 gene analysis in citrin deficiency: sixteen novel mutations in East Asian patients, and the mutation distribution in a large pediatric cohort in China.** *PLoS One* 2013, **8:**e74544.

31. Lee JS, Adler L, Karathia H, Carmel N, Rabinovich S, Auslander N, Keshet R, Stettner N, Silberman A, Agemy L, et al: **Urea Cycle Dysregulation Generates Clinically Relevant Genomic and Biochemical Signatures.** *Cell* 2018, **174:**1559-1570.e1522.

32. Menden MP, Casale FP, Stephan J, Bignell GR, Iorio F, McDermott U, Garnett MJ, Saez-Rodriguez J, Stegle O: **The germline genetic component of drug sensitivity in cancer cell lines.** *Nat Commun* 2018, **9:**3385.

Chapter 5: Discussion

Inferring Germline Variant Status from Tumor Samples and RNA-Sequencing Data

The initial goal of this project was to determine whether or not germline variants contribute to tumor progression. To get at this question, we decided to utilize sequencing data from The Cancer Genome Atlas, as it was one of the largest repositories of publicly available sequencing data from tumors with matched germline samples. While developing our pipeline, we had found that the status of some germline variants could not always be determined using only the whole exome sequenced non-tumor sample due to limited sequencing depth. We hypothesized that the germline variants should also be found in the whole exome sequenced tumor sample and the RNA-sequenced tumor sample, assuming that these germline variants were not somatically mutated in the tumor samples, changed through RNA editing, or suppressed through allele-specific expression in the RNA-sequenced tumor sample. We decided to include these samples in our pipeline to increase the total number of germline variants that we could call. We hypothesized that the inclusion of additional patients would increase our statistical power to detect significant associations with patient outcome. Our approach is detailed in **Chapter 2** and **Chapter 3** [1, 2].

Our method can be applied to whole exome sequenced and RNA sequenced samples to identify the statuses of common germline variants, even in the absence of non-tumor samples. While studies performing whole exome sequencing in oncology typically collect both a tumor and non-tumor sample to identify somatic mutations, some studies only perform RNA sequencing on

tumors, and it is not uncommon to use tumor-only panels in a clinical setting. Our method will enable the study of common germline variants in these studies, even in the absence of a non-tumor sample. Furthermore, our method could also be applied to datasets for which it is not possible to generate a non-tumor sample, such as data generated from cancer cell lines [1-4].

Limitations to our Method and Possible Improvements

There are several limitations to our method and opportunities for future improvement. Most importantly, our method is only able to extract common germline variants and is entirely unable to differentiate between rare pathogenic germline variants and somatic mutations [1, 2]. As the number of paired tumor-normal samples in oncology continues to grow, it may be possible to train machine learning classifiers to distinguish between rare germline variants and somatic mutations in samples lacking a normal sample, as driver somatic mutations are known to occur in certain hotspots and local mutation rates between the tumors and normal healthy tissue are different [5]. Other methods consider features such as allele fraction to distinguish between rare germline variants and somatic mutations in patient samples, though allele fraction would not be able to distinguish between germline variants and highly clonal somatic mutations in samples from clonal cell lines [6]. Methods exist to enrich for somatic mutations by excluding the set of known germline variants, though this approach would not be effective for identifying rare germline variants with high confidence [7].

In **Chapter 3**, we explain how the overlap between the common germline variants pulled out using our approach and somatic mutations is less than 0.02% from tumor samples. We examined the overlap between common germline variants and somatic mutations to get a sense for how often the germline variants that we extracted from tumor sequencing data could have been somatic mutations. Overall, we found the overlap to be quite small, suggesting to us that our method is fairly accurate. We found the overlap between the common germline variants that we pulled out and RNA editing sites to be less than 0.20%. While the potential error rate is quite low for genome wide association studies structured similarly to ours, our approach does not involve estimating the probability that an individual germline variant is not a somatic mutation or is affected by RNA editing. Therefore, our method can be extended by identifying genomic features that can be used to calculate the probability of individually extracted variants actually being germline variants by using paired tumor-normal samples for validation. To do this, we could use the germline variant status from the whole exome sequenced normal sample as the true set of germline variant calls and compare this true set of germline variant calls against the germline variant calls that we extract from the whole exome sequenced tumor and RNA sequenced tumor samples. We could create a model to predict the likelihood of the germline variant calls from the whole exome sequenced tumor and RNA sequenced tumor samples actually being real germline variants based on genomic features, such as allele fraction or population allele frequency. As an example, we would expect allele fraction to be inversely correlated to the

probability of being a real germline variant and population allele frequency to be directly correlated with the probability of being a real germline variant.

Germline Variation is Associated with Tumor Progression Across Cancers

Past studies had identified germline variants in previously characterized cancer driver genes associated with overall survival in individual cancers [8-10]. In **Chapter 2**, I describe the unbiased genome wide association study that we performed in a cohort of approximately 500 lower grade glioma patients to test whether or not germline variation is associated with overall survival. Our analysis identified two germline variants associated with patient outcome, one in the oncogene *GRB2* and the other in the tumor suppressor gene *ANKDD1a*. While much of the research in molecular oncology has been on somatic aberrations, our results from this study suggested that germline variation should be studied not only for understanding risk of cancer but also for understanding cancer progression.

In **Chapter 3**, I discuss the extension of this approach across all cancers included within The Cancer Genome Atlas. Similar to what we found in our study of tumors from patients with lower grade gliomas, we found that germline variation is associated with tumor progression across all cancers for which we were well-powered. We mapped many of the prognostic germline variants to known tumor suppressor genes or oncogenes, suggesting that some of the prognostic germline variants perturb similar pathways as those perturbed by somatic mutations. Our results suggest that some of the germline variants may act as expression quantitative trait loci and a few may perturb protein coding

domains, though experimentation is needed to confirm our hypothesized mechanisms of action for each of the variants.

