

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Incorporating Germline Variants into Cancer Analyses: What Lies Beneath

Permalink

<https://escholarship.org/uc/item/6c12t3q3>

Author

Buckley, Alexandra

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

**Incorporating Germline Variants into Cancer Analyses:
What Lies Beneath**

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Biomedical Sciences

by

Alexandra Ray Buckley

Committee in charge:

Professor Nicholas J. Schork, Chair
Professor Arshad Desai, Co-Chair
Professor Vineet Bafna
Professor Hannah Carter
Professor Olivier Harismendy
Professor Elizabeth Winzeler

2018

Copyright
Alexandra Ray Buckley, 2018
All rights reserved.

The dissertation of Alexandra Ray Buckley is approved,
and it is acceptable in quality and form for publication
on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2018

DEDICATION

To my family, for listening to me endlessly complain

To my friends, for the laughs, food, and beer

EPIGRAPH

And the haters gonna hate, hate, hate, hate, hate

Baby, I'm just gonna shake, shake, shake, shake, shake

I shake it off

Shake it off

TABLE OF CONTENTS

Signature Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	xi
Acknowledgements	xiii
Vita	xvi
Abstract of the Dissertation	xvii
Chapter 1 The Value of Germline Variation in Cancer Research	1
1.1 Background	1
1.1.1 Germline Predisposition to Cancer	4
1.1.2 Heritability of Somatic Phenotypes	7
1.1.3 Heritability of DNA Damage Response	8
1.1.4 Somatic Molecular Phenotypes	9
1.1.5 Tumor Immune Phenotypes, Cell Composition, and Metas- tasis	15
1.1.6 Other Reasons to Consider Germline Variants in Cancer Studies	17
1.2 Overview and Organization of Dissertation	20
1.3 Acknowledgements	22
Chapter 2 Pan-Cancer Analysis Reveals Technical Artifacts in TCGA Germline Variant Calls	23
2.1 Abstract	23
2.2 Background	24
2.3 Methods	27
2.3.1 Cohort	27
2.3.2 Germline Variant Calling	27
2.3.3 PCA and Self-Report Ancestry Validation	28
2.3.4 Annotation and BAM metrics	29
2.3.5 Realignment Comparison	29
2.3.6 WGA Enriched Indels	30

	2.3.7	Homopolymer Indel Analyses	31
	2.3.8	Chimera Read Analysis	31
	2.3.9	Repeated Samples	32
	2.3.10	Indel Filter Methods	32
	2.3.11	Statistical Methods	33
2.4	Results		34
	2.4.1	Technical Heterogeneity in TCGA WXS Data Generation	34
	2.4.2	Impact of Technical Heterogeneity on Loss of Function Variants	37
	2.4.3	Characterizing WGA Artifacts	42
	2.4.4	Filtering Artifactual LOF Variant Calls	48
	2.4.5	Consequences of Technical Artifacts on Genetic Associations	50
2.5	Discussion		53
2.6	Acknowledgements		57
Chapter 3	Exome-Wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in the Cancer Genome Atlas		58
	3.1	Abstract	58
	3.2	Background	59
	3.3	Methods	61
	3.3.1	Data Acquisition	61
	3.3.2	Variant Annotation and Filtering	62
	3.3.3	Somatic Methylation	62
	3.3.4	Loss of Heterozygosity	63
	3.3.5	Gene Set Enrichment Analysis	63
	3.3.6	Mutational Signature Analysis	64
	3.3.7	Statistical Analyses	64
3.4	Results		66
	3.4.1	MMR Pathway is Frequently Affected by Bi-allelic Alteration	66
	3.4.2	Landscape of Germline and Somatic Alteration of DNA Damage Repair Pathways	67
	3.4.3	Individuals in TCGA Exhibit Lynch Syndrome Characteristics	69
	3.4.4	Using MSI-H to Reclassify Variants of Unknown Significance	73
	3.4.5	Missense Alterations Exhibit an Attenuated Lynch Phenotype	77
	3.4.6	Mono-allelic Germline Alteration has Little Effect on Somatic MSI	79
	3.4.7	Methylation of SHPRH Associated with Somatic MSI . .	79
	3.4.8	Mono-allelic Germline Alterations not Associated with Mutational Signatures	82

	3.4.9 Cancer Predisposition Syndromes in TCGA	83
	3.5 Discussion	86
	3.6 Acknowledgements	90
Chapter 4	Rare Variant Phasing Using Paired Tumor:Normal Sequence Data . .	92
	4.1 Abstract	92
	4.2 Background	93
	4.3 Methods	95
	4.3.1 Data Acquisition	95
	4.3.2 Variant Annotation and Filtering	96
	4.3.3 Implementation of VAF Phasing	96
	4.3.4 Comparison To Other Phasing Methods	97
	4.3.5 HMMvar Annotation and Compound Heterozygosity Analysis	98
	4.3.6 Statistical Analyses	99
	4.4 Results	99
	4.4.1 Phasing with Variant Allele Frequency	99
	4.4.2 VAF Phasing is Concordant with Other Methods	106
	4.4.3 Application of VAF Phasing to Cancer Predisposition . .	108
	4.5 Discussion	111
	4.6 Acknowledgements	114
Chapter 5	The Upshot	116
	5.1 Discussion	116
	5.1.1 Batch Effects in Public Datasets	116
	5.1.2 Germline Variants and Tumor Phenotypes in TCGA . . .	117
	5.1.3 Germline Variant Phasing in Cancer Samples	119
	5.1.4 Tumors are Like Onions	120
	5.1.5 Germline Variation, What is it Good For?	122
	5.2 Future Directions	123
	5.2.1 The Upshot	127
Appendix A	Supplemental Material: Pan-Cancer Analysis Reveals Technical Artifacts in TCGA Germline Variant Calls	128
Appendix B	Supplemental Material: Exome-Wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in the Cancer Genome Atlas	150
Appendix C	Supplemental Material: Rare Variant Phasing Using Paired Tumor:Normal Sequence Data	176
Bibliography	201

LIST OF FIGURES

Figure 1.1: The Environment Shapes Growth	3
Figure 2.1: Technical Covariates in TCGA WXS Samples	35
Figure 2.2: WGA Increases LOF Indel Burden	40
Figure 2.3: Characteristics of Variants in WGA Samples	43
Figure 2.4: Comparison of Indel Filters	49
Figure 2.5: Association Between LOF Burden and Cancer Type	52
Figure 3.1: Frequency of Germline and Somatic Alterations in Cancer-Relevant Pathways	68
Figure 3.2: Genetic and Clinical Characteristics of MSI-H Individuals	71
Figure 3.3: Identification of Potential Pathogenic Lynch Syndrome Variants	75
Figure 3.4: Germline, Somatic, and Epigenetic Associations with MSI	80
Figure 3.5: Cancer Predisposition Syndromes in TCGA	84
Figure 4.1: Overview of VAF Phasing Method	101
Figure 4.2: Using Duplicated Normal Samples to Identify SCNAs	105
Figure 4.3: Comparison of Phasing Methods	107
Figure 4.4: Leveraging Phase to Identify Cancer Predisposing Germline Variation	111
Figure 5.1: Tumors are Like Onions	121
Figure A.1: Technical Covariates of Cohort	129
Figure A.2: Number of Processing Workflows	130
Figure A.3: Variant Call Discordance Between NewAlign and OldAlign	131
Figure A.4: Discordance With BAM Realignment	132
Figure A.5: PCA of Common Variants	133
Figure A.6: LOF SNV and Indel Burden	134
Figure A.7: LOF Indel Burden in NewAlign Cohort	135
Figure A.8: Coverage and Read Depth in WGA Samples	135
Figure A.9: Frequently Inserted and Deleted Bases of WGA Indels	136
Figure A.10: Discordance Between Repeated WXS Samples	137
Figure A.11: Proposed Mechanism of Artifactual Indel Generation	138
Figure A.12: Distribution of Indel Sequence BLAST Hits	139
Figure A.13: Individual LOF Indel Burden Across Filtering Methods	140
Figure A.14: LOF Indel Burden in WGA Samples Across Filtering Methods	141
Figure A.15: G/C Homopolymer Content of Genes Shared Between OV and LAML	142
Figure A.16: LOF SNV Logistic Regression Analysis	143
Figure B.1: Calling Somatic Methylation Status	151
Figure B.2: Example LOH Events	152
Figure B.3: Genes Frequently Affected by Germline:Somatic Alteration	153

Figure B.4: Association Between Germline LOF Burden and Cancer Type	154
Figure B.5: Germline and Somatic LOF in <i>PMS2</i>	155
Figure B.6: Mutational Signature Analysis of Germline:Somatic MMR Alteration .	156
Figure B.7: Mono-allelic Germline MMR Variation Not Associated With MSI . . .	157
Figure B.8: Association Testing Between Genomic Alteration and MSI Burden . .	158
Figure B.9: <i>SHPRH</i> Methylation in Uterine Cancer	159
Figure B.10: <i>SHPRH</i> Expression in Normal Tissues	160
Figure B.11: Mutational Signature Analysis of <i>MLH1</i> and <i>SHPRH</i> Methylation . .	161
Figure B.12: Co-Occurrence Testing for <i>SHPRH</i> Methylation	162
Figure B.13: Mutational Signature Analysis of <i>BRCA1/2</i> Carriers	163
Figure B.14: Mutational Signature Analysis of DDR Pathway Alteration	164
Figure B.15: Association Between Age and Damaging Germline Variants	165
Figure C.1: Schematic of Δ VAF Changes in Cancer	177
Figure C.2: Example Δ VAF Data	178
Figure C.3: Example VCF-CBS Data	179
Figure C.4: VAF-CBS Segment Metrics	180
Figure C.5: Δ VAF Data From a Contaminated Sample	181
Figure C.6: Discordance Between Phasing Methods	182
Figure C.7: Sample Metrics that Affect Δ VAF Phasing	183
Figure C.8: Phasing Performance on Rare Variants	184
Figure C.9: Fraction of Phased Variants Visualized by Chromosome	185
Figure C.10: Comparison Between TCGA and VAF-CBS SCNAs	186
Figure C.11: Method Used to Calculate Pairwise Error	187
Figure C.12: Discordance Between VAF and 10X Genomics Phasing	188
Figure C.13: Fraction of Compound Heterozygosity Events Phased	189
Figure C.14: Association Between Age and Damaging Germline Variants	190
Figure C.15: Association Between Age and <i>BRCA1/2</i> Germline Variants	191
Figure C.16: Assumptions of VAF Phasing Model	192

LIST OF TABLES

Table 2.1: ANOVA of LOF Variant Burden	41
Table 2.2: Characteristics of WGA Indels	47
Table 2.3: Variant Filter Metrics	50
Table 3.1: Bi-allelic Germline:Somatic MMR Alteration	78
Table A.1: Composition of the Pan-Cancer Cohort	144
Table A.2: Coverage of the Six TCGA Capture Kits	145
Table A.3: K-means Cluster Membership of HapMap Samples	145
Table A.4: K-means Cluster Membership of TCGA Samples	146
Table A.5: GC Content of the Sequence Surrounding WGA Indels	146
Table A.6: Allele Frequency of Homopolymer Indels	146
Table A.7: Frequency of BLAST Match for WGA Indels	147
Table A.8: ANOVA of LOF Indel Burden Using Different Filters	148
Table A.9: Correlation Between LOF Indels and Homopolymer Tracts	149
Table B.1: Association Between MSI and MMR Alteration	165
Table B.2: Association Between Age and MMR Alteration	166
Table B.3: Germline Variants Pathogenic for Lynch Syndrome	166
Table B.4: Germline Variants of Unknown Significance for Lynch Syndrome	167
Table B.5: Modeling a Germline:Somatic Interaction for L-MMR Genes	169
Table B.6: Association Between MSI and MMR Germline:Somatic Alteration	170
Table B.7: Association Between Age and MMR Germline:Somatic Alteration	170
Table B.8: Association Between MSI and Mono-allelic Germline MMR Variants	171
Table B.9: Association Between MSI and MMR Alteration Including Confounders	172
Table B.10: Association Between MSI and Alterations Correlated With <i>SHPRH</i> Methylation	174
Table B.11: Association Between MSI and <i>SHPRH</i> Expression	174
Table B.12: Association Between Age and Known Predisposing Germline Variants	175
Table B.13: Association Between Age and Predicted Predisposing Germline Variants	175
Table C.1: Possible Contaminated Normal Tissue Samples	193
Table C.2: Discordance Between VAF Phasing and Other Methods	194
Table C.3: Factors That Influence VAF Phasing	195
Table C.4: Discordance Between VAF Phasing with TCGA SCNA and Other Methods	195
Table C.5: Discordance Between VAF and 10X Genomics Phasing	196
Table C.6: Features of VAF Phasing Errors	196
Table C.7: Association Between Age and Compound Heterozygosity	197
Table C.8: Association Between Age and Non-Compensatory Variants	197
Table C.9: Association Between Age and Non-Compensatory Variants, ClinVar Samples Removed	197
Table C.10: Association Between Age and Non-Compensatory Variants in <i>BRCA1/2</i>	198

Table C.11: Germline Non-Compensatory Variants in *BRCA1/2* 199

ACKNOWLEDGEMENTS

They say it takes a village to raise a child, and I feel it took a village to get me through my PhD. First and foremost I'd like to thank my advisor, Nik Schork. He took a chance on me as a runaway grad student and gave me the opportunity and the guidance I needed to metamorphosize from a biologist to a data scientist. My labmates from the Schork lab, for endless help during the aforementioned metamorphosis. Particularly Kris Standish for teaching me everything I know about GATK, R, and baseball and Danjuma Quarless for his perpetually spot-on life advice. Arshad Desai and Tracy Handel, for believing in me when it felt like no one else did. Olivier Harismendy and Hannah Carter for being excellent committee members, collaborators, and mentors. Barry Demchak, for being the best sys admin a grad student could ask for. Rebecca Boumil and Wayne Frankel, for igniting my interest in research and helping me get into grad school. The BMS program and staff for creating such a welcoming and supportive home for me in San Diego. Finally, the two MacBooks that gave their lives for the making of this thesis. Their sacrifice will not be forgotten.

Chapter 1 is being prepared for publication under the title, "Germline Variation in Cancer, What is it Good For?". Authors included Alexandra R. Buckley and Nicholas J. Schork. AB and NJS wrote manuscript. The dissertation author was the primary researcher and author on this manuscript.

Chapter 2 was previously published in *BMC Genomics* in June 2017 under the title, "Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls". Authors included Alexandra R. Buckley, Kristopher A. Standish, Kunal Bhutani, Trey Ideker, Roger S. Lasken, Hannah Carter, Olivier Harismendy, and Nicholas J. Schork. NJS designed and supervised the research. ARB executed pipelines, analyzed data, and drafted the manuscript. KAS helped assemble pipelines and assisted with statistical anal-

ysis. ARB, KAS, and RSL designed figures. ARB, KAS, KB, RSL, OH, and HC designed experiments. NJS, OH, and HC helped write the manuscript. TI set up high performance computing infrastructure. All authors read and approved the final manuscript. The dissertation author was the primary researcher and author on this manuscript.

Chapter 3 is under review for publication in *BMC Genome Medicine* under the title, "Exome-Wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in the Cancer Genome Atlas". Authors included Alexandra R. Buckley, Trey Ideker, Hannah Carter, Olivier Harismendy, and Nicholas J. Schork. NJS designed and supervised the research. NJS designed and supervised the research. ARB performed the statistical analysis, prepared the figures and tables, and drafted the manuscript. ARB, OH, and HC designed experiments. NJS, OH, and HC assisted writing the manuscript. TI set up high performance computing infrastructure. All authors read and approved the final manuscript. We would like to thank Bethany Buckley for her assistance in obtaining ClinVar annotations and interpreting germline variants, and Barry Demchak for his assistance with managing data and setting up analysis pipelines on NRNB. The dissertation author was the primary researcher and author on this manuscript.

Chapter 4 is being prepared for publication under the title, "Rare Variant Phasing Using Paired Tumor:Normal Sequence Data". Authors included Alexandra R. Buckley, Trey Ideker, Hannah Carter, Jonathan Keats, and Nicholas J. Schork. NJS designed and supervised the research. ARB executed pipelines, analyzed data, and drafted the manuscript. ARB and HC designed experiments. NJS and HC helped write the manuscript. TI set up high performance computing infrastructure. All authors read and approved the final manuscript. The dissertation author was the primary researcher and author on this manuscript.

This work was funded in part by the National Institute Of General Medical Sciences

of the National Institutes of Health under Award Number T32GM008666, the Translational Genomics Research Institute (TGen), and a gift from the San Diego Cancer Research Institute. NJS and his lab are also supported in part by National Institutes of Health Grants UL1TR001442 (CTSA), U24AG051129, U19G023122, as well as a contract from the Allen Institute for Brain Science. All computing was done using the National Resource for Network Biology (NRNB) P41 GM103504. All primary data were accessed from The Cancer Genome Atlas Research Network (cancergenome.nih.gov).

VITA

- 2011 - B.S. Bachelor of Science, Neuroscience
University of Scranton, Scranton, PA
- 2018 - Ph.D. Doctor of Philosophy, Biomedical Sciences
Biomedical Sciences Graduate Program
University of California San Diego

PUBLICATIONS

Asinof SK, Sukoff Rizzo SJ, **Buckley AR**, Beyer BJ, Letts VA, Frankel WN, Boumil RM. Independent Neuronal Origin of Seizures and Behavioral Comorbidities in an Animal Model of a Severe Childhood Genetic Epileptic Encephalopathy. *PLoS Genetics*

Buckley AR, Standish KA, Bhutani K, Ideker T, Lasken RS, Carter H, Harismendy O, Schork NJ. Pan-Cancer Analysis Reveals Technical Artifacts in The Cancer Genome Atlas (TCGA) Germline Variant Calls. *BMC Genomics*

Buckley AR, Ideker T, Carter H, Harismendy O, Schork NJ. Exome-wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in The Cancer Genome Atlas (TCGA). *Under Review*

Buckley AR, Keats J, Ideker T, Carter H, Schork NJ. Rare Variant Phasing Using Paired Tumor:Normal Sequence Data. *In Preparation*

ABSTRACT OF THE DISSERTATION

**Incorporating Germline Variants into Cancer Analyses:
What Lies Beneath**

by

Alexandra Ray Buckley

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2018

Professor Nicholas J. Schork, Chair
Professor Arshad Desai, Co-Chair

Cancer results from the progressive accumulation of genetic alterations that drive uncontrolled cell growth. The genetic alterations present in a cancer cell originate from two sources: 1) inherited, or germline, variants present in every cell of the body and 2) acquired, or somatic, mutations specific to tumor cells. These two sources of genetic alterations have largely been studied separately: germline variants for their role in cancer risk and somatic mutations for their role in shaping somatic phenotypes. Only recently have these two fields intersected, most notably by the observation that germline *BRCA1/2* variants not

only predispose to cancer but also influence the mutational profile of the resultant tumors. The degree to which germline variation influences somatic phenotypes in sporadic cancer remains unclear. We propose that similar to how the climate of a region influences the local flora and fauna, germline variation in genes mediating processes such as DNA damage repair, immune response, and drug metabolism, shapes tumor development.

In this work, we study germline variation in 9,099 individuals from the Cancer Genome Atlas (TCGA) with the goal of identifying associations between germline variants and somatic phenotypes and determining what, if any, value is added by integrating germline variants into cancer analyses. A hindrance to this type of study was a lack of publicly available germline variant calls from individuals with cancer. To address this, we developed and implemented a variant calling pipeline to generate a high quality germline variant dataset from TCGA data. Accurately assessing the contribution of germline variants to somatic phenotypes requires models that account for both germline and somatic sources of genetic alterations. We integrated germline variation and somatic mutation, epigenetic modification, and copy number alteration data to identify genetic factors that underlie variation in two somatic phenotypes: microsatellite instability and somatic mutational signatures. We further describe a novel method to phase germline variants that leverages unique properties of paired somatic and germline sequence data, and demonstrate the value of including phase information into germline analyses of cancer. Overall, this study illustrates that integration of germline and somatic data can reveal novel biological and methodological insights.

Chapter 1

The Value of Germline Variation in Cancer Research

1.1 Background

It is recognized that cancer results from a progressive accumulation of damaging genetic alterations that drive cells to grow unchecked [1, 2]. In addition to these acquired somatic mutations that drive to tumorigenesis, all cells of the body, including all cells of the tumor, possess inherited germline genetic variants. In this light, the cancer cell genome can be envisioned as two distinct layers: 1. alterations that are inherited and 2. alterations accumulated during somatic cell replication. These two genomic 'layers' have largely been studied separately: somatic mutations to understand how tumors grow and take on certain molecular characteristics, and germline variants to understand heritable risk for cancer. This disconnect between these two sources of genomic alterations is evident from the fact that it has only recently been recommended that joint analysis of paired tumor:normal sequence data is essential for accurate genomic interpretation [3].

Incorporating germline variants into genomic profiling of cancer patients can aid

in identifying cancers that have a heritable origin and provide a deeper understanding of tumor phenotypes in both inherited and sporadic cancers. As germline variants and somatic mutations are both present in tumor cells, they in theory have the same potential to shape tumor phenotypes. Germline variants have been demonstrated to influence somatic phenotypes both independently and through interactions with other somatic mutations [4, 5]. They can also influence the course of disease in a cell-extrinsic manner by influencing how permissive the host environment is toward tumor growth. Similar to how the climate of a region influences the local flora and fauna, there is potential for heritable variation in genes mediating processes such as angiogenesis, immune response, and drug metabolism, to shape individual tumor development (Figure 1.1).

The ability of germline variants to shape tumor phenotypes has been increasingly recognized by the cancer genomics community. For example, it has recently been proposed that germline variants can act as "co-oncogenes", or genetic alterations that are not sufficient to induce cancer on their own, but can complement acquired somatic mutations [4]. This can be seen as an extension of the Knudson two-hit hypothesis, which focused on combined germline and somatic bi-allelic alterations in the same gene, to include germline and somatic alterations co-occurring not necessarily in the same gene, but rather in the same functional pathways [6, 7]. Here, we will build on this idea and more broadly describe the importance of integrating germline variation into cancer analyses. To gain additional perspective, we summarize the current evidence that germline variation modulates cancer risk and shapes somatic phenotypes with an emphasis on somatic phenotypes derived from next generation sequencing (NGS) data and germline variants in the DNA damage repair pathway.

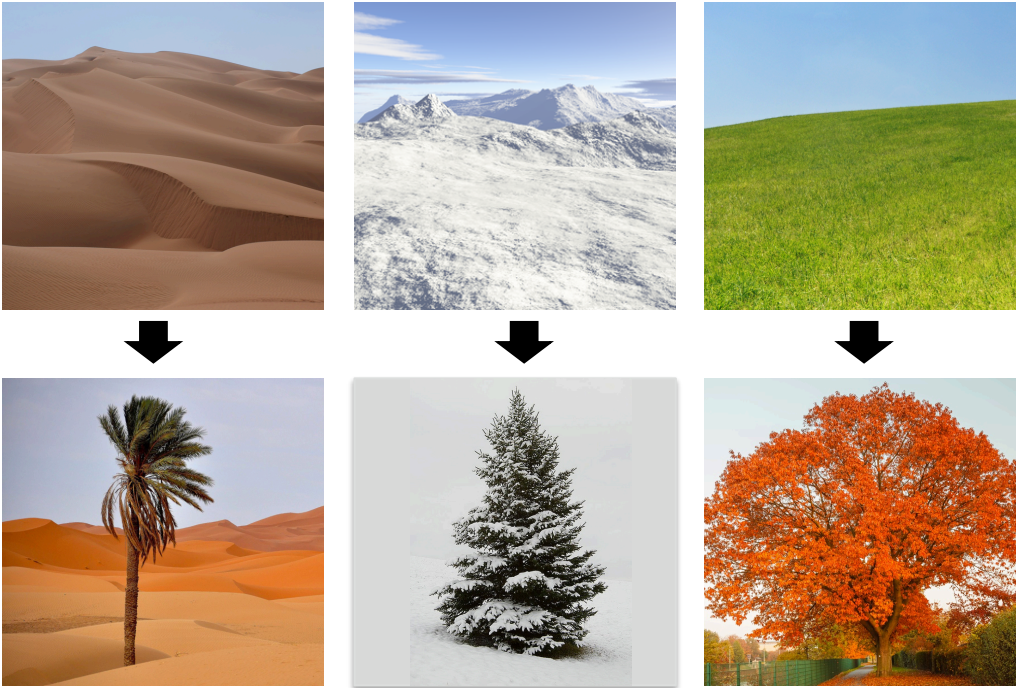
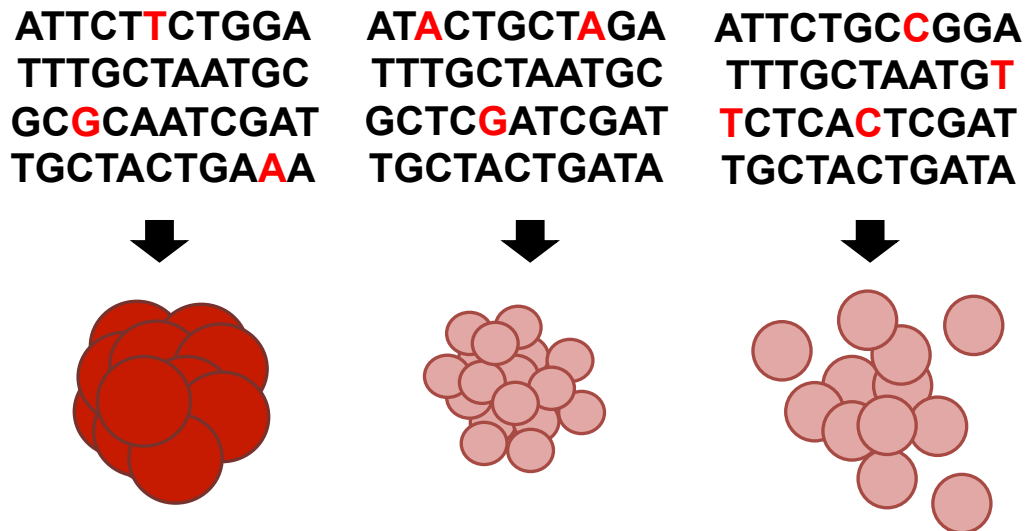
A**B**

Figure 1.1: The Environment Shapes Growth. (A) The climate of a region influences the composition of the local flora and fauna. (B) Similarly, we propose that germline genetic variants shape the host environment and influence tumor characteristics.

1.1.1 Germline Predisposition to Cancer

It has been estimated that 4-7 genetic alterations are required to transform a normal cell to a fully malignant cell [7, 8]. The origins of these genetic alterations have only now been the subject of intense scrutiny. The idea that a germline variant could serve as an initiating genetic alteration was proposed first by Nordling and then Knudson [6, 7]. In Knudson's 'two-hit' model, an individual inherits a single damaging allele ('hit') in a gene, and at a probability related to the somatic mutation rate, acquires a secondary alteration of the remaining wild type allele ('two-hit') in a subset of cells that go on to form a tumor. Genes that follow this 'two-hit' bi-allelic inactivation mechanism typically are haplosufficient: a single genetic 'hit' has no phenotypic effect, dual inactivation is required to drive tumor formation. Inheriting damaging variation in certain genes effectively increases cancer risk by decreasing the number of somatic alterations required to drive a cell to malignancy.

The germline variants observed by Knudson that are known to drive overt risk for cancer are highly damaging and relatively rare. Naturally, highly damaging germline variants that lead to early onset cancer decrease reproductive success and will be selected against in the population [9]. Study of rare and damaging variants that predispose to cancer has led to the description of a number of hereditary cancer predisposition syndromes, each associated with a specific spectrum of cancer types and characteristic clinical features. Cancer predisposition syndromes can result from germline alteration of a single gene, such as Li-Fraumeni and the *TP53* gene, or from germline alteration of a functional pathway, such as Lynch syndrome and the mismatch repair (MMR) pathway [10]. Most of the causal genetic alterations for these syndromes follow the 'two-hit' model of dual germline and somatic inactivation of the predisposing [6]. However, there is also evidence for heritable syndromic predisposition to cancer that doesn't follow this mechanism, such as dominant negative *TP53* alleles [8] or activating mutations seen in RASopathies [11]. While these

syndromes demonstrate a clear association between germline variants and increased cancer risk, they account for only 5-10% of the total incidence of cancer [10].

The ability of germline variants to increase risk outside of these known predisposition syndromes is debated [12, 13]. The estimated heritability of sporadic cancer from twin studies was recently estimated to be 33% [14]. Importantly, there is significant heritability even when excluding highly penetrant germline variants in *BRCA1* and *BRCA2*, which are more common in the population than known cancer syndrome predisposition variants [15, 16]. Damaging germline variation in genes implicated in cancer predisposition is rare in sporadic cancer datasets [17]. Studies estimate 8-11% of adult cancers [18, 19] and 8% of pediatric cancers [20, 21] harbor a likely pathogenic germline variant in a cancer predisposition gene. While this is higher than what was observed in control cohorts, where 1% of patients carried a likely pathogenic germline variant, it still suggests that much of the heritable component of cancer risk is unexplained [21]. It is unlikely that this 'missing heritability' will be largely explained by new high-risk cancer predisposition genes. For example, it has been suggested that it is highly unlikely that more genes exist that increase risk for breast cancer to the same degree as *BRCA1/2* [22].

In terms of genetic factors that underlie susceptibility to cancer, to date genome-wide association studies (GWAS) have identified over 430 associations between common variants and cancer risk [23]. The influence of these loci individually are generally small; however, multiple susceptibility loci can be combined into a polygenic risk score that can identify individuals with high risk [24]. In colorectal cancer, individuals in the top 1% of a polygenic risk score distribution have a 2.9 fold increased risk over the population median [23]. Often the mechanism by which common, non-cancer syndrome-related SNPs identified using GWAS increase cancer risk is unclear, and frequently these SNPs are in noncoding regions. Many noncoding GWAS SNPs, both in the context of cancer and other diseases, have been shown to act as 'eQTLs' that impact the expression of genes that con-

tribute to molecular pathophysiology [25, 26]. For example, multiple cancer susceptibility SNPs increase expression of cancer-relevant genes in normal tissue, such as *PSCA* expression in bladder cancer [27], *SMAD7* expression in colon cancer [28], and *TERT* expression in melanoma [29]. This suggests common variants identified by GWAS can modulate cancer risk through perturbations in expression of genes involved in oncogenic transformation.

The ability of germline variants to influence cancer risk has been shown to be context dependent, both for high-risk cancer predisposition syndrome variants and for low-risk GWAS susceptibility loci. A germline variant may increase cancer risk only in certain tissues, in certain genomic contexts, or following certain carcinogen exposures. Many cancer predisposition genes are integral components of the DNA damage repair pathway, yet they only increase risk for specific cancer types [17]. It has not been fully elucidated how ubiquitously expressed genes can cause tissue-specific patterns of cancer. Similarly, the majority of GWAS loci predispose to a specific cancer type, they don't increase risk for cancer generally [23]. Together, this suggests that the effect of cancer predisposing germline variants varies across tissue types. This type of interaction was demonstrated for the 8q24 loci, which increases risk for prostate, colon, and breast cancer. Cancer-type specific risk is achieved through tissue-specific 3D genome looping interactions with a nearby oncogene *MYC* [30]. The effects of germline variants can be modified by other genetic alteration elsewhere in the genome. Genetic modifier effects that delay cancer onset have been described for variants pathogenic for both Lynch [31, 32] and Li-Fraumeni syndrome [33]. Similarly, a SNP in *RAD51C* has been shown to increase cancer risk in *BRCA1/2* pathogenic variant carriers [34]. Finally, lifestyle factors can modulate the effect of germline variants; for example, polymorphisms in carcinogen metabolizing genes are associated with bladder cancer risk in the context of cigarette smoking [35]. As these examples demonstrate, the association between germline variation and cancer risk can be indirect and modulated by other factors.

1.1.2 Heritability of Somatic Phenotypes

If heritable genetic variation can shape tumor development, it follows that genetically similar individuals should develop phenotypically similar tumors. Two studies have tested this hypothesis using individuals who developed multiple independent cancers. The first examined multiple distinct kidney tumors arising in individuals carrying pathogenic germline variants in the *VHL* gene. They observed that all tumors acquired clonal chromosome 3p loss and somatic mutation of the *PI3K* signaling pathway, but that each tumor acquired different specific genetic alterations, suggesting convergent evolution to a similar somatic phenotype [36]. The second examined high multiplicity squamous cell carcinomas (SCCs) in organ transplant patients and observed more similar somatic copy number profiles in tumors arising in the same individual than in tumors across the cohort [37]. While these studies offer insight into how genetic similarity correlates with somatic phenotypic similarity, both study somewhat unique cases of individuals with an extreme predisposition to cancer.

Studies of monozygotic and dizygotic twins have offered insight into the heritability of cancer risk, but few twin studies have incorporated molecular profiling of tumor characteristics [16, 14]. Somatic phenotypes have been studied in infant twins with concordant leukemia, a rare phenomenon that occurs through intraplacental transfer of tumor-initiating cells between monozygotic twins [38]. Due to this unique method of tumor initiation, both twins share a common driving genetic alteration; however, the development of further genetic alteration in the tumors varied between twins. In one case it was reported that similar copy number alterations occurred in both twins [39], and another that there was little similarity [40]. Thus, evidence is inconclusive, and again confounded by the fact that these studies focus on an atypical form of cancer.

Phenotypic differences due to strain background in genetic mouse models are common [41]. Mouse models of cancer can provide insight into the role of genetic background

in cancer development. TP53 knockout models of Li-Fraumeni show different rates and types of tumors depending on mouse strain background [42]. Similarly, while all strain backgrounds of *APC* mutant models of FAP develop colon polyps, the number varies considerably. This effect was mapped to a strain-specific frameshift insertion in *PLA2G2A* [41, 43]. These examples demonstrate how naturally occurring genetic variation can modify the course of cancer in mouse models. Similarly, engineered genetic variation can alter somatic phenotypes. *NF1* deficient models of neurofibromatosis only develop tumors when *NF1* is deleted in Schwann cells and hemizygous in non-neoplastic cells: mice with *NF1* deletion in Schwann cells and a wild type *NF1* genetic background don't develop cancer [44]. Fewer tumor infiltrating immune cells were observed in mice with a wild type *NF1* background, indicating that genetic differences in the host environment, particular host cells that interact with cancer cells, play a key role in the development of neurofibromas. Both limited studies in humans and extensive work in mouse models demonstrate that genetic background can alter tumor phenotypes and cancer progression.

1.1.3 Heritability of DNA Damage Response

Defects in DNA repair are closely linked with the development of cancer, as evidenced by the fact that many predisposition genes are involved in the DNA damage response (DDR) [45, 17]. A typical somatic cell acquires tens of thousands of DNA lesions per day, giving a cell ample opportunity to acquire transforming mutations should these lesions not be faithfully repaired [46]. Defects in DDR pathways can leave a signature pattern of somatic alterations in a tumor, as evidenced by high levels of microsatellite instability (MSI) in MMR defective tumors [47, 48] and the well-described homologous recombination deficient (HRD) genomic rearrangement pattern seen in *BRCA1/2* defective tumors [49]. Thus, defective DDR can both increase cancer risk and determine the molecular phenotype of a tumor.

The ability to repair DNA lesions has been demonstrated to vary between individuals and is heritable [50, 9, 51]. A common method used to estimate DDR heritability involves isolating lymphocytes from twin pairs and unrelated individuals, exposing them to a DNA damaging agent, and assaying the degree of damage accumulated. Using this approach, multiple DDR phenotypes have been shown to be heritable, including: irradiation (IR)-induced apoptosis and cell cycle delay [52, 53], IR-induced micronuclei burden [54], basal micronuclei burden [54], and bleomycin-induced chromatid breaks [55]. In some instances, specific genetic variants have been associated with DDR defects, but these studies were conducted using a limited number of candidate genetic variants [56].

It has been proposed that defects in DDR could be used as an intermediate phenotype to predict cancer risk [54]. Heritability estimates of DDR are higher than the heritability estimates of cancer, characteristic of an intermediate phenotype. In line with this reasoning, lymphocytes isolated from cancer patients show DDR defects following mutagen exposure [57, 58]. Many of these studies are a retrospective case control design between healthy individuals and cancer patients, thus it is impossible to distinguish if the DDR defect is constitutional or due to the cancer. However, one study included unaffected family members and observed a stepwise increase in DDR defects, as measured by the comet assay and micronuclei burden, between healthy controls, unaffected family members, and affected patients [59]. It is intuitive to think of DDR defects influencing cancer risk and phenotypes, but there is also potential for robust DDR to prevent cancer. For example, 'Super p53' mice with an extra copy of *TP53* under endogenous transcriptional control show decreased cancer incidence [60].

1.1.4 Somatic Molecular Phenotypes

Novel analytical methods have been developed to dissect a number of molecular phenotypes from paired tumor:normal HTS data, such as mutational signatures [61, 62],

MSI [47, 48], and burden of tumor infiltrating immune cells [63, 64]. Many of these phenotypes exhibit a high degree of variability between cancer types and even between individual tumors within a single cancer type. While a large fraction of this variability can be explained by somatic alterations, a number of associations between germline variants and somatic molecular phenotypes have been discovered. Below we will summarize these findings, with an emphasis on DDR pathways and somatic mutation, copy number alteration, and methylation phenotypes.

As somatic mutation of DDR genes has been associated with characteristic patterns of somatic base substitutions and the ability to repair DNA is heritable, it follows logically that germline variation in DDR genes can also influence somatic mutational patterns. Rare germline variants in DDR genes have been associated with an overall increase in the number of somatic mutations [65]. Interestingly, a GWAS approach has also identified a common haplotype at 11q22 that is also associated with increased somatic mutation burden that does not seem to be driven by a DDR gene [19]. The strongest evidence that germline variants can shape the type of somatic mutations a tumor acquires comes from the study of somatic mutational signatures. Mutational signature analysis is a method that uses the profile of somatic base substitutions and flanking bases to identify patterns that reflect an underlying mutational processes, or 'signature' [61, 62]. Pathogenic germline variants in *BRCA1/2* are associated with an increased number of somatic mutations produced by mutational signature 3. However, this association requires bi-allelic germline and somatic *BRCA1/2* alteration [66, 67, 68, 19]. In contrast, mono-allelic damaging germline variants in homologous recombination genes *PALB2*, *FANCD2*, and *FANCM* have been associated with signature 3 [65, 69, 68]. Mono-allelic variants in *PALB2* in particular have been shown to cause DDR defects in cell lines and primary lymphocyte models [70]. In what contexts germline variants require bi-allelic alteration to affect change in somatic mutational profile remains an open question. For example, there is conflicting

evidence that mono-allelic alteration of mismatch repair can increase somatic MSI burden [71, 72].(cite our paper). Outside of DDR genes, associations have been found between common variants in the *APOBEC3* region and signatures 2 and 13 [73, 19], and between rare variants in *MDB4* and signature 1 [19]. An intriguing study showed that germline HLA type can restrict the type of somatic mutations observed both at the individual and population level [74]. Using germline HLA type and predicted neoantigen peptide binding efficiencies, the authors show that the most frequently observed somatic mutations are those that produce a neoantigen epitope predicted to be poorly presented by HLA. While HLA type does not alter somatic mutational processes, it can indirectly influence the final somatic mutational profile of a tumor by determining what somatic mutations are more likely to escape detection by the immune system.

We have discussed somatic mutations in terms of physical changes to DNA and mutational phenotypes as patterns of base pair substitutions across the genome. Somatic mutations can also be analyzed in terms of the genes and functional pathways altered in the tumor. It has been proposed that inherited variation can act as a 'co-oncogene' to complement acquired somatic mutations, both at the single gene level in a 'two-hit' mechanism, and at the functional pathway level [4]. Hanahan and Weinberg have described 'hallmark' functional processes that confer oncogenic potential to cancer cells and are frequently somatically altered [75]. For each 'hallmark' process, there are different paths a tumor can take to acquire oncogenic traits. For example, cells can become resistant to death through somatic upregulation of anti-apoptotic factors or somatic upregulation pro-survival signals. The choice of what path a tumor takes, and therefore what genes are somatically altered, could be influenced by germline variants that alter protein function or expression of 'hallmark' pathway genes. The average individual carries 85 heterozygous and 35 homozygous loss of function (LOF) germline variants [76], and considerably more variants predicted to be damaging or alter gene expression. Should this hypothesis be true,

it is expected that there will be a relationship between the germline variants a person carries and what somatic mutations they acquire. Indeed, co-occurrence and mutual exclusivity between specific germline variants and somatic mutations has been shown [4, 77, 65, 78]. A haplotype at the 19p13.3 locus has been associated with *PTEN* somatic mutations [77], and two common SNPs have been associated with *PIK3CA* mutation, possibly via acting as *cis*-eQTLs and increasing expression of *MAP3K1* and *SETD9* [78]. *ATM* germline truncations and *TP53* somatic mutations were found to be mutually exclusive, supporting the idea that germline dysregulation of the apoptotic pathway obviates the need for somatic mutation of the pathway [65]. Finding meaningful relationships between specific germline and somatic alterations is difficult as most damaging alterations are rare, and there are an incredible number of pairwise hypotheses that could be tested. One approach is to bin alterations by gene or by functional pathway. It has been shown that leveraging known biological network data to smooth somatic mutation profile produces clusters of individuals that are predictive of overall survival [79]. Further, most *BRCA1/2* germline carriers fell within the same 'network-smoothed' cluster. This suggests that germline alterations can influence somatic mutation of both specific genes and the overall profile of pathways that are somatically altered.

Somatic copy number alterations (SCNAs) are common in tumors, with approximately 90% of solid tumors exhibiting some degree of aneuploidy [80]. As mentioned above, SCNA profile has been found to be similar in multiple tumors originating in the same individual or twins, suggesting that inherited variation can shape the pattern of SCNAs a tumor acquires [37, 36, 39, 40]. The most frequently studied SCNA phenotype is 'BRCAness', a somatic phenotype identified in *BRCA1/2* germline pathogenic variant carriers characterized by loss of heterozygosity (LOH) events, large-scale transitions, and a distinct somatic mutation pattern described by mutational signature 3 [49, 62]. Using similar methodology as mutational signature analysis, a rearrangement signature has

been defined that describes the BRCAness phenotype [66]. The term BRCAness is not limited to tumors arising in germline *BRCA1/2* carriers, associations between other somatic alterations in the HR pathway and a BRCAness somatic profile have been identified [49]. There is suggestive evidence that non-*BRCA1/2* hereditary breast cancer, called 'BRCAX', also have a distinct SCNA phenotype; however this study was limited due to small sample size [81]. Chromothripsis is a distinctive SCNA event whereby chromosomes undergo catastrophic 'shattering' that results in massive rearrangements. An association between pathogenic Li-Fraumeni germline variants and somatic chromothripsis was found in pediatric medulloblastoma, with suggestive evidence that the same association exists in other Li-Fraumeni cancers [82]. Thus, there is evidence that both inherited variation as a whole and specific pathogenic germline variants can influence SCNA profile of tumors.

DNA methylation is a common epigenetic mechanism of gene silencing in tumors. The most direct association between germline variants and somatic methylation is constitutional epimutation of *MLH1* seen in some Lynch syndromes cases. Germline SNPs in *MLH1* regulatory elements have been shown to induce mosaic methylation of *MLH1* in somatic tissues [83]. Similarly, a germline *MGMT* promoter SNP is associated with somatic methylation of *MGMT* in colorectal cancers [84]. The exact mechanism of how germline variants in gene regulatory elements can cause aberrant methylation remains to be elucidated. At a broad level, germline *MTHFR* variants known to decrease *MTHFR* enzymatic activity have been associated with the CpG island methylator phenotype (CIMP) in colorectal cancer [85]. The overall landscape of methylation in a tumor has been used to improve classification of brain tumor subtypes over traditional histopathology classification [86]. Using similar methods, new associations between germline variants and methylation patterns may be found, similar to the associations between *BRCA1/2* and SCNA profiles. Thus far it has been shown that germline variation can influence somatic methylation by rendering a specific locus more liable to DNA methylation, or by altering

carbon metabolizing pathways.

A number of somatic gene expression signatures have been identified to characterize tumors, particularly to identify cancer subtypes with differential overall survival. For example, the PAM50 gene expression signature can predict survival in breast cancer [87], and GBM expression subtypes were defined that predict overall survival [88]. There is some evidence that germline variants can alter overall somatic expression profile. Studies of BRCA families have shown that expression-derived tumor subtypes were more similar within families than between unrelated individuals [89]. However, expression-derived signatures are sensitive to tumor heterogeneity, as differing gene expression among heterogeneous cell types is lost in bulk tumor RNA profiling. It has been shown using multiregion tumor sampling that a single tumor can exhibit all GBM expression subtypes [90]. Similar intratumoral heterogeneity of a prognostic expression signature was also shown in multiregion sampling of kidney cancer [91]. Germline variants have been shown to alter somatic gene expression at the single gene level as *cis*-eQTLs [92]. A study of paired tumor and normal expression data in colorectal cancer revealed that heritable inter-individual variation in gene expression is largely conserved between tumor and normal samples [93]. A study in breast cancer estimated that 1.2% of the variation in somatic gene expression is due to germline *cis*-eQTLs, with somatic copy number alterations and methylation explaining another 7.3% and 3.3% respectively [25]. This indicates that germline variants influence global somatic gene expression profiles, in accordance to what was observed in BRCA familial cancers. Further, a subset eQTLs alter gene expression only in the tumor [93]. These tumor-specific eQTLs have the potential to act as germline driver events that are only activated during oncogenic transformation. While gene expression is a challenging somatic phenotype to study, there is evidence that germline variants acting as eQTLs influence individual differences in somatic expression.

Similar to gene expression subtyping, some cancer types are classified into subgroups

based on histopathology and expression of a few key genes. For example, presence or absence of the estrogen receptor differentiates the two main subtypes of breast cancer. It is well described that *BRCA1* pathogenic germline variant carriers are more likely to develop ER- breast cancers whereas *BRCA2* carriers are more likely to develop ER+ [94]. Interestingly, there are no observed histological or molecular differences between *BRCA1* and *BRCA2* carriers in ovarian cancer [94, 95]. While the relationship between germline variants and cancer subtypes is most well understood in breast cancer, two different regions of the 5p15 locus have been associated with two lung cancer subtypes: squamous cell carcinoma and adenocarcinoma [96]. From the study of cancer predisposition syndromes, it is known that germline variation can influence the tissue affected by cancer. This work demonstrates that germline variation can also influence the histopathological subtype.

1.1.5 Tumor Immune Phenotypes, Cell Composition, and Metastasis

The importance of the immune system in cancer is increasing being recognized, as evidenced by the addition of 'avoiding immune destruction' to the hallmarks of cancer [75] and the great interest in immunotherapy [97]. Evidence that the host immune system plays an important role in cancer development comes from studies of immunocompromised individuals, which show that these individuals are at a higher risk for some cancer types [97]. Interestingly, it has been shown that melanoma can be transferred from an organ donor in remission to an immunosuppressed recipient [98]. It is speculated that the donor's immune system can keep the cancer in a dormant state, but once in the immune-depleted recipient environment, the cancer could grow and spread. While in these instances the host's immune state is influenced by immunosuppressive medication, not inherited variation, it suggests that host differences in immune response can affect tumor development.

Methods exist to determine the composition of infiltrating immune cell types using

bulk tumor gene expression profile [64], and to determine infiltrating T cell abundance by quantifying the number of sequencing reads that correspond to rearranged T cell receptors [63]. These methods have allowed for a more robust quantification of the immune cell environment in large public datasets such as the Cancer Genome Atlas (TCGA) [99]. In an extensive study of the immune component of tumors from 33 cancer types, evidence was found that genetic ancestry can influence *PD-L1* expression, a key target of checkpoint immunotherapy, and the type and abundance of tumor infiltrating lymphocytes [99]. While specific germline variants were not implicated in these associations, it suggests there is potential for inherited variation in immune-related pathways to influence the immune cell composition of the tumor. A pair of studies on the rs351855 germline polymorphism in the *FGFR* gene exquisitely highlights how a specific germline variant can alter immune infiltration, and how a single variant can have both cell-intrinsic and cell-extrinsic effects on tumor development [100, 101]. In the first study, it was shown that this polymorphism creates a novel *STAT3* binding site that enhances *STAT3* signaling and cell-intrinsic tumor growth in a transgenic mouse model carrying a homozygous rs351855 polymorphism [101]. In a follow-up study, it was additionally shown that this polymorphism alters the balance of CD8⁺ T cells and T regulatory cells systemically, ultimately resulting in fewer infiltrating T cells in the tumors of transgenic homozygous rs351855 mice [100]. Another interesting avenue of investigation in tumor immunology is to understand the genetic determinants of immunotherapy response. An estimated 9% of patients have accelerated progression in response to immunotherapy, termed 'hyperprogressors' [102]. Thus far there are few genetic markers that can identify which patients may have a negative response to immunotherapy. It would be interesting to examine heritable variation in host immune response as a potential explanatory factor.

Associations between inherited variation and other, less easily quantified tumor characteristics have been reported. It is now well recognized that heterotypic interactions

between tumor cells and host stromal cells play an important role in influencing tumor growth [75]. A fact that is often overlooked when pursuing this line of study is that host stromal cells vary between individuals, and this host:tumor interaction may vary depending on heritable characteristics of the host cells. This has been demonstrated in prostate cancer, where polymorphisms in *ASPN* are associated with an increased risk of metastatic disease [103]. While the exact mechanism is unclear, *ASPN* is highly expressed in cancer associated fibroblasts (CAFs), and a mouse model where CAFs are engineered to overexpress *ASPN* risk alleles showed more metastases. It has been proposed that metastasis is a stochastic process [2]; however, this finding suggests certain host environments may be more hospitable to metastatic cells. Host differences in angiogenesis have also been implicated in tumor development. Mice engineered to lack endogenous inhibitors of angiogenesis show enhanced angiogenesis and faster tumor growth [104]. Interestingly, these mice show no phenotype without tumor induction, demonstrating that germline variation that produces no overt systemic phenotype can influence tumor growth. Finally, a suggestive association between germline variation in the *ADAMTSL1* gene and overall survival have been found in breast cancer, but further study is required to confirm this finding and determine a mechanism [105].

1.1.6 Other Reasons to Consider Germline Variants in Cancer Studies

Incorporating germline variants into the analysis of cancer samples can be informative not only for tumor phenotyping but also for personalized therapy. It is known that germline variation can alter metabolism of common chemotherapeutics and the propensity to have an adverse event in response to therapy [106]. For example, genetic variation in *CYP2D6* alters the metabolism of the prodrug tamoxifen to the active metabolite endoxifen. It has been demonstrated that there are significant differences in plasma concentration

of endoxifen based on *CYP2D6* genotype [107]. Increasing tamoxifen dose in low metabolizers abrogates this difference without a significant change in adverse events. While it has not been shown that increased plasma endoxifen correlates with improved response, these results suggest germline *CYP2D6* genotype may be useful tool when deciding tamoxifen dosing in breast cancer. Study of immortalized lymphoblastoid cell lines from 14 families estimated that the heritability of response to 29 common chemotherapeutic agents ranged from 0.06 - 0.64 [108]. Much like in the case of DDR capacity, this study suggests that it is possible to identify specific germline variants that underlie drug response. The authors suggest the need for large-scale studies of chemotherapy response to identify pharmacologic QTLs (pQTLs).

In a similar vein, germline variants may help inform the choice of chemotherapeutic agent for an individual's cancer. The breast and ovarian cancers of *BRCA1/2* pathogenic germline variant carriers exhibit high sensitivity to drugs that induce replication fork collapse and DNA double strand breaks, such as platinum agents and *PARP* inhibitors [49]. These drugs induce DNA damage that would normally be repaired via HR; however, in HRD *BRCA1/2* cancers this onslaught of genetic mutation overwhelms the cell's capacity to repair leading to apoptosis or general loss of cell fitness. The increased drug sensitivity of *BRCA1/2* breast cancer is also associated with a greater overall survival [109]. It is speculated that the improved survival is due to the fact that *BRCA1/2* carriers harbor clonal *BRCA1/2* haploinsufficiency, making the development of drug-resistant subclones less likely. This idea is supported by the observation that somatic back mutation of germline *BRCA1/2* variants is a common mechanism of platinum resistance [110]. Interestingly, platinum sensitivity was also observed in ovarian cancer patients carrying germline defects in other HR genes, suggesting this phenomenon is not limited to *BRCA1/2* [109]. *PARP* inhibitors are commonly used as a maintenance therapy following platinum-based chemotherapy in germline *BRCA1/2* carriers. The use of *PARP* inhibitors in *BRCA1/2*

carriers represents the first targeted treatment for an inherited cancer disorder [49]. There is also potential to target chemopreventive agents using germline carrier status, such as the use of daily aspirin in Lynch syndrome patients [111]. Tailoring treatment to pathogenic germline variants gives a unique opportunity to target a genetic alteration that is clonal in the tumor, as well as design chemoprevention strategies for those at high cancer risk.

Another often overlooked consideration when studying germline variation is the importance of phase information [112]. Humans have two copies of every chromosome, one inherited maternally and the other paternally. In typical HTS experiments germline variants are not phased, or assigned to a homologous chromosome of origin. It is impossible to fully interpret the pathogenicity of multiple heterozygous variants within a gene region without resolving what variants lie in the same gene copy (*cis*) vs. those in opposite copies (*trans*). For example, deleterious germline variation in a single copy of a MMR gene results in Lynch syndrome and adult onset cancer [113]; however, deleterious germline variation in both copies of a MMR gene is known as bi-allelic mismatch repair deficiency (bMMRD) and leads to pediatric cancer [114]. Phase information is undoubtedly important in situations where two highly damaging variants are present in the same gene region; however, it can also be important for variants that are not obviously pathogenic. For example, if an individual carries a single pathogenic variant and a *cis*-eQTL that affects the expression of the relevant gene in the same gene region, two scenarios could occur: the individual overexpresses the altered gene copy, or the individual overexpresses the WT gene copy. The difference could have important biological implications and would be overlooked without phase information. Further, in the situation where multiple missense variants exist in the same gene, which is common for many large DNA damage repair proteins such as *BRCA1/2*, having certain combinations of variants *in cis* may have a different functional effect than would be expected from individual variant scores. Currently few tools exist to computationally model these types of genetic interactions [115]. It is important to include

phase information when incorporating germline variants into cancer studies, particularly given the fact that many cancer susceptibility loci have been shown to modulate expression of local genes.

1.2 Overview and Organization of Dissertation

In this work we aim to further the understanding of how genetic germline variants shape the course of cancer development and the development of tumor phenotypes. While there have been numerous studies on the relationship between germline variants and somatic phenotypes in the context of cancer predisposition syndromes, mainly cancers associated with pathogenic *BRCA1/2* alleles, this relationship is less understood in sporadic cancers. We note that since the initiation of this project in 2014, interest in germline variation in cancer has grown rapidly, resulting in a number of publications that we described in the previous section.

To address this question, we utilized the Cancer Genome Atlas (TCGA). TCGA currently represents the largest dataset containing paired tumor:normal sequence data from cancer patients, with data from over 10,000 individuals representing 33 cancer types [116]. While raw germline sequence data is available from TCGA, germline variant calls are not, largely due to patient privacy concerns. Therefore, in order to study the germline we first had to call germline variants from the raw sequence data. We selected a cohort of 9,099 individuals with paired tumor:normal whole exome sequencing (WXS) data for our study. We describe our experience calling germline variants and our subsequent discovery of a technical artifact in chapter 2. At the time of publication, our data represented the largest set of germline variant calls in a cancer cohort [117].

After stringent quality control of the germline variants calls, we next tested for associations between rare, damaging germline variants and somatic phenotypes. We chose

to focus on microsatellite instability (MSI) and somatic mutational signatures [61] for a number of reasons: 1) these phenotypes are quantifiable and easily extracted from tumor sequencing data, 2) these phenotypes are highly variable within and between cancer types, 3) there is strong underlying biological mechanism between DNA damage repair pathway defects and manifestation of these phenotypes. In chapter 3 we explore the relationship between germline, somatic, and epigenetic alteration of DNA damage repair (DDR) genes and these somatic phenotypes. We surprisingly found evidence of heritable cancer predisposition syndromes in TCGA, a dataset widely thought to represent sporadic adult-onset cancer. We describe individuals in TCGA that exhibit characteristics of Lynch syndrome and identify novel potentially pathogenic Lynch syndrome variants.

In the course of identifying loss of heterozygosity (LOH) events, I made the observation that unique properties of paired tumor:normal sequence data could be exploited to phase germline variants. Briefly, changes in variant allele frequency (VAF) between the normal and tumor sample in regions of somatic copy number alteration can be used to assign germline variants to their homologous chromosome of origin. Chapter 4 describes this approach, which we call VAF phasing. We benchmarked VAF phasing against other phasing methods and performed a phase-informed analysis of germline variants in cancer predisposition genes.

We conclude in chapter 5 with a discussion of our results in the context of the current knowledge of the role of germline variants in determining somatic phenotypes. We identify limitations of the datasets currently used to investigate these questions and propose directions for future research in the field.

1.3 Acknowledgements

Chapter 1 is being prepared for publication under the title, "Germline Variation in Cancer, What is it Good For?". Authors included Alexandra R. Buckley and Nicholas J. Schork. AB and NJS wrote manuscript. The dissertation author was the primary researcher and author on this manuscript.

Chapter 2

Pan-Cancer Analysis Reveals Technical Artifacts in TCGA Germline Variant Calls

2.1 Abstract

Background: Cancer research to date has largely focused on somatically acquired genetic aberrations. In contrast, the degree to which germline, or inherited, variation contributes to tumorigenesis remains unclear, possibly due to a lack of accessible germline variant data. Here we called germline variants on 9,618 cases from The Cancer Genome Atlas (TCGA) database representing 31 cancer types.

Results: We identified batch effects affecting loss of function (LOF) variant calls that can be traced back to differences in the way the sequence data were generated both within and across cancer types. Overall, LOF indel calls were more sensitive to technical artifacts than LOF Single Nucleotide Variant (SNV) calls. In particular, whole genome amplification of DNA prior to sequencing led to an artificially increased burden of LOF

indel calls, which confounded association analyses relating germline variants to tumor type despite stringent indel filtering strategies. The samples affected by these technical artifacts include all acute myeloid leukemia and practically all ovarian cancer samples.

Conclusions: We demonstrate how technical artifacts induced by whole genome amplification of DNA can lead to false positive germline-tumor type associations and suggest TCGA whole genome amplified samples be used with caution. This study draws attention to the need to be sensitive to problems associated with a lack of uniformity in data generation in TCGA data.

2.2 Background

Cancer research to date has largely focused on genetic aberrations that occur specifically in tumor tissue. This is not without reason as tumor formation is driven to a great degree by somatically-acquired changes [2]. However, the degree to which germline, or inherited, DNA variants contribute to tumorigenesis is unknown. While it has been clearly demonstrated that germline variation increases cancer risk in overt and rare familial cancer predisposition syndromes, the contribution of germline to more common and sporadic cancer risk is unclear and highly debated [2, 10]. It is likely that inherited germline variation in fundamental molecular processes, such as DNA repair, can create a more permissive environment for tumorigenesis and shape tumor growth in some individuals [55, 50, 51]. It is also likely that variation in the host germline genome can act synergistically with acquired somatic mutations to shape the way in which tumors grow and ultimately manifest.

There is a growing interest in better understanding the contribution of germline variation to cancer risk and tumor phenotypes [65, 21]. The most extensive pan-cancer germline study to date identified associations between deleterious germline variation in known cancer predisposing genes and both age of onset and somatic mutation burden [65].

Lu et. al demonstrated that inherited variants can increase risk of developing cancer, as well as influence tumor growth and overall phenotypic features. Similar results were found in a study of biallelic mismatch repair deficiency (bMMRD). It is known that bMMRD predisposes to childhood cancer, but it was further demonstrated that acquisition of somatic mutations in polymerase genes (*POLE*, *POLD1*) led to a hypermutated phenotype in childhood brain tumors [118]. This demonstrates a synergistic interaction between germline variation and somatic mutation. A comprehensive study of breast cancer whole genomes identified a somatic copy number profile signature associated with *BRCA1* inactivation [66]. Interestingly, this profile was associated with either inactivation of *BRCA1* in the tumor via mutation or promoter hypermethylation, or via inherited germline variants. This shows that somatic mutation and germline variation can both influence tumor phenotype.

We chose to use the whole exome sequence (WXS) data from TCGA to investigate the role of germline variation in shaping tumor phenotypes. TCGA is an attractive dataset for this purpose as there are paired tumor normal data for many cancer types. We took a pan-cancer approach for two reasons: 1. increased sample size and therefore increased power to detect associations of small effect size; and 2. cancers of disparate origin may share common features which would be overlooked in a cancer type-specific analysis [119]. For example, germline mutations in *BRCA1/2* are most commonly studied in breast and ovarian cancer, but have also been shown to increase risk for stomach and prostate cancer [120]. Further, germline *BRCA2* mutations have been associated with a distinct somatic mutational phenotype and an overall increased somatic mutation burden in both prostate and breast cancer [121, 65, 66]. To our knowledge, a comprehensive germline analysis of all cancer types available in TCGA has not been performed. Thus other cross-cancer germline associations likely remain to be discovered.

In an ideal dataset, a single protocol should be used for processing all samples.

Unfortunately, this is unrealistic in large public datasets like TCGA in which samples are collected over time and across many data centers. Since its inception in 2005, TCGA has collected data on 11,000 patients from 20 collaborating institutions and generated sequence data from 3 sequencing centers [116]. Differences in sample collection and processing across centers could lead to batch effects, or variation in the data due to a technical factor that masks relevant biological variation [122]. Problems with batch effects can be amplified when analyzing samples across TCGA, since the number of methods used to collect samples increases with the number of cancer types. The Pan-Cancer Analysis Project has recognized this and aims to generate a high quality dataset of 12 TCGA cancer types, taking care to identify and minimize technical artifacts [119].

While extensive curated somatic data are available from TCGA, germline information is currently only available in raw form, under controlled access. Therefore, we first had to develop and execute a variant calling pipeline on the raw normal tissue sequence data. As a main goal of our variant calling analysis is to create a cohesive, pan-cancer dataset, we chose to use the Genome Analysis Toolkit (GATK) joint calling approach [123, 124]. Joint calling is a strategy for variant calling in which read data are shared across samples, in contrast to single sample calling where genotype decisions are made based on reads from a single sample only. There are three major advantages of this approach: the ability to distinguish sites that are homozygous reference vs. those that have insufficient data to make a call, increased sensitivity to detect variant sites that are poorly covered in any individual sample but well covered when the cohort is considered as a whole, and the ability to use GATK’s statistical modeling approach to variation filtration, known as ‘variant quality score recalibration’ (VQSR).

Here we describe our experience calling germline variants from a large cohort of TCGA normal tissue WXS samples spanning 31 cancer types. Specifically, we were interested in cataloguing sources of heterogeneity in sample preparation, identifying batch

effects in our variant calls, and determining methods to reduce or control for technical noise. Our finding reveals a critical artifact introduced by preparation of DNA samples through whole genome amplification, leading to false positive LOF indels. The study therefore highlights the importance of quality control at all stages of the variant calling process and suggest that pan-cancer analysis with TCGA data be approached with caution.

2.3 Methods

2.3.1 Cohort

Approval for access to TCGA case sequence and clinical data were obtained from the database of Genotypes and Phenotypes (dbGaP). We selected a total of 9,618 normal tissue DNA samples with whole exome sequence data (Supplementary Table A.1). We limited analysis to samples sequenced with Illumina technology and aligned to the GRCh37/hg19 reference genome.

2.3.2 Germline Variant Calling

Aligned sequence data for normal samples in BAM file format and the accompanying metadata was downloaded from CGHub [125]. Individual samples were matched with the target regions for the exome capture kit used to generate the sequence data, and variant calling was limited to these target regions +/- 100 bp. SNVs and small indels were identified using the GATK v.3.5/v.3.4 best practices pipeline and a joint calling approach [123, 124]. The GATK pipeline includes two preprocessing steps to improve the quality of the BAM file. Local realignment of reads is performed in regions containing indels, and base quality scores are recalibrated to minimize known sources of score bias. 'HaplotypeCaller' was run on individual samples in gVCF output mode, producing an intermediate single sample gVCF to be used for joint genotyping. Running this pipeline on

a single BAM from CGhub took approximately 15 compute hours and produced a 100MB gVCF. Individual gVCFs were combined in groups of 100 and the final joint genotyping step was performed by chromosome on all 9,618 samples as a single cohort. Following this joint genotyping step, all future analysis was limited to the intersection of all exome kit capture regions. The intersection of the kits covered 27 MB and 97.7% of Gencode v19 exons (Supplementary Table A.2) [126]. GATK VQSR was run separately for SNVs and indels. VQSR learns from variant quality annotations using variants overlapping with vetted resources such as dbSNP and 1000 genomes as a truth set. VQSR filters are defined by the percentage of truth variants that pass filter, termed truth sensitivity (TS). For the initial analysis, SNVs were filtered at VQSR TS 99.5% and indels at VQSR TS 99.0%, as suggested by GATK documentation.

2.3.3 PCA and Self-Report Ancestry Validation

PCA was performed jointly on the filtered pan-cancer VCF and HapMap genotype data from 1,184 individuals using PLINK v1.90b3.29 [127, 128]. Multiallelic sites, rare variants ($< 1\%$ AF), and sites with missing values were excluded from the pan-cancer VCF. A final variant set of 4,376 SNPs was obtained by taking the union of the pan-cancer and HapMap variant calls, requiring 100% genotyping rate across all samples. To assess accuracy of self-report ancestry from TCGA clinical data, principle component (PC) loadings of TCGA samples and HapMap samples were compared. HapMap samples were clustered on PC 1 and PC 2 using the R package 'flexclust' and K-means clustering with $k=4$ to roughly approximate the four major TCGA self-reported ancestry categories (White, Asian, Black, and Hispanic) (Supplementary Table A.3) [129]. TCGA samples were assigned to one of these four clusters using the predict function and PC 1 and PC 2 loadings (Supplementary Table A.4). Comparing self-reported ancestry to HapMap cluster membership showed 4% of TCGA samples had inaccurate self-reported ancestry.

2.3.4 Annotation and BAM metrics

Putative LOF variants, defined here as stop-gained, nonsense, frameshift, and splice site disrupting, were identified using the LOFTEE plugin for VEP and Ensembl release 85 [130]. LOFTEE assigns confidence to loss of function annotations based on position of variant in the transcript, proximity to canonical splice sites, and conservation of the putative LOF allele across primates. For our analysis we used default LOFTEE filter setting and only included high confidence predicted LOF variants. A variant was called LOF if it received a high confidence LOF prediction in any Ensembl transcript.

Predicted variant effects were obtained using Annovar v.2014Jul14 [131]. Annovar returns a single prediction for each variant position, collapsing across transcripts and reporting the most damaging variant prediction.

Allele frequencies were obtained from ExAC v0.3.1 and used for comparison to our cohort [76].

We quantified capture efficiency in this analysis as the percentage of capture target area covered by at least 20 X read depth (denoted C20X). Sequence depth information was obtained on BAMs downloaded from CGhub using GATK 'DepthOfCoverage' and the corresponding exon capture bed file to define coverage intervals. Gene level read depth information was obtained from a 5113 BAM files using GATK 'DepthOfCoverage' and a RefSeq exon coordinate file obtained from UCSC's table browser [132, 133]. For the gene level depth analysis, files were downloaded from GDC legacy archive to preserve the original sequence alignment [134].

2.3.5 Realignment Comparison

To assess the effect of heterogeneous alignment protocols on variant calls, we realigned the raw sequence data for a subset of our cohort. We chose 345 samples to represent

a large range of sample preparation variation present in the TCGA BAM files. Reads were stripped from the BAM to generate a FASTQ file using samtools v.0.1.18 bam2fq [135]. The FASTQ was realigned to GRCh37 using BWA MEM v.0.7.12 (with parameters -t 3 -p -M) and duplicates were marked using Picard v.1.131 [135, 136]. From this point the realigned BAM file was processed through the same GATK pipeline described above to produce individual gVCFs. To directly compare the effect of realignment, we generated a VCF for the 345 realigned samples (NewAlign) and for the same 345 samples processed without the realignment step (OldAlign). We were unable to run GATK indel VQSR on a cohort of this size, thus we filtered both VCFs with GATK SNV VQSR TS 99.5 and GATK indel hardfilters (settings QD > 2, FS < 200, ReadPosRankSum > -20). We calculated discordance between alignment pipelines as the percent discordant variant calls: $1 - (\text{intersection of variant calls} / \text{union of variant calls})$. Variant calls were matched by position and alternate base, disregarding zygosity.

2.3.6 WGA Enriched Indels

Indel allele counts were obtained for $n=614$ WGA and $n=9,004$ DNA samples separately. For each indel site, we obtained a contingency table of the number observed alternate allele counts vs number reference allele counts in DNA vs WGA samples. Reference allele counts were calculated as $(2 * \text{the number of samples}) - \text{alternate allele count}$. A one-way Fisher's exact test was used to define indels with allele counts enriched in WGA samples. A threshold of $p < 0.063$ was used to define WGA enrichment. This cutoff corresponds to the p value of a one-way Fisher's exact test for a singleton present only in WGA samples. Using this method we define $n=5,654$ WGA-enriched and $n=34,880$ non-enriched indels.

2.3.7 Homopolymer Indel Analyses

To determine if indels occurred within homopolymer sequences, we obtained the GRCh37 reference sequence ± 10 base pairs from each indel start position. The only indels considered for homopolymer analysis were those that were single base insertions or deletions or multi base insertions or deletions of the same base. All indels used for homopolymer analysis were < 15 bp in length. An indel was labeled as a homopolymer + indel if a sequential repeat of the inserted/deleted base/s occurred within ± 1 bp of the indel start position. Using this method we labeled every indel in the pan-cancer VCF as homopolymer \pm . The GC content of the region ± 10 bp of each indel was additionally determined as number G,C bases/ total number of bases.

Homopolymer content by gene was determined using RefSeq coding exon definitions and the GRCh37 reference sequence [133]. For this analysis a homopolymer region was defined as four or more sequential repeats of a single base pair. For each gene, the sequence of all coding exon regions was scanned for homopolymer sequences. Sum totals of number of homopolymers of each type (A,T,C,G) were obtained. G/C and A/T homopolymers were considered together by summing single base homopolymer counts. To compare homopolymer content across genes of different sizes, these counts were divided by the total number of base pairs in the gene's coding region to obtain the homopolymer count per exonic basepair.

2.3.8 Chimera Read Analysis

We define large indels as those with an inserted or deleted sequence ≥ 15 base pairs in length. We identify $n=1,418$ WGA-enriched and $n=2,301$ non-enriched large indels. The inserted or deleted sequence for each indel was aligned to the GRCh37 reference genome using ncbi-blast-2.6.0+ (with parameters -reward 1 -outfmt 6 -num_alignments

1 -max_hsps 3) [137]. For insertions, the match with the highest predicted similarity was retained. For deletions, the best match excluding the actual deleted reference sequence was retained. For all indels with a BLAST hit, the distance between the start position BLAST hit and the indel start position was determined. Indels with BLAST hits > 10 kB away from the indel start position were excluded from this analysis, as MDA chimera artifacts act predominantly within a 10kB proximal region [138].

2.3.9 Repeated Samples

A subset of individuals in our cohort have multiple germline DNA WXS samples. This cohort of 9,618 samples represents 9,099 unique individuals; 1,012 of the normal WXS samples were obtained from 492 individuals (2-5 samples per individual). The repeated samples all represent germline DNA from the individual, but differ in terms of sample preparation, sequencing, and processing. Percent discordance between repeated samples was calculated as described above. One sample (TCGA-BH-A0BQ) was removed from future analysis due to a high discordance between two high coverage DNA samples. We suspect a sample label mismatch. For association testing, we selected one the sample with the highest coverage that was not whole genome amplified, leaving 9,098 samples.

2.3.10 Indel Filter Methods

To assess different indel filtering methods, indels were extracted from the raw pan-cancer VCF using GATK 'SelectVariants'. Multiallelic sites containing both SNPs and indels were included in the indel VCF. Four filter methods were tested on the pan-cancer indel VCF: GATK VQSR TS 90.0, TS 95.0, TS 99.0, and GATK Hardfilter. GATK VQSR and Hardfilter filters were applied using the modules 'ApplyRecalibration' and 'VariantFiltration' respectively (Hardfilter settings $QD > 2$, $FS < 200$, $ReadPosRankSum > -20$). Indels were additionally identified using Varscan v.2.3.9 (with parameters $-p$ -value

0.1 `-strand-filter 1`) on BAMs downloaded directly from CGHub with no preprocessing [139]. Single sample indel VCFs were generated using Varscan for all 9,618 samples in our cohort.

2.3.11 Statistical Methods

To detect contribution of technical factors to LOF variant burden Type II ANOVA was performed using the R package 'car' [140]. To determine the percent variance explained by technical factors the sum of squared error for each factor was divided by the total sum of squared error. To create 95% confidence intervals for non-normally distributed data, we used the R package 'boot' [141]. The mean for each of 1,000 bootstrap samples was calculated and a confidence interval was constructed using the `boot.ci` function with type set to 'basic'.

To detect association between germline gene LOF status and cancer type, we used an 'one vs. rest' approach. For each cancer type, a binary ('dummy') vector was created indicating whether each individual had the given cancer type (1) or another cancer type (0). For sex specific cancers, only individuals of the same gender were compared. LOF variants with $AF < 0.05$ were binned by individual by gene to generate an individual LOF variant count for each gene. Genes were only included in our analysis if at least two individuals in the cohort had germline LOF variants in the gene. For each cancer type and each gene we used a logistic regression to test association between germline LOF variant burden and cancer type. Our regression model took the form: `glm(cancer type indicator ~ variant burden + race + age)`. To discover significant gene-cancer type associations we obtained the p value of the β coefficient for the variant burden term and used a Bonferroni cutoff of 1.61×10^{-7} to account for multiple testing (31 cancer types x $\sim 10,000$ genes).

2.4 Results

2.4.1 Technical Heterogeneity in TCGA WXS Data Generation

We obtained TCGA WXS data from CGHub in the form of reads aligned to the human reference genome (BAM files) [125]. From the BAM files and available metadata we identified seven technical sources of variation in the way the sequence data were generated: tissue source of normal DNA, exome capture kit, whole genome amplification of DNA prior to sequencing (WGA), sequencing center, sequencing technology, BWA version, and capture efficiency (C20X) (Supplementary Figure A.1). We found substantial variation existed within and between cancer types with respect to these technical factors (Figure 2.1). Some of these technical factors were found to be highly associated with cancer type, such as use of Illumina Genome Analyzer II and ovarian cancer (OV), while others exhibited no clear relationship with cancer type, such as use of solid normal tissue as opposed to blood as a source of normal DNA. Relationships existed between pairs of technical factors as well, such as the Broad Institute's exclusive use of a custom Agilent exome capture kit. All possible combinations of the first six technical factors produce 1,152 unique workflows, of which only 44 were used to generate the TCGA data. This further demonstrates that relationships exist between technical factors. Of the 31 cancer types examined, only uveal melanoma (UVM) and testicular germ cell tumors (TCGT) had a uniform workflow for all samples (Supplementary Figure A.2). These observations highlight the substantial heterogeneity in data generation across TCGA and importantly even within cancer types.

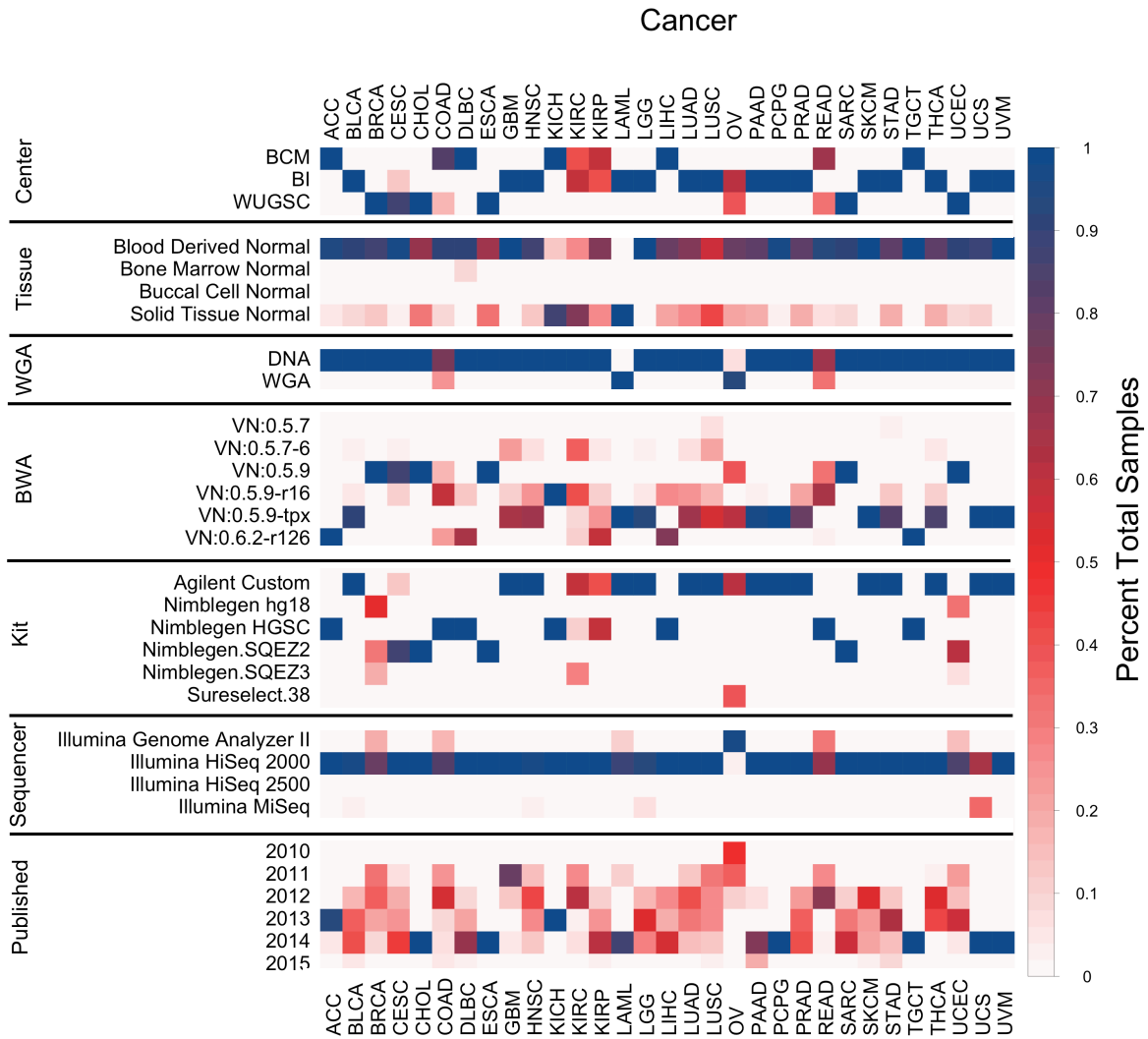


Figure 2.1: Technical Covariates in TCGA WXS Samples. For each covariate and cancer type, color represents the fraction of total samples. Fraction of total samples sums to 1 for each covariate and cancer type. Red indicates higher heterogeneity. Year first published included for context.

The technical factors can ultimately be divided into two categories: those that can be modified during processing of the sequence data (BWA version, target regions of a capture kit), and those that cannot be modified computationally (source of normal DNA, WGA, center, technology, capture efficiency). Six exome capture kits ranging in size from 33-64 MB were used to capture normal DNA for sequencing (Supplementary Table A.2). As the goal of our variant calling pipeline was obtain a uniform set of variants across samples, we chose to restrict analysis to the intersection of the capture regions. The area hereby excluded consists largely of exon flanking regions. The intersection covers 97.7% of Gencode exons, thus for the purposes of studying protein-coding variation using the intersection of the kits leads to minimal loss of data (Supplementary Table A.2) [126]. It has been shown that differences in capture efficiency and sample preparation protocols between exome kits can affect variant calls, even in regions common between kits [142]. Therefore, despite using the common capture region, the use of multiple capture kits may still introduce artifacts.

To assess the effect of heterogeneous BWA alignments on variant calls, we called variants on 345 of the TCGA normal samples either using the provided BAM (OldAlign) or stripping and realigning reads to GRCh37 using BWA MEM v.0.7.12 (NewAlign). The overall raw discordance rates between the two sets of variants was 5%, which is in the expected range for different alignment protocols (Supplementary Figure A.3) [143]. Indel calls were noticeably more discordant, consistent with the specific challenges and notorious variability of indel calling [144]. Interestingly, the discordance rate was correlated with BWA version used to generate the BAM file in CGHub, with older versions displaying more discordance. This effect can largely be reduced by applying VQSR filters, which decreases overall discordance from 5% to 3% (Supplementary Figure A.4). Greater discordance between variant calling pipelines has been observed in repetitive regions of the genome, and in accordance with this we reduce overall discordance to 1.7% with the removal of

repetitive regions from analysis (Supplementary Figure A.3) [145]. As no set of true positive variants exists for TCGA samples, we cannot determine whether realigning BAM files produces more accurate calls. Given the computational cost of realignment, and that discordance can be mitigated by filtering variants and masking repetitive regions of the genome, we proceeded with variant calling using the provided BAM files.

Functional annotation of the 1,093,501 variants in the final VCF predicted 625,365 missense; 371,754 silent; 24,455 nonsense; 2,968 splice site; 553 stoploss; 46,280 frameshift indels and 22,126 in-frame indels in 9,618 samples. For initial quality control we performed principal component analysis (PCA) to identify the most significant sources of variation in the variant calls. PCA on common variants showed that the first two principal components stratified samples by self-reported race and ethnicity, indicating that the largest source of variation is ethnic background and not technical factors (Supplementary Figure A.5). To assess the quality of the calls, we measured the fraction of variants also present in the ExAC database [76]. We expect a high degree of overlap between our calls and ExAC, as the ExAC v0.3.1 dataset includes germline variants from 7,601 TCGA individuals. Overall 88.56% of the variant calls were present in ExAC, with SNVs showing higher overlap than indels (89.91% vs. 53.94%). Based on these results, we concluded the variant calls were free of overt technical artifacts and proceeded to the next stage of analysis.

2.4.2 Impact of Technical Heterogeneity on Loss of Function Variants

There is great interest in understanding how inherited impaired functionality of cancer-relevant pathways shapes tumor phenotypes, as has been previously demonstrated for bMMRD and *BRCA1* germline mutations [65, 66, 118]. To identify germline variation likely to disrupt function of genes, we used VEP and LOFTEE to predict LOF variants in this cohort [130]. We observed a median 150 LOF per sample across our entire cohort,

consistent with the ExAC findings (Figure 2.2A). [76]. However, two cancer types, acute myeloid leukemia (LAML) and OV deviate significantly from this expected value, with individuals with these cancers having up to 500 LOF germline variants. This suggests an artifact was manifesting in rare LOF variants that was not identified by PCA on common variants. Notably this effect is specific to LOF indels, in contrast to LOF SNVs that are distributed more uniformly across cancer types (Supplementary Figure A.6).

We used Analysis of Variance (ANOVA) to assess the contribution of each technical factor to individual LOF variant burden. Initial analysis showed that source of normal control DNA and sequencing technology were not significantly associated with LOF variant burden, and that capture kit was highly collinear with sequencing center. Therefore, we limited subsequent analysis to sequencing center, BWA version, WGA, and C20X. It is known that LOF variant burden varies between ethnic groups, thus we include self-reported race as a covariate in this analysis as a reference point for expected variation [76]. All technical factors combined explain less than 1% of the variance in LOF SNV burden, indicating SNVs are largely unaffected by technical variation. In contrast, 59% of variation in LOF indel burden was explained by technical factors, with WGA alone explaining over 50% (Table 2.1).

WGA samples have a higher LOF variant burden with a median 201 LOF variants per WGA sample. Four cancer types contain samples that underwent WGA: colon adenocarcinoma (COAD) (26% WGA), rectum adenocarcinoma (READ) (33% WGA), OV, (92% WGA) and LAML (100% WGA) (Figure 2.1). Analyzing cancer types containing both amplified and non-amplified DNA samples, we observed that WGA samples had a significantly higher LOF variant burden (Figure 2.2B), further suggesting that WGA rather than cancer type is the main source of bias. The cohort contains 13 individuals with both amplified and non-amplified DNA samples. We observed a 1.5 fold increase in LOF variant burden in amplified samples relative to non-amplified samples from the same

individuals ($p = 0.0002$ by paired Wilcoxon Signed Rank test) (Figure 2.2C), suggesting that WGA prior to sequencing leads to an artificially inflated number of predicted LOF variants.

Table 2.1: ANOVA of LOF Variant Burden. Variance in LOF SNV and indel burden explained by technical covariates. Sum. Sq., Sum of Squares; Df, Degrees of Freedom; % Var. Exp., Percent variance explained by each factor (factor Sum. Sq./total Sum. Sq.)