Given our findings in **Chapter 2** and **Chapter 3**, it is likely that germline variation contributes to tumor progression in most cancers. The study of germline variation in the context of tumor progression will likely become more commonplace in the future, as long-term outcome data for large cohorts of cancer patients becomes more readily available. Future studies will likely replicate the findings that we report in **Chapter 2** and **Chapter 3** by identifying germline variants associated with tumor progression that are not solely found in driver genes characterized in that particular cancer. Our study sets the groundwork for the study of germline variation in the context of tumor progression and can be extended in several ways.

Identification of Prognostic Pathogenic and Rare Germline Variants

In **Chapter 3**, I discussed our study of common germline variants with a population allele frequency greater than five percent across the 33 cancers included within The Cancer Genome Atlas (TCGA). We had identified germline variants associated with patient outcome across all cancers for which we were sufficiently powered to detect variants with moderate effect size [1, 2].

Numerous studies in the area of cancer risk have reported a negative correlation between population allele frequency and cancer risk [11]. We also reported a similar finding in our study of variants associated with tumor progression [2]. This finding suggests that the pathogenic (typically extremely rare germline variants that are predicted to functionally perturb genes associated

with known human diseases) and rare (variants with population allele frequencies lower than five percent) variants may have larger effect sizes and could therefore substantially augment clinical outcome model predictions.

These rare and pathogenic germline variants are challenging to study. In relatively modest sized cohorts such as TCGA, we were unable to study these variants individually due to lack of statistical power. While one possible solution is to group these variants together, this solution introduces two new problems:

(1) In grouping variants together, some variants may have functional consequences, whereas others may not. If the effect sizes of the variants with functional consequences are small or the proportion of non-functional variants is high, our ability to detect significant associations would be low.

(2) Variants in the same gene could have effect sizes in opposite directions. If we used a statistical approach similar to the one employed in **Chapter 3**, the effects of these variants could cancel each other out, resulting in us being unable to detect significant associations.

(3) In grouping rare germline variants, the germline variants that are more common may have more influence on the statistical test than germline variants that are very rare because these germline variants are found in more individuals. This is likely undesirable, as this would decrease our statistical power to detect significant associations since we observed a negative correlation between allele frequency and effect size.

Other branches of genomics have proposed several solutions to these problems which could be repurposed for analyses of tumor progression in cancer genomics [12, 13]:

(1) We could enrich for functional variants by restricting our analysis only to pathogenic germline variants or variants predicted to have functional consequences based on another metric (such as CADD score, SIFT score, and PhyloP score) [14-16].

(2) In **Chapter 3**, we described how we had tested germline variants individually using Cox regression models. For analyses of groups of variants, we could test for associations using variance-based tests such as the sequenced kernel association test (SKAT). The Cox regression models that we utilized in Chapter 2 and Chapter 3 were burden-based tests that tested whether or not a group of patients with the germline variant did significantly better or significantly worse as a set than patients without the germline variant. Although this approach works well for individual germline variants, it may not work well for testing sets of variants. When considering sets of variants, if half of the variants are associated with favorable outcome and half of the variants are associated with poor outcome and the magnitude of effects are similar, then on net the differences will cancel each other out and the test will not detect a significant difference from the control group. Variance-based tests such as the sequenced kernel association test (SKAT) could be designed to examine the dispersion of outcomes in the test group compared to the

control group. If the dispersion is larger in the test group, the variance-based test would yield a significant difference. This would allow us to detect associations even in cases in which the directions of the effects of individual variants are in opposite directions [17, 18]. The major disadvantage to variance-based tests is that it would not be clear which variants are associated with favorable outcome and which would be associated with poor outcome without further post-hoc testing [19].

(3) To deal with more common germline variants potentially having too much influence on the statistical test, we could adjust the weights of the germline variants included within the tested sets. For example, we could consider decreasing the weight of more common variants and increasing the weight of rarer variants. Similarly, we may weigh the variants based on other metrics, such as CADD score, SIFT score, and PhyloP score [20].

Ideally, these approaches will enable us to identify additional sets of prognostic germline variants that could further improve clinical outcome model predictions. Algorithms such as backward elimination can be used to further prioritize the variants in these sets based on their probability to be causal [19]. These algorithms work by removing variants for which the elimination of the variant results in a decreased p-value when performing a sequenced kernel association test. Furthermore, these approaches would enable us to test whether the pathogenic and rare variants in well-known oncogenes and tumor suppressor genes that contribute to increased risk for cancer also contribute to an increased rate of cancer progression [21-26]. We provide an approach to grouping variants

based on pathway in **Chapter 4** that could be applied to the study of rare germline variants associated with differences in patient outcome.

Understanding how Prognostic Germline Variants Vary Across Different Races

Our study of germline variants associated with differences in patient outcome described in **Chapter 2** and **Chapter 3** was performed using data from The Cancer Genome Atlas [1, 2]. Although The Cancer Genome Atlas is a rich multi-omic resource for genomic studies, most patients from The Cancer Genome Atlas are of European ancestry [27-29]. As a result, the prognostic germline variants that we reported were discovered in a cohort of patients primarily of European descent. The genetic ancestry of all TCGA patients has been reported in The Cancer Genome Ancestry Atlas [29]. From their report of the genetic ancestry of TCGA patients, the cohort is primarily of European ancestry. BRCA, GBM, BLCA, LGG, HNSC, THCA, PCPG, COAD, KIRP/KIRC, PRAD, OV, and UCEC have greater than 20 patients each that are of African-American descent. BRCA, ESCA, LIHC, STAD, THCA, CESC, and UCEC are cancers that have 20 or more patients of Asian descent. Based on these numbers, it is feasible to investigate the contribution of germline variation to patient outcome in these cancers, stratifying the analysis by race. In our own analysis, we found that calculated race was a significant predictor of patient outcome for ACC, CESC, CHOL, COAD, HNSC, KIRC, LIHC, LUSC, OV, PAAD, SKCM, STAD, TGCT, and UCEC, suggesting that these cancers may have significant differences at the level of germline variants or that outcome may be confounded by socioeconomic factors tied to race in these cancers.