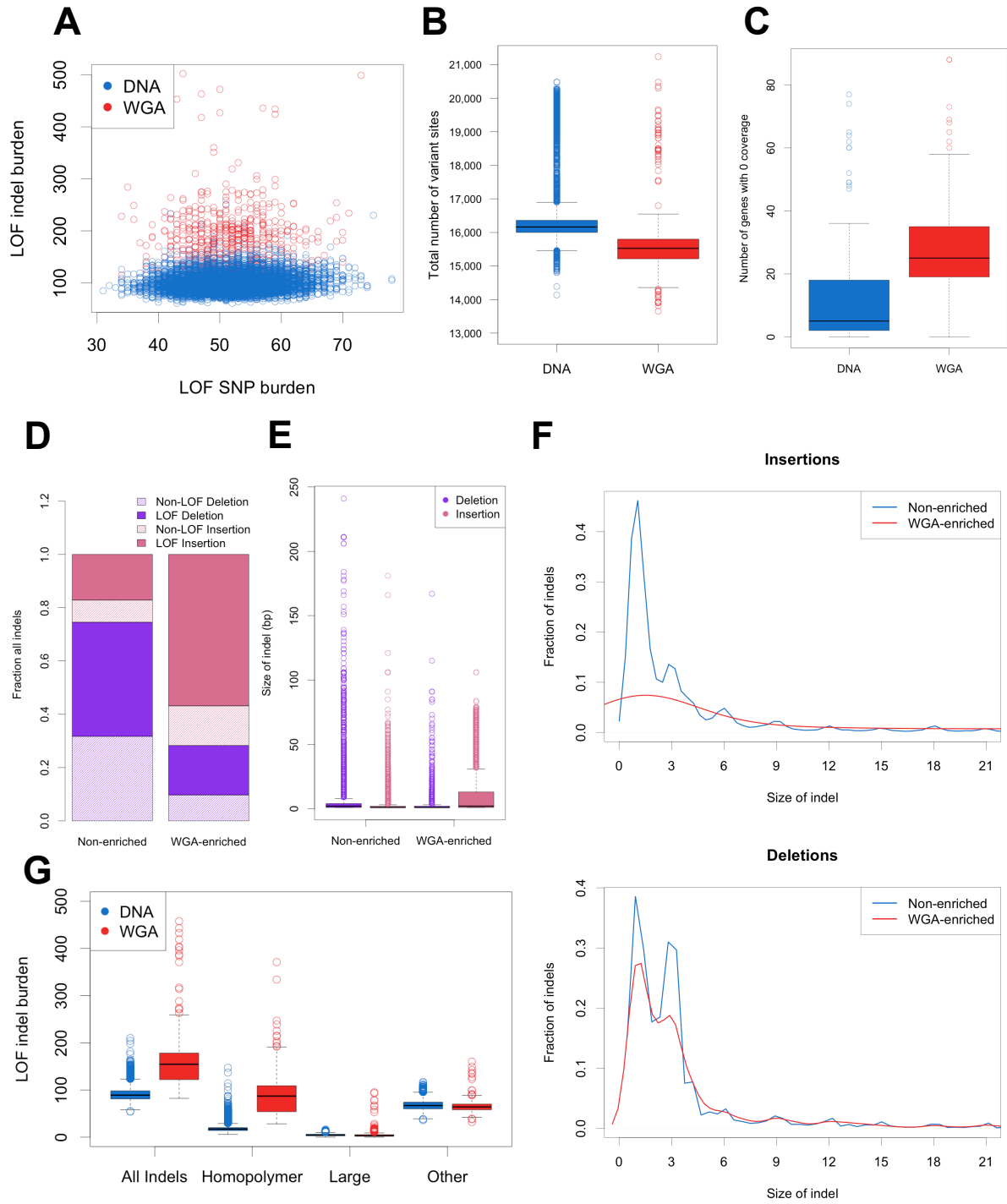
LOF SNV					
	Sum. Sq.	Df	F value	P value	% Var. Exp.
C20X	1785.49	1	52.95	$3.72e^{-13}$	0.0056
WGA	156.89	1	4.65	$3.10e^{-02}$	0.0005
Center	716.08	2	10.61	$2.48e^{-05}$	0.0023
BWA	79.59	5	0.47	$7.97e^{-01}$	0.0003
RACE	30698.90	5	182.10	$1.33e^{-184}$	0.0973
Residuals	281966.85	8363			0.8940
LOF Indel					
	Sum. Sq.	Df	F value	P value	% Var. Exp.
C20X	52930.43	1	153.90	$4.95e^{-35}$	0.0072
WGA	3744887.28	1	10888.62	0.0000	0.5080
Center	383585.43	2	557.65	$4.53e^{-228}$	0.0520
BWA	169507.9	5	98.57	$2.76e^{-101}$	0.0229
RACE	146904.86	5	85.42	$7.59e^{-88}$	0.0199
Residuals	2876257.21	8363			0.3900

To determine whether our choice not to realign BAM files contributed to the observed WGA effect, we calculated LOF variant burden in our NewAlign and OldAlign cohort using the same protocol. Realignment of the sequence data with BWA MEM increased the number of LOF calls per individual but overall LOF burden was highly correlated (Pearson $R^2 = 0.95$) (Supplemental Figure A.7). WGA explained a significant amount of variance in LOF variant burden in both NewAlign and OldAlign samples (Supplemental Figure A.7). Thus we can conclude that realignment does not remove WGA artifacts observed in our variant calling pipeline.

2.4.3 Characterizing WGA Artifacts

Having demonstrated that WGA is associated with increased LOF variant burden, we sought to characterize WGA samples more deeply. We observe that WGA samples have an excess of LOF indels while LOF SNV burden appears unaffected, as expected from the ANOVA results (Figure 2.3A). Interestingly, WGA samples had fewer variants overall, due more variable coverage depth over the capture regions (Figure 2.3B, Supplementary Figure A.8). Read depth was highly variable across genes in WGA samples with an average depth of 165 X and standard deviation of 140 X (Supplementary Figure A.8). As a consequence of this variable coverage, an average of 27 genes per sample had 0 coverage in WGA samples (Figure 2.3C).

Figure 2.3: Characteristics of Variants in WGA Samples. (A) Individual LOF indel burden vs. individual LOF SNV burden. Color indicates WGA status. (B) Total number of variant calls plotted by WGA status. (C) Number of genes with 0 read depth across 16,824 genes. (D) Fraction of insertions and deletions in $n=5,654$ WGA-enriched and $n=34,880$ non-enriched indels. Shading indicates LOF status. (E) Size in base pairs of WGA-enriched and non-enriched indels. (F) Density plot showing distribution of insertion and deletion size for WGA-enriched and non-enriched indels. (G) Individual burden of LOF indels for all indels, homopolymer + indels, indels 15 base pairs or longer, and other indels. Color indicates WGA status. Indel burden calculated using GATK VQSR TS99 filter.



As indel variant calls are the source of inflated LOF variant burden in WGA samples, we next determined which indels are enriched in WGA samples using a one-way Fisher’s exact test. While it is impossible to distinguish errors from true indels definitively at this scale, indels that are found at a significantly higher frequency in WGA samples relative to DNA samples are good candidates to be errors. The majority of WGA-enriched indels are insertions, and the ratio of insertions to deletions is skewed relative to non-enriched indel sites (Figure 2.3D). Further, 75% of WGA-enriched indels are LOF relative to 60% of non-enriched indels (Figure 2.3D). Upon examining the size of the indels in base pairs, we noticed that WGA-enriched insertions were larger than non-enriched insertions and their size distribution deviated from what is expected for coding indels (Figure 2.3E,F). The length of indels in coding regions is frequently a multiple of three base pairs, due to natural selection acting to maintain the reading frame [146]. WGA-enriched insertions did not show this expected distribution, and thus are more likely to be LOF frameshift indels. As previously reported, LOF variants are enriched for sequencing errors, supporting our hypothesis that the excess LOF indels in WGA samples are technical artifacts [147].

We observe that the local sequence context surrounding WGA-enriched insertions has a higher GC content, and that G and C insertions are twice as frequent in WGA-enriched insertions than non-enriched insertions (Supplementary Figure A.9 and Table A.5). This observation prompted us to look for homopolymer repeats in the sequence surrounding WGA-enriched indels. WGA-enriched indels occur in homopolymer repeats more frequently than non-enriched indels (Table 2.2). Further, indels that occur in homopolymer regions had an increased allele frequency in WGA samples relative to indels not in homopolymer regions, indicating that homopolymer indels are also more recurrent in WGA samples (Supplementary Table A.6). We observe that WGA-enriched indels are larger on average and are frequently in homopolymer regions, but that these two charac-

teristics are mutually exclusive. To better resolve the contribution of each of these indel types to WGA technical artifacts, we define three distinct categories of indels: homopolymer +, large, and all other indels (Table 2.2). Calculating individual LOF indel burden for each of these categories shows that the increased LOF indel burden observed in WGA samples is due to an excess of LOF homopolymer + indels (Figure 2.3G).

The pan-cancer cohort contains 492 individuals with multiple germline WXS samples. Presumably, variants that are not concordant between repeated samples on the same individual are errors, and thus we used genotype discordance as a surrogate measure for variant calling error. In addition to the 13 individuals with paired normal WXS samples with and without amplification (denoted WGA:DNA), 44 individuals have paired normal WXS samples where both samples have been amplified (denoted as WGA:WGA) and 435 are paired samples without amplification (denoted DNA:DNA). We calculated genotype discordance between all repeated samples for SNVs and indels separately and observed a stepwise increase in discordance with amplification of one or both samples. This effect was most apparent in indels, with a median 59.9% indel discordance between repeated WGA:WGA samples (Supplementary Figure A.10). Calculating indel discordance using the indel categories previously defined reveals that discordance between WGA samples is highest for homopolymer + indels, lower for large indels, and similar to DNA samples for other indels (Supplementary Figure A.10). This demonstrates that WGA errors manifest as small indels in homopolymer regions and large indels with no clear sequence context bias.

Table 2.2: Characteristics of WGA Indels. Fraction of WGA-enriched and non-enriched indels in three indel categories. Homopolymer indels: indels with a 4 or more single base repeat directly proximal to the indel; Large indels: indels with 15 or more inserted or deleted bases. Other indels: all indels that don't fit one of the previous criteria.

	% Other Indels	% Homopolymer Indels	% Large Indels
WGA-enriched	47.78	27.13	25.07
Non-enriched	83.52	9.63	6.83

WGA by multiple displacement amplification (MDA) is known to create chimeric DNA rearrangements, which manifest in the sequence data as reads with sequence from noncontiguous portions of the reference genome (Supplementary Figure A.11).[138]. To determine if chimeric reads were responsible for the large indels in WGA samples, we used BLAST to align the inserted and deleted sequences of large indels to the reference genome [137]. We observe that 86% of WGA-enriched large insertion sequences have a BLAST match, whereas only 10% WGA-enriched large deletions and non-enriched large indels have a BLAST match (Supplementary Table A.7). Further, the BLAST matches for WGA-enriched insertions were predominantly within 2 kb of the indel start position which is in accordance with the mechanism of MDA chimeric rearrangements (Supplementary Figure A.12). Thus, the large indels we observe in WGA samples can be explained by known MDA artifacts (Supplementary Figure A.11). Small indels in homopolymer regions may occur by the same mechanism, as it has been shown that the majority of MDA chimeric junctions occur in regions of short complimentary sequence [138]. The small homopolymer indel errors may also be due to known difficulties of calling indels in homopolymer regions,

which is exacerbated with amplification [148].

2.4.4 Filtering Artifactual LOF Variant Calls

We next sought an appropriate filter to remove artifactual LOF variant calls in WGA samples. As SNV calls were largely robust to technical artifacts, we focused on filtering indels specifically (Supplementary Figure A.6). We used two strategies available from GATK: 1) Statistical model filtering using VQSR with increasing stringency cutoffs (99%, 95%, 90%), and 2) Heuristic filtering (Hardfilter) based on fixed thresholds ($QD > 2$, $FS < 200$, $ReadPosRankSum > -20$), for a total of four filtering approaches [124]. The four filters varied in stringency, resulting in a median individual LOF indel burden ranging from 53-98 across methods (Figure 2.4A and Supplementary Figure A.13). To assess the efficiency of each filter to remove technical artifacts, we performed an ANOVA analysis as described in Figure 2.2 for each filtering approach, including the initial filter (GATK VQSR 99) as a reference (Figure 2.4B). VQSR 90 and VQSR 95 reduced technical artifacts to a similar degree, whereas VQSR 99 and Hardfilters performed poorly (Supplementary Figure A.14A and Table A.8).

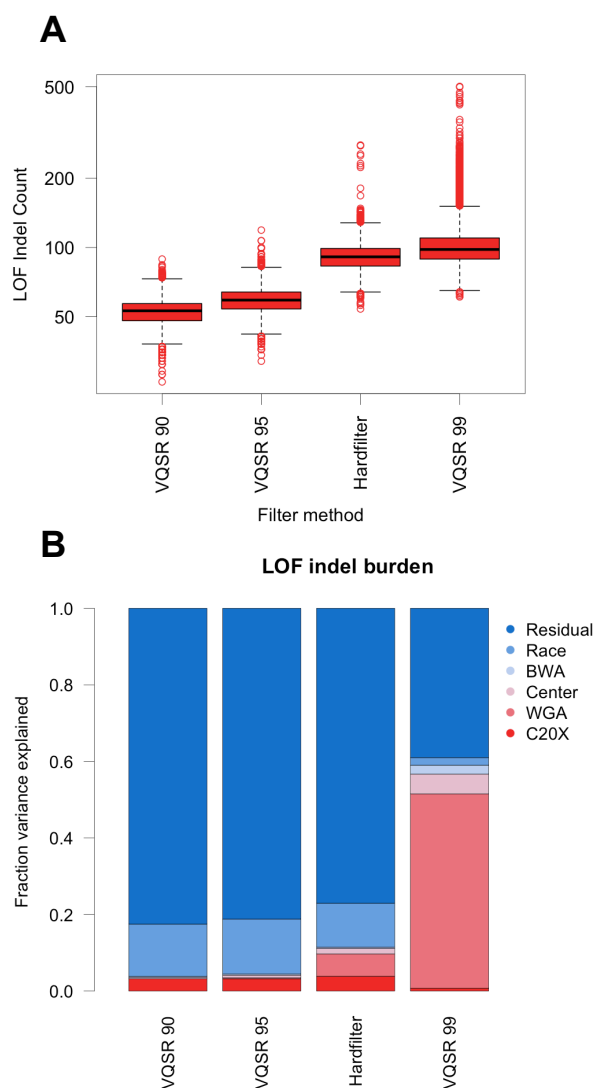


Figure 2.4: Comparison of Indel Filters. (A) Individual LOF indel burden for all indel filter methods in order of decreasing stringency. (B) Percent of variation in individual LOF indel burden explained by technical covariates for each filter method.

Variant filtering is a balance between removing likely false positive signal while retaining true positive signal. Using VQSR 99 we observe an individual LOF variant burden similar to that reported in the ExAC database, while all other methods produce lower LOF burden than expected (Supplementary Figure A.14B) [76]. Therefore, while more stringent filtering approaches can reduce technical artifacts, they do so at the cost

of losing likely true positive indels. Without a way to manually validate a large number of rare indel variant calls, it is impossible to exactly measure false positives rates for our filter approaches. Instead, we once again used the repeated samples in our cohort to identify likely true positives (indels concordant between repeated samples) and likely false positives (indels discordant between repeated samples). We assessed filter quality using three measures: the fraction of discordant indels removed by the filter, the fraction of concordant indels removed by the filter, and the fraction of indels overlapping the ExAC database. The stringency of each filter was measured as the total number of LOF indel sites and the median individual indel LOF burden when each filter was applied (Table 2.3).

Table 2.3: Variant Filter Metrics.

Filter	LOF indel sites	Median LOF indel burden	Fraction discordant indels removed	Fraction concordant indels removed	Indel overlap with ExAC
VQSR 90	6212	53	0.8667	0.4514	0.7079
VQSR 95	9177	59	0.8064	0.3760	0.6776
Hardfilter	24212	91	0.3600	0.0210	0.3527
VQSR 99	26134	98	0.2763	0.1100	0.5394

2.4.5 Consequences of Technical Artifacts on Genetic Associations

To determine how sensitive association results are to filtering method, we tested for association between germline LOF variant burden and cancer type using different filtering approaches. We took an 'one vs. rest' approach with our samples using all cancers except

the cancer of interest as a control. Thus, we tested for enrichment of LOF germline variants in one cancer type as compared to other cancers, which is different than other studies that have used control cohorts [65]. Our rationale for using this approach was to minimize heterogeneity that would be introduced by including control samples collected in different studies. We chose to highlight the results only from OV for two reasons. First, it is established that *BRCA1/2* germline variants are enriched in OV so the OV - *BRCA1/2* association can be used as a positive control, and second virtually all OV samples have been amplified and are confounded with WGA artifacts [149, 65, 150].

Quantile-quantile plots from logistic association tests for three indel filter methods are shown in Figure 2.5A. It was immediately apparent that our initial filtering approach (VQSR 99) produced an excess of significant associations even above a strict Bonferroni multiple hypothesis correction (Figure 2.5B). True associations are mixed with false associations due to WGA artifacts in LOF indel calls. Increasing the stringency of indel filtering reduced noise due to technical artifacts while retaining a putative true positive *BRCA1/2* association signal. Stringent filtering removes noise at the cost of reducing potential signal, as evidenced by the decreased number of genes that can be tested for association. This inflation in significant associations was only observed in cancers containing WGA samples, and persisted, albeit to a far lesser extent, even with the most stringent filter (Figure 2.5B). Supporting the idea that some of the associations in WGA cancer types are false, only two of the significant genes (*BRCA1/2*) in OV and none in LAML are genes where germline variation is known to be associated with cancer risk [17].

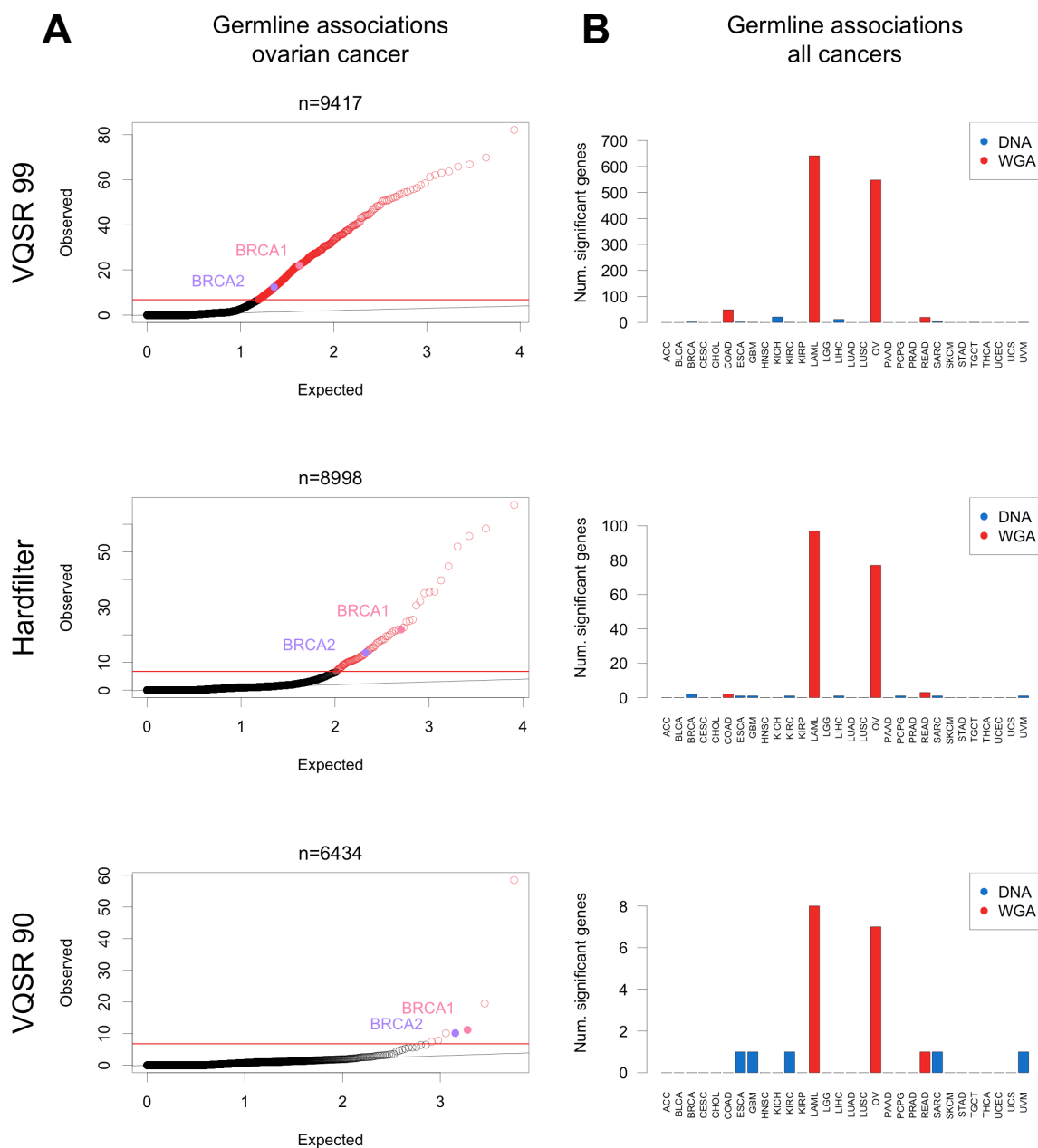


Figure 2.5: Association Between LOF Burden and Cancer Type.

(A) Quantile-quantile plots from logistic regression association testing between germline LOF burden and ovarian cancer for three indel filter methods. n=number of genes tested. Red line indicates significant cutoff and red points indicate associations significant $p < 1.61 \times 10^{-7}$. *BRCA1/2* associations highlighted.

(B) Number of significant cancer type - gene associations in each cancer type for three indel filter methods. Color indicates cancer types with WGA samples.

We observe that an unusually high fraction of significantly associated genes were shared between LAML and OV, with 69%, 55%, and 25% of significant genes shared for VQSR filters TS99, TS95, and TS90, respectively. Having demonstrated that LOF indels occur at a high allele frequency in homopolymer regions in WGA samples, we calculated the number of homopolymer regions in these shared genes. We observe that shared genes have a higher G/C homopolymer content compared to all genes tested (Supplementary Figure A.15). Further we see a stronger correlation between LOF indel burden and homopolymer content in WGA samples than in DNA samples (Supplementary Table A.9). Taken together, we can conclude that the high fraction of shared genes between LAML and OV is driven by high allele frequency LOF indels in homopolymer regions. LOF indel calls are more prone to batch effects than LOF SNVs, therefore we repeated the association test limiting to LOF SNVs only. While this reduces the excess number of significant associations, the analysis was underpowered to detect the true positive *BRCA1/2* - OV association (Supplementary Figure A.16). These results demonstrate that technical artifacts can lead to spurious associations and highlight the difficulty of correcting for artifacts in a pan-cancer analysis when technical factors are highly correlated with the phenotype being tested (Figure 2.1).

2.5 Discussion

We identified sources of technical variation in LOF variant calls from TCGA germline WXS data. Overall SNV calls were more robust to technical factors than indel calls. We found the strongest association between amplification of DNA prior to sequencing and an excess of LOF indel calls. Other factors tested were found to be significantly associated with both LOF SNV and LOF Indel burden, but explain little of the total variance in LOF variant burden when appropriate filters are applied (Table 2.1 and Fig 2.4B). The

factor explaining the most technical variation in total LOF variant calls after filtering is capture efficiency (C20X). It is likely that poor coverage over common capture regions, perhaps due to the different capture technologies used, decreased the ability to assign genotypes in some samples. Joint calling distinguishes sites with insufficient coverage to make a genotype call from those with adequate coverage for calling a homozygous reference genotype. Therefore, while C20X is a significant factor in the simple burden analyses performed here, a more sophisticated burden testing approach that can accommodate missing genotype values should mitigate this technical artifact.

Difficulty producing reliable variant calls in WGA exome samples has been previously reported [142, 151]. Inaccurate read alignment has been identified as a main contributor to spurious calls in WGA samples. However, even with an alignment protocol optimized for WGA samples it is still estimated that 7% of variant calls in WGA samples are artifactual [142]. Previous work comparing amplified and non-amplified DNA obtained from the same biological sample report higher variant call discordance in indels compared to SNVs, similar to what we observe [151]. These studies conclude that overall concordance between amplified and non-amplified samples is satisfactory; however, neither examined the impact of WGA on deleterious variants. Here we have demonstrated that errors introduced by WGA manifest as rare frameshift indels that are difficult to distinguish from true rare deleterious variation. We further demonstrated that the WGA indel errors we observe are in accordance with known errors and biases that occur due to MDA, and provide a mechanism by which MDA chimeric reads lead to erroneous indel calls (Supplemental Figure A.11). In addition to drawing attention to batch effects in TCGA sequence data, our study also provides valuable insight into potential pitfalls of calling indels in sequence data generated from MDA.

Simultaneous to our investigation, the genomic data commons (GDC) has called somatic mutations on TCGA tumor sequence data using four different pipelines and dis-

covered an excess of insertion mutations in tumor samples with amplified DNA [152]. This validates our findings in the orthogonal process of somatic mutation calling. Further, GDC only reports this observation for the MuTect2 pipeline, which combines aspects of the original MuTect algorithm and GATK's 'HaplotypeCaller' [153]. As WGA artifacts have thus far only been observed in GATK-derived variant callers, it is possible that these artifacts are specific to the GATK pipeline. An alternate method of variant calling could reduce or eliminate WGA errors, but this issue is still problematic as GATK is one of the most commonly used variant callers for large datasets such as ExAC and gnomAD [76].

While joint calling is the approach recommended by GATK, with the exception of one paper from our lab exploring the impact of genetic background on joint calling, to our knowledge there has not been a published systematic comparison of joint calling vs. single sample calling with GATK on a gold standard dataset to quantify the advantages of joint calling [154]. GATK's joint calling approach is not without problems. Greater accuracy for the group as a whole comes at the cost of loss of singleton variants from any given sample. Another complicating factor unique to joint called samples are multi-allelic sites, or sites where multiple alternate alleles are found in the population genotyped. Relatively few sites in our VCF were multi-allelic (3%, or 30,620 sites), but these sites contain 4,947 high-confidence LOF variants (11% of all LOF variants), indicating the importance of correct multi-allelic site parsing. Multi-allelic sites additionally pose a problem when filtering reliable from unreliable variants. With current tools for filtering VCFs, it is only possible to filter at the site level, meaning at multi-allelic sites all alleles will either be included or excluded by the filter. Further, in the version of GATK used for this analysis (v3.5), quality annotations for a site are calculated using all alternate reads without distinguishing between alleles. Therefore it is possible for low quality alternate alleles to pass filter at multi-allelic sites if high quality alternate alleles are present at the same site.

Our work shows that amplification of DNA prior to sequencing resulted in an excess

of predicted damaging indel variants. In our dataset, we find that using VQSR TS90 can eliminate the significant association between WGA and LOF indel burden, but it appears false associations persist in our association analyses (Figure 2.5B, Supplementary Table A.8). Thus, we find removal of WGA samples to be the only option to fully eliminate batch effects in our dataset. It is possible WGA indel artifacts could be eliminated in WGA samples using a different variant calling approach perhaps sensitive to MDA induced errors. The GDC has worked to optimize MuTect2 parameters for WGA samples, and their methods could potentially be applied to germline variant calling [152]. We suggest that variant calling in these samples should be handled with extra care.

TCGA is often thought of as a single dataset, but due to differences in sample collection and processing across the participating sites, should be thought of as a collection of studies. While we focused on the germline WXS sequence data, it is likely that batch effects are present in other data types. This has been recognized by the Pan-Cancer TCGA effort, although it is less often acknowledged in papers published on one or few cancer types [119]. There is heterogeneity even within cancer types in terms of sample preparation, such as in COAD and READ where roughly a third of germline WXS samples were prepared using WGA. Batch effects present in TCGA data can potentially confound even single cancer type analyses if not properly addressed. In terms of pan-cancer analysis, the correlation between certain technical factors and cancer types confounds analyses that use cancer type as the phenotype of interest, as we demonstrated in Figure 2.5. We note that since the initiation of our analysis, the raw TCGA sequence data have moved to the GDC [134]. The GDC has realigned the sequence to the current reference genome (GRCh38 .d1.vd1) using a standardized pipeline to harmonize the BAM file. Although this will eliminate one source of variation (BWA version), it only serves to remind researchers how sensitive data analyses might be to non-standardized data collection protocols, especially in the context of the TCGA data, as our study makes clear. Analyses of large, extant data

sets will continue to grow and impact biomedical research, with many in the community committed to pointing out the need for care in interpreting the results and impact of those analyses [155, 156, 122].

2.6 Acknowledgements

Chapter 2 was previously published in *BMC Genomics* in June 2017 under the title, "Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls". Authors included Alexandra R. Buckley, Kristopher A. Standish, Kunal Bhutani, Trey Ideker, Roger S. Lasken, Hannah Carter, Olivier Harismendy, and Nicholas J. Schork. NJS designed and supervised the research. ARB executed pipelines, analyzed data, and drafted the manuscript. KAS helped assemble pipelines and assisted with statistical analysis. ARB, KAS, and RSL designed figures. ARB, KAS, KB, RSL, OH, and HC designed experiments. NJS, OH, and HC helped write the manuscript. TI set up high performance computing infrastructure. All authors read and approved the final manuscript. The dissertation author was the primary researcher and author on this manuscript. ARB is supported in part by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number T32GM008666, the Translational Genomics Research Institute (TGen), and a gift from the San Diego Cancer Research Institute. NJS and his lab are also supported in part by National Institutes of Health Grants UL1TR001442 (CTSA), U24AG051129, U19G023122, as well as a contract from the Allen Institute for Brain Science. All computing was done using the National Resource for Network Biology (NRNB) P41 GM103504. All primary data were accessed from The Cancer Genome Atlas Research Network (cancergenome.nih.gov).

Chapter 3

Exome-Wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in the Cancer Genome Atlas

3.1 Abstract

Background: Damaging germline *BRCA1/2* variants do not influence tumor mutation profile unless the remaining copy of *BRCA1/2* is somatically altered. Whether combined germline and somatic bi-allelic alterations are universally required for germline variation to influence tumor mutational profile is unclear. Here we performed an exome-wide analysis of the frequency and functional effect of bi-allelic alterations in The Cancer Genome Atlas (TCGA).

Methods: We integrated germline variant, somatic mutation, somatic methylation, and somatic copy number loss data from 7,790 individuals from TCGA to identify germline and somatic bi-allelic alterations in all coding genes. We used linear models to test for association between mono- and bi-allelic alterations and somatic microsatellite

instability (MSI) and somatic mutational signatures.

Results: We discovered significant enrichment of bi-allelic alterations in mismatch repair (MMR) genes, and identified 6 bi-allelic carriers with elevated microsatellite instability (MSI), consistent with Lynch syndrome. In contrast, we find little evidence of an effect of mono-allelic germline variation on MSI. Using MSI burden and bi-allelic alteration status, we reclassify two variants of unknown significance in *MSH6* as potentially pathogenic for Lynch syndrome. Extending our analysis of MSI to a set of 127 DNA damage repair (DDR) genes, we identified a novel association between methylation of *SHPRH* and MSI burden.

Conclusions: We find that bi-allelic alterations are infrequent in TCGA, but most frequently occur in *BRCA1/2* and MMR genes. Our results support the idea that bi-allelic alteration is required for germline variation to influence tumor mutational profile. Overall, we demonstrate that integrating germline, somatic, and epigenetic alterations provides new understanding of somatic mutational profiles.

3.2 Background

In rare familial cancer, inherited variation can both increase cancer risk and influence the molecular landscape of a tumor. For example, Lynch syndrome is characterized by an increased cancer risk and increased burden of somatic microsatellite instability (MSI) [10, 72]. The study of this phenomenon has been recently extended to sporadic cancers. For example, carriers of pathogenic mutations in *BRCA1/2* have both increased cancer risk and molecular evidence of homologous recombination deficiency in their tumors [15, 62]. Novel sequencing and analytical methods can be used to reveal a myriad of molecular phenotypes in the tumor, such as mutational signatures, rearrangement signatures, MSI, and infiltrating immune cell content [61, 47, 48, 64, 66]. A number of novel associations

between these molecular somatic phenotypes and germline variants have recently been discovered. Rare variants in *BRCA1/2* have been associated with mutational signature 3, a novel rearrangement signature, and an overall increased mutational burden [65, 66, 69, 67]. Common variants in the *APOBEC3* region have been associated with the corresponding APOBEC deficient mutational signature, and a haplotype at the 19p13.3 locus has been associated with somatic mutation of *PTEN* [77, 73]. In addition, interestingly, distinct squamous cell carcinomas (SCCs) arising in the same individual have a more similar somatic copy number profile than SCCs that occur between individuals [37]. Taken together, these results demonstrate that both common and rare germline variation can influence the somatic phenotype of sporadic cancers.

Similar to the two-hit mechanism of inactivation of tumor suppressor genes in familial cancer syndromes described by Nordling and then Knudson decades ago, germline and somatic bi-allelic alteration of *BRCA1/2* is required to induce somatic mutational signature 3, a single germline 'hit' is not sufficient [6, 7, 69, 67]. Whether a secondary 'hit' is universally required for germline variation to influence somatic phenotype is currently unclear. Here we address this question using the Cancer Genome Atlas (TCGA) dataset. TCGA is the most comprehensive resource of germline and somatic variation to enable this analysis, as it contains paired tumor and normal sequence data and a number of other molecular somatic phenotypes for 33 cancer types [119]. In contrast with previous studies of TCGA germline variation that focused on specific cancer types or candidate genes, we performed an exome-wide analysis to identify genes affected by both germline and somatic alterations (referred to as bi-allelic alteration) and study their association with somatic phenotypes [77, 65, 69, 67, 88]. Specifically, we conducted an integrated study of all genetic factors that contribute to somatic MSI burden and identified 6 individuals with characteristics consistent with Lynch syndrome: bi-allelic alteration of a MMR gene, elevated somatic MSI and an earlier age of cancer diagnosis.

3.3 Methods

3.3.1 Data Acquisition

Approval for access to TCGA case sequence and clinical data were obtained from the database of Genotypes and Phenotypes (project #8072: Integrated analysis of germline and somatic perturbation as it relates to tumor phenotypes). WXS germline variant calls from 8,542 individuals were obtained using GATK v3.5 as described previously [117]. Samples prepared using whole genome amplification (WGA) were excluded from analysis due to previous identification of technical artifacts in both somatic and germline variant calls in WGA samples [117, 152]. Somatic mutation calls obtained using MuTect2 were downloaded from GDC as mutation annotation format (MAF) files [134]. Raw somatic sequence data was downloaded from the genomic data commons (GDC) in BAM file format aligned to the hg19 reference genome. Normalized somatic methylation beta values from the Illumina 450 methylation array for the probes most anti-correlated with gene expression were downloaded from Broad Firehose (release stddata__2016_01_28, file extension: min_exp_corr). A total of 7,790 samples and 28 cancer types had germline, somatic, and methylation data available.

Segmented SNP6 array data were downloaded from Broad Firehose (release stddata__2016_01_28, file extension: segmented_scna_hg19). Segments with an estimated fold change value ≤ 0.9 , which corresponds to a single chromosome loss in 20% of tumor cells, were considered deletions. RNAseq RSEM abundance estimates normalized by gene were downloaded from Broad Firehose (release 2016_07_15, file extension: RSEM_genes_normalized). MSI calls from 5,931 TCGA WXS samples were obtained from previous work done by Hause et. al [48]. MSI is expressed as the percentage of microsatellite regions that display somatic instability. Aggregate allele frequencies and allele frequencies in 7 ancestry groups (African, Admixed American, East Asian, Finnish,

non-Finnish European, South Asian, and other) were obtained from ExAC v3.01 [76]. Gene-level expression data from normal tissues was downloaded from the GTEx portal (V7, file extension: RNASEQCv1.1.8_gene_tpm) [157].

3.3.2 Variant Annotation and Filtering

Raw variant calls were filtered using GATK VQSR TS 99.5 for SNVs and TS 95.0 for indels. Additionally, indels in homopolymer regions, here defined as 4 or more sequential repeats of the same nucleotide, with a quality by depth (QD) score < 1 were removed.

Putative germline and somatic loss-of-function (LOF) variants were identified using the LOFTEE plugin for VEP and Ensembl release 85 [130]. LOFTEE defines LOF variants as stop-gained, nonsense, frameshift, and splice site disrupting. Default LOFTEE settings were used and only variants receiving a high confidence LOF prediction were retained. It was further required that LOF variants have an allele frequency < 0.05 in all ancestry groups represented in ExAC. For somatic mutations, LOFTEE output with no additional filters was used. Gene level, CADD score, and ClinVar annotations were obtained using ANNOVAR and ClinVar database v.20170905 [158]. A germline variant was determined to be pathogenic using ClinVar annotations if at least half of the contributing sources rated the variant "Pathogenic" or "Likely Pathogenic". Li-Fraumeni variant annotations were obtained from the IARC-TP53 database [159, 160, 131]. Pfam protein domain annotations used in lollipop plots were obtained from Ensembl biomaart [161, 162].

3.3.3 Somatic Methylation

For each gene, the methylation probe that was most anti-correlated with gene expression was obtained from Broad firehose and used for all subsequent analyses. Methylation calls were performed for each gene and each cancer type independently. For each gene, the beta value of the chosen methylation probe was converted to a Z-score within each

cancer type. Individuals with a Z -score ≥ 3 were considered hyper methylated ($M=1$) and all others were considered non-methylated ($M=0$). To determine if methylation calls were associated with reduced somatic gene expression, a linear model of the form $\log_{10}(E_{ij}) \sim C_i + M_{ij}$ was used, where E_{ij} denotes expression of gene j in tumor i , C_i denotes cancer type of sample i , and M_{ij} denotes binary methylation status of gene j in sample i . Only genes where methylation calls were nominally associated with decreased gene expression were retained. Using this process we identified 863,798 methylation events affecting 11,744 genes.

3.3.4 Loss of Heterozygosity

To assess loss of heterozygosity (LOH) for a given heterozygous germline variant, the somatic allele frequency of the germline variant was obtained from the somatic BAM files using samtools mpileup v1.3.1 (SNPs) or varscan v2.3.9 (indels) [139, 135]. Any germline variant that was not observed in the tumor was excluded from further analysis. A one-way Fisher's exact test comparing reference and alternate read counts was performed to test for allelic imbalance between the normal and tumor sample. Only sites with a nominally significant ($p < 0.05$) increase in the germline allelic fraction were retained. To confirm that the observed allelic imbalance was due to somatic loss of the WT allele and not due to somatic amplification of the damaging allele, we required that the region be deleted in the tumor based on TCGA CNV data (fold change value ≤ 0.9). Loci that had a significant Fisher's exact test but were not located in a somatic deletion were considered 'allelic imbalance' (AI). Using this method, we observed 3,418 LOH events in 1,672 genes.

3.3.5 Gene Set Enrichment Analysis

Gene set enrichment analysis was performed using the fgsea R package and the following parameters: minSize=3, maxSize=500, nperm=20,000 and the canonical pathways

gene set from MsigDB (c2.cp.v5.0.symbols.gmt) [163, 164]. Genes were ranked according to the fraction of germline LOF variants that acquired a second somatic alteration (number bi-allelic alterations/number germline LOF variants). Genes with fewer than 3 germline LOF variants in the entire cohort were excluded from this analysis to reduce noise.

3.3.6 Mutational Signature Analysis

To identify somatic mutational signatures, counts for each of 96 possible somatic substitutions +/- 1 bp context were obtained for all tumor samples. For each sample, mutational signatures were identified using the DeconstructSigs R package, which uses a non-negative least squares regression to estimate the relative contributions of previously identified signatures to the observed somatic mutation matrix [165]. DeconstructSigs was run with default normalization parameters and relative contributions were estimated for the 30 mutational signatures in COSMIC [166].

To estimate significance of association between germline variants and somatic mutational signature burden we employed both a pan-cancer Wilcoxon rank sum test and a permutation-based approach to ensure that significance was due to germline variant status and not cancer type. For the permutation approach, the pairing between germline variant status and mutational signature profile was shuffled 10,000x. A Wilcoxon rank sum test was run for each permutation to obtain a null distribution for the test statistic. P-values were determined for each signature as the fraction of permutations with a Wilcoxon test statistic greater than or equal to the observed data.

3.3.7 Statistical Analyses

Principal Component Analysis (PCA) was performed on common (allele frequency > 0.01) germline variants using PLINK v1.90b3.29 and the first two principal components obtained from this analysis were used to control for ancestry in all of the regression models

we fit to the data [127]. G*Power 3.1 was used to perform a power calculation for the contribution of damaging germline variants to somatic MSI [167]. The following parameters were used: α error probability = 0.05, power = 0.80, effect size = $6.83e^{-4}$, number of predictors = 20. To assess potential co-occurrence of *SHPRH* methylation with alterations in other genes, individuals were grouped according to presence (+) or absence (-) of *SHPRH* methylation. A one-way Fisher's exact test was used to test for an abundance of another alteration of interest in *SHPRH* methylation positive individuals vs. *SHPRH* methylation negative individuals. Individuals with $> 5,000$ somatic mutations were excluded from these analyses to exclude potential confounding due to somatic hypermutation.

To test for association between genetic alteration and somatic MSI burden, a linear model of the form $\log_{10}(M_i) \sim G_{ij} + S_{ij} + Me_{ij} + X_i$ was used, where M_i denotes somatic MSI burden of sample i , G_{ij} , S_{ij} , and Me_{ij} are binary indicators for germline, somatic, and methylation alteration status of gene j in sample i , and X_i represents a vector of covariates for sample i (cancer type, PC1, PC2). All analyses using somatic MSI data were performed on a maximum of $n = 4,997$ individuals. To test for association between germline alteration and age of diagnosis a linear model of the form $A_i \sim G_{ij} + X_i$ was used where A_i denotes age of diagnosis for sample i , G_{ij} , is a binary indicator for germline alteration status of gene j in sample i , and X_i represents a vector of covariates for sample i (cancer type, PC1, PC2). All analyses using age of diagnosis were performed on a maximum of $n = 8,913$ individuals.

3.4 Results

3.4.1 MMR Pathway is Frequently Affected by Bi-allelic Alteration

To find events most likely to influence a somatic phenotype, we limited our analysis to alterations predicted to be highly disruptive. We therefore only considered loss-of-function (LOF) germline variants, LOF somatic mutations, epigenetic silencing of genes via DNA hyper-methylation, and somatic loss of heterozygosity (LOH) events that select for a germline LOF allele (see methods and Supplementary Figure B.1 and B.2). In total, we analyzed 7,790 individuals with germline variant, somatic mutation, and methylation data available, corresponding to 95,601 germline LOF variants, 225,257 somatic LOF mutations, and 863,798 somatic methylation events (Figure 3.1). Using this data, we were able to determine the frequency of three types of germline bi-allelic alterations: 1) germline LOF and somatic LOF (germline:somatic), 2) germline LOF and somatic epigenetic silencing (germline:methylation), and 3) germline LOF with somatic LOH.

Surprisingly, we found a low incidence of bi-allelic alterations, with only 4.0% of all germline LOF variants acquiring a secondary somatic alteration via any mechanism. We observed 198 germline:somatic events (0.02% of all germline LOF), 433 germline:methylation events (0.04%), and 3,279 LOH events (3.4%). To determine whether bi-allelic alterations affect specific biological processes, we ranked genes by the frequency of bi-allelic alteration and performed a gene set enrichment analysis (GSEA) using 1,330 canonical pathway gene sets [163, 164]. The only association significant beyond a multiple hypothesis correction was an enrichment of germline:somatic alterations in the KEGG mismatch repair (MMR) pathway ($q = 0.0056$) (Supplementary Figure B.3). To ensure that the lack of enriched pathways wasn't due to our strict definition of somatic damaging events, we repeated the analysis including all somatic mutations with a CADD score ≥ 20 . Though this increased

the number of germline:somatic alterations (376, 0.039%), no additional significantly enriched pathways were found. Similarly we repeated the analysis using a less restrictive definition of LOH, referred to as 'allelic imbalance' (AI), that accommodates other mechanisms such as copy neutral LOH (see methods). We again observed more AI events (7,920, 8.2%), but no additional pathways were significantly enriched.

3.4.2 Landscape of Germline and Somatic Alteration of DNA Damage Repair Pathways

Having shown that MMR genes frequently harbor bi-allelic alterations, we next investigated the frequency of germline, somatic, and epigenetic alterations in a panel of 210 DNA damage repair (DDR) genes. While germline variation in DDR genes has previously been studied, only a few studies have considered specific DDR pathway information. DDR genes were assigned to 8 gene sets using pathway information: direct repair, translesion synthesis, mismatch repair, Fanconi anemia, non-homologous end joining, base excision repair, homologous recombination, and nucleotide excision repair [45]. We also examined 3 additional cancer-relevant gene sets: oncogenes, tumor suppressors, and cancer predisposition genes [17, 2]. For each gene set and cancer type, we calculated the fraction of individuals with bi-allelic, germline, somatic, or epigenetic alteration of any gene in the gene set (Figure 3.1).