Many studies have reported genomic differences in cancers from patients of different races, suggesting that the germline variants that contribute, along with the strength of their contribution to cancer progression likely varies based on race [30-44]. These results suggest that the study of prognostic germline variants should be extended to cohorts of patients of non-European ancestry.

In **Chapter 2**, we had identified two germline variants predictive of outcome in patients with lower grade gliomas and tested those germline variants in an independent population of patients with lower grade gliomas of Chinese ancestry. We found that one of the germline variants was not found in any of the patients in the Chinese cohort, which was consistent with the reported allele frequency from a study of thousands of individuals [45]. We found the other germline variant to have nearly the same effect in the Chinese cohort as the cohort of patients from The Cancer Genome Atlas. If these results are consistent with future studies in non-European cohorts, then we can reasonably expect some of the prognostic germline variants to be shared across races and for some other prognostic germline variants to be specific to individual races.

Discovery of Germline Variants with Lower Effect Sizes

In our analyses described in **Chapter 2** and **Chapter 3**, we were able to detect associations in individual cancers beginning at a hazard ratio of about 2, based on our power analysis. Future studies in larger cohorts will likely be able to detect germline variants with lower effect sizes. Alternatively, our original analysis on the TCGA cohort could be reperformed after removing low frequency germline variants. We had initially tested germline variants found in 15 or more

individuals across cancers because we found this to be the optimal threshold for correlating the allele frequency of germline variants extracted from tumor sequencing data with the known population allele frequency. If we had used a custom threshold across cancers, we may have been able to detect additional germline variants with lower effect sizes.

Translation to Clinical Practice

While our studies have provided us with insight into which genetic loci are associated with patient outcome in each of the individual cancers, additional work is necessary to translate these findings into clinical practice.

(1) Firstly, the associations described in **Chapter 3** need to be tested in other cohorts to attain a better understanding as to which groups of patients these germline variants could be useful in.

(2) The prognostic germline variants need to be more rigorously integrated with clinical information, along with other genomic data types, to build models with maximal predictive power. Although The Cancer Genome Atlas is a rich resource of multi-omic data, the annotation of clinical data is far less rigorous than what would be available to a practicing clinician from an electronic medical record. It is essential to create models taking into account the wealth of clinical information available to a physician along with the multi-omic data from studies such as TCGA to best individualize a patient's care. We had shown that germline variation provides additional information about patient outcome not captured by clinical information alone in **Chapter 3**. To be useful in clinical practice, standardized clinical

models need to be generated using the same clinical information and genomic information. One of the current challenges in cancer genomics is that while most datasets offer similar types of genomic data, the clinical data that accompany these datasets vary. The clinically rich datasets with multi-omic data that will likely be generated in the future will be quite useful for generating integrated models that could be used in clinical practice.

(3) Further discussions with expert clinicians who treat each tumor type could help clarify the circumstances in which additional insight about a patient's prognosis could be clinically valuable. Prognostic models for cancers for which most patients have a very favorable prognosis may be less useful compared to cancers in which there is much more heterogeneity in patient outcome.

Interaction Between Germline Variation and the Landscape of Somatic Aberrations

Numerous studies in cancer genomics have now suggested that there is an interaction between germline variants and the landscape of somatic aberrations, suggesting that knowledge of germline variation can be used to predict future somatic events [46-52]. Our results discussed in **Chapter 2**, **Chapter 3**, and **Chapter 4** also support this idea.

In **Chapter 2**, I discussed our discovery of a germline variant in the 3'UTR of the oncogene *GRB2*, an adaptor protein in the *Ras* signaling pathway, associated with poor patient outcome. This germline variant was associated with widespread upregulation of downstream genes in the *Ras* signaling pathway and

was associated with increased incidence of *CIC* mutations and 1p/19q co-deletions. *CIC* is a tumor suppressor gene located on 19q that downregulates the *Ras* signaling pathway. These results suggest that the *GRB2* variant may serve as the first "hit" to the *Ras* signaling pathway and that a second "hit" by a somatic event to the *Ras* signaling event, perhaps to *CIC*, may be responsible for the widespread upregulation of genes involved in *Ras* signaling that we found in patients with lower grade gliomas [1]. In **Chapter 3**, I described how we had found that the prognostic germline variants associated with poor outcome were more likely to be associated with somatic mutations in driver genes. I also provided examples of how tumors from patients with germline variants in *MAP2K3* and *BIRC5* exhibited expected transcriptomic differences in their respective pathways [2]. Experimental work or more complex network-based approaches is necessary to better understand the molecular underpinnings of this association.

In **Chapter 4**, I described our approach to identifying pathogenic germline variants associated with elevated tumor mutational burden. The results described in this analysis clearly suggest that the somatic aberrations present in tumors that arise in patients with and without pathogenic germline variants are clearly not equivalent. Generally speaking, we found tumors from patients with pathogenic germline variants in genes related to DNA repair and cell cycle pathways to be associated with higher tumor mutational burden. When looking at the somatic mutation signatures and the transcriptomic changes in these patients, we observed changes consistent with the expected effects of the

pathogenic germline variants based on previously published experimental work, suggesting that these germline variants shape the somatic mutation landscape.

The Need for an Unbiased Analysis of the Interaction Between Germline Variants and the Landscape of Somatic Aberrations

Most studies, including ours, analyzing the interaction between germline variation and the landscape of somatic aberrations, such as somatic mutations, gene duplications, gene deletions, methylation changes, and transcriptomic dysregulation, have tested specifically for somatic aberrations that would be expected to be associated with particular germline variants [10, 21, 24, 49, 50, 52, 53]. For example, in **Chapter 2** we tested whether or not the prognostic germline variant in *GRB2*, an oncogene in the Ras signaling pathway, was associated with differences in Ras signaling due to somatic mutations in genes like *CIC* in that pathway. In **Chapter 3**, we tested for the transcriptomic changes that we expected to be present in patients with germline variants in *MAP2K3* and *BIRC5*, based on previously published experimental data. In **Chapter 4**, we looked for differences in the somatic mutational profiles and transcriptome that would be consistent with the field's understanding of the pathogenic germline variants that we were studying.