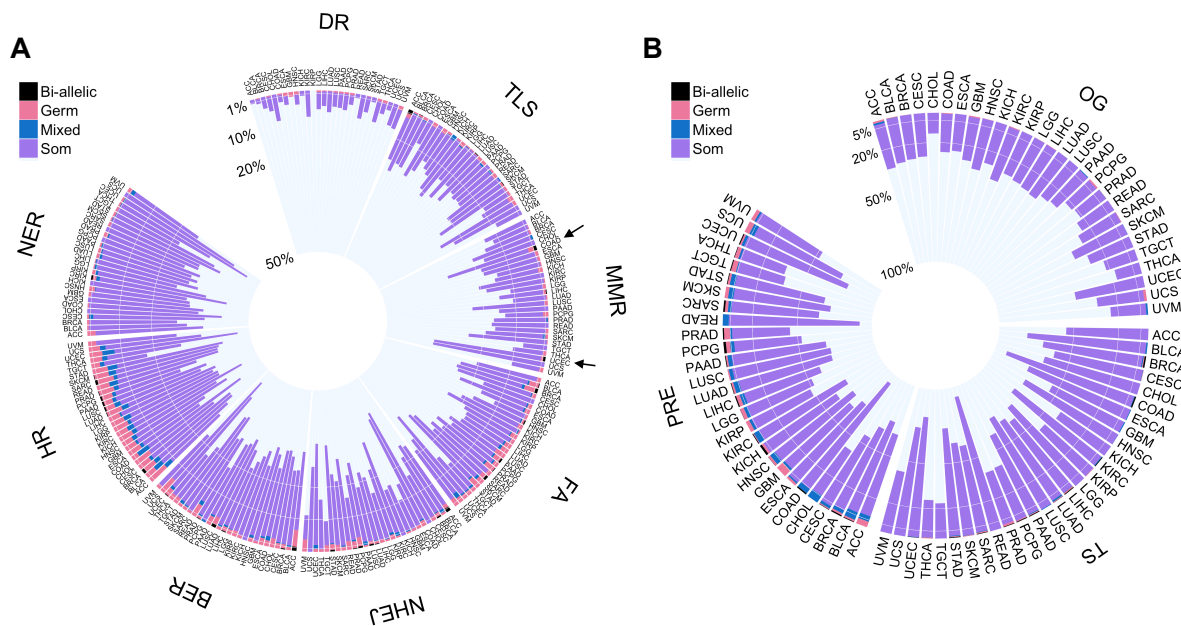


Figure 3.1: Frequency of Germline and Somatic Alterations in Cancer-Relevant Pathways. (A-B) Circos plots displaying the individual-level frequency of alterations for each cancer type in DNA damage repair pathways (A) or oncogenes, tumor suppressors, and cancer predisposition genes (B). Individuals were grouped into four mutually exclusive categories based on the type of alterations observed in the gene set: Bi-allelic: combined germline and somatic alteration of the same gene; Mixed: germline and somatic alteration of different genes in the set; Germ: germline alterations only; Som: somatic alterations only (mutation or methylation). The height of each bar represents the fraction of individuals in each alteration category. The black arrows highlight cancer types with bi-allelic mismatch repair alterations. Gene sets are ranked according to size moving clockwise. Pathway abbreviations and size: DR: direct repair (N = 3 genes); TLS: translesion synthesis (N = 19); MMR: mismatch repair (N = 27); FA: Fanconi anemia (N = 34); NHEJ: non-homologous end joining (N = 37); BER: base excision repair (N = 43); HR: homologous recombination (N = 53); NER: nucleotide excision repair (N = 70); OG: oncogenes (N = 54); TS: tumor suppressors (N = 71); PRE: predisposition genes (N = 144). There are 382 unique genes total and gene sets are not mutually exclusive.

Consistent with previous studies, the fraction of individuals carrying germline LOF was low for both DDR genes and cancer-relevant gene sets (Figure Figure 3.1) [65]. Overall, 16% of individuals carried a germline LOF in any of the genes interrogated, with 5% carrying a germline LOF in a known predisposition gene. For each gene set, we tested for

overabundance of germline LOF carriers in each cancer type vs. all other cancer types. We discovered associations between breast cancer and germline alteration of the Fanconi anemia and tumor suppressor gene set, which are likely driven by *BRCA1/2* germline variants (Supplementary Figure B.4A). We expanded our analysis to include known pathogenic missense variants from the ClinVar database and discovered additional significant associations between pheochromocytoma and paraganglioma (PCPG) and both the predisposition and oncogene sets (Supplementary Figure B.4B) [158]. This association is driven by the known PCPG predisposition genes *SDHB* and *RET* that have been previously reported [168]. Loss of heterozygosity in these PCPG individuals was frequently observed (77% of *SDHB* germline carriers), consistent with *SDHB* acting via a tumor suppressor mechanism [169]. We conclude that there is no cancer type in TCGA that harbors an excess of damaging germline variants in DDR or cancer-relevant genes, with the exception of the well-described predisposition syndrome genes *BRCA1/2*, *SDHB*, and *RET*.

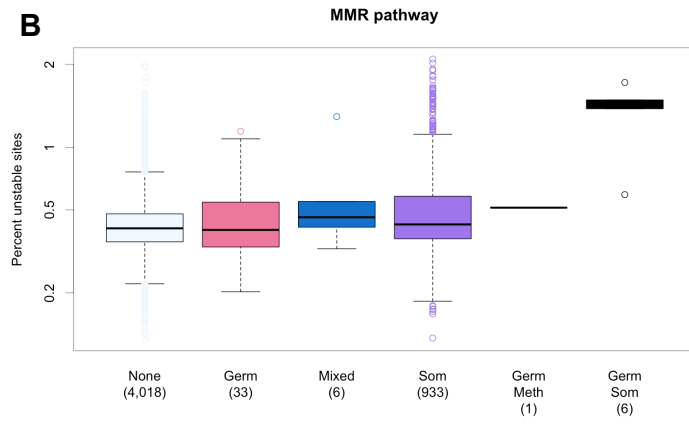
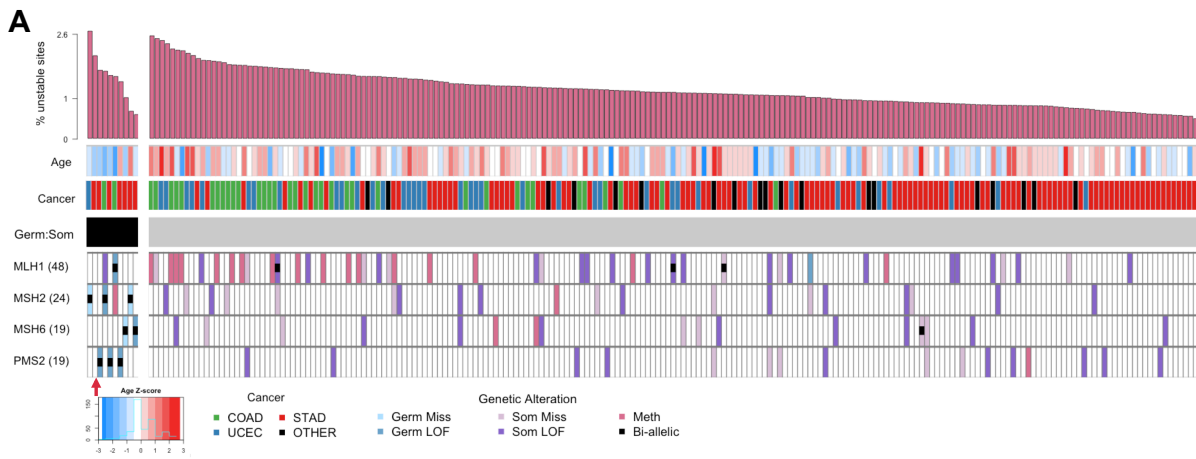
3.4.3 Individuals in TCGA Exhibit Lynch Syndrome Characteristics

We found that the MMR pathway was significantly enriched for germline:somatic alterations. This association was driven by six individuals who carry a germline:somatic alteration of a MMR gene. In five individuals, the gene affected was a known Lynch syndrome gene (*MLH1*, *MSH2*, *MSH6*, and *PMS2*), which we will refer to as L-MMR genes [72]. The remaining individual carried a germline:somatic alteration of *MSH5* (Figure 3.2A, red arrow). While *MSH5* is not known to be a Lynch syndrome gene, we included this individual in further analyses of MMR germline:somatic alteration carriers. Four of the germline:somatic alteration carriers have uterine cancer (UCEC) and two have colon cancer (COAD), cancer types characteristic of Lynch syndrome (Figure 3.1A, arrows) [113]. This prompted us to investigate the molecular and clinical phenotype of germline:somatic

alteration carriers to determine if they are consistent with Lynch syndrome characteristics. While germline:somatic alteration of MMR genes in TCGA has been previously described, detailed somatic phenotyping of these individuals has not been performed [47]. Using previously published MSI data, we investigated the fraction of microsatellite loci that exhibit instability in the tumor (somatic MSI burden) of individuals carrying alterations in MMR genes [48]. Figure 3.2A shows germline, somatic, and epigenetic alteration status of L-MMR genes for all individuals classified as MSI high (MSI-H) by Hause et.al, with bi-allelic mutation carriers grouped to the left. Interestingly, only 76% of MSI-H individuals have an alteration (germline LOF, somatic LOF, or hyper-methylation) of an MMR gene, indicating that some of the variation in somatic MSI is not explained by the genetic alterations investigated.

Figure 3.2: Genetic and Clinical Characteristics of MSI-H Individuals.

(A) CoMut plot displaying germline, somatic, and epigenetic events in L-MMR genes (bottom 4 rows - number of affected individuals in parenthesis) for 217 MSI-H individuals (columns). The top histogram represents MSI burden expressed as the fraction of possible microsatellite sites that are unstable. Age of diagnosis was converted to a Z-score using the mean and standard deviation age for each cancer type. Cancer types with fewer than 5 MSI-H individuals are labeled "Other" and include bladder, head and neck, kidney, glioma, lung, liver, prostate, stomach, and rectal cancer. The type of genetic alteration is indicated by color and bi-allelic events are indicated by a black box. Individuals with bi-allelic (germline:somatic) MMR mutations are grouped to the left. The red arrow highlights an individual with bi-allelic alteration in MSH5 (not an L-MMR gene). (B) Somatic MSI burden in 4,997 TCGA individuals grouped by type of MMR pathway alteration. Categories are the same as those described in Figure 3.1: Bi-allelic: combined germline and somatic alteration of the same gene; Mixed: germline and somatic alteration of different genes in the set; Germ: germline alterations only; Som: somatic alterations only (mutation or methylation). Individuals with bi-allelic alteration occurring via germline:somatic and germline:methylation mechanisms are displayed separately. The number of individuals in each category is indicated in parentheses.



Using a linear model controlling for cancer type, we found that the six individuals with germline:somatic MMR alterations were diagnosed on average 14 years earlier ($p = 0.0041$) and have 2.8 fold higher somatic MSI ($p = 3.95e^{-15}$) than individuals with any other type of MMR pathway alteration (Figure 3.2B, Supplemental Tables B.1 and B.2). Of the five individuals with germline:somatic alteration of a L-MMR gene, four carried a germline LOF variant that is known to be pathogenic for Lynch syndrome, and one carried a LOF variant *MSH6* (p.I855fs) not present in ClinVar (Supplementary Table B.3). This *MSH6* variant is located in the same exon as another known pathogenic frameshift variant, suggesting it likely predisposes to Lynch syndrome (Supplementary Table B.4). While a diagnosis of Lynch syndrome requires clinical family history data not available in TCGA, the carriers were diagnosed at an earlier age and exhibit increased somatic MSI characteristic of Lynch syndrome. We note that this result would have gone unnoticed in an analysis of somatic MSI using interaction terms to model bi-allelic alteration at the single gene level, highlighting the value of grouping genes by biological pathway (Supplementary Table B.5). Interestingly, we observed the identical nonsense mutation in *PMS2* (p.R628X) in two individuals, once as an inherited variant and once as an acquired somatic mutation (Supplementary Figure B.5). This overlap between clinically-relevant germline variants and somatic mutations suggests that, in some instances, the origin of a mutation is less important than its functional effect.

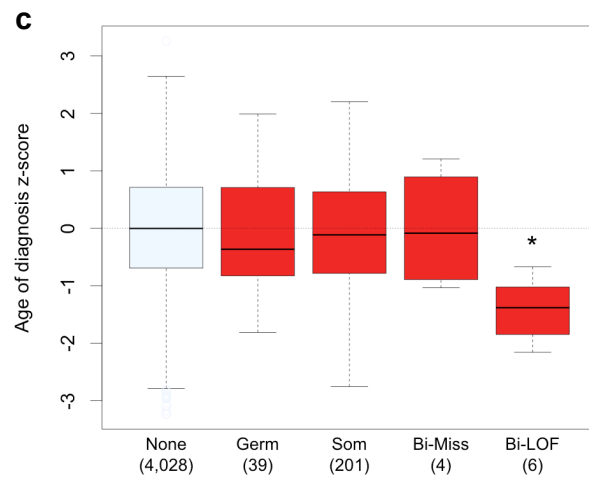
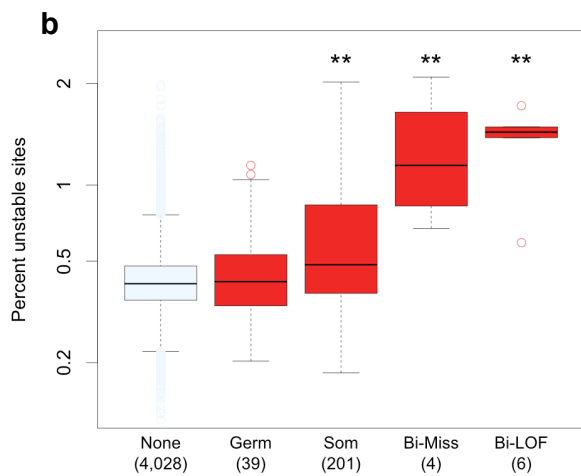
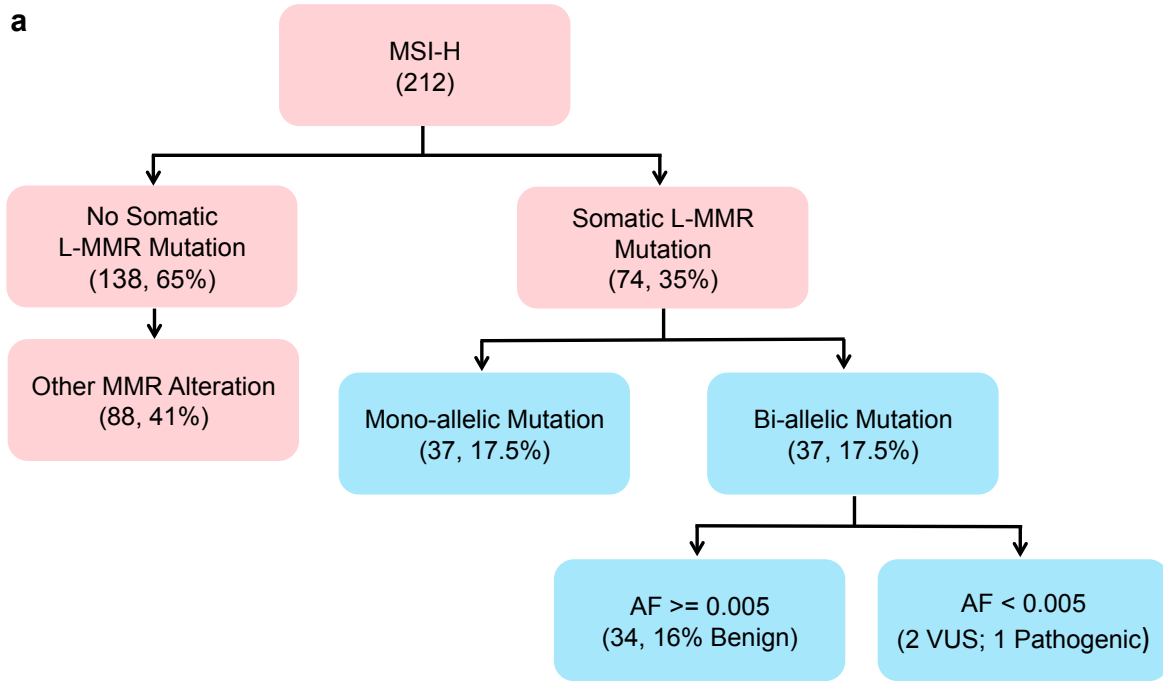
3.4.4 Using MSI-H to Reclassify Variants of Unknown Significance

Given the large effect of germline:somatic LOF mutations on somatic MSI, we next asked whether germline:somatic missense mutations produced a similar phenotype. We expanded our analysis to include missense variants known to be pathogenic for Lynch syndrome from ClinVar. We identified one individual with bi-allelic alteration of *MSH2*

involving a pathogenic missense germline variant (p.S554N) and a somatic LOF mutation (Supplementary Table B.3). Including missense somatic mutations with a CADD score ≥ 20 led to the identification of one individual with bi-allelic alteration of *PMS2* involving a germline LOF variant (p.R563X) and a secondary somatic missense mutation (Supplementary Table B.4).

We observed a number of missense germline variants in L-MMR genes not present in ClinVar, which we consider variants of unknown significance (VUS). We reasoned that the phenotype of elevated somatic MSI and germline:somatic L-MMR mutation could be used to identify germline VUS likely to be pathogenic for Lynch syndrome. Using 212 individuals classified as MSI-H, we identified 74 individuals with a damaging somatic mutation in a L-MMR gene (Figure 3.3A) [48]. Of the individuals with L-MMR somatic mutations, 37 have a germline missense variant in the somatically mutated gene. To identify variants most likely to be damaging, we retained only those with a minor allele frequency < 0.005 in all ancestry groups represented in ExAC. Three individuals met the criteria of having an MSI-H phenotype and a bi-allelic L-MMR mutation involving a likely damaging missense germline variant. One was the previously identified *MSH2* p.S554N variant carrier, the others carried two VUS: *MSH2* (p.P616R) and *MSH6* (p.F432C) (Supplementary Table B.4).

Figure 3.3: Identification of Potential Pathogenic Lynch Syndrome Variants. (A) Analysis workflow: 212 individuals with MSI-H classification were dichotomized based on presence of germline:somatic mutation of a L-MMR gene. Individuals carrying germline:somatic mutations were further subdivided by allele frequency of the candidate germline variant in ExAC. Pink boxes indicate the use of somatic data, blue boxes integrate somatic and germline data. Numbers in parenthesis refer to number of individuals that fulfill the box criteria. Individuals that carry bi-allelic alterations are labeled according to ClinVar significance of the germline variant. VUS: variant of unknown significance. (B-C) Somatic MSI burden (B) and age of diagnosis (C) of individuals who carry germline:somatic mutations in a MMR gene. Individuals were grouped by MMR gene mutation type: None: no alteration; Germ: germline LOF variants only; Som: somatic LOF mutations only; Bi-Miss: bi-allelic alteration including a missense mutation; Bi-LOF, bi-allelic alteration via dual LOF mutations. Age was converted to a Z-score using the mean and standard deviation age of diagnosis for each cancer type. ** = $p < 0.001$, * = $p < 0.01$; p-values were determined using a linear model to predict somatic MSI burden while accounting for cancer type.



Closer investigation of the *MSH6* p.F432C variant showed that other amino acid substitutions at the same residue were classified as pathogenic in ClinVar (Supplementary Table B.4). Further, the individual carrying the *MSH6* p.F432C variant had an earlier age of diagnosis ($Z = -1.03$), suggesting this variant is pathogenic. In contrast, the individual carrying the *MSH2* p.P616R variant had an older age of diagnosis ($Z = 1.20$), suggesting this variant is not pathogenic. While validation is required to confirm pathogenicity of this variant as well as the previously mentioned *MSH6* p.I855fs, we offer genetic and clinical evidence that these variants may predispose to Lynch syndrome, as well as show evidence suggesting that *MSH2* p.P616R is likely benign.

3.4.5 Missense Alterations Exhibit an Attenuated Lynch Phenotype

Taken together, we have identified ten individuals with germline:somatic MMR alterations, six of which carry a germline variant that is known to be pathogenic for Lynch syndrome (Table 3.1). With this in mind, we asked whether individuals with germline:somatic LOF mutations have a more severe phenotype than those with combined LOF and missense mutations. Bi-allelic alteration carriers were divided into two groups: those with germline and somatic LOF mutations (Bi-LOF, $n = 6$) and those with missense germline variants or missense somatic mutations (Bi-Miss, $n = 4$). We found that both Bi-LOF ($p = 2.78e^{-15}$) and Bi-Miss ($p = 1.01e^{-10}$) groups have significantly elevated MSI (Figure 3.3B and Supplementary Table B.6). Bi-Miss and Bi-LOF have a median 1.50 and 2.35 fold higher somatic MSI compared to individuals with somatic MMR alteration alone, demonstrating a synergistic effect between germline variants and somatic mutations. Similarly, both Bi-LOF and Bi-Miss groups had significantly higher contribution of mutational signature 6, a signature associated with mismatch repair defects (Supplementary Figure B.6) [61]. In contrast, only Bi-LOF individuals were diagnosed at an earlier age (Figure

3.3C and Supplementary Table B.7). These results show that any damaging bi-allelic MMR alterations are sufficient to induce a MSI-H tumor phenotype, but only bi-allelic alterations via dual LOF mutation are associated with an earlier age of diagnosis.

Table 3.1: Bi-allelic Germline:Somatic MMR Alteration. Number of individuals affected by three types of germline:somatic alterations in MMR genes. LOF = Loss of function variant, MISS = missense variant, * = individual carries a ClinVar pathogenic germline variant

Gene	germline LOF somatic LOF	germline LOF somatic MISS	germline MISS somatic LOF
<i>MLH1</i>	1*		
<i>MSH2</i>	1*		1,1*
<i>MSH6</i>	1		1
<i>PMS2</i>	2*	1*	
<i>MSH5</i>	1		

3.4.6 Mono-allelic Germline Alteration has Little Effect on Somatic MSI

Having shown that combined germline LOF and missense somatic mutations are sufficient to cause elevated MSI, we hypothesized that damaging germline variation in the absence of somatic mutation could also increase somatic MSI. To maximize power we expanded our analysis to include all MMR genes as well as two different categories of damaging germline variation: known (ClinVar) and predicted (CADD ≥ 30) pathogenic. Individuals with any somatic alterations in MMR genes were excluded from this analysis to get an accurate estimate of the effect of damaging germline variation alone. There were no significant association between damaging germline variation in the MMR pathway and somatic MSI burden (Supplementary Figure B.7 and Table B.8). Known variants showed the strongest effect (0.02 fold increase in MSI burden), and this was largely driven by *MLH3* p.V741F, a variant with conflicting reports of pathogenicity that is carried by 195 individuals. From this we conclude that the effect of damaging germline variation without concomitant somatic mutation on somatic MSI is small.

3.4.7 Methylation of SHPRH Associated with Somatic MSI

We observe that 24% of MSI-H individuals have no alteration of an MMR gene, suggesting that there is variation in somatic MSI burden due to factors outside of known MMR genes (Figure 3.3B) [170]. To investigate this further, we extended the search to all DDR genes. We separately assessed the contribution of germline LOF, somatic LOF, and somatic methylation to somatic MSI burden using a gene level linear model. Somatic LOF frameshift mutations that overlap with microsatellite loci were removed from this analysis, as we were unable to determine the direction of causality between these mutations and overall MSI burden (Supplementary Figure B.8 and Table B.9). Additionally, the MMR bi-

allelic alteration carriers were excluded from this analysis to obtain an accurate assessment of mono-allelic germline variation. The results of this analysis are summarized in Figure 3.4. Consistent with the lack of association between damaging MMR germline variants and somatic MSI, we found no significant association at the single gene level between germline LOF and somatic MSI (Figure 3.4A).

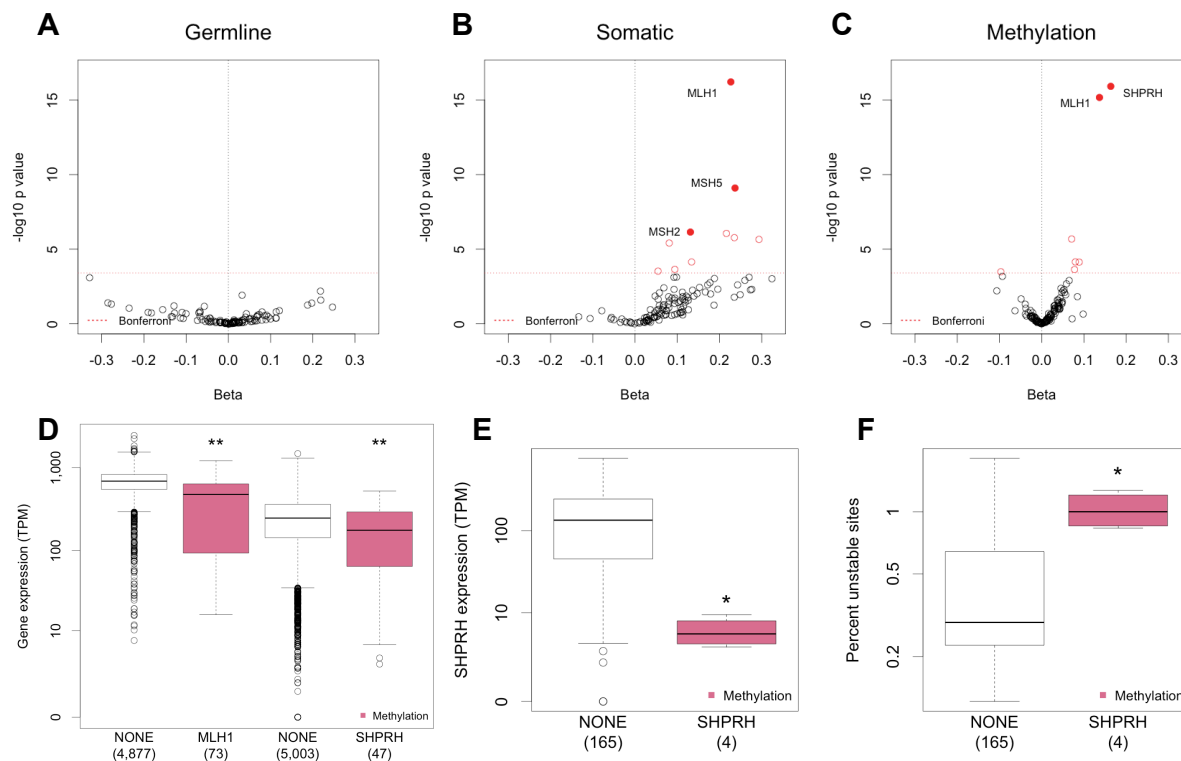


Figure 3.4: Germline, Somatic, and Epigenetic Associations with MSI. (A-C) Volcano plots of gene-level association testing between germline LOF (A) somatic LOF (B) and somatic methylation (C) and somatic MSI burden. A total of 127 DDR genes were tested in 4,987 individuals. Red dotted line represents Bonferroni significance cutoff. (D) Somatic expression of *MLH1* and *SHPRH* in individuals with somatic methylation. ** = $p < 0.001$ as determined using a linear model to predict gene expression while accounting for cancer type. (E-F) Somatic *SHPRH* expression is significantly reduced (E, Wilcox $p = 0.0018$) and somatic MSI is significantly increased (F, Wilcox $p = 0.0067$) in uterine tumors with *SHPRH* methylation. TPM = transcripts per million. The number of individuals in each category is indicated in parentheses.

We found that somatic mutation of *MLH1* and *MSH2* and somatic methylation of

MLH1 were associated with increased MSI burden, confirming what has been previously reported (Figure 3.4B,C) [170]. In addition, we discovered a novel association between methylation of *SHPRH* and elevated somatic MSI ($p = 1.19e^{-16}$) (Figure 3.4C). *SHPRH* is a E3 ubiquitin-protein ligase and a member of the translesion synthesis pathway, a pathway that enables DNA replication to traverse regions of DNA damage via specialized polymerases [171]. Methylation of *SHPRH* was associated with a 16% decrease in gene expression in a pan-cancer analysis (Figure 3.4D). We observed that methylation of *SHPRH* has the strongest effect both on *SHPRH* expression and somatic MSI burden in uterine cancer (Figure 3.4E,F and Supplementary Figure B.9). Interestingly, *SHPRH* expression is highest in normal ovarian and uterine tissues among 23 tissues examined, suggesting a specific function for *SHPRH* in these organs (Supplementary Figure B.10) [157]. Methylation of *MLH1* and *SHPRH* are both associated with mutational signature 6, with a stronger association in uterine cancer (Supplementary Figure B.11).

To confirm that *SHPRH* methylation is the likely causal factor influencing somatic MSI, we performed a co-occurrence analysis to find other somatic events correlated with *SHPRH* methylation (Supplementary Figure B.12). There were a large number of somatic events significantly correlated with *SHPRH* methylation, including somatic MMR mutations; however, we found that *SHPRH* methylation remains a significant determinant of somatic MSI even after accounting for other somatic MMR alterations (Supplementary Table B.10). Furthermore, we found a significant, albeit weaker, association between somatic expression of *SHPRH* and MSI burden, indicating that *SHPRH* methylation likely affects MSI burden via silencing of *SHPRH* (Supplementary Table B.11).

3.4.8 Mono-allelic Germline Alterations not Associated with Mutational Signatures

It was previously reported that bi-allelic alteration of *BRCA1/2* is associated with somatic mutational signature 3 [69]. As a result, we hypothesized that bi-allelic alterations in other DDR pathways may also be associated with known mutational signatures. We first attempted to replicate the *BRCA1/2* association, but surprisingly found high levels of mutational signature 3 in individuals carrying mono-allelic damaging germline *BRCA1/2* variation. However, when we considered AI events to be bi-allelic alterations, we no longer found a significant association between mono-allelic *BRCA1/2* alterations and somatic mutational signature 3 (Supplementary Figure B.13). In contrast to individuals with *BRCA1/2* LOH, we suspect that individuals with AI have sub-clonal *BRCA1/2* loss, which would explain the lower levels of signature 3 observed. Thus, we demonstrate that variability in LOH calling method can lead to conflicting results.

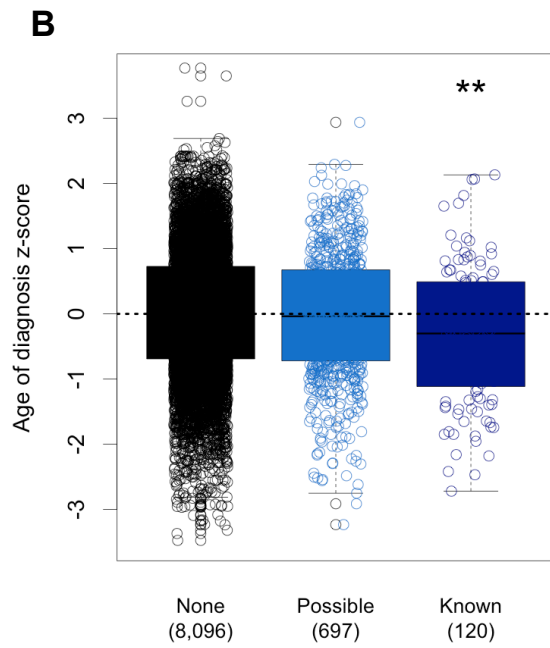
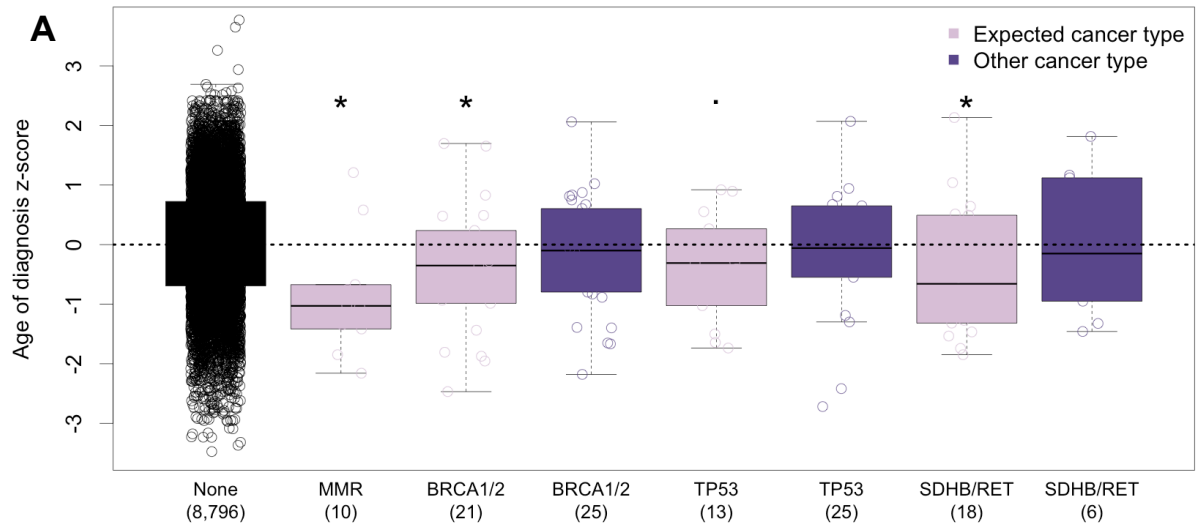
We next tested for association between 30 somatic mutational signatures from COSMIC and germline bi-allelic alteration in six DDR pathways with more than five individuals carrying bi-allelic alteration (FA, MMR, HR, BER, NHEJ, and TLS) (Supplementary Figure B.14A) [166]. The only significant association uncovered (FDR < 15%) was between Fanconi anemia and signature 3, which was driven by the known association between *BRCA1/2* alterations and signature 3. We found that when we include all bi-allelic alterations in MMR genes, there was not a significant association with signature 6. This was due to the inclusion of germline:methylation events. Limiting our analyses to germline:somatic events led to an association that was statistically significant after multiple hypothesis correction (Supplementary Figure B.6). This suggests that the mechanism of secondary somatic alteration modulates the effect of germline variation on somatic phenotype. We repeated this analysis expanding to include individuals with mono-allelic germline alteration

in DDR pathways (Supplementary Figure B.14B). We found no significant associations, consistent with the idea that bi-allelic alteration is required for the germline to alter somatic mutational phenotypes.

3.4.9 Cancer Predisposition Syndromes in TCGA

While TCGA is generally thought to represent sporadic adult onset cancers, our work as well as that of others has shown evidence suggesting that some individuals in TCGA have hereditary cancer predisposition syndromes. Known pathogenic variation in *SDHB/RET*, *BRCA1/2* and MMR genes is thought to be responsible for a subset of pheochromocytoma and paraganglioma, breast, ovarian, colon, and uterine cancers in TCGA [47, 168, 172, 69]. Another relatively common cancer syndrome that predisposes to cancer types found in TCGA is Li-Fraumeni syndrome (LFS), which arises due to inherited variation in *TP53* [10]. Using the IARC-TP53 variant database, we identified 38 individuals carrying a potential LFS variant. Interestingly, aside from bi-allelic MMR alteration, we observed that pathogenic germline variation in cancer predisposition genes was not associated with an earlier age of diagnosis. To explore this further, we divided individuals into two groups: those who developed the cancer type expected given the predisposition gene altered and those with another cancer type. Using this approach, we found significant associations between germline alteration status and age of diagnosis for the expected cancer type (Figure 3.5A and Supplementary Table B.12). This suggests that predisposition syndromes can lead to an earlier age of onset in a specific spectrum of cancers, but have no significant effect on other cancer types.

Figure 3.5: Cancer Predisposition Syndromes in TCGA. (A) Age of diagnosis for MMR germline:somatic alteration carriers and individuals carrying ClinVar pathogenic or LOF germline variation in *BRCA1*, *BRCA2*, *TP53*, *SDHB*, and *RET*. Age was converted to a Z-score using the mean and standard deviation age of diagnosis for each cancer type. The expected cancer types for each gene set are: MMR, colon, uterine, and stomach; *BRCA1/2*, breast cancer; *TP53*, adrenal cortical carcinoma, glioma, glioblastoma, breast cancer, and sarcoma; *SDHB/RET*, pheochromocytoma and paraganglioma. All MMR germline:somatic alteration carriers have the expected cancer type. The number of individuals in each category is displayed in parentheses. (B) Age of diagnosis for individuals carrying ClinVar pathogenic or LOF germline variation in genes described in (A) ('known') compared to a set of 75 other cancer predisposing genes ('possible'). ** $p < 0.001$, * = $p < 0.05$, . $p < 0.1$. p values were determined using a linear model to predict age of onset while accounting for cancer type.



To determine if damaging germline variation in other predisposition genes was associated with earlier age of diagnosis, we examined 75 cancer predisposition genes not included in the previous analysis. We found no significant association between germline alteration status and age of diagnosis in any of these additional genes (Supplementary Figure B.15 and Table B.13). To increase power, we examined these additional genes in aggregate as a gene set ("possible") and compared this gene set to the genes we examined previously ("known", *BRCA1*, *BRCA2*, *MLH1*, *MSH2*, *MSH5*, *MSH6*, *PMS2*, *SDHB*, *RET*, and *TP53*). The known gene set was associated with an earlier age of diagnosis, but the possible gene set was not (Figure 3.5B). It is possible that using biological knowledge to group genes or cancer types in a meaningful way could increase power and find new associations. However, we believe much of the variation in age of diagnosis due to germline variation lies in genes associated with prevalent cancer predisposition syndromes.

3.5 Discussion

We present an analysis of cancer exomes that integrates germline variation, somatic mutation, somatic LOH, and somatic methylation. To our knowledge, our study is the first exome-wide analysis of the prevalence of bi-allelic alterations across the full spectrum of cancer types represented in TCGA, and one of the first to integrate somatic methylation data for a large number of genes. Of all gene sets and bi-allelic alteration mechanism examined, we only discovered a significant enrichment of combined germline and somatic LOF mutations in the MMR pathway. Bi-allelic alteration of the MMR pathway has been previously reported; however, the individuals harboring these alterations weren't studied in detail [47]. While a diagnosis of Lynch syndrome cannot be made without a family history, we identified 10 individuals with bi-allelic alteration in an MMR gene, elevated somatic MSI burden, and in individuals with bi-allelic LOF mutations, earlier age of cancer

diagnosis.

The genes harboring bi-allelic alterations by our analyses are predominantly those that are less frequently mutated in Lynch syndrome: *MSH6* and *PMS2*. Similarly, only 20% of the proposed Lynch individuals have colon cancer, the classic Lynch presentation. Thus, it is possible that what we observe is not bona fide Lynch syndrome, but an attenuated form of the disease [31, 113]. The median age of cancer onset in TCGA is 60, thus the individuals in TCGA carrying cancer predisposing variants may have genetic modifier mechanisms that delay cancer onset and severity. Interestingly, proposed mechanisms of genetic compensation delaying cancer onset have been described previously both for Lynch syndrome and Li-Fraumeni syndrome [33, 32]. We observed six individuals carrying a potentially pathogenic germline variant in a L-MMR gene (2 ClinVar pathogenic, 4 LOF) who did not acquire a second somatic mutation and don't have elevated somatic MSI burden. This is not unexpected as the penetrance of Lynch syndrome variants is often incomplete [72]. We observed that any damaging germline:somatic alteration is sufficient to induce elevated somatic MSI, but only individuals with Bi-LOF mutation have an earlier age of diagnosis. This observation is consistent with the previously proposed idea that bi-allelic MMR mutation is likely not the tumor-initiating event but instead acts to accelerate tumor growth (Figure 3.3B,C) [72]. Given our observations, we propose that the less damaging Bi-Miss mutations could lead to slower tumor growth than Bi-LOF mutations.

Recently, Polak et. al demonstrated that somatic mutational signature 3 and *BRCA1/2* LOH bi-allelic inactivation could be used to reclassify *BRCA1/2* germline variants that were previously considered VUS [69]. Here we provide another example of how somatic phenotype data can be used to reclassify germline VUS. We identify two novel potentially damaging Lynch syndrome variants in *MSH6*. Of note, the ClinVar pathogenic Lynch predisposing *MSH2* variant was not present in the ANNOVAR ClinVar database despite being reported in ClinVar, highlighting the importance of manual curation of po-

tentially pathogenic variants. Further experimental validation of these variants is required. Germline MMR variants can be used to guide therapy and monitoring for patients at risk. For example, the risk of colorectal cancer can be reduced in individuals carrying pathogenic germline MMR variants using a daily aspirin regimen [111, 17]. Distinguishing between sporadic cancer and cancer driven by inherited variation is important both for treatment of the individual as well as for informing relatives who may carry the same inherited predisposition. The novel variants we discovered could increase the knowledge base of variants that predispose to cancer.

A large portion of population-level variation in MSI is not easily explained by germline, somatic, or epigenetic alteration in DDR genes. This could be due to our modeling approach, our strict criteria for defining damaging events, copy number events we did not analyze, measurement error in the evaluation of the MSI phenotype, or the limited focus on DDR genes. Despite these constraints, we successfully identified a novel association between methylation of *SHPRH* and somatic MSI burden, with a particularly strong effect in uterine cancer where *SHPRH* methylated individuals exhibit a 2.4 fold increase in somatic MSI burden. This finding is particularly interesting as outside of *MLH1*, there is little evidence of other epigenetic alterations associated with somatic MSI burden [173, 174]. Knockdown of *SHPRH* in yeast has previously been shown to increase DNA breaks and genomic instability [175]. To our knowledge, *SHPRH* has not been directly associated with MSI and therefore should motivate further biological validation of this result.

The lack of significant GSEA hits from the exome-wide bi-allelic alteration analysis suggests that there are few novel genes to be found using TCGA that fit the two-hit inactivation model proposed by Nording and Knudson [6, 7]. However, we show how even with the same raw data, differences in methodology used to determine bi-allelic alteration, specifically calling LOH events in *BRCA1/2*, can lead to conflicting results. Therefore, it

is possible that more sophisticated methods may discover novel genes frequently affected by bi-allelic alteration. Outside of bi-allelic alteration, we find that mono-allelic damaging germline variation has little effect on somatic MSI burden. This is not entirely surprising, as there is conflicting evidence on the effect of MMR haploinsufficiency on mutation rates [71, 113]. Using the effect size of known pathogenic MMR variants, we performed a power calculation and estimated that 11,482 individuals (6,485 more than our analysis) would be required to detect the association between mono-allelic damaging germline MMR variants and somatic MSI (see methods). We further found no significant association between mono-allelic damaging germline variants and somatic mutational signatures. Our analysis suggests that the contribution of mono-allelic germline variation to somatic mutational phenotypes is likely to be small.

In addition to individuals with potential Lynch syndrome, we identified individuals who carry germline variants that reportedly predispose to Li-Fraumeni spectrum cancers as well as pheochromocytoma and paraganglioma. While the number of individuals who carry these variants is small, in some cases their phenotype is extreme enough to confound analyses, as we saw with somatic MSI (Supplementary Figure B.8B and Table B.9). It is important that studies using TCGA as a sporadic cancer control remove potential confounding cases [176]. These individuals may have escaped previous notice due to the fact that many did not develop the cancer type expected based on their germline predisposition. This confirms what is known about predisposition syndromes: that a variant can predispose to one cancer type but have no significant effect on the course of disease of another cancer type [17]. Variable penetrance could explain why some individuals will not acquire the cancer type they are predisposed toward, but 'bad luck' or environmental exposures may lead them to develop another sporadic cancer [13, 12].

The goal of this study was to assess the ability of germline mono-allelic and germline and somatic combined bi-allelic alterations to alter somatic molecular phenotypes. We

observed that combined germline and somatic alteration of MMR genes had a synergistic effect on somatic MSI burden, but germline alteration alone showed no effect. We later showed that germline variation in known cancer predisposition genes only led to an earlier age of diagnosis only in a subset of cancer types. From these observations, we conclude that germline variation has the ability to influence both somatic phenotypes and cancer development, but often this ability is dependent on other somatic alterations or tissue type specific processes. Our work highlights the importance of integrating germline and somatic data to identify bi-allelic alterations when testing for associations between germline variants and somatic phenotypes.

In this study we intended to characterize sporadic adult-onset cancers, but in the course of our analyses, we identified individuals that likely have rare cancer predisposition syndromes. Our results and observations shed important light on the issue of incidental findings, not just in the TCGA, but with any dataset with paired germline variant and phenotype data. We have taken care to be sensitive in our reporting of the data for patient privacy and followed precedents set by others using the TCGA germline data. We believe it will be important moving forward to have a set standard for reporting germline variation, especially given the recent surge of interest in germline variation in cancer.