Although this approach is a reasonable starting point, the work by Carter et al. suggests that germline variants can shape somatic events in genes outside of the immediate pathway that the germline variant is found in [46]. Molecularly, this finding certainly seems plausible given the large amount of cross-talk between pathways. This finding therefore raises the need for unbiased analyses between germline variants and the landscape of somatic changes. Network

based or experimental follow-up of these findings may reveal new avenues for cross-talk between pathways and would improve our understanding as to why a germline variant in one gene would predispose tumor cells to somatic events in a different seemingly unrelated gene.

Mechanisms of Action of the Prognostic Germline Variants

Experimental study of the prognostic germline variants could further reveal the mechanisms by which germline variants may affect tumor progression. Although many of the germline variants discussed in **Chapter 2** and **Chapter 3** are candidates for experimental study, I will discuss the germline variants in *GRB2* and *ANKDD1a* below.

In **Chapter 2**, we report a germline variant in the 3' UTR *GRB2* to be associated with poor patient outcome in patients diagnosed with lower grade gliomas. *GRB2* is an adaptor protein near the beginning of the *Ras* signaling pathway [54]. We found this germline variant to be associated with upregulation of *Ras* signaling and to be associated with an increased risk for *CIC* somatic mutations and 1p/19q co-deletions. The variant we identified in *GRB2* is genetically linked to four other germline variants and requires genetic perturbation to identify the causal variant:

(1) If the causal variant is the variant we identified in **Chapter 2**, then the causal variant may act by disrupting a miRNA binding site, resulting in elevated GRB2 protein and subsequent increased *Ras* signaling. If this is the case, then a luciferase experiment in which the wild type and mutant *GRB2* 3' UTRs have been inserted into luciferase constructs may show

increased fluorescence in cells transfected with the mutant construct compared to the wild type construct.

(2) If there is no difference in the luciferase assay results then the variant may not act through disruption of the 3' UTR or one of the other four genetically linked variants may be causal. To identify the causal variants, mutants of each of the variants could be created through CRISPR/Cas9.

After creating mutants through CRISPR/Cas9, the mutants could be screened relative to the control (wild type at all sites) for several different phenotypes:

(1) Increased tumor aggressiveness – I would expect cell proliferation, invasion, migration, and soft agar colony formation in a cell culture system and the rate of tumor expansion in a mouse xenograft experiment to be elevated in the cell line with the causal variant.

(2) Evidence of increased Ras signaling activity – I would expect increased phosphorylation of MEK, ERK, and Elk-1 on Western blot and upregulation of Elk-1 targets on RT-PCR and RNA sequencing in cell lines with the causal variant.

(3) Increased frequency of *CIC* somatic mutations and 1p/19q co-deletions – I would expect increased frequency of *CIC* somatic mutations and 1p/19q co-deletions following long-term culture of the cell lines.

In **Chapter 2**, we discovered a germline variant in the tumor suppressor gene *ANKDD1a* to be associated with poor patient outcome in a cohort of American patients and a cohort of Chinese patients. *ANKDD1a* is a tumor

suppressor gene that promotes the activity of FIH-1 and results in the degradation of HIF-1α. By doing so, ANKDD1a downregulates the hypoxia induced response and decreases the ability of tumor cells to proliferative in their hypoxic microenvironment. The variant that we identified results in an amino acid change from positively charged lysine to negatively charged glutamic acid. This germline variant could be studied through the following experiments:

(1) Overexpress wild type and mutant *ANKDD1a* constructs to test whether the mutant is associated with more aggressive phenotypes compared to the wild type form (increased cell proliferation, invasion, migration, and soft agar colony formation in a cell culture system and increased rate of tumor expansion in a mouse xenograft model). Repeat with glioma cell lines that have been edited through CRISPR/Cas9.

(2) Test whether or not the mutant form of *ANKDD1a* is associated with upregulation of HIF-1α responsive genes indirectly using a luciferase reporter and directly through RT-PCR and RNA sequencing.

(3) Test whether the mutant form of ANKDD1a has lower affinity for FIH-1 than the wild type form through Western blotting. If there is no difference in affinity, perform mass spectrometry following immunoprecipitation of ANKDD1a to identify possible binding partners for ANKDD1a. Validate these binding partners through Western blotting.

Germline Variation Informs Therapeutic Decisions

Historically, therapeutic decisions in oncology have been based on tumor location, grade, and stage. The explosion of next generation sequencing data has suggested that therapeutic decisions in oncology should also be based on somatic aberrations. Recent studies now suggest that germline variation also contributes to drug sensitivity as well and that the contribution of germline variation to drug sensitivity may actually be greater than the contribution from somatic aberration for some drugs [55-62].

In **Chapter 4**, we identify sets of germline variants associated with differences in tumor mutational burden. Tumor mutational burden is a strong predictor of response to treatment with immune checkpoint inhibitors. Most of the pathogenic germline variants associated with differences in tumor mutational burden are found in genes with known functions in DNA repair, mitosis, or cell cycle regulation. Our results suggest that germline variation could potentially be used to predict whether or not a patient will respond to treatment with immune checkpoint inhibitors.

The Need for Large Datasets with Better Clinical Annotation

Using germline variation for making treatment decisions is arguably one of the most clinically promising applications of studying germline variation. Currently, these studies have been very challenging due to the lack of treatment and response data in large cohorts of cancer patients. Substantial efforts by consortia are now underway to generate multi-omic datasets with detailed clinical

annotation. The future study of these datasets will likely result in the discovery of many germline variants that predict response to therapy.