3.6 Acknowledgements

Chapter 3 is under review for publication in *BMC Genome Medicine* under the title, "Exome-Wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in the Cancer Genome Atlas". Authors included Alexandra R. Buckley, Trey Ideker, Hannah Carter, Olivier Harismendy, and Nicholas J. Schork. NJS designed and supervised the research. NJS designed and supervised the research. ARB performed the statistical analysis, prepared the figures and tables, and drafted the manuscript. ARB, OH, and HC designed

experiments. NJS, OH, and HC assisted writing the manuscript. TI set up high performance computing infrastructure. All authors read and approved the final manuscript. The dissertation author was the primary researcher and author on this manuscript. ARB is supported in part by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number T32GM008666 and as a TGen scholar at the University of California, San Diego. NJS and his lab are also supported in part by National Institutes of Health Grants UL1TR001442 (CTSA), U24AG051129, and U19G023122. (note that the content of this manuscript is solely the responsibility of the authors and does not necessarily represent the official views of the NIH). All computing was done using the National Resource for Network Biology (NRNB) P41 GM103504. All primary data were accessed from The Cancer Genome Atlas Research Network (cancergenome.nih.gov). We would like to thank Bethany Buckley for her assistance in obtaining ClinVar annotations and interpreting germline variants, and Barry Demchak for his assistance with managing data and setting up analysis pipelines on NRNB.

Chapter 4

Rare Variant Phasing Using Paired Tumor:Normal Sequence Data

4.1 Abstract

Background: In standard high throughput sequencing analysis, genetic variants are not assigned to a homologous chromosome of origin. Haplotype phasing can reveal information important for understanding the relationship between genetic variants and biological phenotypes. For example, in genes that carry multiple heterozygous missense variants, phasing resolves whether one or both gene copies are altered. Here, we present a novel approach to phasing variants that takes advantage of unique properties of paired tumor:normal sequencing data from cancer studies.

Results: VAF phasing uses changes in variant allele frequency (VAF) between tumor and normal data in regions of somatic chromosomal gain or loss to phase germline variants. We apply multiple phasing methods to 6,180 samples from the Cancer Genome Atlas (TCGA) and demonstrate that VAF phasing is highly concordant with other standard phasing methods, and can phase an average of 33% more variants than other read-

backed phasing methods. Using variant annotation tools designed to score gene haplotypes, we find a suggestive association between carrying multiple missense variants in a single copy of a cancer predisposition gene and earlier age of cancer diagnosis.

Conclusions: VAF phasing exploits unique properties of tumor genetics and paired tumor:normal sequence data to increase the number of germline variants that can be phased over standard read-backed methods in paired tumor:normal samples. Our phase-informed association testing results call attention to the need to develop more tools for assessing the joint effect of multiple genetic variants.

4.2 Background

Humans have two copies of every chromosome, one inherited maternally and the other paternally. Studying genetic variants in the context of their haplotype, termed diplomics, can yield important biological insights [112]. Assigning genetic variants to their homologous chromosome of origin is called phasing. There are three main strategies for phasing unrelated individuals using next generation sequencing (NGS) data: population-based, which relies on population linkage disequilibrium structure, laboratory-based, which relies on physical isolation of homologous chromosome segments, and read-backed, which relies on paired-end sequencing reads that span multiple heterozygous loci [177, 112]. Each method comes with a cost: population-based methods perform poorly on rare and *de novo* variants and at phasing distances greater than a haplotype block, laboratory-based methods require sample preparation which can be costly or impractical depending on the source of input DNA, and read-backed methods generally can only phase a fraction of possible variants at distances limited by read and insert size.

Here we present VAF phasing, a method that uses changes in variant allele frequency (VAF) between paired tumor and normal samples in regions of somatic chromosomal copy

loss or gain to phase germline variants. Similar to read-backed approaches, VAF phasing is unaffected by population allele frequency and can be run on existing NGS data. Unlike read-backed approaches, VAF phasing is not limited by read and insert size, and can phase over long distances including whole chromosomes. VAF phasing is limited to regions of somatic copy number alteration (SCNA); however, SCNAs are widespread in cancer and approximately 90% of solid tumors exhibit some degree of aneuploidy [80]. The concept of using allelic ratios of heterozygous germline variants to infer somatic copy number changes is the basis for many SCNA detection algorithms [178, 179]. However, using this data to infer the phase of germline variants has not been widely implemented. While a similar method of using SCNAs to phase variants exists, VAF phasing is a more simple approach that can be run without training data using only read counts from paired tumor:normal NGS [180].

There is growing interest in the role of germline variation in increasing cancer risk and influencing molecular tumor phenotypes [77, 20, 65, 69]. Many cancer-relevant genes, such as DNA damage repair genes, are large and often contain multiple missense variants [45, 17]. Phasing damaging heterozygous variants in these genes is important to determine whether an individual carries variants in both homologous copies of the gene, termed compound heterozygosity, or carries multiple variants in a single copy. The biological consequences of compound heterozygosity is exemplified by cancer predisposition syndromes involving deleterious germline alteration of the mismatch repair (MMR) genes [10]. Germline compound heterozygosity of a MMR gene is associated with bi-allelic mismatch repair deficiency (bMMRD) and childhood onset cancer [114, 181], whereas mono-allelic germline alteration is associated with Lynch syndrome and adult onset cancer [113]. In the event a gene harbors multiple missense variants in a single copy, it is possible the combined effect of these variants on protein structure and function is different than the predicted effect of each variant independently. Further, non-coding variants can act

as eQTLs and alter expression of a single gene copy [157, 93]. Resolving which gene copy is under regulation by proximal eQTLs can provide important information, particularly if one gene carries inactivating or dominant negative alleles [112]. Therefore, resolving the phase of both coding and non-coding variants in a gene region can provide important insight into the biological consequences of germline variation.

We apply VAF phasing to 6,180 whole exome sequencing (WXS) samples from the Cancer Genome Atlas (TCGA), and benchmark VAF phasing against two read-backed methods: HapCUT2 and phASER, one population-based method: SHAPEIT, and one laboratory-based method: 10X Genomics sequencing [182, 183, 184, 185]. VAF phasing is highly concordant with all phasing methods assessed up to at distances of 10 Mb. We demonstrate the value of phase information by testing for association between germline variation in cancer predisposition genes and age of cancer diagnosis. We find suggestive evidence that carrying sets of non-compensatory missense variants in the same gene copy is associated with an earlier age of cancer diagnosis.

4.3 Methods

4.3.1 Data Acquisition

Approval for access to TCGA case sequence and clinical data were obtained from the database of Genotypes and Phenotypes (project #8072: Integrated analysis of germline and somatic perturbation as it relates to tumor phenotypes). WXS germline variant calls from 8,542 individuals were obtained using GATK v3.5 as described previously [117]. Samples prepared using whole genome amplification (WGA) were excluded from analysis due to previous identification of technical artifacts in both somatic and germline variant calls in WGA samples [117]. Raw somatic WXS sequence data and somatic RNA-seq data was downloaded from the legacy archive of the genomic data commons (GDC) in

BAM file format aligned to the hg19 reference genome [134]. Segmented SNP6 array data were downloaded from Broad Firehose (release stddata__2016_01_28, file extension: segmented_scna_hg19). Aggregate allele frequencies and allele frequencies in 7 ancestry groups (African, Admixed American, East Asian, Finnish, non-Finnish European, South Asian, and other) were obtained from ExAC v3.01 [76]. Clinical biospecimen histology slide data for tumor purity measurements was downloaded from GDC.

4.3.2 Variant Annotation and Filtering

Raw variant calls were filtered using GATK VQSR TS 99.5 for SNVs and TS 95.0 for indels. Putative germline loss-of-function (LOF) variants were identified using the LOFTEE plugin for VEP and Ensembl release 85 [130]. Only germline LOF variants with an AF < 0.05 in all ancestry groups represented in ExAC were used in the age of diagnosis association analyses. Gene, CADD score, and ClinVar annotations were obtained using ANNOVAR and ClinVar database v.20170905[158]. A germline variant was determined to be pathogenic using ClinVar annotations if at least half of the contributing sources listed the variant "Pathogenic" or "Likely Pathogenic".

4.3.3 Implementation of VAF Phasing

Somatic reference and alternate read counts for germline variants were obtained from the germline VCF and somatic BAM files using samtools mpileup v1.3.1 (SNPs) or varscan v2.3.9 (indels) [139, 135]. Germline variants not present in the somatic sequence data were excluded from further analysis. A two-way Fisher's exact test comparing reference and alternate read counts was performed on variants within significant segments to test for deviation in VAF between the normal and tumor sample. Only sites with a nominally significant ($p < 0.05$) change in VAF between tumor and normal sample were considered for phasing. Circular binary segmentation (CBS) was performed on absolute Δ

VAF values, calculated as $\text{abs}(\text{somatic VAF} - \text{germline VAF})$, of all heterozygous germline variants using the R package 'PSCBS', a process we refer to as VAF-CBS [186]. Smoothing of gaps between heterozygous sites was implemented using the function 'findLargeGaps' and setting 'minLength' to the values of 0.5, 1, 2, or 3 Mb. For all segments containing > 1 variants, the mean absolute Δ VAF of all variants was calculated. Significant segments were determined either directly using a region-specific or a hard cutoff null model. In the region-specific model, the absolute Δ VAF of each segment identified by VAF-CBS is compared to the absolute Δ VAF of the same genomic region in $n = 416$ paired normal replicate samples. Segments with an absolute Δ VAF in the 90th percentile were considered significant and to represent true SCNA. In the cutoff model, a hard cutoff of absolute Δ VAF ≥ 0.14 was used to identify significant segments. Within each segment, significant variants were assigned to a chromosome of origin using the sign of Δ VAF, such that all variants with an increasing VAF are assigned to one chromosome and all variants with a decreasing VAF are assigned to the other. For analyses using TCGA SCNA calls, segments with an estimated fold change value < 0.9 or > 1.1 , which corresponds to a single chromosome loss or gain in 20% of tumor cells, were considered significant. VAF phasing was applied to a total of 6,180 TCGA samples with tumor WXS, normal WXS, somatic RNA-seq, and evidence of SCNA burden > 0 .

4.3.4 Comparison To Other Phasing Methods

HapCUT2 was run with default parameters using germline WXS BAM files from GDC and single sample VCFs of germline variant calls generated as described previously [184]. PhASER was run with the parameters `-mapq 255`, `-baseq 10`, and `-paired_end 1` [182]. The HLA region was blacklisted with the `-blacklist` option and indels were excluded from analysis. Phaser was run on somatic RNA-seq BAM files and single sample germline VCFs. For SHAPEIT phasing, the germline VCF from the full cohort of 8,542 individ-

uals from TCGA was converted to PLINK bed format, excluding multiallelic sites [183]. SHAPEIT was run with default parameters on the full cohort with the genetic HapMap phase II recombination map provided by SHAPEIT specified with the -M parameter.

To determine overall discordance between two methods, phase blocks in common between both methods were found. Within a common block, the number of variants with disagreeing phase orientation by the two methods as well as the total number of variants phased in common were counted. Discordance was calculated as: $1 - (\text{the number of concordant phased variants} / \text{number of phased variants in common})$ (Supplemental Figure C.11). To obtain pairwise discordance and features of individual phase pairs, all unique pairwise combinations of variants were identified within each common phase block. Pairs with disagreeing phase orientation were considered discordant (Supplemental Figure C.11). Additional features calculated for each phase pairs were: Minimum Read Depth = lowest read depth of phase pair, Segment Size = size of VAF-CBS segment in base pairs, Segment Abs. Δ VAF = absolute Δ VAF of the VAF-CBS segment, $\Delta \Delta$ VAF = difference in Δ VAF between the phase pair variants, Pair Distance = distance between phase pair variants in base pairs, Δ Allele Frequency = difference in allele frequency between phase pair variants, Minimum Allele Frequency = lowest allele frequency of phase pair variants.

4.3.5 HMMvar Annotation and Compound Heterozygosity Analysis

HMMvar v.1.1.0 was used to jointly assess the functional effect of multiple cis-phased nonsynonymous variants [115]. HMMvar is a method that uses a hidden markov model computed from multiple sequence alignment of homologous proteins to predict the effect of multiple nonsynonymous coding variants in a gene based on amino acid conservation of the variant set. For each gene and each individual, a gene variant set was constructed using phased heterozygous and homozygous nonsynonymous variants. For

each gene, the RefSeq standard transcript or the longest coding transcript was used to calculate HMMvar scores. HMMvar scores were calculated for individual variants and for variant sets. Compensatory variant sets were defined as those with a set score $\leq \min(\text{individual variant scores}) - 1.5 * (\max(\text{individual variant scores}) - \min(\text{individual scores}))$, non-compensatory variant sets were defined as those with a set score $\geq \max(\text{individual variant scores}) + 1.5 * (\max(\text{individual variant scores}) - \min(\text{individual scores}))$.

For identifying compound heterozygosity events, variants with a CADD score ≥ 15 were considered damaging. Compound heterozygosity events were defined at the gene level as possessing two damaging variants in trans configuration (one variant in each copy). Cis damaging events were defined as possessing two damaging variants in cis configuration (two variants in one copy).

4.3.6 Statistical Analyses

Principal Component Analysis (PCA) was performed on common ($AF > 0.01$) germline variants using PLINK v1.90b3.29 and the first two principal components obtained from this analysis were used to control for ancestry in all of the regression models we fit to the data [127]. To test for association between germline alteration and age of diagnosis a linear model of the form $A \sim G_{ij} + X_i$ was used where A denotes age of diagnosis, G_{ij} , is a binary indicator for germline alteration status of gene j in sample i , and X_i represents a vector of covariates for sample i (cancer type, PC1, PC2).

4.4 Results

4.4.1 Phasing with Variant Allele Frequency

In regions of SCNA sequencing reads will be skewed toward the homologous chromosome that is amplified, or in the case of deletions, the chromosome that is retained

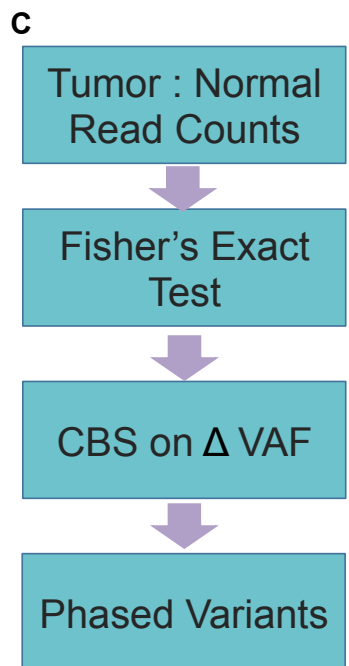
(Supplementary Figure C.1). It follows that the VAF of heterozygous germline variants in the tumor will also deviate from the expected value of 0.5 dependent on the homologous chromosome of origin. The difference in VAF between tumor and normal sequencing samples, which we refer to as Δ VAF, can be used to phase germline variants in regions of SCNA (Figure 4.1A). Germline variants that lie on the same homologous chromosome (*cis* phase) will have Δ VAF values of similar magnitude and direction, whereas variants that lie on opposite homologous chromosomes (*trans* phase) will have Δ VAF values of similar magnitude but opposite direction. Importantly, phasing with Δ VAF requires identifying regions of contiguous SCNA. Therefore, the VAF phasing method has two steps: 1) determine which heterozygous germline variants have significantly deviant Δ VAF to confidently phase and 2) identify the coordinates of SCNAs (Figure 4.1B).

To identify variants with deviant Δ VAF, a Fisher's exact test was performed for each germline heterozygous variant comparing reference and alternate read counts between tumor and normal samples. Only variants with a nominally significant p-value were considered for phasing. A number of methods exist to detect SCNAs from SNP array or NGS data, many of which use differences in signal intensity or read depth between normal and tumor samples to identify SCNA regions [179, 187]. Similarly, we reasoned that absolute Δ VAF could be used to identify SCNAs, as within a single SCNA the absolute Δ VAF of heterozygous germline variants should be contiguous and of a similar magnitude (Supplementary Figure C.2). While this approach does not distinguish amplifications from deletions, for the purposes of phasing germline variants only the coordinates of SCNAs are of interest. We applied circular binary segmentation (CBS), a method to partition the genome into segments with similar values, using absolute Δ VAF as input, a method we refer to as VAF-CBS [186]. While any SCNA calling method could be used to identify SCNAs coordinates for VAF phasing, we sought to provide a method that could be run entirely on paired tumor:normal reference and alternate read count data.

Figure 4.1: Overview of VAF Phasing Method. (A) The left panel illustrates two heterozygous germline SNVs in trans phase with the chromosome carrying SNV1 somatically amplified. In the normal sample, both SNVs have a VAF of 0.5. In the tumor sample, SNV1 is overrepresented in the sequence data (VAF = 0.75) and SNV2 is underrepresented (VAF = 0.25). The difference in VAF between the tumor and normal sample, which we refer to as Δ VAF, indicates that the VAF of SNV1 is increased (Δ VAF = 0.25) and that the VAF of SNV2 is decreased (Δ VAF = -0.25) in the somatic sample. For a pair of variants, somatic changes VAF in opposite directions suggest that the variants lie on different homologous chromosomes. (B) The right panel illustrates two heterozygous germline SNVs in cis phase with the chromosome carrying both SNV1 and SNV2 somatically amplified. In this case, both variants have an increased VAF (Δ VAF = 0.25). For a pair of variants, somatic changes VAF in the same direction suggest that the variants lie on the same homologous chromosome. (C) The VAF phasing pipeline has two steps: a Fisher’s exact test to identify sites with significant Δ VAF, and circular binary segmentation (CBS) on Δ VAF values to identify SCNA regions.

A = SNV 1
G = SNV 2

	A Normal		A Tumor		B Normal		B Tumor	
Copy Number								
Sequencing Reads								
Reference Reads	2	2	1	3	2	2	1	1
Alternate Reads	2	2	3	1	2	2	3	3
VAF	0.5	0.5	0.75	0.25	0.5	0.5	0.75	0.75
Δ VAF	$A = 0.75 - 0.5 = 0.25$ $G = 0.25 - 0.5 = -0.25$				$A = 0.75 - 0.5 = 0.25$ $G = 0.25 - 0.5 = 0.25$			
Phase	A G Trans Phase				A G Cis Phase			



Identifying SCNA breakpoints using WXS data is difficult due to the sparse coverage of the genome, and this problem is exacerbated when only using heterozygous variants as informative data points [179]. In an effort to account for this known difficulty, we tested multiple values of a smoothing parameter that allows the CBS algorithm to join distant data points with similar Δ VAF values (see methods, Supplementary Figure C.3). Increasing the smoothing distance resulted in longer predicted SCNA segments (Supplementary Figure C.3). This allows for longer range phasing, however, it also carries the risk of missing SCNA breakpoints in regions not covered by exome capture. Increased smoothing distance also resulted in fewer segments carrying a single heterozygous variant and more variants able to be phased overall (Supplementary Figure C.4). To balance the assumptions made by smoothing with the increased phasing capacity, we used a value of 1 Mb for future analyses.

Changes in VAF between normal and tumor samples may be due to a biased read sampling or low read depth, not a physical change in chromosomal copy number in the tumor. To determine a threshold to identify true SCNA segments above background noise, we utilized duplicated normal WXS samples. A subset of individuals in TCGA have multiple normal WXS samples, typically a blood and normal tissue sample. As there should be no CNAs in duplicated normal samples from the same individual, we used these samples to derive a null distribution of read sampling noise (Figure 4.2A). Interestingly, we identified seven duplicated normal samples with strong evidence of CNAs (Supplementary Figure C.5). Given that the CNA regions observed in paired normal:normal samples were also observed in paired tumor:normal samples, we suspect this observation is due to tumor contamination of normal tissue and excluded these samples from further analysis (Supplementary Table C.1). We ran VAF-CBS on duplicated normal samples from 416 individuals and observed 95% of segments identified have a mean absolute Δ VAF value < 0.14 (Figure 4.2B). Therefore, we expect using a hard cutoff mean absolute Δ VAF of

0.14 to call SCNA segments would result in a 5% error rate. In an alternate approach, the mean absolute Δ VAF of each segment identified by VAF-CBS was compared to the mean absolute Δ VAF of the same genomic region in the duplicated normal samples, generating a null distribution for that specific genomic region. While this method has the advantage of accounting for region-specific read sampling noise, it is likely only applicable for samples within the TCGA cohort. We refer to these methods as "hard cutoff" and "region-specific" and use the region-specific null model for future analyses.

Similarly, we used duplicated normal samples to confirm that the assumptions of the Fisher exact test were not violated. Indeed, the Fisher p-values for all heterozygous germline variants in duplicated normal samples followed the expected distribution, with a median 6% of heterozygous loci significant at a $p < 0.05$ cutoff (Figure 4.2C). In contrast, a median 17% of heterozygous loci were significant in paired tumor:normal samples (Figure 4.2D,E). By requiring that a variant both have a nominally significant p-value and be in a VAF-CBS SCNA region to be considered for phasing, we further reduce false positives due to read sampling noise. Applying VAF phasing with the hard cutoff null model to the duplicated normal samples, we observe only 0.3% of variants erroneously meet criteria for phasing.

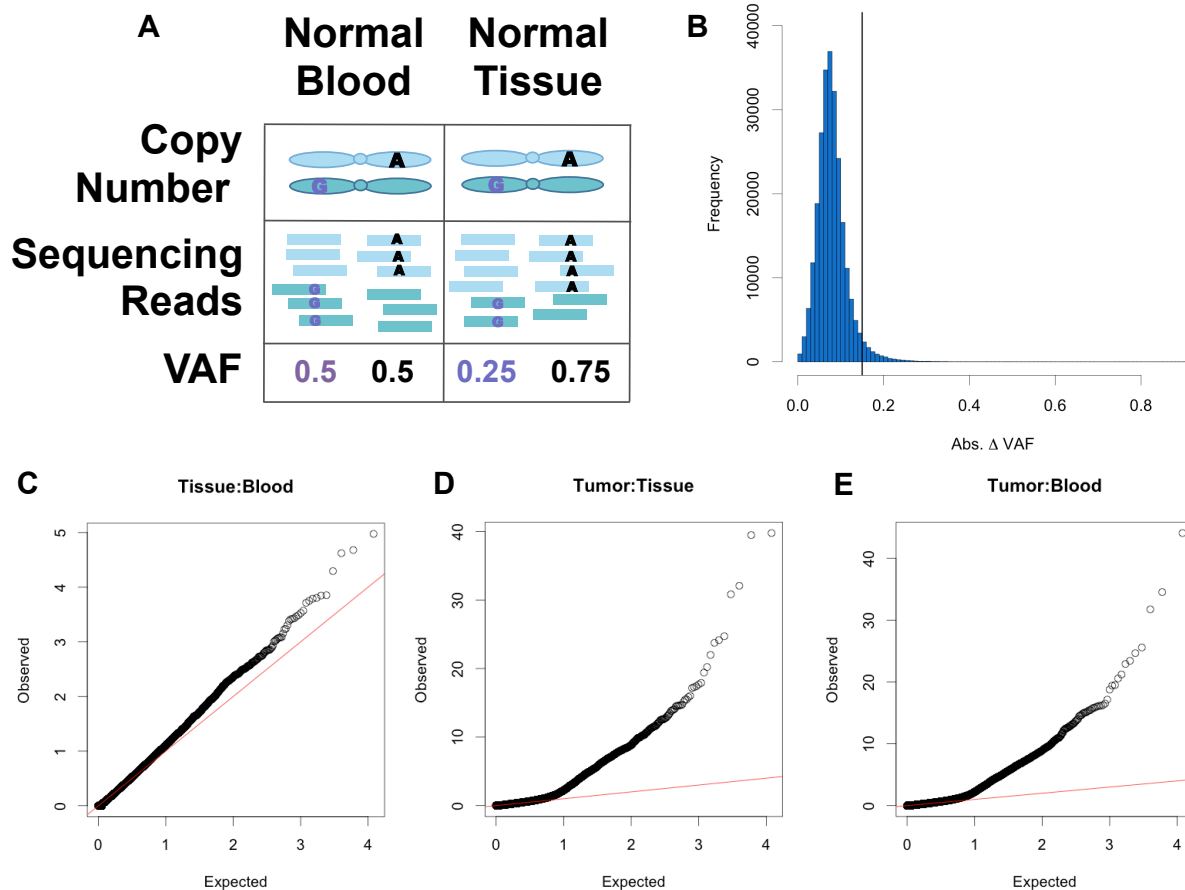


Figure 4.2: Using Duplicated Normal Samples to Identify SCNAs. (A) The expectation in diploid regions is that the VAF of heterozygous SNVs will be 0.5; however, due to read sampling error, VAF greater or less than 0.5 is frequently observed. Duplicated normal samples in TCGA can be used as a null model to estimate how often read sampling error resembles an SCNA event by chance. (B) Distribution of mean segment absolute Δ VAF for 249,471 segments identified from $n = 416$ duplicated normal samples. Segments were identified using VAF-CBS with a smoothing parameter of 1 Mb. The solid line represents the 95% percentile (absolute Δ VAF = 0.14). QQ plots showing p-values obtained from a Fisher's exact test on tumor and normal read counts for an example sample (A) paired tumor:normal tissue, (B) tumor:normal blood, (C) normal tissue:normal blood.

4.4.2 VAF Phasing is Concordant with Other Methods

We ran VAF phasing on 6,180 TCGA samples using a range of smoothing parameters and both region-specific and hard cutoff null model approaches for SCNA identification. As there is no gold standard phasing dataset with paired tumor:normal sequence data, we assessed accuracy of our phase calls by comparing to TCGA germline phase calls generated by HapCUT2, phASER, and SHAPEIT (see methods) [182, 183, 184]. We observed a median 99% concordance between VAF phasing and both HapCUT2 and phASER (Supplementary Table C.2). Samples with poor concordance were largely those with few variants phased in common between methods (Supplementary Figure C.6). Choice of smoothing parameter did not have a large effect on concordance, however concordance was lower when using the hard cutoff null model (Supplementary Table C.2).

There is considerable SCNA burden in TCGA samples, allowing VAF phasing to phase a median 1,276 variants per sample (Supplementary Figure C.7). The addition of VAF phasing to HapCUT2 and phASER increased the cumulative number of variants phased by 33% on average, and VAF phasing phased a median 942 variants not accessible to other methods (Figure 4.3 A,B). We observed similar results when restricting to rare variants (Supplementary Figure C.8). The number of variant phased by VAF phasing is variable between samples and across genomic regions (Supplementary Figure C.9). We performed linear regression to identify factors underlying the performance of VAF phasing and found that the number of variants phased by VAF phasing is largely determined by CNV burden and estimated tumor sample purity (Supplementary Table C.3). We compared the performance of VAF phasing using SCNA calls derived from VAF-CBS vs. SCNA calls from TCGA SNP6 array data. A median 63% of variants were phased using both methods of SCNA identification (Supplementary Figure C.10). However, the variants uniquely phased using VAF-CBS had higher concordance with HapCUT2 and phASER, suggesting that VAF-CBS provides more accurate phase calls (Supplementary Table C.4).

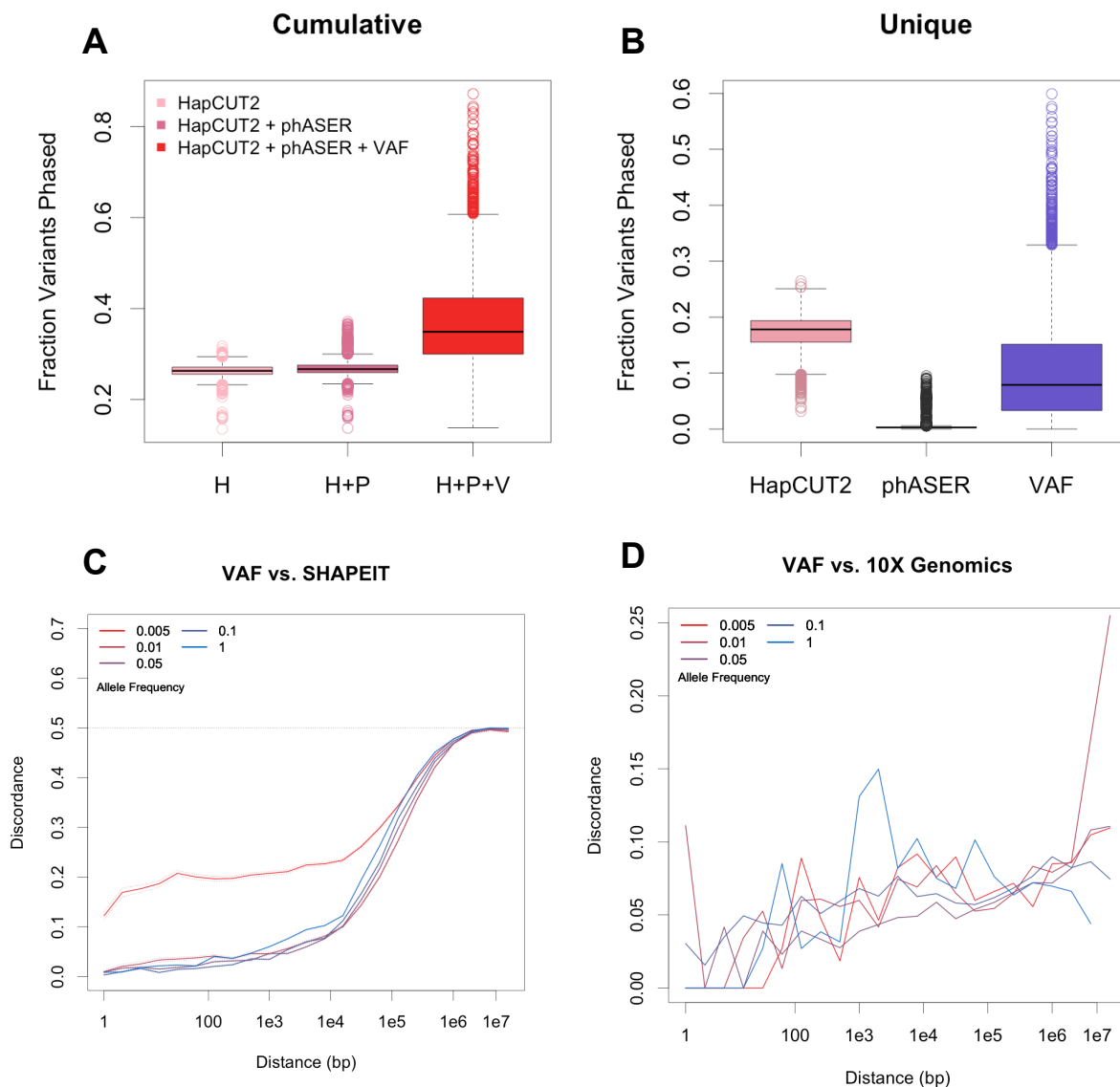


Figure 4.3: Comparison of Phasing Methods. Comparison of VAF phasing to read backed, population based, and laboratory phasing methods. (A) The fraction of germline heterozygous variants phased by HapCUT2 alone, HapCUT2 and phASER, and by HapCUT2, phASER, and VAF in $n = 6,180$ samples. (B) The fraction of germline variants phased that are unique to each method. (C) Pairwise discordance between VAF phasing and SHAPEIT for $n = 6,263$ samples as a function of distance and allele frequency. Pairs of variants were binned according to distance between the variants in base pairs and binned according to minimum allele frequency of the variant pair. Colors represent allele frequency bins. Solid lines represent the mean discordance, dotted lines are mean \pm 2 s.e.m. (D) Pairwise discordance between VAF phasing and 10X Genomics phasing for the COLO829 cell line as a function of distance and allele frequency.

We measured long range phasing performance using SHAPEIT phase calls and a pairwise approach to measuring phase accuracy (see methods, Supplementary Figure C.11). VAF phasing and SHAPEIT are highly concordant up to approximately 10kb (Figure 4.3C). At distances larger than 10kb discordance between VAF phasing and SHAPEIT sharply increases, likely due to the fact that median haplotype block size in humans is 45 kb [188]. As expected, discordance was also higher for very rare and singleton variants, which are not amenable to phasing using population-based methods. To validate VAF phasing in a separate dataset, we performed 10X Genomics phasing on COLO829, a tumor:normal pair of cell lines generated from a melanoma patient [189, 185]. Overall concordance between VAF phasing and 10X Genomics phasing was 96% (Supplementary Table C.5). Pairwise discordance was largely unaffected by allele frequency and distance up to approximately 10 Mb (Figure 4.3D). Supporting our previous finding that choice of VAF-CBS smoothing parameter doesn't significantly impact phasing accuracy, we observe that discordance between VAF phasing and 10X Genomics phasing is similar for smoothing values between 0.5-2 Mb (Supplementary Figure C.12, Supplementary Table C.5). Finally, to assess what genomic features of variants are associated with VAF phasing errors, we examined all phased variant pairs in the COLO829 sample. We observe that read depth and magnitude of Δ VAF change between variants are the features most strongly associated with phase errors (Supplementary Table C.6).

4.4.3 Application of VAF Phasing to Cancer Predisposition

In genes that carry multiple variants, phase information disambiguates whether a single copy or both copies of the gene are altered. To demonstrate the value of phasing for biological analysis, we performed a phase-informed analysis relating germline variants in a set of 114 cancer predisposition genes to age of cancer diagnosis [17]. We first identified compound heterozygosity events, which we defined as carrying a variant with a CADD

score ≥ 15 in both copies of a gene [160]. Using all read backed phasing methods combined, we were able to phase resolve 50% of all possible compound heterozygosity events exome-wide, and identified a total of 54,284 compound heterozygosity events in 4,873 genes (Supplementary Figure C.13). As we found few compound heterozygosity events for any single gene, we categorized individuals into four hierarchical mutually exclusive groups based on type of alteration in the predisposition gene set: those carrying a compound heterozygosity event (Trans), those with two or more phased CADD damaging variants in the same gene copy (Cis), those carrying mono-allelic ClinVar pathogenic or loss-of-function variants (ClinVar/LOF), and those carrying mono-allelic CADD damaging variants (CADD). We found no association between carrying a compound heterozygosity event in a cancer predisposition gene and age of cancer diagnosis (Figure 4.4A, Supplementary Table C.7).

We next asked whether carrying multiple missense variants in a single gene copy may be more deleterious than carrying a single variant. Variant scoring tools such as CADD scores are not designed to address this question, as they score variants independently [160]. Instead we used HMMvar, a variant scoring tool that assesses the collective effect of multiple missense variants and identifies sets of variants that collectively have a different score than expected based on single variant scores [115]. HMMvar identifies both compensatory variant sets, which collectively are less damaging than independently, and non-compensatory variant sets, which collectively are more damaging than independently. Similar to the previous analysis, we categorized individuals into four hierarchical mutually exclusive groups based on type of alteration in the predisposition gene set: those carrying a non-compensatory variant set (Non-Compensatory), those carrying a compensatory variant set (Compensatory), those carrying mono-allelic ClinVar pathogenic or loss-of-function variants (ClinVar/LOF), and those carrying mono-allelic CADD damaging variants (CADD). We found a significant association between carrying a non-compensatory

variant set in a predisposition gene and an earlier age of cancer diagnosis (Figure 4.4B, Supplementary Table C.8). However, this may be in part due to six individuals who carry both a non-compensatory variant set and a ClinVar/LOF germline variant in different predisposition genes. Removing these samples reduces this association below nominal significance (Supplementary Figure C.14, Table C.9). *BRCA1/2* is one of the most frequently studied cancer predisposition genes [190, 10, 15]. Limiting analysis to *BRCA1/2* identified three non-compensatory variant sets and a suggestive, but not significant, association between carrying a *BRCA1/2* non-compensatory variant set and earlier age of diagnosis (Supplementary Figure C.15, Table C.10). Interestingly, all variants in the predicted *BRCA1/2* non-compensatory variant sets are individually predicted to be benign in ClinVar (Supplementary Table C.11). Our results are suggestive that multiple missense variants that appear benign based on individual variant scores may collectively contribute to cancer predisposition.

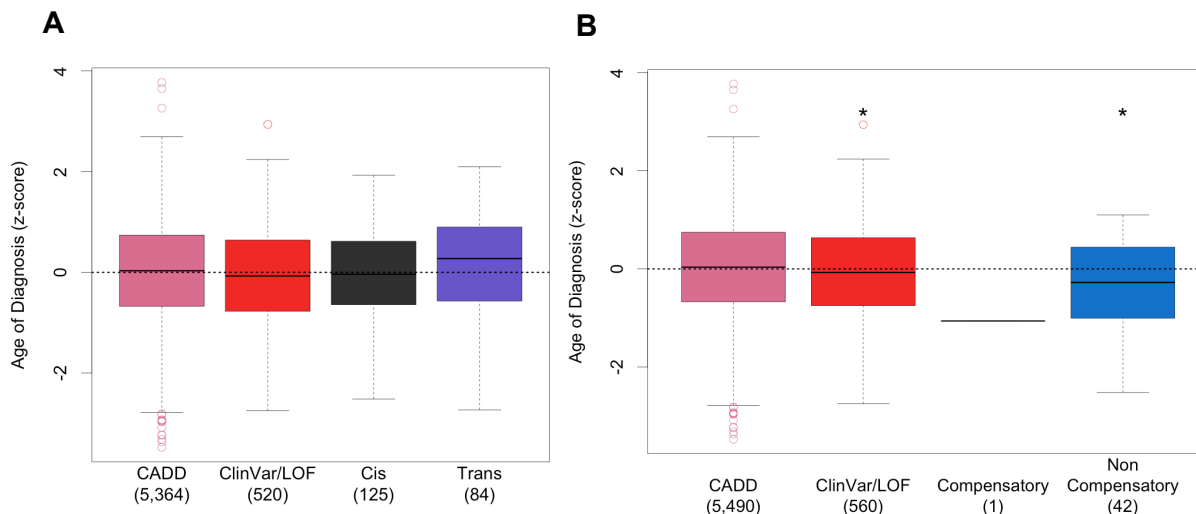


Figure 4.4: Leveraging Phase to Identify Cancer Predisposing Germline Variation. Association between germline compound heterozygosity events and non-compensatory cis variants and age of diagnosis. (A-B) Age of cancer diagnosis Z-score in $n = 6,093$ TCGA individuals grouped by type of germline alteration in a set of 144 cancer predisposition genes. (A) Individuals were grouped into four hierarchical mutually exclusive groups based on compound heterozygosity status: those carrying a compound heterozygosity event (Trans), those with two or more phased CADD damaging variants in the same gene copy (Cis), those carrying mono-allelic ClinVar pathogenic or loss-of-function variants (ClinVar/LOF), and those carrying mono-allelic CADD damaging variants (CADD). (B) Individuals were grouped using HMMvar cis variant scores: those carrying a non-compensatory variant set (Non-Compensatory), those carrying a compensatory variant set (Compensatory), those carrying mono-allelic ClinVar pathogenic or loss-of-function variants (ClinVar/LOF), and those carrying mono-allelic CADD damaging variants (CADD). The number of samples is shown in parentheses. * = $p < 0.05$; p-values were determined using a linear model to predict age of diagnosis while accounting for cancer type.

4.5 Discussion

A major assumption of VAF phasing is that SCNAs detected using VAF-CBS originate from a single germline homologous chromosome. This assumption is based off previous work showing somatic amplifications are predominantly mono-allelic [180, 191]. If VAF-CBS SCNA calls represent copy number alterations of similar magnitude from both

homologs, phase switch errors will occur (Supplementary Figure C.16). Given our high concordance with 10X Genomics long-range phasing, we believe our mono-allelic SCNA assumption is largely valid. Due to the sparse genomic coverage of the WXS data used in this study, we primarily apply VAF phasing to phase variants within a single gene. However, VAF phasing could be applied to paired tumor:normal whole genome sequencing (WGS) data, such as the 2,800 WGS samples in PCAWG, to potentially phase up to entire chromosomes [19]. While the goal in developing VAF phasing was to create a straightforward and specific approach, our model could be improved to incorporate uncertainty and increase sensitivity. Heterozygous variants in SCNA regions with a non-significant Fisher's exact test could be given an estimated phase confidence score based on read count and population haplotype data. From a sample preparation perspective, VAF phasing could be improved by better isolation of tumor from normal tissue and with deeper sequencing depth.

VAF phasing can be used to extract more value from the numerous existing paired tumor:normal datasets. Phasing germline variants from individuals with cancer is not only of interest to understanding cancer predisposition, as we demonstrated, but also to population genetics as a whole. Phased germline variants obtained from cancer data can serve as a reference dataset of phased gene haplotypes. The human leukocyte antigen (HLA) locus is of great importance to many diseases, including autoimmune diseases, infection, and cancer [192]. The complex nature and high degree of polymorphism of this region makes phasing difficult [193]. In the TCGA samples we examined, 3,651 individuals have SCNA of chromosome 6p spanning the major histocompatibility locus (MHC) region, including 602 ethnically-diverse samples. VAF phasing could potentially be incorporated into existing HLA phasing methods to facilitate phasing of this region and increase the knowledge base of known HLA haplotypes. HLA typing of cancer patients has become increasingly important in personalized immunotherapy [194]. VAF phasing could potentially

also be leveraged for patients with chromosome 6 SCNA to better estimate individual HLA haplotypes.

Using phase information we identified compound heterozygosity events and sets of missense variants in the same homologous gene copy predicted to negatively interact. We found no significant association between carrying a compound heterozygosity event in a cancer predisposition gene and age of cancer diagnosis; however, we found that individuals carrying non-compensatory missense variant sets had a significantly earlier age of diagnosis. While it seems counterintuitive that alteration of both gene copies is less deleterious than alteration of a single copy, it's likely that our definition of compound heterozygosity included missense variants that don't fully disrupt gene function. Using a more strict threshold to identify damaging variants, we observe few individuals carrying compound heterozygosity events, presumably because dual inactivation of cancer predisposing genes would result in childhood onset cancer [114]. We noted that different predisposition genes were preferentially affected by compound heterozygosity vs. those affected by non-compensatory missense events, which may be in part due to selection against dual alteration of specific genes key for survival. Further, using VAF phasing we are unable to resolve all possible compound heterozygosity events, thus we likely underestimated of the effect of compound heterozygosity.

We identified 42 missense variant sets in 20 predisposition genes predicted to have a non-compensatory effect on protein function. We investigated non-compensatory *BRCA1/2* missense variants in detail and noted that all were independently annotated as benign in ClinVar. This could indicate that there is a negative interaction between variants or that some of the variants are miss-annotated as benign in ClinVar. While there are a tremendous number of tools aimed at predicting the functional effect of individual missense variants, few methods exist that predict the effect of multiple missense variants simultaneously [195, 160, 196, 115]. There is some evidence in cardiovascular disorders

that multiple missense variants in a gene are more deleterious than single variants [197]. However, as it is not routine to assess the potential for negative interactions between multiple missense variants in a gene, the importance is likely underestimated. High throughput in vitro assays have been used to predict the effect of 2,000 amino acid substitutions on *BRCA1* E3 ubiquitin ligase activity [198]. Similar approaches could be used to assess the joint effect of multiple missense variants in a protein.

In this study we present a simple method to phase germline variants in preexisting tumor:normal sequencing datasets. We demonstrate VAF phasing is highly concordant with two read-backed and one laboratory-based phasing method, and that the addition of VAF phasing to existing read-backed methods increased the number of variants phased by an average of 33%. VAF phasing performs well on common and rare variants and at long distances, with the potential to phase up to entire chromosome lengths with WGS data. We identified individuals from TCGA that carry multiple missense variants in a single gene copy predicted to collectively be more deleterious than independently, and show that carrying one of these non-compensatory variant sets in a cancer predisposition gene is associated with an earlier age of cancer diagnosis. Our work demonstrates the biological relevance of phasing germline variants in cancer and highlights the need for better scoring tools to account for multiple variants in a single gene copy.

4.6 Acknowledgements

Chapter 4 is being prepared for publication under the title, "Rare Variant Phasing Using Paired Tumor:Normal Sequence Data". Authors included Alexandra R. Buckley, Trey Ideker, Hannah Carter, Jonathan Keats, and Nicholas J. Schork. NJS designed and supervised the research. ARB executed pipelines, analyzed data, and drafted the manuscript. ARB and HC designed experiments. NJS and HC helped write the

manuscript. TI set up high performance computing infrastructure. All authors read and approved the final manuscript. ARB is supported in part by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number T32GM008666, the Translational Genomics Research Institute (TGen), and a gift from the San Diego Cancer Research Institute. NJS and his lab are also supported in part by National Institutes of Health Grants UL1TR001442 (CTSA), U24AG051129, U19G023122, as well as a contract from the Allen Institute for Brain Science. All computing was done using the National Resource for Network Biology (NRNB) P41 GM103504. All primary data were accessed from The Cancer Genome Atlas Research Network (cancergenome.nih.gov).

Chapter 5

The Upshot

5.1 Discussion

5.1.1 Batch Effects in Public Datasets

While the TCGA is an incredible resource, it is not without problems. TCGA samples were collected over the course of five years using a variety of sample collection and preparation techniques. For example, the germline WXS samples were collected using 44 different sequencing workflows. This renders TCGA data sensitive to batch effects, as we demonstrated in chapter 2 with our discovery of a technical artifact in germline indel calls due to whole genome amplification (WGA) of DNA samples prior to sequencing. While the TCGA pan-cancer working group is aware of such issues and together with the genomic data commons (GDC) is attempting to ameliorate batch effects, to our knowledge, our paper was one of the first to draw attention to the magnitude of variability in sample preparation procedures and possible implications of batch effects in TCGA [119, 134, 117]. Despite an in-depth analysis of the characteristics and possible mechanisms underlying WGA indel artifacts, we were unable to identify a method that could satisfactorily filter out artifacts while retaining true variation. As frameshift coding indels are a major source

of pathogenic variation, this is a severe limitation on how WGA samples from TCGA can be utilized. While we only examined germline sequence data, as noted, GDC has also observed similar artifacts in the somatic mutation data [152]. Virtually all ovarian cancer samples were prepared using WGA, making deleterious germline variation in these samples difficult to interpret. Despite this, papers have been published using the TCGA ovarian cancer data [149].

In our work, we describe methods to reduce errors due to technical artifacts when working with public data. First, we describe a method of group-calling for WXS germline variants that minimizes batch effects and demonstrate the effectiveness of different variant filtering approaches provided by GATK at reducing technical artifacts [123]. Second, we describe a 'one vs. rest' approach for pan-cancer analysis. TCGA is frequently used in a 'case vs. control' study design where variant allele frequencies from TCGA are compared to a non-cancer dataset such as ExAC [76]. Using germline variant calls from another dataset only serves to introduce more heterogeneity and risk of artifactual findings. Overall, this work draws attention to the difficulty of working with heterogeneous public datasets, and demonstrates the types of errors that can occur if batch effects aren't accounted for in association studies. I hope that our work will prompt other researchers using the TCGA data to carefully assess how they utilize the germline data, and motivate more homogenous sample collection for future genomics datasets.

5.1.2 Germline Variants and Tumor Phenotypes in TCGA

Our study of the effect of germline variation on somatic phenotypes in chapter 3 led to unanticipated discoveries. The most striking finding was evidence that at least six individuals in TCGA likely have Lynch syndrome. We drew this conclusion based on the fact that these individuals have ClinVar pathogenic or predicted deleterious germline variants in known Lynch syndrome genes, secondary bi-allelic somatic mutations, earlier

age of cancer diagnosis, and elevated somatic microsatellite instability (MSI). This finding prompted us to investigate the frequency of bi-allelic alteration exome-wide. It seems probable that bi-allelic alteration would occur in sporadic cancer, perhaps not as the driver event as is seen in cancer predisposition syndromes, but as a compliment to other somatic drivers. We found bi-allelic alteration to be extremely rare, and enriched in known pathways (mismatch repair genes and *BRCA1/2*), suggesting that novel genes acting through a 'two-hit' mechanism won't be discovered using TCGA data. We went on to categorize other pathogenic germline variants and identified a total of ~60 individuals likely to have a cancer predisposition syndrome. This was surprising to me, as TCGA is often thought of, and utilized as [176], a sporadic cancer dataset.

We found no evidence that mono-allelic germline variation could alter somatic MSI or somatic mutational signatures. As detailed in the introduction, other studies have shown associations between mono-allelic germline variants and somatic phenotypes. Our lack of significant findings is likely due to a number of factors. First, our study design was limited to highly deleterious loss of function (LOF) germline variants. From preliminary testing on the known association between deleterious *BRCA1/2* germline variants and age of cancer diagnosis, we found that we were unable to detect the effects of non-LOF variants. While LOF variants are more likely to have a large effect, they are rare. It is possible that associations between mono-allelic germline variants and somatic phenotypes will be discovered by incorporating both rare and common germline variants. However, using variant scoring tools to identifying which common variants are functional is still an difficult problem. Second, it is possible that MSI is less regulated by germline variation than other somatic phenotypes. Third, it seems probable that germline variants will play a larger role in pediatric cancer, as pediatric tumors typically have fewer somatic mutations [20]. This also facilitates study of the effect of germline variants on somatic phenotypes, as fewer confounding somatic mutations need to be accounted for in the model. Finally, as

detailed in chapter 2, TCGA is heterogeneous dataset. It is possible that some true signal is lost amidst technical noise in both the germline and somatic sequence data.

In the course of this analysis we stumbled across an interesting ethical dilemma. In datasets with paired germline and phenotypic data, it is near possible to diagnosis individuals with genetic diseases, as we demonstrated with Lynch syndrome. While the TCGA data is entirely de-identified to researchers, it would be possible for the individuals who contributed their data to TCGA to identify themselves based on data reported in a publication, especially if the reported genetic disease is rare. Further, family members of TCGA participants may question whether their relative, and possibly themselves, carry genetic risk variants. We have taken great care in our reporting of our findings to anonymize the data; however, our experience draws attention to the need for set standards for reporting germline findings. With the surge of interest in germline variation in cancer, and genomics in general, it is becoming increasingly apparent that the balance between protecting patient privacy while also collecting the type of paired genotype:phenotype data needed to advance genomic medicine is a difficult problem that needs to be addressed.

5.1.3 Germline Variant Phasing in Cancer Samples

In chapter 4 we describe a novel method for phasing germline variants in cancer samples with paired tumor:normal sequence data. While a similar approach has been proposed, this approach is more complicated and requires more data to estimate parameters [180]. Further, it was not implemented on a large dataset or benchmarked against other phasing approaches. We demonstrate that VAF phasing is highly concordant with standard read-backed, population-based, and laboratory-based phasing methods, and can phase an average of 33% more variants than read-backed approaches. We propose VAF phasing be used as an add-on to cancer informatics pipelines to extract more potentially useful information out of pre-existing datasets.

With WXS sequence data, phase information allows for the discovery of compound heterozygosity events and genes that carry multiple missense variants. We used HMMvar, a variant scoring method that jointly scores multiple missense variants occurring in the same gene, to identify sets of variants that produce a more deleterious effect on gene function than would be expected from their independent scores (non-compensatory sets) [115]. We found suggestive evidence that carrying a non-compensatory germline variant set in a cancer predisposition gene is associated with an earlier age of cancer diagnosis. While the functional analyses of this study were limited, it suggests that with better tools to jointly predict the effect of multiple variants within a gene region, novel biological associations will be uncovered.

5.1.4 Tumors are Like Onions

From my extensive analysis of the TCGA genomics data, I have come to think of a tumor like an onion (Figure 5.1). At the core are germline variants, present in every cell of the body including every cell in the tumor. As the tumor develops, layers of complexity in the form of other genetic alterations, like somatic mutations, methylation, or copy number changes, are added. It is only through integrating all of these layers of genetic information that the origin, phenotype, and trajectory of a tumor will be fully understood.

In most instances, somatic genetic alterations likely have a greater impact on somatic phenotypes than germline variants. Therefore, it is important to build models that account for somatic alterations when studying the relationship between germline variants and somatic phenotypes in order to avoid misattributing significance to germline variants. Incorporating somatic alterations into these models is particularly difficult when studying a phenotype such as overall somatic mutation burden, as the probability of having any specific gene mutated will be correlated with overall mutation burden. We encountered this problem when looking for relationships between somatic frameshift mutations and

MSI burden, and attempted to reduce this correlation between predictor and response by filtering indels in microsatellite regions. Future studies would benefit from more sophisticated models that can account for 'chicken or the egg' type relationships between somatic mutations and somatic phenotypes.

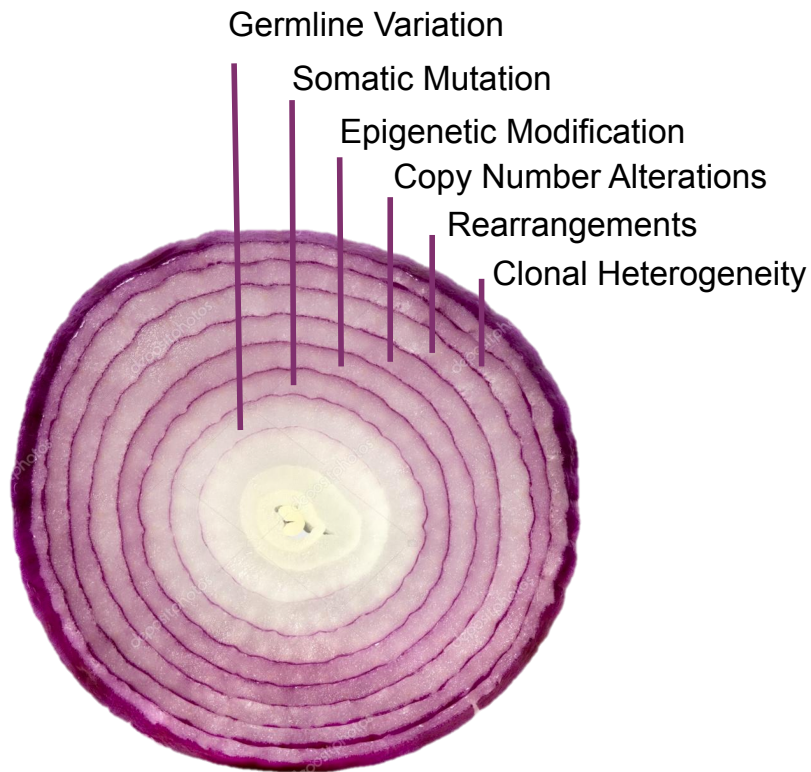


Figure 5.1: Tumors are Like Onions. Multiple layers of genomic alterations create the final phenotype of a tumor. As germline variants are present in all cells of the tumor, they represent the core of the onion. Other genetic alterations acquired during tumorigenesis add layers of complexity. Only by integrating all genomic perturbations will the etiology, phenotype, and trajectory of a tumor be fully understood.

A recent study of TCGA germline data described an association between deleterious germline variants in *PMS2* and earlier age of cancer diagnosis, and an association between variants in *MSH6* and somatic mutation burden [65]. I believe this is a motivating example of how an incorrect model can lead to misattribution of significance. In my own analysis of the TCGA, I identified an association between combined germline and somatic bi-allelic alteration of *PMS2* and *MSH6* and age of diagnosis and MSI burden, likely due to Lynch syndrome. I suspect this study did not integrate somatic and germline information to identify bi-allelic alterations. As a result, they claim there is a direct association between germline variants and somatic phenotypes, whereas I believe this association is not direct, but mediated through a secondary somatic bi-allelic alteration.

5.1.5 Germline Variation, What is it Good For?

It has recently been suggested that heritable factors play a small role in the development of sporadic cancer, and that the majority of cancer incidence can be attributed to replicative mutations that occur during cell division [12, 13]. Even in cancers that show a strong heritable component, the risk of developing cancer in a specific organ is influenced by tissue turnover rates. For example, in familial adenomatous polyposis (FAP) the risk of colorectal is much higher than the risk of duodenal cancer, which correlates with the much higher rate of cell division in the colon vs. the duodenum [13]. While this may seem to discount the importance of heritable factors in the development of cancer, this finding is not incongruent with what is known about context-dependent associations between germline variants and cancer risk. It seems likely that germline variants do contribute to risk of sporadic adult-onset cancer, but that their contribution is modulated by factors such as tissue turnover, environmental factors, and the genetic background (i.e., ancestry and polygenic profile) of an individual. Moreover, germline variants not only modulate cancer risk, but can also shape tumor phenotypes. Even if the development of cancer is

determined entirely by chance mutations and 'bad luck', other factors such as tumor progression, metastatic potential, druggability, and response to therapy may be influenced by germline variants. As we have made clear, it is quite likely that inherited genetic variation can also influence the course of tumor development and the phenotypic features exhibited by tumors.

Germline variants have been shown to influence a number of somatic phenotypes including: mutation profile, copy number profile, methylation, gene expression, immune cell infiltration, and metastatic potential. Both rare germline variants that disrupt gene function and common variants that modulate gene expression have been implicated in shaping tumor phenotypes. Often these germline variant:somatic phenotype associations depend on specific contexts, such as a secondary somatic bi-allelic mutation, the cell of origin, or tumor specific changes in transcriptional regulation. The mechanisms by which the germline can alter somatic phenotypes have been shown to be cell-intrinsic by increasing mutagenic or oncogenic potential of the cancer cells themselves, or cell-extrinsic by altering host stromal or immune cell interaction with the tumor. Many of the associations have been found in breast cancer and involved *BRCA1/2* and other HR genes. It is unclear whether germline variants in other pathways and other cancer types have a smaller effect, or if they are simply less thoroughly studied, although we believe the latter.

5.2 Future Directions

This study focused on coding variation from the TCGA WXS data. TCGA also contains roughly 2,000 paired tumor:normal whole genome sequencing (WGS) samples. It would be relatively straightforward to apply our GATK germline variant calling pipeline and our VAF phasing method to these samples to obtain phased WGS germline variants. Using WGS data, one could perform a phase-informed analysis of germline variants that

integrates eQTLs and missense variants at the gene level. Combining noncoding variants that regulate gene expression and coding variants that perturb gene function allows for the identification of genes with allele-specific expression favoring either the WT or perturbed allele. This additional information about how germline variants function as a haplotype may uncover novel mechanisms by which germline variants function in cancer. For example, certain missense germline variants may only promote a somatic phenotype when *in cis* with an eQTL that increases their expression. In addition to TCGA, other large public cancer datasets are now available. As mentioned in the previous discussion, studying the relationship between germline variants and somatic phenotypes in a pediatric cancer dataset may yield more results. It would be interesting to investigate this question in a cohort of both pediatric and adult cancers. GATK's latest pipeline makes group genotyping rapid and scalable, such that our germline variant calls from TCGA could easily be combined with germline calls from a pediatric dataset while minimizing batch effects [123].

In this study we investigated MSI and mutational signatures as somatic phenotypes that may be modulated by germline variation. There are a large number of other phenotypes that can be extracted from tumor sequencing and array data. We have discussed many in the introduction; however, I believe immune phenotypes are the most exciting direction for future research. Specifically, the type and abundance of infiltrating immune cells as well as markers of immune response. Immunotherapy is a promising new avenue of cancer treatment, and a better understanding of what tumors are vulnerable to immune attack would be immensely beneficial to patients. The recently published work on immuno phenotyping of TCGA samples identified possible relationships between genetic background and different tumor immune characteristics [99]; however, specific germline variants were not implicated. Recently work in transgenic mouse models suggests that a germline polymorphism common in the human population can regulate immune cell infiltration [100]. Together these results suggest that germline variants can shape immune

phenotypes, and novel biology remains to be discovered.

Network based stratification of tumors using somatic mutations has produced meaningful clustering of patients that can predict survival [79]. It may be possible to improve this stratification by incorporating salient germline variants into the network smoothing algorithm. As we demonstrated with pathogenic variants in *PMS2*, the same functional pathogenic alteration can occur as a somatic mutation or an inherited variant (Supplementary Figure B.5). For some genetic alterations, the need to distinguish between somatic vs. germline origin, as is currently dogma in cancer genomics, may be unnecessary. While germline variants are largely studied as mediators of cancer risk, they are genetic variants present in all tumor cells with the same potential to affect somatic phenotypes as somatic mutations. Overall, I believe better integration of the germline and somatic 'layers' of the cancer genome will be beneficial to understanding of molecular somatic phenotypes.

The analyses discussed above could be performed on pre-existing datasets. A problem with the current public cancer datasets is a relative dearth of clinical and phenotype data. For example, in TCGA many samples lack information about important clinical covariates such as smoking history and clinical outcomes such as response to treatment. As we demonstrated in our work in chapter 3, a number of patients appear to have a heritable predisposition to cancer, but the TCGA clinical data available on the samples had no record of cancer predisposition genetic testing. Some tumor phenotypes of interest, for example the degree of vascularization or metastatic spread of a tumor, are difficult to quantify from HTS data. Further, most public cancer datasets generate data from a single tumor biopsy, which will not accurately represent the full repertoire of intra-tumoral heterogeneity that exists within many cancer types. The ability to detect associations between the germline and somatic characteristics depends on accurate quantification of somatic phenotypes and important clinical covariates. Therefore, novel discoveries may require deeper phenotyping of tumor samples and better integration of medical records

into public cancer datasets.

Large-scale analyses have identified a number of genetic variants associated with variation in gene expression (eQTLs) or protein levels (pQTLs) [157]. As shown in the introduction, the ability to repair DNA in response to damaging agents is heritable [51]. Similar QTL analyses could be performed to identify the specific germline variants that underlie heritable variation in DNA damage response (DDR) activity, which could be called drQTL. Identifying drQTLs would require paired sequencing and DDR response measurements from a diverse population of individuals, similar to the GTEx project [157]. A possible source of this information would be the lymphoblastoid cell lines available from HapMap samples, which have germline variant data available and could be subjected to a high-throughput DDR assay [128]. If common drQTL are identified, DDR could serve as an intermediate phenotype for association analyses similar to TWAS [199]. With only germline variants, DDR could be imputed in cohorts of patients where DDR was not measured and used as to predict cancer risk or specific somatic phenotypes.

Monozygotic and dizygotic twins have been extensively used to determine the heritability of cancer risk [14, 16]; however, these studies generally don't include somatic phenotyping of the tumors. If germline variation does influence the growth of a tumor and the resultant somatic phenotype, cancers that arise in concordant twins should be more phenotypically similar than those arising in unrelated individuals. In other words, twin studies could also be utilized to estimate the heritability of somatic phenotypes. To our knowledge, no study has been undertaken to formally test this hypothesis. While obtaining sufficient numbers of twins concordant for cancer would be difficult, tools like GCTA could be applied to data from large public cancer datasets to estimate the heritability of somatic phenotypes [200].

5.2.1 The Upshot

While our work did not identify novel associations between germline variants and somatic phenotypes, we provide methods and suggestions for future research on germline variation in cancer. We provide a best practices approach to analyzing the germline variant data in TCGA, identified potential confounding Lynch syndrome samples in TCGA, describe an approach to modeling relationships between germline variants and somatic phenotypes that accounts for other somatic alterations, and developed a method for phasing variants in that takes advantage of unique properties of paired tumor:normal data. It has been clearly demonstrated that germline variants can influence somatic phenotypes, the question remaining for future research is to what degree and in what contexts. We believe that both creative analysis of existing datasets as well as the generation of larger, more homogenous cancer datasets will yield novel insights.

Appendix A

Supplemental Material: Pan-Cancer Analysis Reveals Technical Artifacts in TCGA Germline Variant Calls

This appendix contains supplemental figures for the work described in Chapter 2 of this document: "Pan-Cancer Analysis Reveals Technical Artifacts in TCGA Germline Variant Calls". Each figure is referred to in the main text of the chapter and a brief description of each figure is given here.

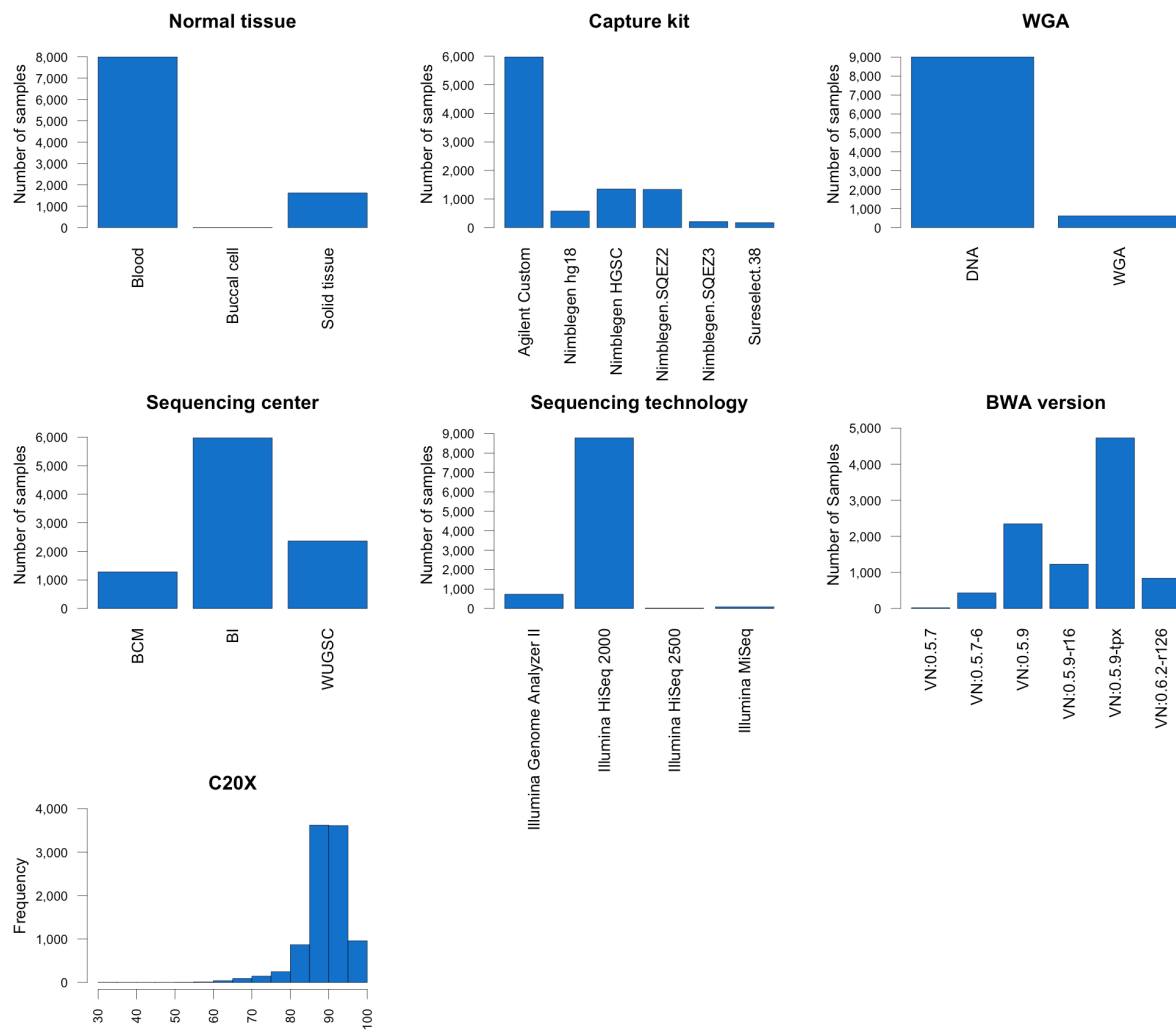


Figure A.1: Technical Covariates of Cohort. The distribution of the seven identified technical covariates for $n=9618$ TCGA WXS samples. Capture efficiency is measured as percentage of capture target area covered by at least 20 X read depth (denoted C20X).

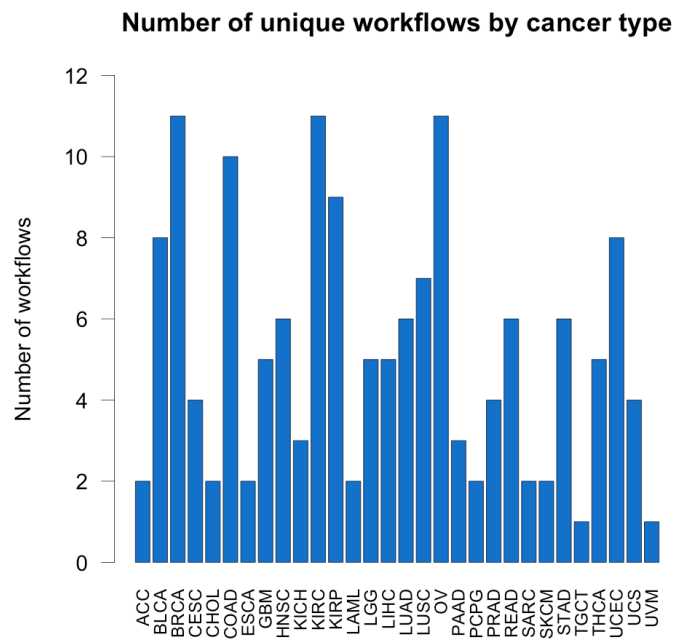


Figure A.2: Number of Processing Workflows. The number of unique combinations of six technical factors (sequencing center, normal tissue, WGA, BWA version, capture kit, and sequencing technology) per cancer type.

Discordance between alignment pipelines

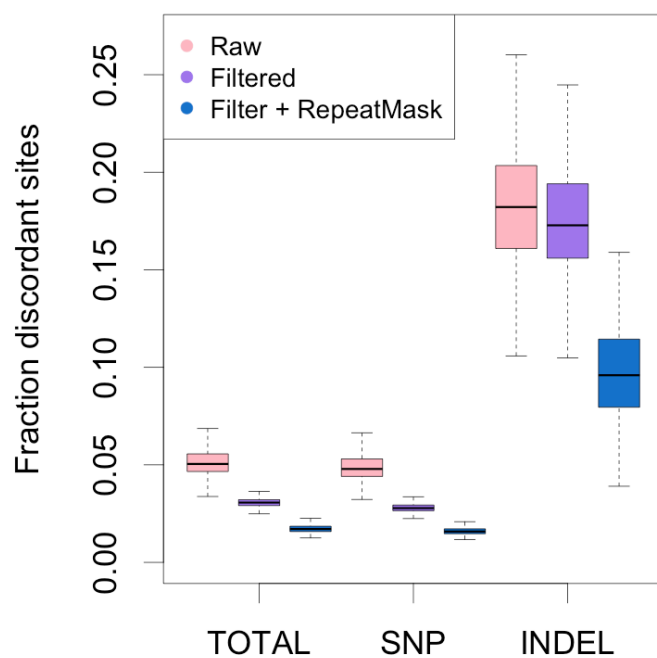


Figure A.3: Variant Call Discordance Between NewAlign and OldAlign ($n=345$). For filtered condition SNPs were filtered using GATK VQSR TS 99.5 and indels using GATK hardfilter. For filtered + RepeatMask condition variants in UCSC tracks RepeatMasker and Segmental Dups were excluded.

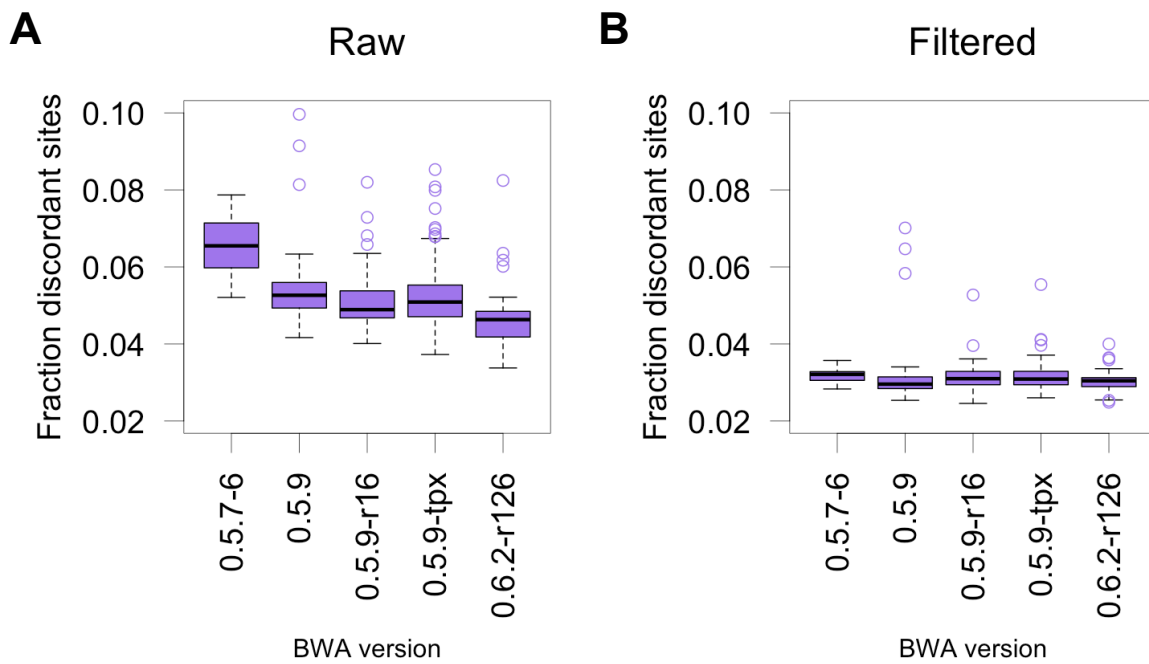


Figure A.4: Discordance With BAM Realignment. (A) Raw VCF discordance between NewAlign and OldAlign samples plotted by BWA version. (B) Filtered VCF discordance between NewAlign and OldAlign samples plotted by BWA version. SNVs were filtered at GATK VQSR TS 99.5, indels with GATK Hardfilter.

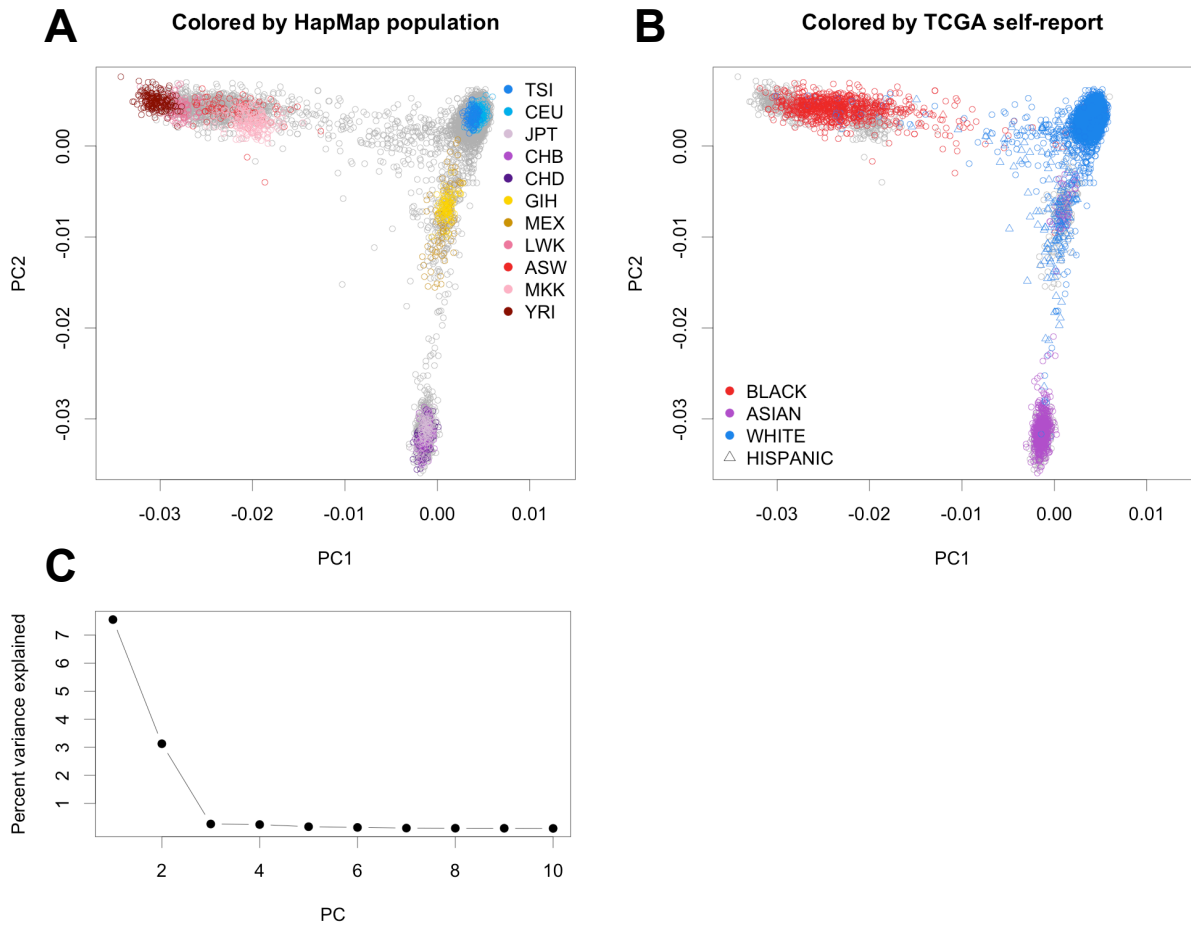


Figure A.5: PCA of Common Variants.(A) Principal components (PC) 1 and 2 from joint pan-cancer and HapMap analysis, HapMap samples are colored by population. (B) Same data as A, TCGA samples are colored by self-report ancestry. (C) Percent total variance explained by the top 10 PCs.

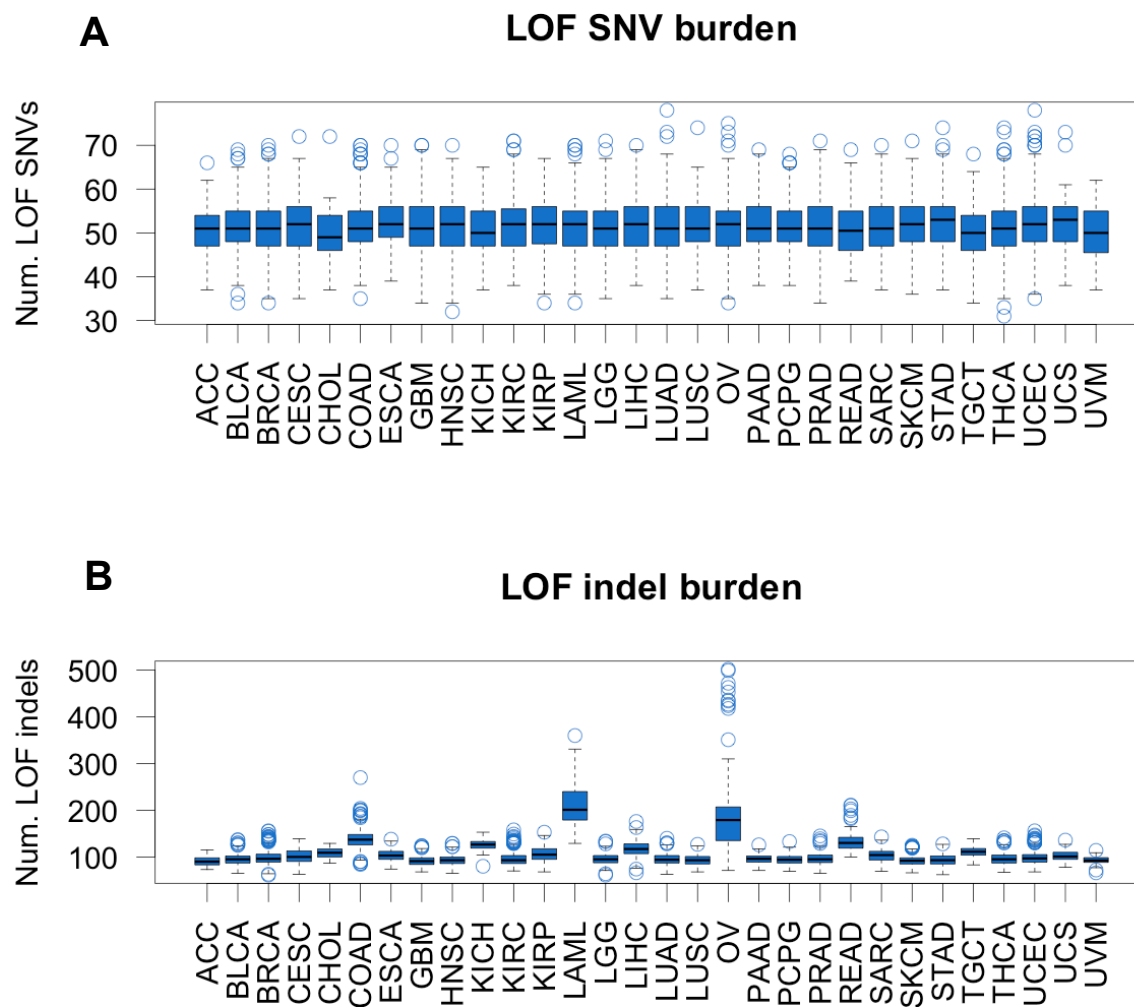


Figure A.6: LOF SNV and Indel Burden. (A) Individual LOF SNV burden plotted by cancer type. (B) Individual LOF indel burden plotted by cancer type.

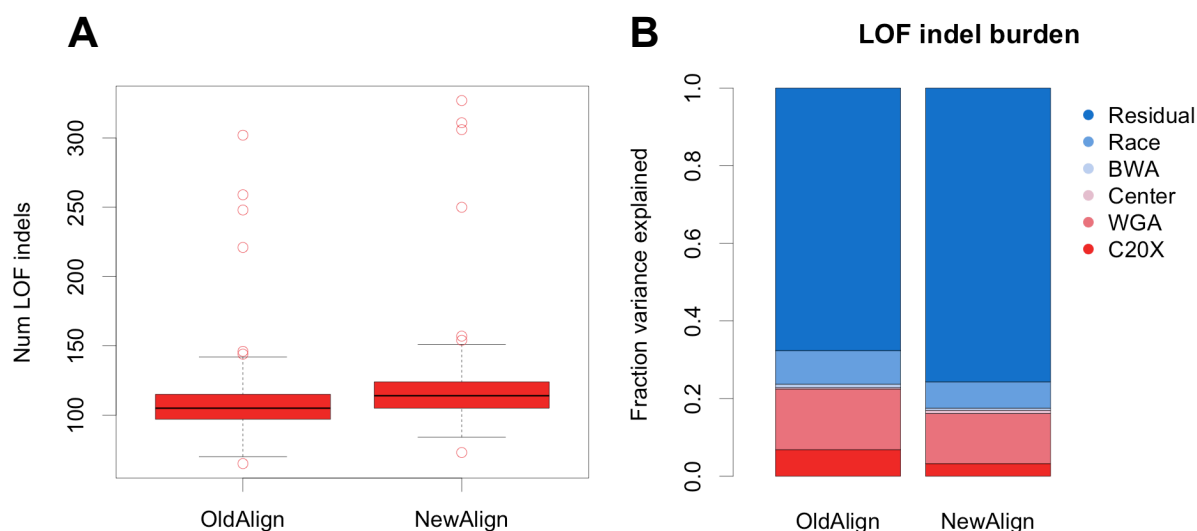


Figure A.7: LOF Indel Burden in NewAlign Cohort. (A) Number of LOF indels per individual in NewAlign and OldAlign pipelines. There were a median 8 more LOF indels in the NewAlign pipeline. Overall individual LOF indel burden was highly correlated between pipelines (Pearson $R^2 = 0.947$). (B) Percent of variation in individual LOF indel burden explained by technical covariates as assessed by ANOVA.

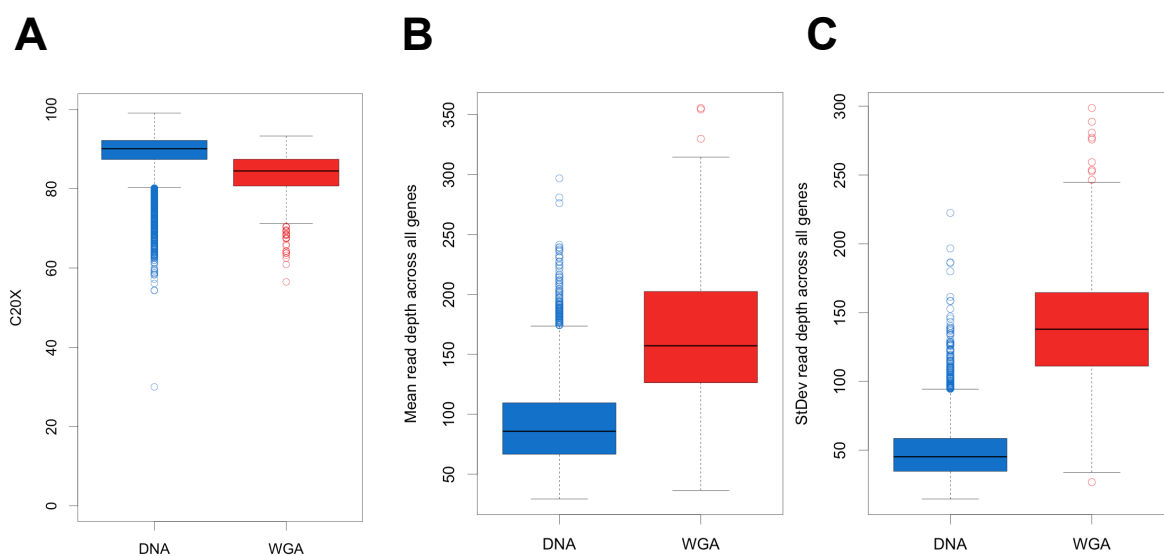


Figure A.8: Coverage and Read Depth in WGA Samples. (A) C20X plotted by WGA status. (B) Mean read depth per individual across 16,824 genes for $n=446$ WGA samples and $n=4,667$ DNA samples. (C) Standard deviation in read depth per individual across 16,824 genes.