Germline Variants Predicting Response to Therapy and Outcome in Diseases Other than Cancer

Genome wide association studies have been largely focused on the risk of acquiring disease. These studies are beneficial for identifying patients that should be screened for diseases earlier. However, several studies have suggested that germline variation contributes to the progression of other diseases as well, such as HIV/AIDS, systemic mastocytosis, and major depression [63-65]. The literature published in the area of pharmacogenomics suggests that germline variation influences response and toxicity to a variety of drugs, including codeine, tramadol, antidepressents, warfarin, phenytoin, simvastatin, and tacrolimus [66]. Although studying germline variation in the context of disease progression and treatment response is more challenging and expensive due to the need for long-term follow-up, the results presented in this thesis along with the growing body of work on this topic in the literature suggest that germline variation could play a substantial role in personalizing the clinical management of a large number of diseases.

Conclusion

Germline variation has a rich history of being studied in the context of risk for cancer. Emerging studies in the area of cancer genomics now suggest that germline variation contributes to the landscape of somatic aberrations in cancer, affects tumor progression, and informs treatment sensitivity and toxicity. The work described in this dissertation touches on and supports each of these

emerging areas. In **Chapter 2** and **Chapter 3**, we find that germline variation is associated with patient outcome across most cancers and that the notion that germline variants affect tumor progression is likely a fundamental principle of cancer genomics. In **Chapter 4**, we show that the tumors of patients with pathogenic germline variants in certain genes are substantially different from the tumors of patients who do not have pathogenic germline variants in these genes and that this difference can likely be exploited through the use of immune checkpoint inhibitors to improve the care of patients with these germline biomarkers. Overall, this work suggests that germline variation warrants deeper study in clinical oncology as germline variation likely has untapped potential for improving the care of patients with cancer.

References

1. Chatrath A, Kiran M, Kumar P, Ratan A, Dutta A: **The Germline Variants rs61757955 and rs34988193 Are Predictive of Survival in Lower Grade Glioma Patients.** *Mol Cancer Res* 2019, **17:**1075-1086.

2. Chatrath A, Przanowska R, Kiran S, Su Z, Saha S, Wilson B, Tsunematsu T, Ahn J-H, Lee KY, Paulsen T, et al: **The Pan-Cancer Landscape of Prognostic Germline Variants in 10,582 Patients.** *medRxiv* 2019**:**19010264.

3. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al: **COSMIC: the Catalogue Of Somatic Mutations In Cancer.** *Nucleic Acids Res* 2019, **47:**D941-d947.

4. Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER, 3rd, Barretina J, Gelfand ET, Bielski CM, Li H, et al: **Next-generation characterization of the Cancer Cell Line Encyclopedia.** *Nature* 2019, **569:**503-508.

5. Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, Lawrence MS, Getz G: **Passenger Hotspot Mutations in Cancer.** *Cancer Cell* 2019, **36:**288-301.e214.

6. Sun JX, He Y, Sanford E, Montesion M, Frampton GM, Vignot S, Soria JC, Ross JS, Miller VA, Stephens PJ, et al: **A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal.** *PLoS Comput Biol* 2018, **14:**e1005965.

7. Hiltemann S, Jenster G, Trapman J, van der Spek P, Stubbs A: **Discriminating somatic and germline mutations in tumor DNA samples without matching normals.** *Genome Res* 2015, **25:**1382-1390.

8. Le Page C, Rahimi K, Rodrigues M, Heinzelmann-Schwarz V, Recio N, Tommasi S, Bataillon G, Portelance L, Golmard L, Meunier L, et al: **Clinicopathological features of women with epithelial ovarian cancer and double heterozygosity for BRCA1 and BRCA2: A systematic review and case report analysis.** *Gynecol Oncol* 2019.

9. Guerrini-Rousseau L, Dufour C, Varlet P, Masliah-Planchon J, Bourdeaut F, Guillaud-Bataille M, Abbas R, Bertozzi AI, Fouyssac F, Huybrechts S, et al: **Germline SUFU mutation carriers and medulloblastoma: clinical characteristics, cancer risk, and prognosis.** *Neuro Oncol* 2018, **20:**1122-1132.

10. Baretta Z, Mocellin S, Goldin E, Olopade OI, Huo D: **Effect of BRCA germline mutations on breast cancer prognosis: A systematic review and meta-analysis.** *Medicine (Baltimore)* 2016, **95:**e4975.

11. Park JH, Gail MH, Weinberg CR, Carroll RJ, Chung CC, Wang Z, Chanock SJ, Fraumeni JF, Jr., Chatterjee N: **Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants.** *Proc Natl Acad Sci U S A* 2011, **108:**18026-18031.

12. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X: **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:**224-237.

13. Lee J, Kim YJ, Lee J, Kim BJ, Lee S, Park T: **Gene-set association tests for next-generation sequencing data.** *Bioinformatics* 2016, **32:**i611-i619.

14. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M: **CADD: predicting the deleteriousness of variants throughout the human genome.** *Nucleic Acids Res* 2019, **47:**D886-d894.

15. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A: **Detection of nonneutral substitution rates on mammalian phylogenies.** *Genome Res* 2010, **20:**110-121.

16. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4:**1073-1081.

17. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare-variant association testing for sequencing data with the sequence kernel association test.** *Am J Hum Genet* 2011, **89:**82-93.

18. Chen H, Lumley T, Brody J, Heard-Costa NL, Fox CS, Cupples LA, Dupuis J: **Sequence kernel association test for survival traits.** *Genet Epidemiol* 2014, **38:**191-197.

19. Lin WY: **Beyond Rare-Variant Association Testing: Pinpointing Rare Causal Variants in Case-Control Sequencing Study.** *Sci Rep* 2016, **6:**21824.

20. Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Masson G, Holm H, Kong A, Thorsteinsdottir U, Sulem P, et al: **Weighting sequence variants based on their annotation increases power of whole-genome association studies.** *Nat Genet* 2016, **48:**314-317.

21. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al: **Pathogenic Germline Variants in 10,389 Adult Cancers.** *Cell* 2018, **173:**355-370.e314.