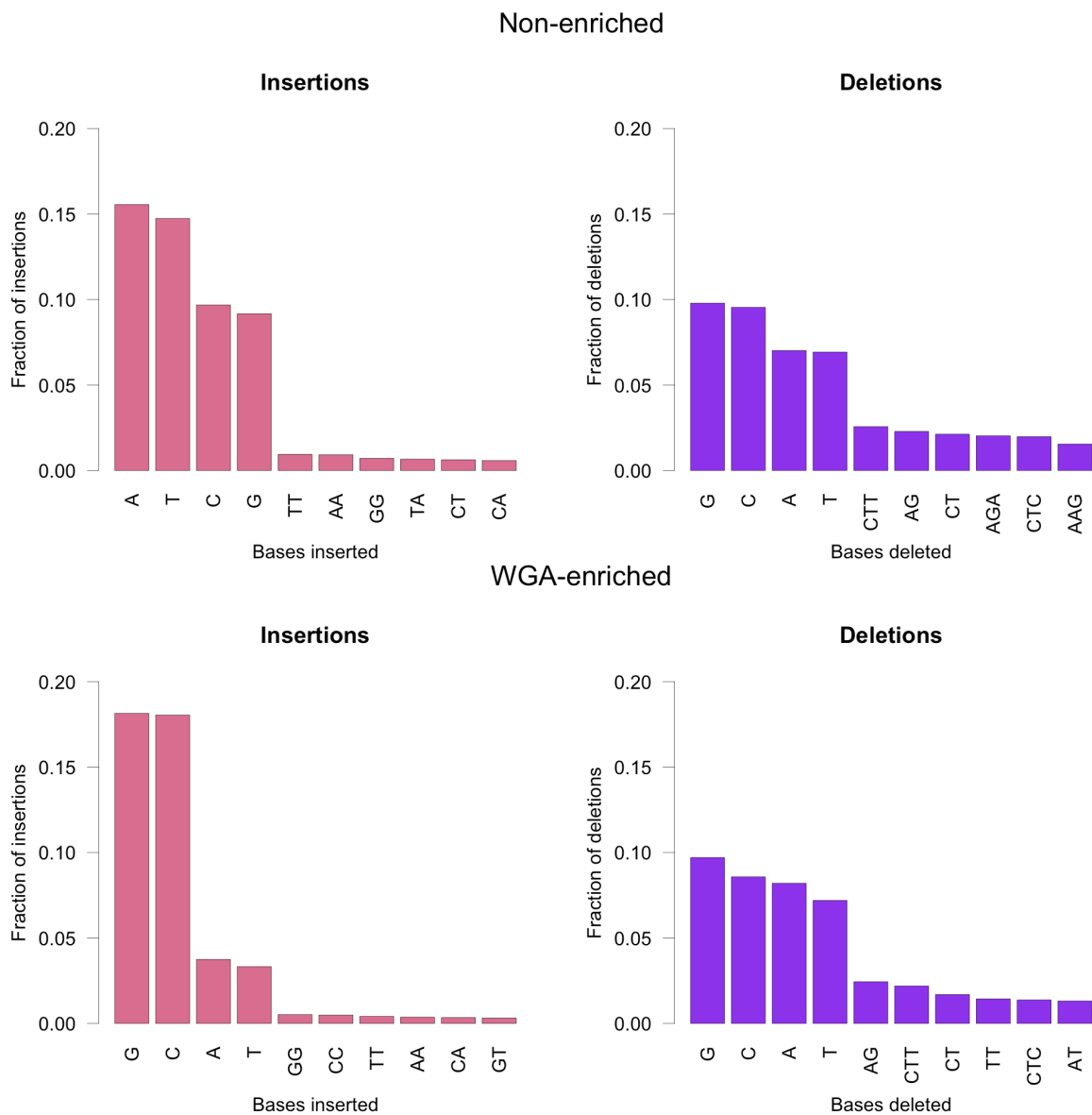


Figure A.9: Frequently Inserted and Deleted Bases of WGA Indels. The ten most frequent inserted or deleted base pairs for WGA-enriched and non-enriched indels. The height of the bar indicates frequency of each insertion or deletion relative to all insertions or deletions in that indel set.

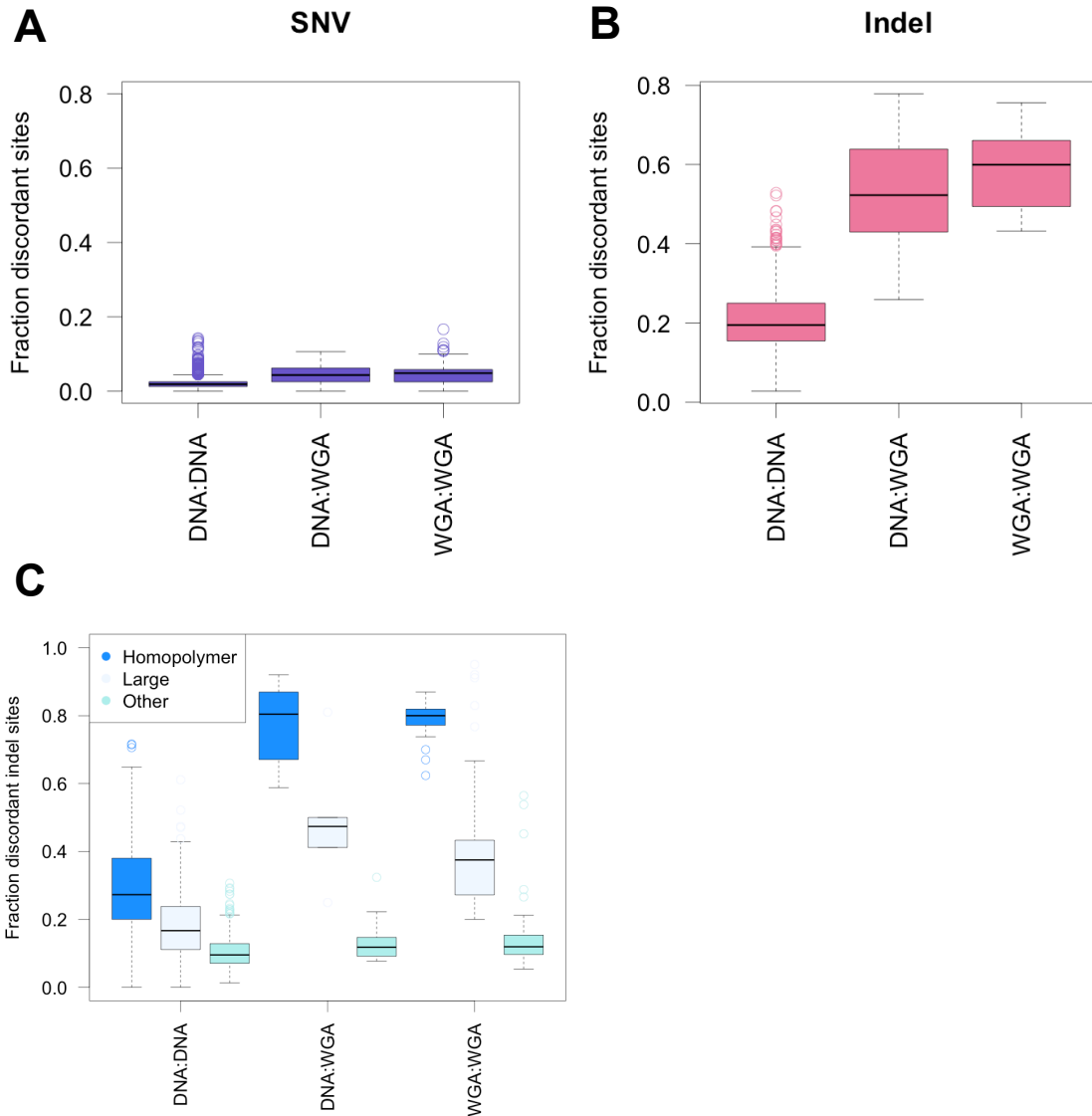


Figure A.10: Discordance Between Repeated WXS Samples. Discordance between repeated samples of $n=492$ individuals plotted by WGA status. DNA:DNA = all samples are DNA, WGA:DNA = at least one sample is WGA, WGA:WGA= all samples are WGA. Discordance was calculated separately for SNVs (A) and indels(B). (C) Indel discordance on the same samples calculated separately for homopolymer + indels, indels 15 base pairs or longer, and other indels.

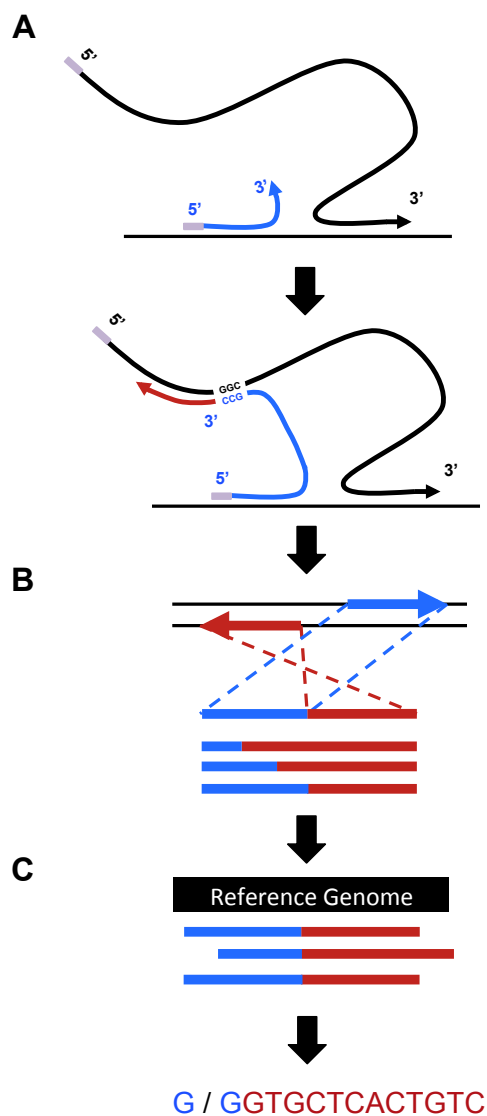


Figure A.11: Proposed Mechanism of Artifactual Indel Generation. (A) Branching during MDA creates free 3' termini that can anneal to proximal 5' strands with complementary sequence, generating chimera events. These events most frequently occur within a 10 kB window. (B) Chimera events manifest in sequence data as reads containing sequence from two noncontinuous regions of the reference genome. Here we demonstrate a chimera read formed by an inverted rearrangement with a deletion. (C) Chimeric reads can be discarded during multiple stages of variant calling, including initial alignment of reads to the genome, GATK's indel realignment step, or GATK's 'HaplotypeCaller' pairHMM realignment. We observe that chimeric reads that persist to the final stages of variant calling resemble insertions of varying sizes.

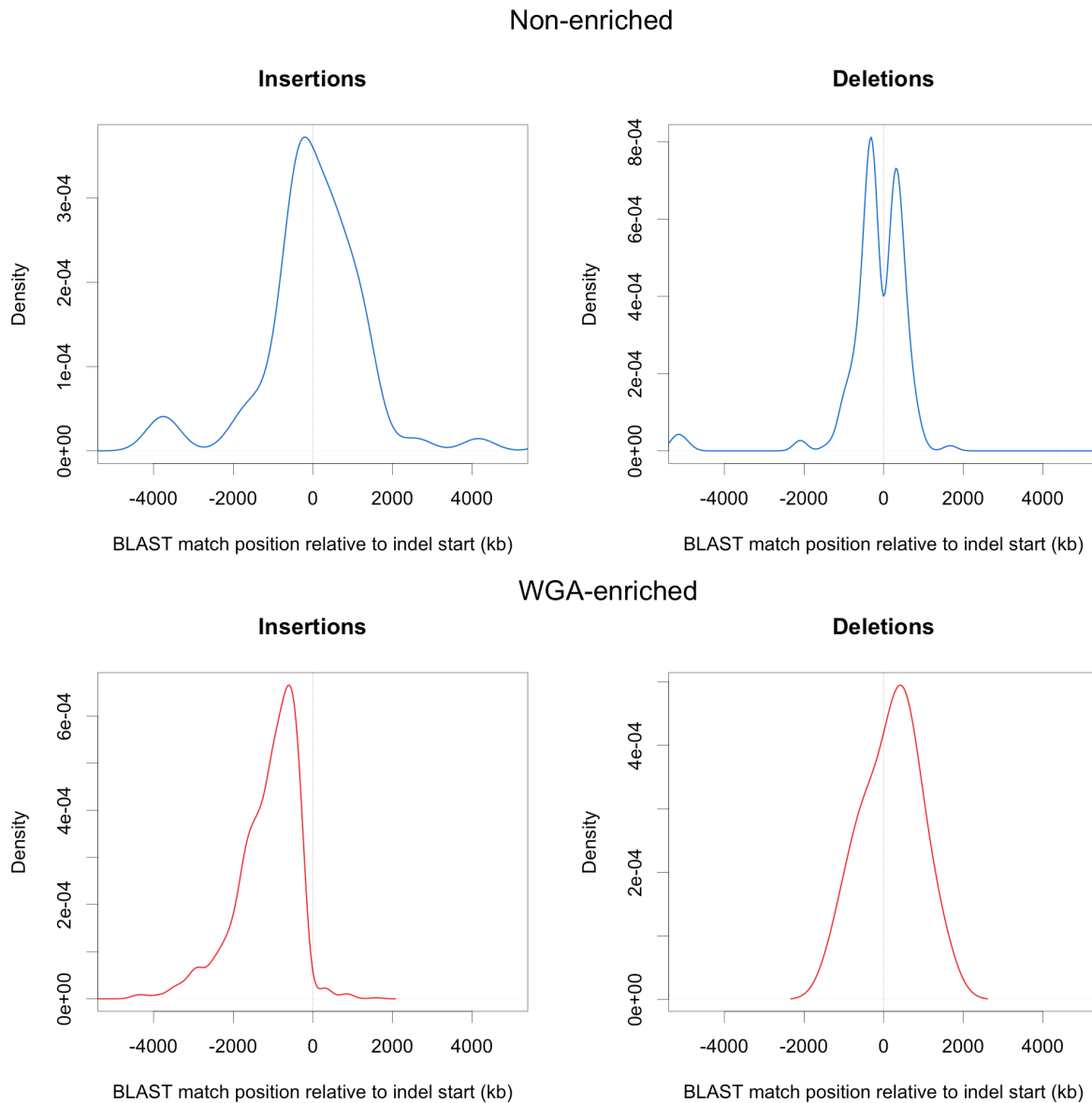


Figure A.12: Distribution of Indel Sequence BLAST Hits. For WGA-enriched and non-enriched large insertions and deletions with BLAST matches, the location of BLAST matches are shown. Indel start position is 0, $n=1,113$ WGA-enriched insertions, $n=11$ WGA-enriched deletions, $n=69$ Non-enriched insertions, $n=175$ Non-enriched deletions.

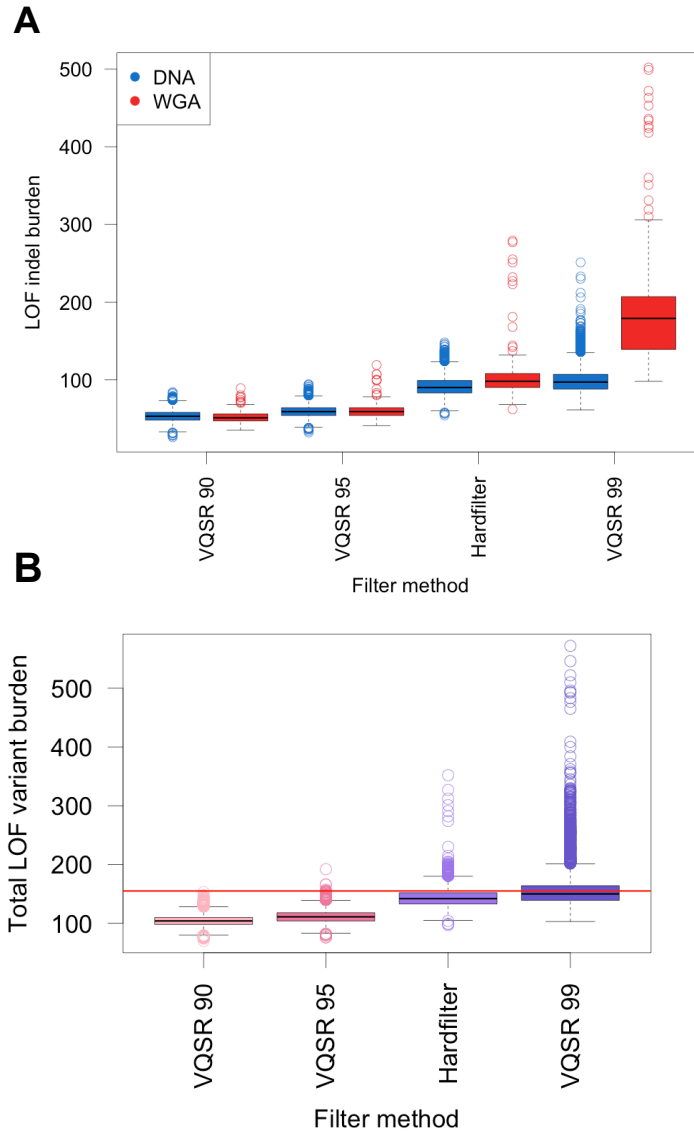


Figure A.14: LOF Indel Burden in WGA Samples Across Filtering Methods. (A) Individual LOF indel burden of WGA and DNA samples for each filter. (B) LOF variant count includes both SNV and indels. The red line indicates expected LOF burden from ExAC (155).

LAML / OV Shared Significant Genes

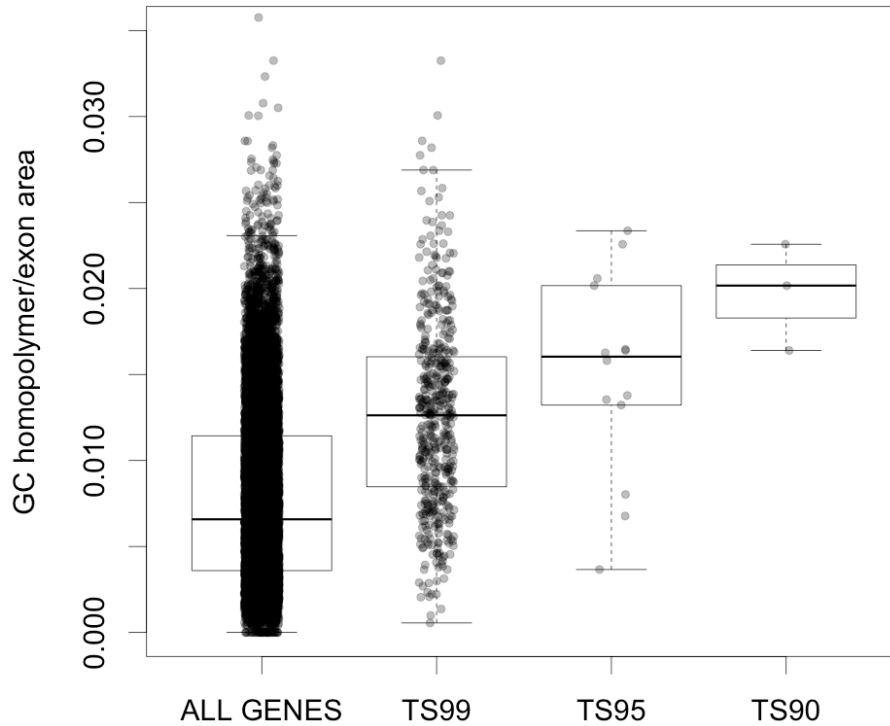


Figure A.15: G/C Homopolymer Content of Genes Shared Between OV and LAML. The number of G/C homopolymer regions normalized by coding exon length in base pairs plotted for all genes and for genes that were significant $p < 1.61 \times 10^{-7}$ by logistic regression for both OV and LAML. Significant genes shared between OV and LAML under three different indel filter conditions (VQSR TS99, TS95, TS90) are plotted for comparison.

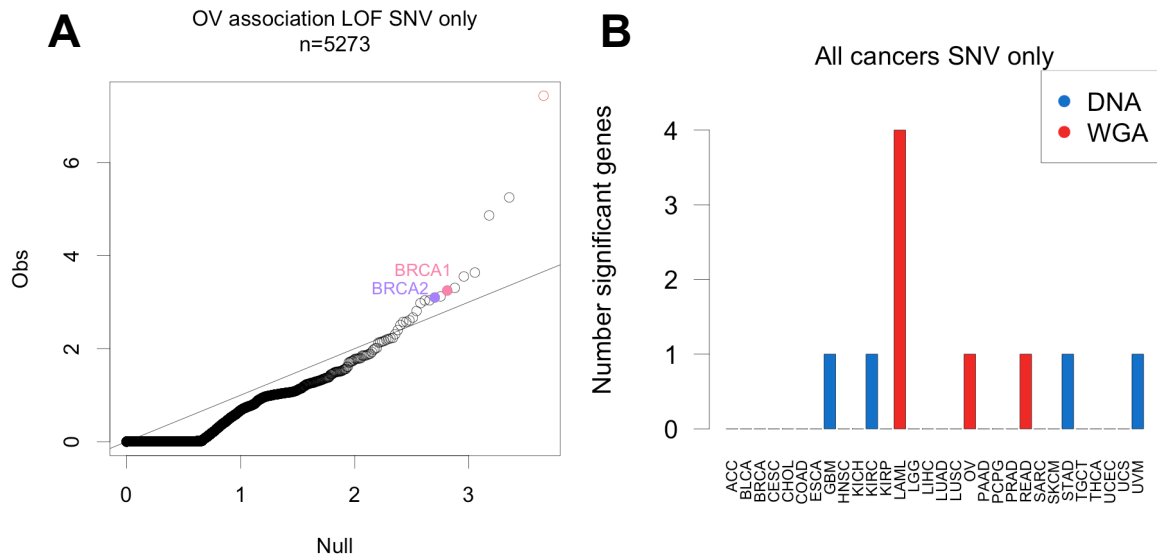


Figure A.16: LOF SNV Logistic Regression Analysis. (A) Quantile-quantile plot from logistic regression association testing between germline LOF SNV burden and OV. n=number of genes tested. Red line indicates significant cutoff and red points indicate associations significant $p < 1.61 \times 10^{-7}$. *BRCA1/2* associations highlighted. (B) Number of genes significant $p < 1.61 \times 10^{-7}$ by logistic regression for all cancer types. Color indicates cancer types containing WGA samples.

Table A.1: Composition of the Pan-Cancer Cohort.

ACC	Adrenocortical carcinoma	89	LUAD	Lung adenocarcinoma	575
BLCA	Bladder Urothelial Carcinoma	416	LUSC	Lung squamous cell carcinoma	327
BRCA	Breast invasive carcinoma	849	OV	Ovarian serous cystadenocarcinoma	399
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	308	PAAD	Pancreatic adenocarcinoma	188
CHOL	Cholangiocarcinoma	49	PCPG	Pheochromocytoma and Paraganglioma	182
COAD	Colon adenocarcinoma	325	PRAD	Prostate adenocarcinoma	510
ESCA	Esophageal carcinoma	190	READ	Rectum adenocarcinoma	114
GBM	Glioblastoma multiforme	315	SARC	Sarcoma	259
HNSC	Head and Neck squamous cell carcinoma	585	SKCM	Skin Cutaneous Melanoma	472
KICH	Kidney Chromophobe	66	STAD	Stomach adenocarcinoma	485
KIRC	Kidney renal clear cell carcinoma	275	TGCT	Testicular Germ Cell Tumors	149
KIRP	Kidney renal papillary cell carcinoma	319	THCA	Thyroid carcinoma	529
LAML	Acute Myeloid Leukemia	123	UCEC	Uterine Corpus Endometrial Carcinoma	487
LGG	Brain Lower Grade Glioma	516	UCS	Uterine Carcinosarcoma	57
LIHC	Liver hepatocellular carcinoma	380	UVM	Uveal Melanoma	80

Table A.2: Coverage of the Six TCGA Capture Kits. Size and overlap with Gencode exons for the six capture kits used to collect TCGA normal DNA samples.

Capture Kit	Size (MB)	Fraction Overlap With Gencode Exons	Number Samples
Agilent Custom	33	0.993	5793
Nimblegen SQEZ v2	36	0.996	1337
Nimblegen hg18	36	0.992	577
Nimblegen HGSC	37	0.997	1350
Nimblegen SQEZ v3	39	0.999	210
SureSelect 38	64	0.982	171
Intersection	27	0.977	

Table A.3: K-means Cluster Membership of HapMap Samples.

K-means Cluster	ASW	LWK	MKK	YRI	CEU	TSI	GIH	MEX	CHB
1	83	90	171	167	0	0	0	0	0
2	0	0	0	0	165	88	0	5	0
3	0	0	0	0	0	0	88	72	0
4	0	0	0	0	0	0	0	0	84

Table A.4: K-means Cluster Membership of TCGA Samples.

K-means Cluster	Black	White	Hispanic	Asian	NA
1	777	25	11	0	103
2	22	6611	129	9	963
3	6	76	152	39	65
4	0	11	6	536	50

Table A.5: GC Content of the Sequence Surrounding WGA Indels. Mean GC content of the sequence surrounding WGA-enriched and non-enriched indels. C.I. = Confidence interval for mean estimate derived from 1,000 bootstrap samples.

	Insertions	Deletions
WGA-enriched	Mean 0.628 (95% C.I. 0.624 - 0.633)	Mean 0.510 (95% C.I. 0.502 - 0.517)
Non-enriched	Mean 0.539 (95% C.I. 0.536 - 0.542)	Mean 0.515 (95% C.I. 0.514 - 0.518)

Table A.6: Allele Frequency of Homopolymer Indels. Mean allele frequency of indels in homopolymer regions in DNA and WGA samples. C.I. = Confidence interval for mean estimate derived from 1,000 bootstrap samples.

	Homopolymer -	Homopolymer +
DNA	Mean 0.0038 (95% C.I. 0.0034 - 0.0043)	Mean 0.0028 (95% C.I. 0.0019 - 0.0036)
WGA	Mean 0.0160 (95% C.I. 0.0141 - 0.0180)	Mean 0.0236 (95% C.I. 0.0213 - 0.0257)

Table A.7: Frequency of BLAST Match for WGA Indels. Total number of large indels and fraction of large indels with a BLAST hit in WGA-enriched and non-enriched indel sets. Large indels are indels ≥ 15 base pairs. BLAST hits are defined as a BLAST match ± 10 kB from the indel start position.

	Num. Insertions	% Insertions with BLAST hit	Num. Deletions	% Deletions with BLAST hit
WGA-enriched	1,310	86.49	108	10.18
Non-enriched	634	10.88	1,667	10.49

Table A.8: ANOVA of LOF Indel Burden Using Different Filters. Variance in LOF indel burden explained by technical covariates for each indel filtering approach. Sum. Sq., Sum of Squares; Df, Degrees of Freedom.

GATK Indel VQSR TS 90.0				
	Sum Sq	Df	F value	Pr(>F)
C20X	1.31E+04	1	323.8629669	4.52E-71
WGA	2.39E+00	1	0.05899371	8.08E-01
Center	1.49E+03	2	18.33944372	1.13E-08
BWA	1.19E+03	5	5.86624688	2.04E-05
Race	5.60E+04	5	276.5583032	1.43E-274
Residuals	3.39E+05	8363		
GATK Indel VQSR TS 95.0				
	Sum Sq	Df	F value	Pr(>F)
C20X	15219.389	1	327.944016	6.31E-72
WGA	1361.419	1	29.335557	6.26E-08
Center	3258.38	2	35.105429	6.57E-16
BWA	1507.617	5	6.497157	4.89E-06
Race	68648.485	5	295.844466	2.06E-292
Residuals	388114.262	8363		
GATK Hardfilter				
	Sum Sq	Df	F value	Pr(>F)
C20X	50615.091	1	419.836036	4.45E-91
WGA	76075.977	1	631.025972	2.60E-134
Center	18981.98	2	78.724735	1.34E-34
BWA	4239.435	5	7.032952	1.44E-06
Race	150187.094	5	249.150811	6.51E-249
Residuals				
GATK Indel VQSR TS 99.0				
	Sum Sq	Df	F value	Pr(>F)
C20X	52930.43	1	153.90042	4.95E-35
WGA	3744887.28	1	10888.62716	0.00E+00
Center	383585.43	2	557.65614	4.53E-228
BWA	169507.9	5	98.57217	2.76E-101
Race	146904.86	5	85.42806	7.59E-88
Residuals	2876257.21	8363		

Table A.9: Correlation Between LOF Indels and Homopolymer Tracts. Spearman correlation between LOF indel burden and homopolymer content. LOF indel allele counts were calculated for each gene separately for WGA and DNA samples. Both allele counts and homopolymer region counts were normalized by gene by dividing by coding exon length in base pairs.

	WGA LOF indel AC/ exon area	DNA LOF indel AC/ exon area
A/T Homopolymer/ exon area	-0.16	-0.11
G/C Homopolymer/ exon area	0.19	0.03

Appendix B

Supplemental Material: Exome-Wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in the Cancer Genome Atlas

This appendix contains supplemental figures for the work described in Chapter 3 of this document: "Exome-Wide Analysis of Bi-allelic Alterations Identifies a Lynch Phenotype in the Cancer Genome Atlas". Each figure is referred to in the main text of the chapter and a brief description of each figure is given here.

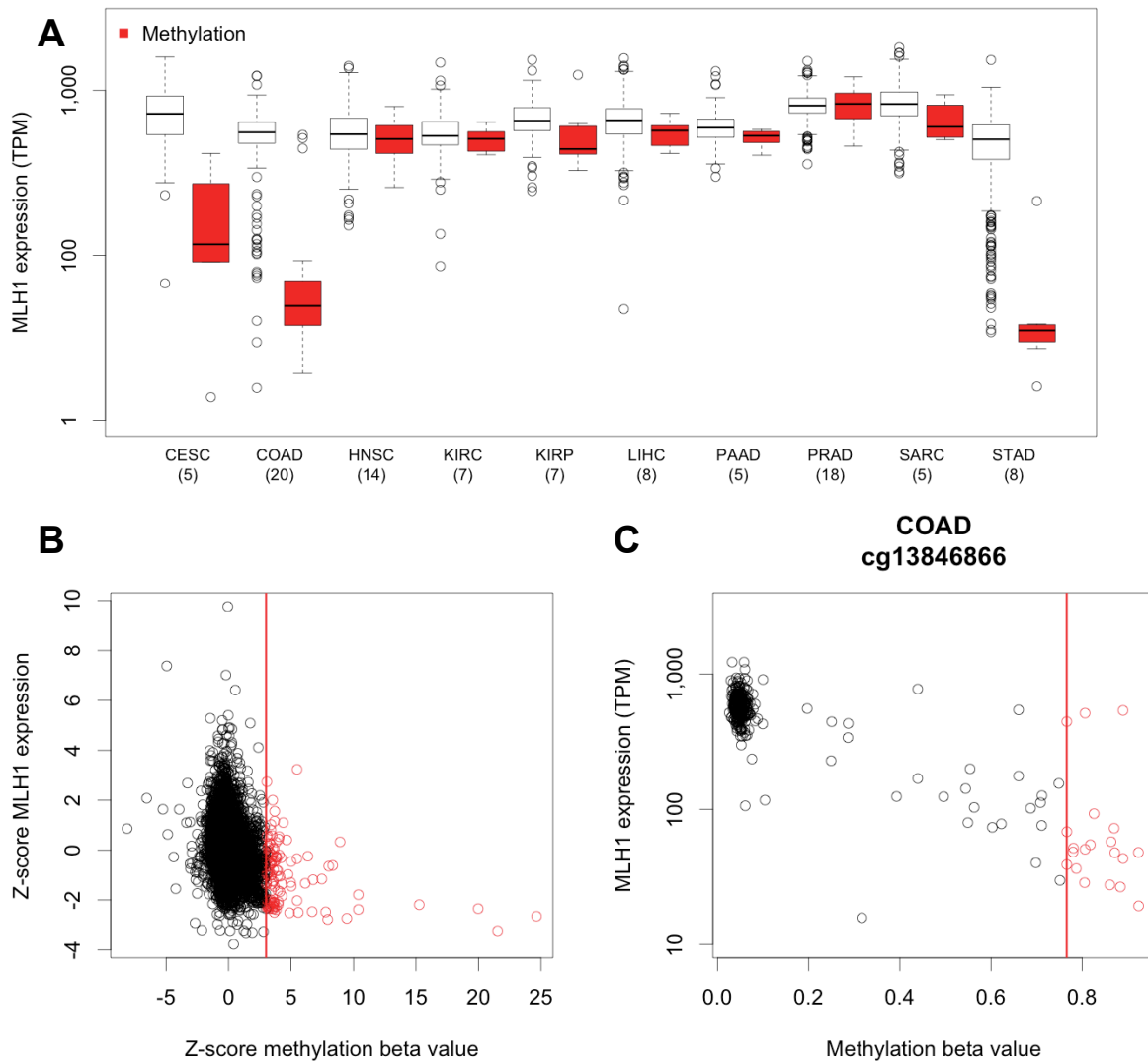


Figure B.1: Calling Somatic Methylation Status. (A) *MLH1* expression is decreased in methylated samples. Cancer types with 5+ methylated samples are shown. TPM = transcripts per million. (B) Expression of *MLH1* vs. methylation beta value. Both expression values (in transcripts per million) and methylation beta values were converted to Z-scores using the mean and standard deviation for each cancer type. The red line indicates the cutoff used to call methylation. (C) Expression of *MLH1* vs. methylation beta value in colon cancer samples only. Beta values are from the methylation probe cg13846866, which was most anti-correlated with *MLH1* expression in colon cancer. The red line indicates the cutoff used to call methylation.

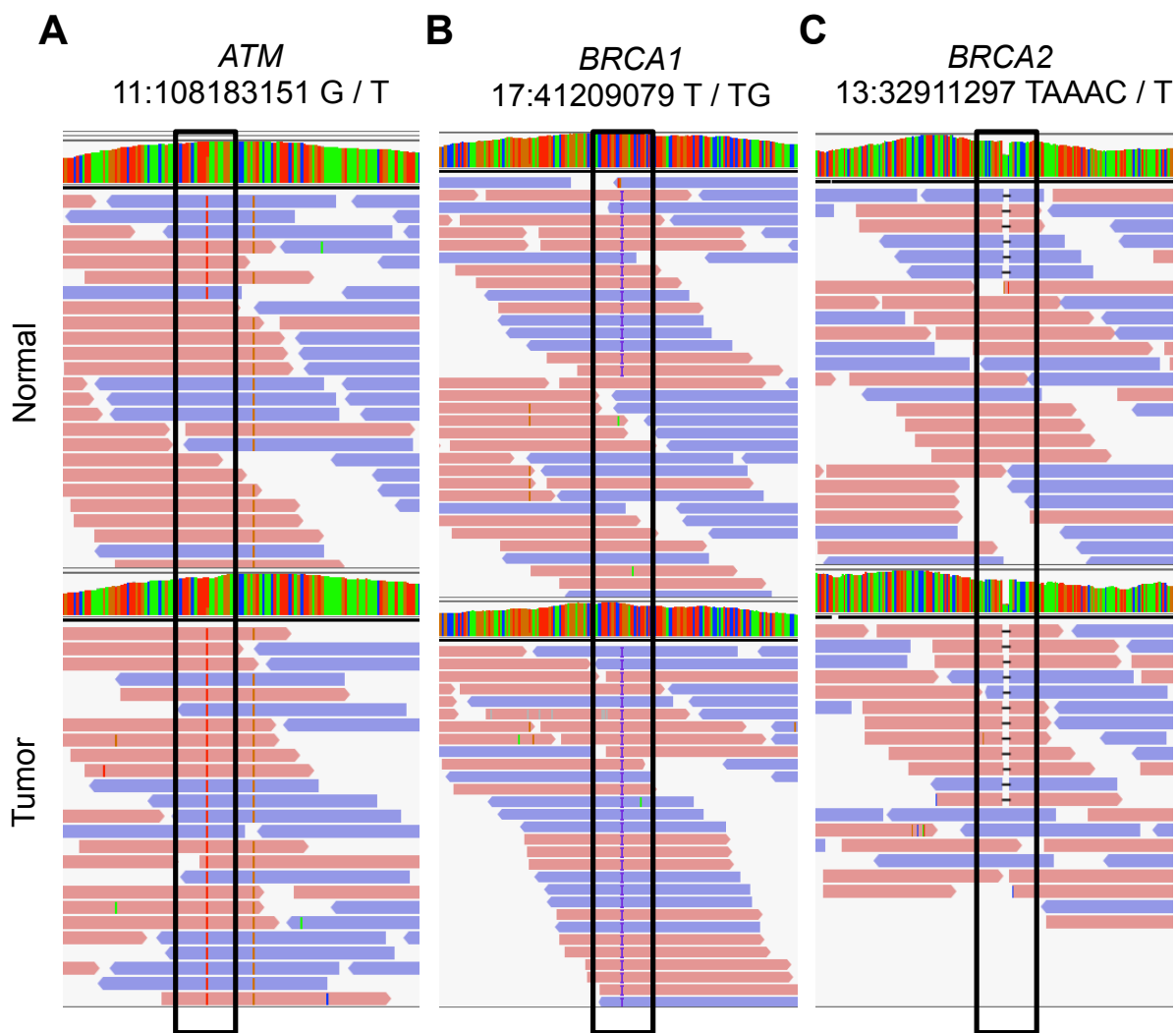


Figure B.2: Example LOH Events. IGV snapshots of paired tumor:normal BAM files with a somatic LOH event for *ATM* (A), *BRCA1* (B), and *BRCA2* (C). The germline locus subject to LOH is highlighted with a box. Screenshots were taken with downsampling settings of max read count 20 per 200 base window.

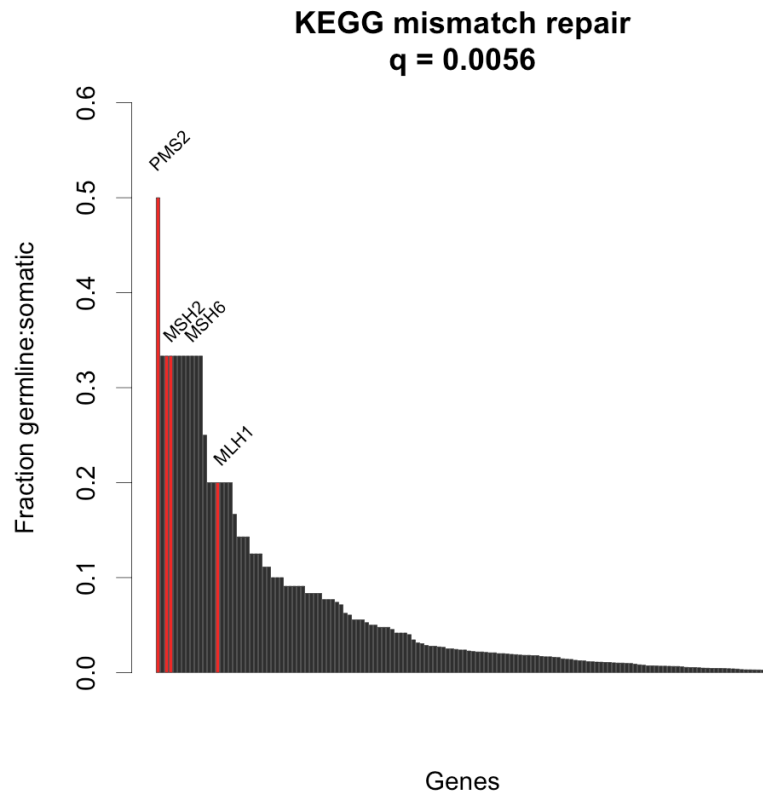


Figure B.3: Genes Frequently Affected by Germline:Somatic Alteration. Barplot showing gene level frequency of germline:somatic alteration with KEGG mismatch repair pathway genes highlighted in red. Fraction germline:somatic alteration was calculated for each gene as number of germline:somatic alterations/number of germline LOF variants. Only genes with > 2 germline LOF variants in the cohort were included in the analysis. Significance was calculated using fgsea and is adjusted for multiple hypothesis testing.

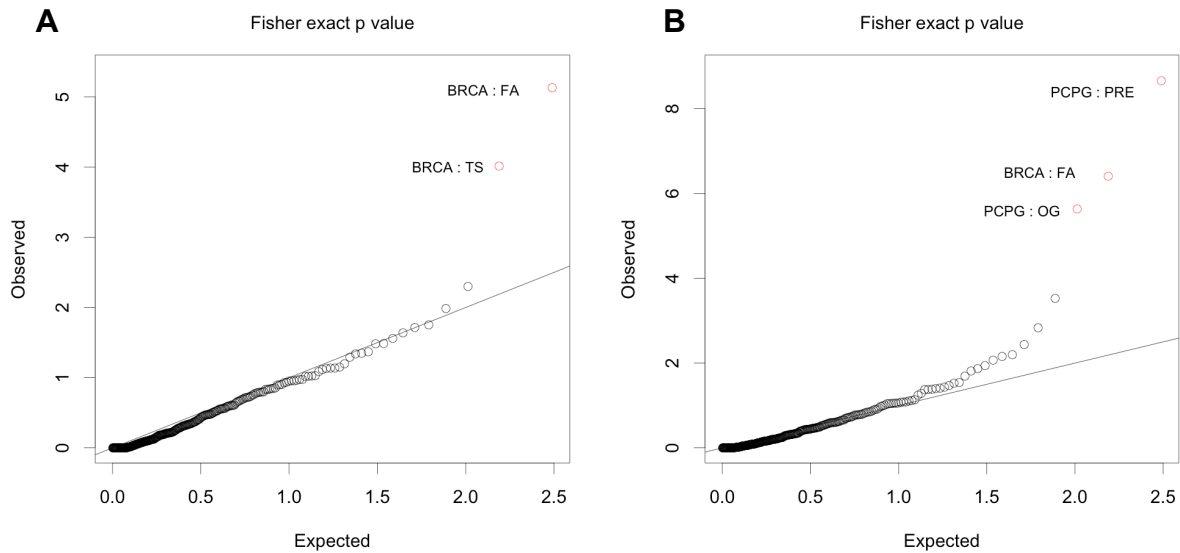


Figure B.4: Association Between Germline LOF Burden and Cancer Type. QQ plot of Fisher exact p-values of the association between 11 DDR pathways and 28 cancer types using LOF variants only (A) or LOF and ClinVar pathogenic variants (B). Only samples determined to be of European descent using PCA were used ($n = 7,734$). Red indicates significance above a Bonferroni threshold. (A) Significant hits: breast cancer:tumor suppressor pathway ($p = 9.69e^{-5}$), breast cancer:Fanconi anemia pathway ($p = 7.42e^{-6}$). (B) Significant hits: pheochromocytoma and paraganglioma:cancer predisposition genes ($p = 1.18e^{-9}$), breast cancer:Fanconi anemia pathway ($p = 3.89e^{-7}$), pheochromocytoma and paraganglioma:oncogenes ($p = 2.30e^{-6}$).

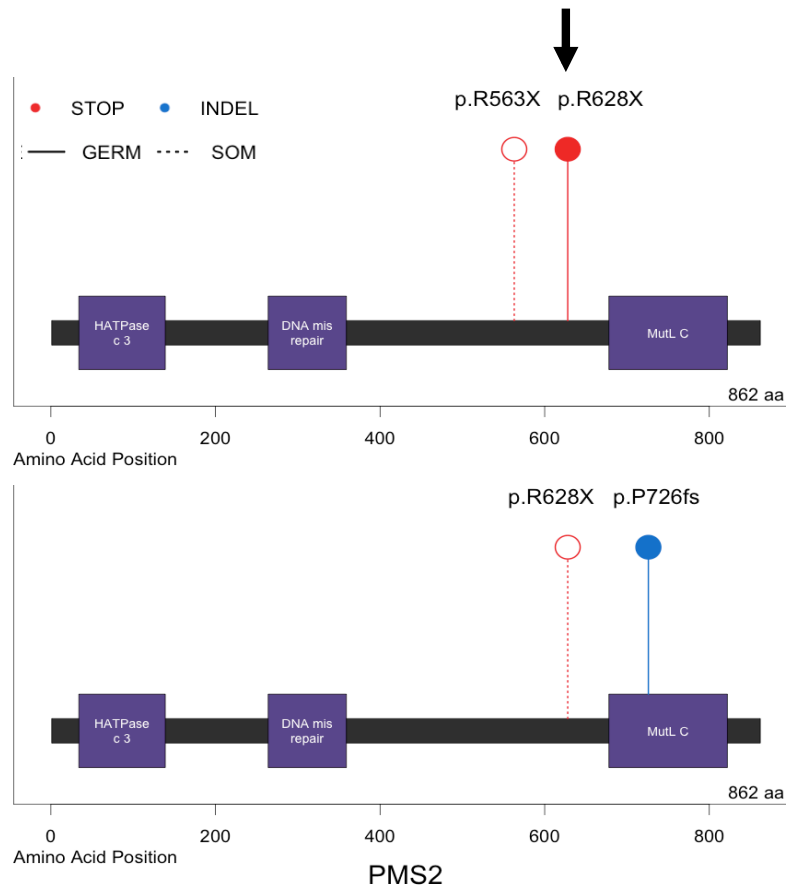


Figure B.5: Germline and Somatic LOF in *PMS2*. Both germline and somatic LOF mutations can alter the same position. Lollipop plot indicating the amino acids altered in two samples with *PMS2* bi-allelic alteration. Germline LOF variants are represented by solid lines, somatic LOF mutations by dashed lines. The stopgain mutation highlighted by an arrow, p.R628X, is inherited in one individual and acquired in the other. PFAM domain abbreviations: HATPase c 3 = Histidine kinase, DNA gyrase B, and HSP90-like ATPase; MutL C = MutL C terminal dimerization domain.