22. Bodmer W, Tomlinson I: **Rare genetic variants and the risk of cancer.** *Curr Opin Genet Dev* 2010, **20:**262-267.

23. Rahman N: **Realizing the promise of cancer predisposition genes.** *Nature* 2014, **505:**302-308.

24. Lu C, Xie M, Wendl MC, Wang J, McLellan MD, Leiserson MD, Huang KL, Wyczalkowski MA, Jayasinghe R, Banerjee T, et al: **Patterns and functional implications of rare germline variants across 12 cancer types.** *Nat Commun* 2015, **6:**10086.

25. Southey MC, Goldgar DE, Winqvist R, Pylkas K, Couch F, Tischkowitz M, Foulkes WD, Dennis J, Michailidou K, van Rensburg EJ, et al: **PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS.** *J Med Genet* 2016, **53:**800-811.

26. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, et al: **Germline Mutations in Predisposition Genes in Pediatric Cancer.** *N Engl J Med* 2015, **373:**2336-2346.

27. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al: **An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics.** *Cell* 2018, **173:**400-416.e411.

28. Huang FW: **Towards Greater Inclusion in Cancer Genomics Studies.** *Cancer Res* 2018, **78:**6726-6727.

29. Yuan J, Hu Z, Mahal BA, Zhao SD, Kensler KH, Pi J, Hu X, Zhang Y, Wang Y, Jiang J, et al: **Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers.** *Cancer Cell* 2018, **34:**549-560.e549.

30. Huang FW, Mosquera JM, Garofalo A, Oh C, Baco M, Amin-Mansour A, Rabasha B, Bahl S, Mullane SA, Robinson BD, et al: **Exome Sequencing of African-American Prostate Cancer Reveals Loss-of-Function ERF Mutations.** *Cancer Discov* 2017, **7:**973-983.

31. Petrovics G, Li H, Stumpel T, Tan SH, Young D, Katta S, Li Q, Ying K, Klocke B, Ravindranath L, et al: **A novel genomic alteration of LSAMP associates with aggressive prostate cancer in African American men.** *EBioMedicine* 2015, **2:**1957-1964.

32. Powell IJ, Dyson G, Land S, Ruterbusch J, Bock CH, Lenk S, Herawi M, Everson R, Giroux CN, Schwartz AG, Bollig-Fischer A: **Genes associated with prostate cancer are differentially expressed in African American and European American men.** *Cancer Epidemiol Biomarkers Prev* 2013, **22:**891-897.

33. Wang BD, Ceniccola K, Hwang S, Andrawis R, Horvath A, Freedman JA, Olender J, Knapp S, Ching T, Garmire L, et al: **Alternative splicing promotes tumour aggressiveness and drug resistance in African American prostate cancer.** *Nat Commun* 2017, **8:**15921.

34. Ademuyiwa FO, Tao Y, Luo J, Weilbaecher K, Ma CX: **Differences in the mutational landscape of triple-negative breast cancer in African Americans and Caucasians.** *Breast Cancer Res Treat* 2017, **161:**491-499.

35. Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, Yoshimatsu TF, Pitt JJ, Hoadley KA, Troester M, et al: **Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas.** *JAMA Oncol* 2017, **3:**1654-1662.

36. Keenan T, Moy B, Mroz EA, Ross K, Niemierko A, Rocco JW, Isakoff S, Ellisen LW, Bardia A: **Comparison of the Genomic Landscape Between Primary Breast Cancer in African American Versus White Women and the Association of Racial Differences With Tumor Recurrence.** *J Clin Oncol* 2015, **33:**3621-3627.

37. Loo LW, Wang Y, Flynn EM, Lund MJ, Bowles EJ, Buist DS, Liff JM, Flagg EW, Coates RJ, Eley JW, et al: **Genome-wide copy number alterations in subtypes of invasive breast cancers in young white and African American women.** *Breast Cancer Res Treat* 2011, **127:**297-308.

38. Guda K, Veigl ML, Varadan V, Nosrati A, Ravi L, Lutterbaugh J, Beard L, Willson JK, Sedwick WD, Wang ZJ, et al: **Novel recurrently mutated genes in African American colon cancers.** *Proc Natl Acad Sci U S A* 2015, **112:**1149-1154.

39. Araujo LH, Timmers C, Bell EH, Shilo K, Lammers PE, Zhao W, Natarajan TG, Miller CJ, Zhang J, Yilmaz AS, et al: **Genomic Characterization of Non-Small-Cell Lung Cancer in African Americans by Targeted Massively Parallel Sequencing.** *J Clin Oncol* 2015, **33:**1966-1973.

40. Campbell JD, Lathan C, Sholl L, Ducar M, Vega M, Sunkavalli A, Lin L, Hanna M, Schubert L, Thorner A, et al: **Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations.** *JAMA Oncol* 2017, **3:**801-809.

41. Kytola V, Topaloglu U, Miller LD, Bitting RL, Goodman MM, D. Agostino RB J, Desnoyers RJ, Albright C, Yacoub G, Qasem SA, et al: **Mutational Landscapes of Smoking-Related Cancers in Caucasians and African Americans: Precision Oncology Perspectives at Wake Forest Baptist Comprehensive Cancer Center.** *Theranostics* 2017, **7:**2914-2923.

42. Schumacher SE, Shim BY, Corso G, Ryu MH, Kang YK, Roviello F, Saksena G, Peng S, Shivdasani RA, Bass AJ, Beroukhim R: **Somatic copy number alterations in gastric adenocarcinomas among Asian and Western patients.** *PLoS One* 2017, **12:**e0176045.

43. Deng J, Chen H, Zhou D, Zhang J, Chen Y, Liu Q, Ai D, Zhu H, Chu L, Ren W, et al: **Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations.** *Nat Commun* 2017, **8:**1533.

44. Krishnan B, Rose TL, Kardos J, Milowsky MI, Kim WY: **Intrinsic Genomic Differences Between African American and White Patients With Clear Cell Renal Cell Carcinoma.** *JAMA Oncol* 2016, **2:**664-667.

45. Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: **A global reference for human genetic variation.** *Nature* 2015, **526:**68-74.

46. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, et al: **Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer.** *Cancer Discov* 2017, **7:**410-423.

47. **Pan-cancer analysis of whole genomes.** *Nature* 2020, **578:**82-93.

48. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, Wendl MC, Zhang Q, Koboldt DC, Xie M, Kandoth C, et al: **Integrated analysis of germline and somatic variants in ovarian cancer.** *Nat Commun* 2014, **5:**3156.

49. Jones AV, Chase A, Silver RT, Oscier D, Zoi K, Wang YL, Cario H, Pahl HL, Collins A, Reiter A, et al: **JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms.** *Nat Genet* 2009, **41:**446-449.

50. Kilpivaara O, Mukherjee S, Schram AM, Wadleigh M, Mullally A, Ebert BL, Bass A, Marubayashi S, Heguy A, Garcia-Manero G, et al: **A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms.** *Nat Genet* 2009, **41:**455-459.

51. Campbell PJ: **Somatic and germline genetics at the JAK2 locus.** *Nat Genet* 2009, **41:**385-386.

52. Olcaydu D, Harutyunyan A, Jager R, Berg T, Gisslinger B, Pabinger I, Gisslinger H, Kralovics R: **A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms.** *Nat Genet* 2009, **41:**450-454.

53. Tung NM, Garber JE: **BRCA1/2 testing: therapeutic implications for breast cancer management.** *Br J Cancer* 2018, **119:**141-152.

54. Margolis B, Skolnik EY: **Activation of Ras by receptor tyrosine kinases.** *J Am Soc Nephrol* 1994, **5:**1288-1299.

55. Longley DB, Harkin DP, Johnston PG: **5-fluorouracil: mechanisms of action and clinical strategies.** *Nat Rev Cancer* 2003, **3:**330-338.

56. Ma JY, Yan HJ, Gu W: **Association between BIM deletion polymorphism and clinical outcome of EGFR-mutated NSCLC patient with EGFR-TKI therapy: A meta-analysis.** *J Cancer Res Ther* 2015, **11:**397-402.

57. Zhao M, Zhang Y, Cai W, Li J, Zhou F, Cheng N, Ren R, Zhao C, Li X, Ren S, et al: **The Bim deletion polymorphism clinical profile and its relation with tyrosine kinase inhibitor resistance in Chinese patients with non-small cell lung cancer.** *Cancer* 2014, **120:**2299-2307.

58. Nie W, Tao X, Wei H, Chen WS, Li B: **The BIM deletion polymorphism is a prognostic biomarker of EGFR-TKIs response in NSCLC: A systematic review and meta-analysis.** *Oncotarget* 2015, **6:**25696-25700.

59. Sargent DJ, Marsoni S, Monges G, Thibodeau SN, Labianca R, Hamilton SR, French AJ, Kabat B, Foster NR, Torri V, et al: **Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer.** *J Clin Oncol* 2010, **28:**3219-3226.

60. Tokunaga E, Oda S, Fukushima M, Maehara Y, Sugimachi K: **Differential growth inhibition by 5-fluorouracil in human colorectal carcinoma cell lines.** *Eur J Cancer* 2000, **36:**1998-2006.

61. Menden MP, Casale FP, Stephan J, Bignell GR, Iorio F, McDermott U, Garnett MJ, Saez-Rodriguez J, Stegle O: **The germline genetic component of drug sensitivity in cancer cell lines.** *Nat Commun* 2018, **9:**3385.

62. Daly AK: **Pharmacogenetics: a general review on progress to date.** *Br Med Bull* 2017, **124:**65-79.

63. McLaren PJ, Carrington M: **The impact of host genetic variation on infection with HIV-1.** *Nat Immunol* 2015, **16:**577-583.

64. Munoz-Gonzalez JI, Jara-Acevedo M, Alvarez-Twose I, Merker JD, Teodosio C, Hou Y, Henriques A, Roskin KM, Sanchez-Munoz L, Tsai AG, et al: **Impact of somatic and germline mutations on the outcome of systemic mastocytosis.** *Blood Adv* 2018, **2:**2814-2828.

65. Baune BT, Hohoff C, Berger K, Neumann A, Mortensen S, Roehrs T, Deckert J, Arolt V, Domschke K: **Association of the COMT val158met variant with antidepressant treatment response in major depression.** *Neuropsychopharmacology* 2008, **33:**924-932.

66. Relling MV, Evans WE: **Pharmacogenomics in the clinic.** *Nature* 2015, **526:**343-350.

Appendix: Scientific Contributions to Other Studies From the Dutta Lab

In addition to the contributions described in the preceding chapters, I have also made contributions to the publications described below [1-3].

1. Kiran M, Chatrath A, Tang X, Keenan DM, Dutta A: **A Prognostic Signature for Lower Grade Gliomas Based on Expression of Long Non-Coding RNAs.** *Mol Neurobiol* 2019, **56:**4786-4798.

    - I analyzed the RNA-sequencing data from the caners included in The Cancer Genome Atlas to determine whether or not the long non-coding RNAs that make up our published prognostic signature ("UVA8") are associated with outcome in other cancers and helped with critically revising the manuscript.

2. Saha S, Kiran M, Kuscu C, Chatrath A, Wotton D, Mayo MW, Dutta A: **Long Noncoding RNA DRAIC Inhibits Prostate Cancer Progression by Interacting with IKK to Inhibit NF-kappaB Activation.** *Cancer Res* 2020, **80:**950-963

    - Dr. Manjari Kiran computationally showed that low expression of the long non-coding RNA DRAIC is associated with NF-kappaB activity in prostate cancer. I analyzed RNA-sequencing data from other cancers to show that this finding was true in several other cancers besides prostate cancer. I also generated figures for the publication based on these results and helped with critically revising the manuscript.

3. Kumar P, Kiran S, Saha S, Su Z, Paulsen T, Chatrath A, Shibata Y, Shibata, E, Dutta A: **ATAC-seq identifies thousands of extrachromosomal circular DNA in cancers and cell lines.** *Science Advances.*

    - I developed a methodology to test whether or not changes in copy number from extrachromosomal circular DNAs would likely be detected using SNP genotyping arrays through standard copy number analyses. I also tested whether the genes contained on the extrachromosomal circular DNAs were enriched for certain pathways or functions through gene ontology analysis across the cancers included in The Cancer Genome Atlas. Finally, I identified and reported the oncogenes found on the extrachromosomal circular DNAs that may be driving tumorigenesis in those caners. I generated figures for each of these analyses which are included in the publication and helped with critically revising the manuscript.

A Prognostic Signature for Lower Grade Gliomas Based on Expression of Long Non-Coding RNAs

Manjari Kiran, Ajay Chatrath, Xiwei Tang, Daniel M Keenan, Anindya Dutta

Adapted From:

Kiran M, Chatrath A, Tang X, Keenan DM, Dutta A: **A Prognostic Signature for Lower Grade Gliomas Based on Expression of Long Non-Coding RNAs.** *Mol Neurobiol* 2019, **56:**4786-4798.

Abstract:

Diffuse low-grade and intermediate-grade gliomas (together known as lower grade gliomas, WHO grade II and III) develop in the supporting glial cells of brain and are the most common types of primary brain tumor. Despite a better prognosis for lower grade gliomas, 70% of patients undergo high-grade transformation within 10 years, stressing the importance of better prognosis. Long non-coding RNAs (lncRNAs) are gaining attention as potential biomarkers for cancer diagnosis and prognosis. We have developed a computational model, UVA8, for prognosis of lower grade gliomas by combining lncRNA expression, Cox regression, and L1-LASSO penalization. The model was trained on a subset of patients in TCGA. Patients in TCGA, as well as a completely independent validation set (CGGA) could be dichotomized based on their risk score, a linear combination of the level of each prognostic lncRNA weighted by its multivariable Cox regression coefficient. UVA8 is an independent predictor of survival and outperforms standard epidemiological approaches and previous published lncRNA-based predictors as a survival model. Guilt-by-association studies of the lncRNAs in UVA8, all of which predict good outcome, suggest they have a role in suppressing interferon-stimulated response and epithelial to mesenchymal

transition. The expression levels of eight lncRNAs can be combined to produce a prognostic tool applicable to diverse populations of glioma patients. The 8 lncRNA (UVA8) based score can identify grade II and grade III glioma patients with poor outcome, and thus identify patients who should receive more aggressive therapy at the outset.

Long Noncoding RNA DRAIC Inhibits Prostate Cancer Progression by Interacting with IKK to Inhibit NF-kappaB Activation

Shekhar Saha, Manjari Kiran, Cana Kuscu, Ajay Chatrath, David Wotton, Marty W Mayo, Anindya Dutta

Adapted From:

Saha S, Kiran M, Kuscu C, Chatrath A, Wotton D, Mayo MW, Dutta A: **Long Noncoding RNA DRAIC Inhibits Prostate Cancer Progression by Interacting with IKK to Inhibit NF-kappaB Activation.** *Cancer Res* 2020, **80:**950-963

Abstract:

DRAIC is a 1.7 kb spliced long noncoding RNA downregulated in castration-resistant advanced prostate cancer. Decreased DRAIC expression predicts poor patient outcome in prostate and seven other cancers, while increased DRAIC represses growth of xenografted tumors. Here, we show that cancers with decreased DRAIC expression have increased NF-κB target gene expression. DRAIC downregulation increased cell invasion and soft agar colony formation; this was dependent on NF-κB activation. DRAIC interacted with subunits of the IκB kinase (IKK) complex to inhibit their interaction with each other, the phosphorylation of IκBα, and the activation of NF-κB. These functions of DRAIC mapped to the same fragment containing bases 701-905. Thus, DRAIC lncRNA inhibits prostate cancer progression through suppression of NF-κB activation by interfering with IKK activity.

SIGNIFICANCE: A cytoplasmic tumor-suppressive lncRNA interacts with and inhibits a major kinase that activates an oncogenic transcription factor in prostate cancer.

ATAC-seq identifies thousands of extrachromosomal circular DNA in cancers and cell lines

Pankaj Kumar, Shashi Kiran, Shekhar Saha, Zhangli Su, Teressa Paulsen, Ajay Chatrath, Yoshiyuki Shibata, Etsuko Shibata, Anindya Dutta

Adapted From:

Kumar P, Kiran S, Saha S, Su Z, Paulsen T, Chatrath A, Shibata Y, Shibata, E, Dutta A: **ATAC-seq identifies thousands of extrachromosomal circular DNA in cancers and cell lines.** *Science Advances.*

Abstract:

Extrachromosomal circular DNAs (eccDNAs) are usually somatically mosaic and a source of intercellular heterogeneity in normal and tumor cells. Because short eccDNAs are poorly chromatinized, we hypothesized that they are sequenced by tagmentation in ATAC-seq experiments, without any enrichment of circular DNA, and thus identified thousands of eccDNAs. The eccDNAs identified in cell lines were validated by inverse PCR on DNA that survives exonuclease digestion of linear DNA, and by metaphase FISH. ATAC-seq in Gliomas and Glioblastomas identify hundreds of eccDNAs, including one containing the well-known EGFR gene amplicon from chr7. Over 18,000 eccDNAs, many carrying known cancer driver genes, are identified in a pan-cancer analysis of 360 ATAC-seq libraries from 23 tumor types. Because of somatic mosaicism, eccDNAs are identified by ATAC-seq even before amplification of the locus is recognized by genome-wide copy number variation measurements. Thus, standard ATAC-seq is a sensitive method to detect eccDNA present in a subset of tumor cells, ready to be amplified under appropriate selection, as during therapy.