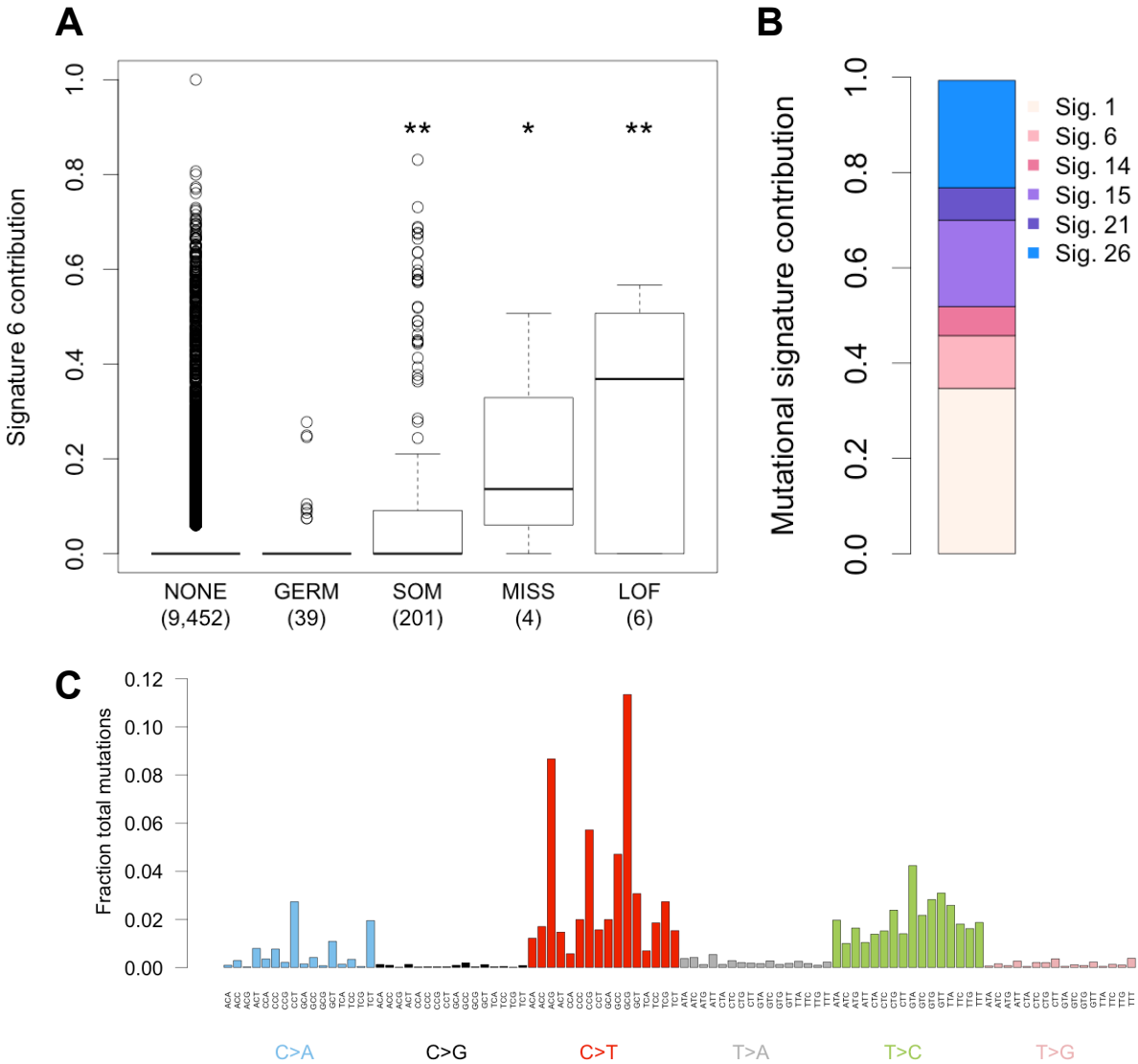


Figure B.6: Mutational Signature Analysis of Germline:Somatic MMR Alteration. (A) Fraction of mutations attributed to mutational signature 6 plotted by type of germline:somatic MMR alteration. Individuals were grouped by MMR gene mutation type: NONE, no alteration; GERM, germline LOF variants only; SOM, somatic LOF mutations only; MISS, bi-allelic alteration including a missense mutation; LOF, bi-allelic alteration via dual LOF mutation. Wilcox $p = 0.00023, 0.0063,$ and 0.00096 ; permutation $p = 0.0002, 0.011,$ and 0.002 for SOM, MISS, and LOF respectively. (B) Fraction of combined germline:somatic alteration carriers mutational profile attributed to the mutational signatures available in COSMIC. (C) Mutational profile of all germline:somatic alteration carriers combined.

Germline MMR pathway burden

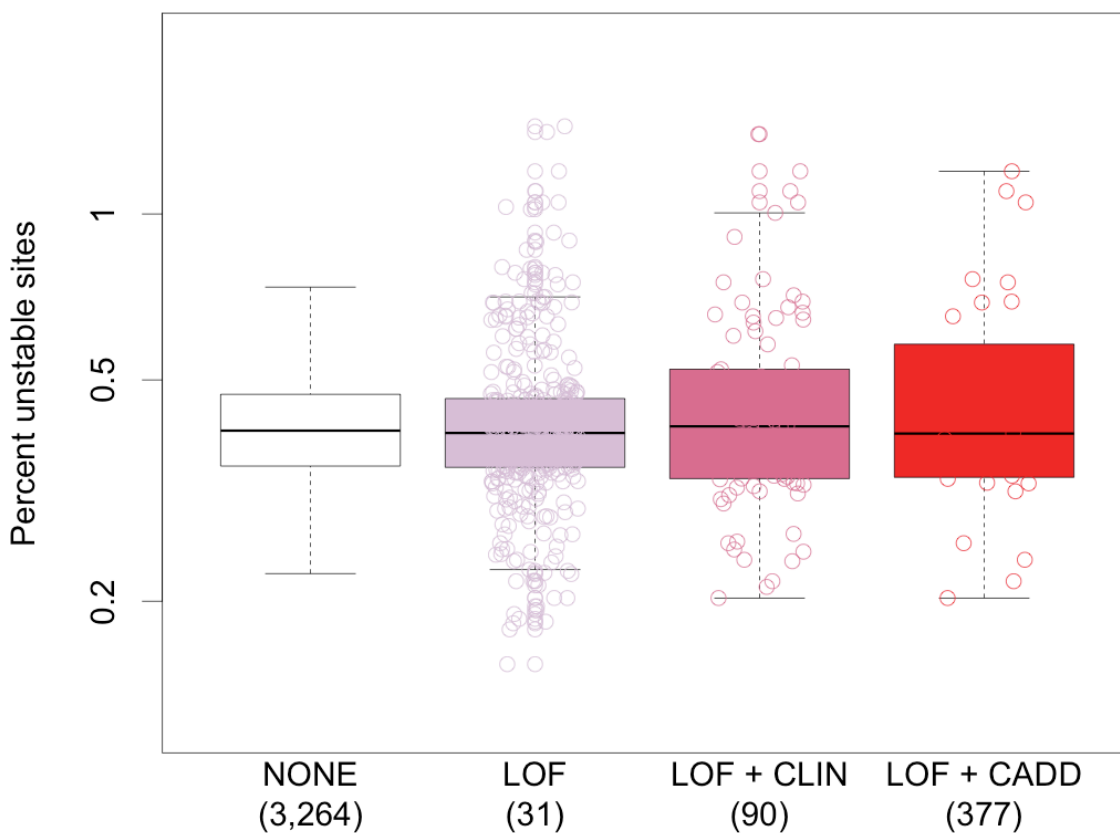


Figure B.7: Mono-allelic Germline MMR Variation Not Associated With MSI. Somatic MSI burden for individuals with MMR germline variants but no somatic alteration of the MMR pathway. NONE = no alteration of the MMR pathway, LOF = germline LOF variant in an MMR gene, LOF + CLIN = germline LOF variant or ClinVar pathogenic variant in an MMR gene, LOF + CADD = germline LOF variant or variant with a CADD score ≥ 30 in an MMR gene. The number of samples in each category is displayed in parentheses.

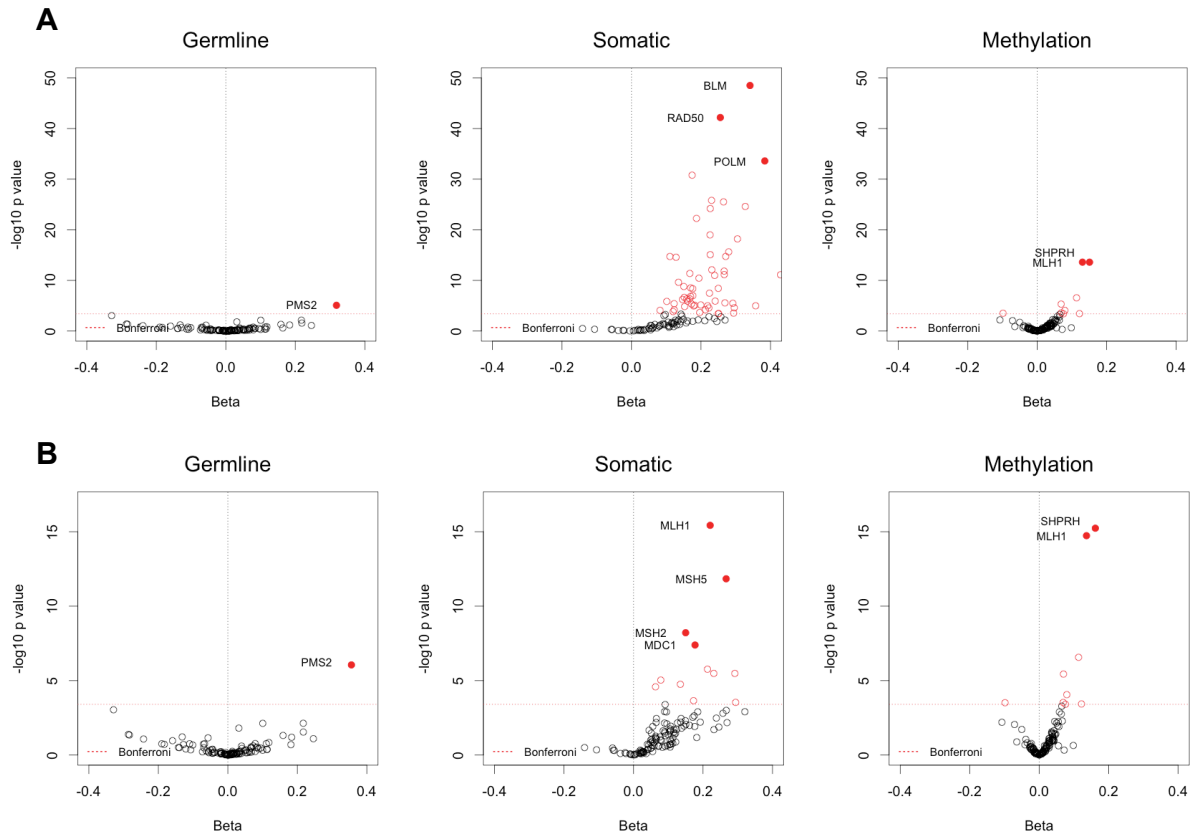


Figure B.8: Association Testing Between Genomic Alteration and MSI Burden. (A) The same analysis as Figure 3.4, but including all somatic LOF mutations. (B) The same analysis as Figure 3.4, but including germline:somatic MMR alteration carriers.

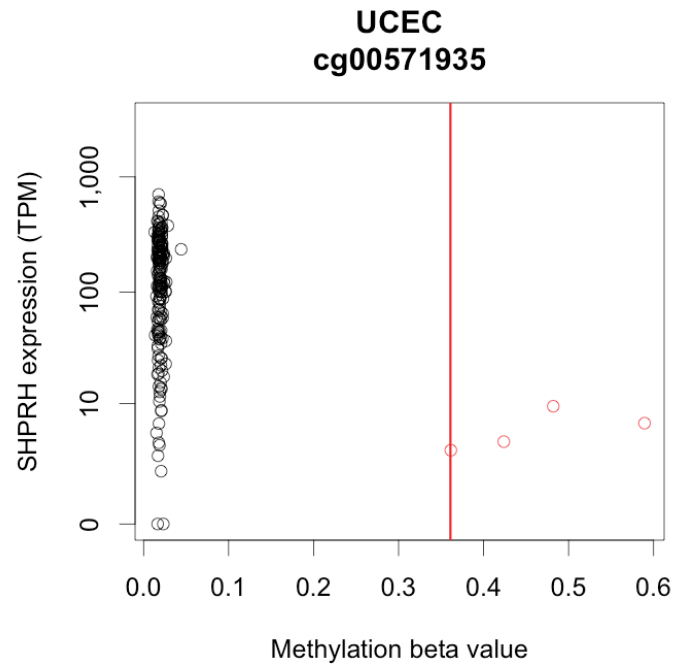


Figure B.9: *SHPRH* Methylation in Uterine Cancer. Expression of *SHPRH* vs. methylation beta value in uterine cancer samples only. Beta values are from the methylation probe cg00571935, which was most anti-correlated with *SHPRH* expression in uterine cancer. The red line indicates the cutoff used to call methylation.

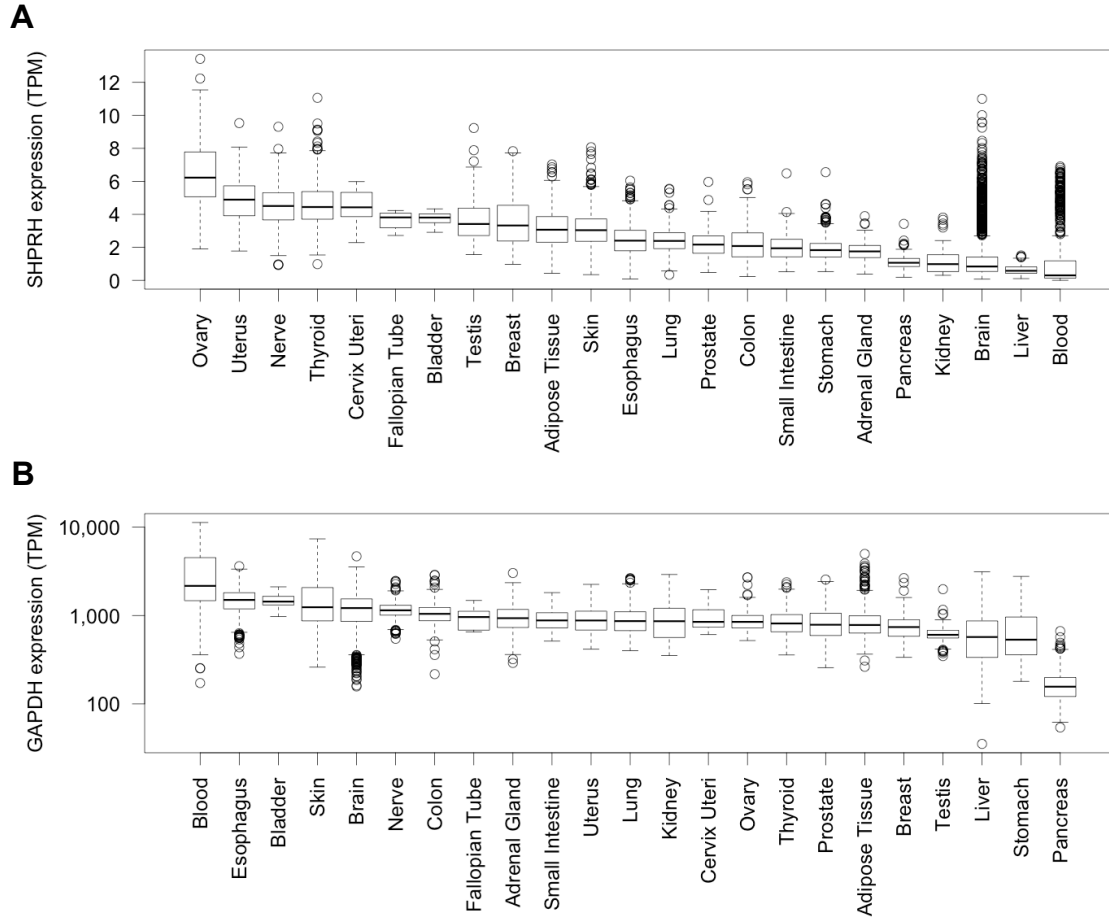


Figure B.10: *SHPRH* Expression in Normal Tissues. Expression of *SHPRH* (A) or *GAPDH* (B) in 23 normal tissue types represented in GTEx. TPM = transcripts per million.

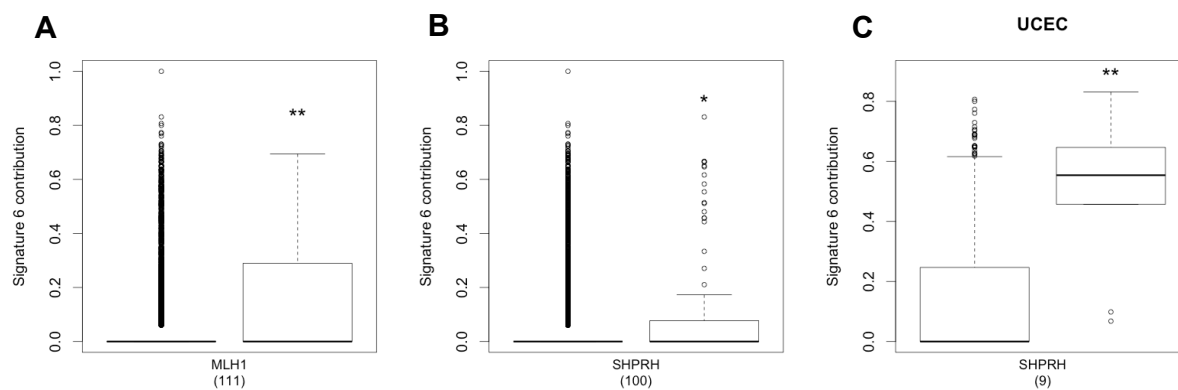


Figure B.11: Mutational Signature Analysis of *MLH1* and *SHPRH* Methylation. (A) Fraction of mutations attributed to mutational signature 6 plotted by *MLH1* methylation status. Wilcox $p = 3.882e^{-15}$, permutation $p < 1e^{-4}$. (B) Fraction of mutations attributed to mutational signature 6 plotted by *SHPRH* methylation status. Wilcox $p = 0.041$, permutation $p = 0.0264$. (C) Fraction of mutations attributed to mutational signature 6 plotted by *SHPRH* methylation status in uterine cancer samples only. Wilcox $p = 3.405e^{-5}$, permutation $p < 1e^{-4}$.

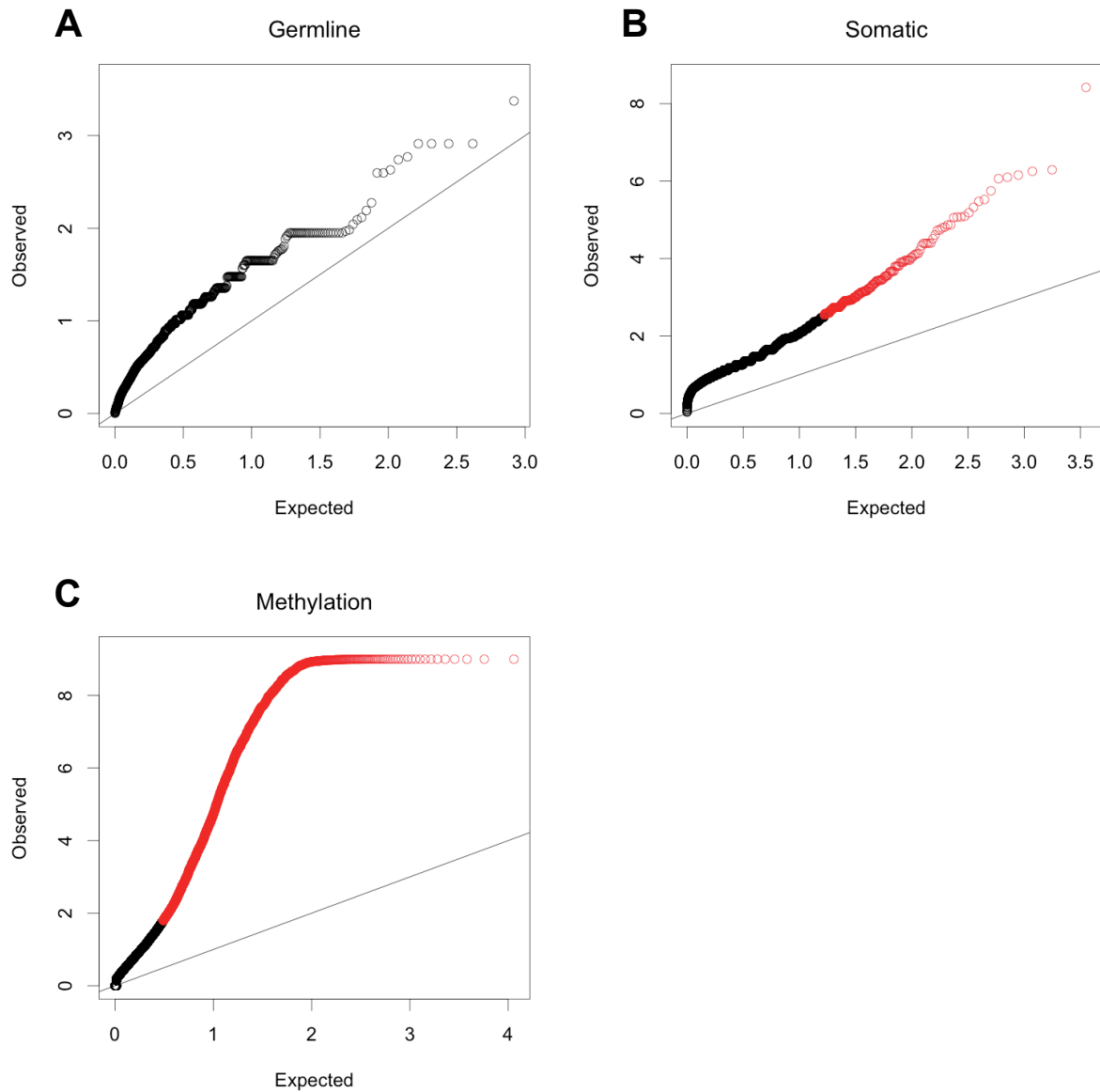


Figure B.12: Co-Occurrence Testing for *SHPRH* Methylation. QQ plots of co-occurrence testing between *SHPRH* methylation and germline LOF variants (A), somatic LOF mutations (B), and somatic methylation (C) in 9,484 genes and $n = 8,087$ samples. Red indicates 5% FDR. Somatic LOF mutations in 213 genes and somatic methylation of 3,694 genes significantly co-occur with *SHPRH* methylation. Of these co-occurring genes, 8 are MMR pathway genes (*MLH3*, *PCNA*, *POLD3*, *RFC1*, *RFC5*, *RPA3*, *MLH3*, and *PMS2*).

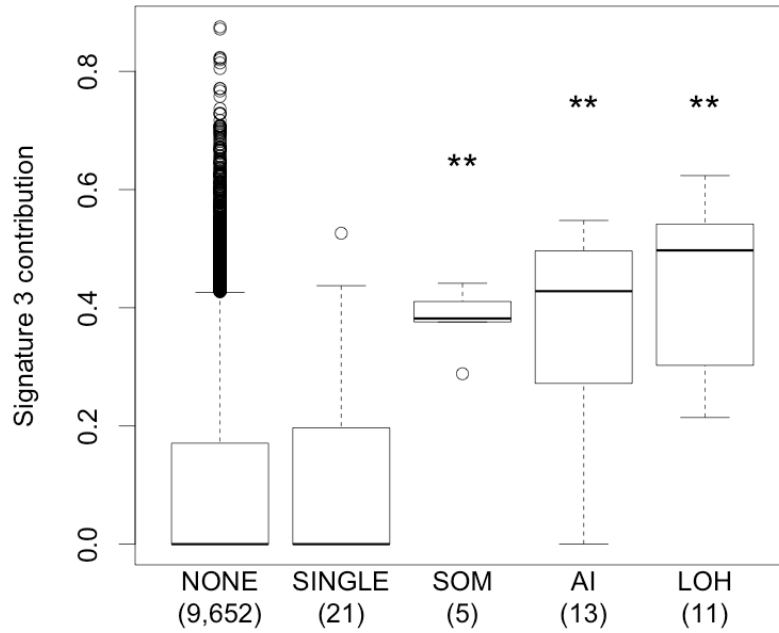


Figure B.13: Mutational Signature Analysis of *BRCA1/2* Carriers. Fraction of mutations attributed to mutational signature 3 plotted by *BRCA1/2* status. Single = germline LOF or ClinVar pathogenic variant only, SOM = bi-allelic alteration via somatic mutation, AI = allelic imbalance, LOH = loss of heterozygosity. Wilcoxon rank sum test $p = 2.246e^{-4}$, $6.981e^{-6}$, $2.343e^{-7}$; permutation $p < 1e^{-4}$, for SOM, AI, and LOH respectively.

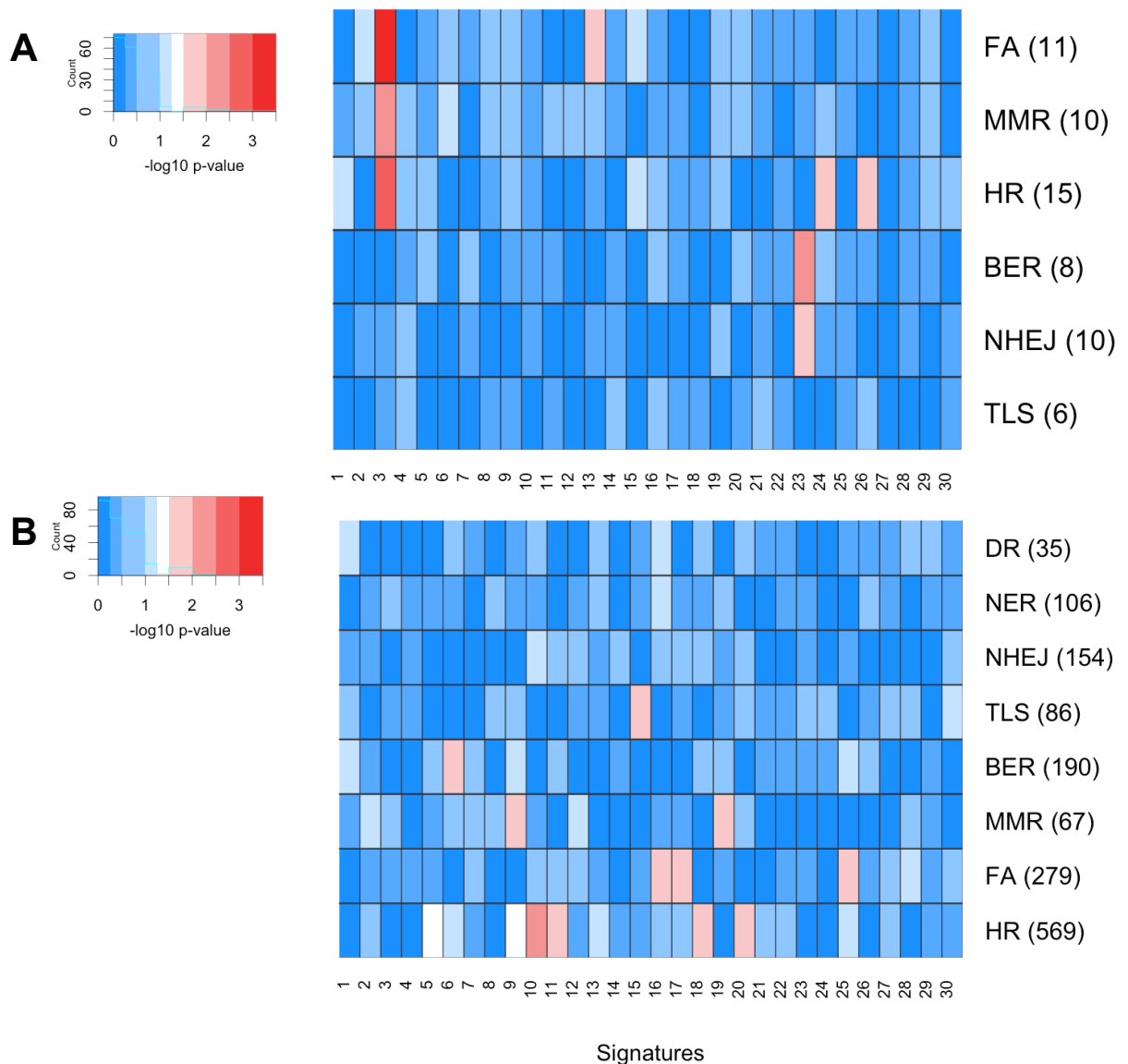


Figure B.14: Mutational Signature Analysis of DDR Pathway Alteration. A Wilcoxon rank sum test was used to test for differences in somatic mutational signature burden between individuals carrying DDR alterations vs. those without. (A) Heatmap of p values for the association between bi-allelic alteration in 6 DDR pathways and 30 mutational signatures in COSMIC. (B) Heatmap of p values for the association between mono-allelic alteration in 8 DDR pathways and 30 mutational signatures in COSMIC.

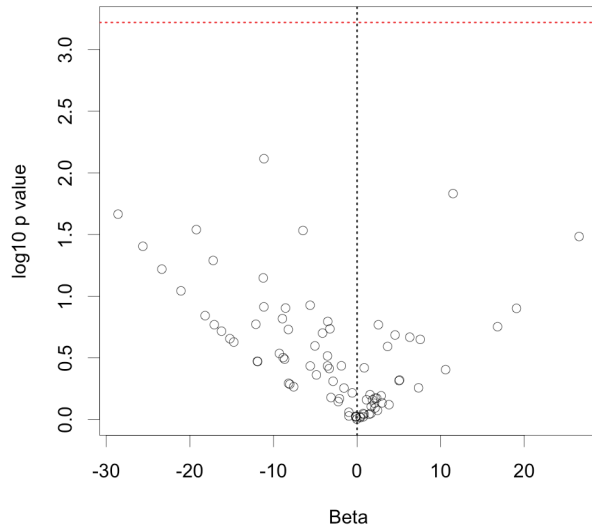


Figure B.15: Association Between Age and Damaging Germline Variants. Volcano plot of association testing between germline LOF and pathogenic ClinVar variant carrier status and age of diagnosis for 85 cancer predisposition genes in $n = 8,913$ samples.

Table B.1: Association Between MSI and MMR Alteration. P-values, β estimates, and standard errors determined using a linear model.

	p value	Beta	Std. Error	t value
Germ	0.756	0.007	0.024	0.310
Som	$1.53e^{-20}$	0.104	0.011	9.331
Meth	$5.44e^{-05}$	0.022	0.005	4.039
Mixed germ	0.0382	0.116	0.056	2.073
Mixed	$2.33e^{-14}$	0.167	0.021	7.653
B-allelic Som	0.574	0.044	0.079	0.561
Germline:Methylation	0.285	0.146	0.137	1.068
Germline:Somatic	$3.95e^{-15}$	0.444	0.056	7.881

Table B.2: Association Between Age and MMR Alteration. P-values, β estimates, and standard errors determined using a linear model.

	p value	Beta	Std. Error	t value
Germ	0.448	-2.251	2.967	-0.758
Som	0.698	-0.536	1.384	-0.387
Meth	0.226	0.831	0.686	1.211
Mixed germ	0.082	12.040	6.935	1.736
Mixed	0.827	-0.590	2.703	-0.218
B-allelic Som	0.921	-0.975	9.832	-0.099
Germline:Methylation	0.861	2.966	16.981	0.174
Germline:Somatic	0.0418	-14.164	6.957	-2.035

Table B.3: Germline Variants Pathogenic for Lynch Syndrome. ClinVar annotations: P = pathogenic, Num. Sub. = number of sources submitting to ClinVar.

Source	Type	Chr	Pos	Ref	Alt	Gene	AA change	ExAC AF	ClinVar	Num. Sub.
TCGA	LOF	3	37053589	C	T	<i>MLH1</i>	p.R226X	singleton	P	8
TCGA	LOF	2	47703538	C	T	<i>MSH2</i>	p.R680X	singleton	P	10
TCGA	LOF	7	6026709	G	A	<i>PMS2</i>	p.R563X	1.65E-05	P	4
TCGA	LOF	7	6026514	G	A	<i>PMS2</i>	p.R628X	3.30E-05	P	5
TCGA	MISS	2	47693947	G	A	<i>MSH2</i>	p.S554N	singleton	P	1

Table B.4: Germline Variants of Unknown Significance for Lynch Syndrome. ClinVar annotations: P = pathogenic, NA = not in ClinVar, Num. Sub. = number of sources submitting to ClinVar.

Source	Type	Chr	Pos	Ref	Alt	Gene	TX	AA change	ExAC AF
TCGA	LOF	2	48027685	AT	A	<i>MSH6</i>	NM_000179	p.I855fs	singleton
ClinVar	LOF	2	48027690	TGATT	T	<i>MSH6</i>	NM_000179	p.D857fs	NA
TCGA	MISS	2	48026417	T	G	<i>MSH6</i>	NM_0001792	p.F432C	singleton
ClinVar	MISS	2	48026417	T	C	<i>MSH6</i>	NM_0001792	p.F432S	NA
ClinVar	MISS	2	48026418	T	G	<i>MSH6</i>	NM_0001792	p.F432L	NA
TCGA	LOF	3	37092010	A	T	<i>MLH1</i>	NM_000249	p.K713X	singleton
ClinVar	LOF	3	37092009	G	A	<i>MLH1</i>	NM_000249	p.W712X	NA
TCGA	MISS	2	47702251	C	G	<i>MSH2</i>	NM_000251	p.P616R	4.12E-05

Table B.5: Modeling a Germline:Somatic Interaction for L-MMR Genes. P-values, β estimates, and standard errors determined using a linear model.

Gene	Alteration	p value	Beta	Std. Error	t value
<i>MLH1</i>	Germ	0.043	0.197	0.097	2.018
<i>MLH1</i>	Som	$3.54e^{-16}$	0.223	0.027	8.180
<i>MLH1</i>	Meth	$1.91e^{-15}$	0.136	0.017	7.972
<i>MLH1</i>	Germ * Som	0.488	-0.119	0.171	-0.693
<i>MSH2</i>	Germ	0.201	0.181	0.142	1.276
<i>MSH2</i>	Som	$6.10e^{-09}$	0.15006	0.025	5.824
<i>MSH2</i>	Meth	0.036	0.0443	0.021	2.094
<i>MSH2</i>	Germ * Som	NA	NA	NA	NA
<i>MSH5</i>	Germ	0.138	-0.0689	0.046	-1.482
<i>MSH5</i>	Som	$1.71e^{-09}$	0.234	0.038	6.034
<i>MSH5</i>	Meth	0.019	-0.0501	0.021	-2.345
<i>MSH5</i>	Germ * Som	$9.29e^{-4}$	0.502	0.151	3.312
<i>MSH6</i>	Germ	0.523	0.0896	0.140	0.638
<i>MSH6</i>	Som	0.019	0.0401	0.030	1.300
<i>MSH6</i>	Meth	0.161	0.0289	0.020	1.400
<i>MSH6</i>	Germ * Som	0.990	-0.00250	0.200	-0.012
<i>PMS2</i>	Germ	0.014	0.241	0.098	2.44
<i>PMS2</i>	Som	$03.33e^{-3}$	0.0797	0.037	2.128
<i>PMS2</i>	Meth	0.841	0.00514	0.025	0.200
<i>PMS2</i>	Germ * Som	0.091	0.2440	0.1445	1.688

Table B.6: Association Between MSI and MMR Germline:Somatic Alteration. P-values, β estimates, and standard errors determined using a linear model.

MMR Category	p value	Beta	Std. Error	t value
NONE	$5.13e^{-05}$	-0.0224	0.005	-4.053
GERM	0.738	-0.007	0.022	-0.335
SOM	$5.70e^{-16}$	0.089	0.011	8.122
MISS	$1.01e^{-10}$	0.446	0.068	6.479
LOF	$2.78e^{-15}$	0.445	0.056	7.925

Table B.7: Association Between Age and MMR Germline:Somatic Alteration. P-values, β estimates, and standard errors determined using a linear model.

MMR Category	p value	Beta	Std. Error	t value
NONE	0.250	-0.388	0.337	-1.149
GERM	0.315	-2.018	2.008	-1.004
SOM	0.151	-1.34	0.934	-1.435
MISS	0.949	-0.398	6.232	-0.063
LOF	$1.12e^{-03}$	-16.603	5.093	-3.259

Table B.8: Association Between MSI and Mono-allelic Germline MMR Variants. P-values, β estimates, and standard errors determined using a linear model.

LOF			
p value	Beta	Std. Error	t value
0.337	0.021	0.021	0.959
CLINVAR + LOF			
p value	Beta	Std. Error	t value
0.125	0.020	0.013	1.913
CADD 30 + LOF			
p value	Beta	Std. Error	t value
0.831	0.001	0.006	0.212

Table B.9: Association Between MSI and MMR Alteration Including Confounders. MSI linear model results using unfiltered somatic mutations and with germline:somatic MMR alteration carriers included. P-values, β estimates, and standard errors determined using a linear model.

Perturbation	Gene	All Samples, MSI filter				All Samples, no MSI filter				Lynch-like samples removed, MSI filter			
		p value	Beta	Std. Error	t value	p value	Beta	Std. Error	t value	p value	Beta	Std. Error	t value
Germ	<i>PMS2</i>	8.82e ⁻⁰⁷	0.355	0.072	4.922	8.19e ⁻⁰⁶	0.318	0.071	4.464	0.830	-0.029	0.138	-0.214
Som	<i>MDC1</i>	4.08e ⁻⁰⁸	0.177	0.032	5.495	3.61e ⁻¹¹	0.194	0.029	6.634	7.13e ⁻⁰⁵	0.133	0.033	3.973
Som	<i>MSH2</i>	6.10e ⁻⁰⁹	0.150	0.025	5.824	4.22e ⁻¹²	0.167	0.024	6.946	2.77e ⁻⁰⁷	0.133	0.026	4.963
Som	<i>MSH5</i>	1.45e ⁻¹²	0.266	0.037	7.097	1.45e ⁻¹²	0.266	0.037	7.097	7.93e ⁻¹⁰	0.236	0.038	6.157
Som	<i>MLH1</i>	3.73e ⁻¹⁶	0.220	0.026	8.174	1.02e ⁻¹⁹	0.226	0.024	9.125	6.29e ⁻¹⁷	0.226	0.027	8.391
Meth	<i>MLH1</i>	1.85e ⁻¹⁵	0.136	0.017	7.976	2.49e ⁻¹⁴	0.130	0.017	7.645	6.60e ⁻¹⁶	0.136	0.016	8.101
Meth	<i>SHPRH</i>	5.82e ⁻¹⁶	0.161	0.019	8.120	2.50e ⁻¹⁴	0.15	0.019	7.644	1.19e ⁻¹⁶	0.163	0.019	8.309

Table B.10: Association Between MSI and Alterations Correlated With *SHPRH* Methylation. P-values, β estimates, and standard errors determined using a linear model.

Perturbation	Gene	p value	Beta	Std. Error	t value
Meth	<i>MLH3</i>	0.949	-0.001	0.019	-0.064
Meth	<i>PCNA</i>	0.0447	0.036	0.018	2.007
Meth	<i>POLD3</i>	0.0400	-0.069	0.033	-2.054
Meth	<i>RFC1</i>	0.223	0.027	0.022	1.218
Meth	<i>RFC5</i>	0.339	0.016	0.016	0.956
Meth	<i>RPA3</i>	0.155	0.025	0.017	1.423
Meth	<i>SHPRH</i>	$3.61e^{-16}$	0.164	0.020	8.180
Som	<i>PMS2</i>	0.0101	0.098	0.038	2.574
Som	<i>MLH3</i>	0.0570	0.060	0.031	1.903

Table B.11: Association Between MSI and *SHPRH* Expression. P-values, β estimates, and standard errors determined using a linear model.

	p value	Beta	Std. Error	t value
<i>SHPRH</i> expression	0.000152	$-4.487e^{-05}$	$1.184e^{-05}$	-3.791

Table B.12: Association Between Age and Known Predisposing Germline Variants. Association between MMR, *BRCA1/2*, *SDHB/RET*, and *TP53* germline variant carrier status and age of diagnosis. P-values, β estimates, and standard errors determined using a linear model.

Gene Set	Cancer Type	p value	Beta	Std. Error	t value
MMR	Expected	0.011	-9.995	3.950	-2.530
<i>BRCA1/2</i>	Expected	0.030	-5.939	2.745	-2.163
<i>BRCA1/2</i> NON	Other	0.455	-1.697	2.275	-0.746
<i>TP53</i> CANCEERTYPE	Expected	0.073	-6.198	3.466	-1.788
<i>TP53</i> NON	Other	0.513	-1.920	2.935	-0.654
<i>SDHB/RET</i> CANCEERTYPE	Expected	0.012	-7.738	3.085	-2.508
<i>SDHB/RET</i> NON	Other	0.780	-1.096	3.934	-0.279

Table B.13: Association Between Age and Predicted Predisposing Germline Variants. P-values, β estimates, and standard errors determined using a linear model.

Gene Set	p value	Beta	Std. Error	t value
Other Predisposition	0.271	-0.541	0.492	-1.099
Known	4.52e ⁻⁰⁵	-4.710	1.154	-4.081

Appendix C

Supplemental Material: Rare Variant Phasing Using Paired Tumor:Normal Sequence Data

This appendix contains supplemental figures for the work described in Chapter 3 of this document: "Rare Variant Phasing Using Paired Tumor:Normal Sequence Data". Each figure is referred to in the main text of the chapter and a brief description of each figure is given here.

A = SNV


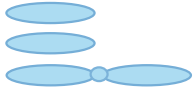





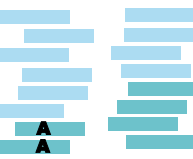

	A	B	C
	Normal	Amp	Del
Maternal			
Paternal			
Sequencing Reads			
Ratio M:P	0.5 0.5	0.75 0.5	1 0.5
SNV VAF	0.5	0.25	0

Figure C.1: Schematic of Δ VAF Changes in Cancer. In regions of SCNA, somatic sequencing reads are skewed toward the chromosomes that are physically more abundant in the sample. Here we illustrate the maternal chromosome in blue and the paternal in green. (A) The expectation in diploid regions is that the sequencing reads will originate from the maternal and paternal chromosome in a 0.5 ratio and that the VAF of heterozygous SNVs will be 0.5. (B) The expectation in regions of amplification is that sequencing reads will be skewed toward the amplified chromosome and that the VAF of heterozygous germline SNVs will change from 0.5. In this illustration the SNV lies on the non-amplified chromosome and therefore the VAF decreases to 0.25. (C) The expectation in regions of deletion is that sequencing reads will be skewed toward the non-deleted chromosome. In this illustration the SNV lies on the deleted chromosome and therefore the VAF decreases to 0.

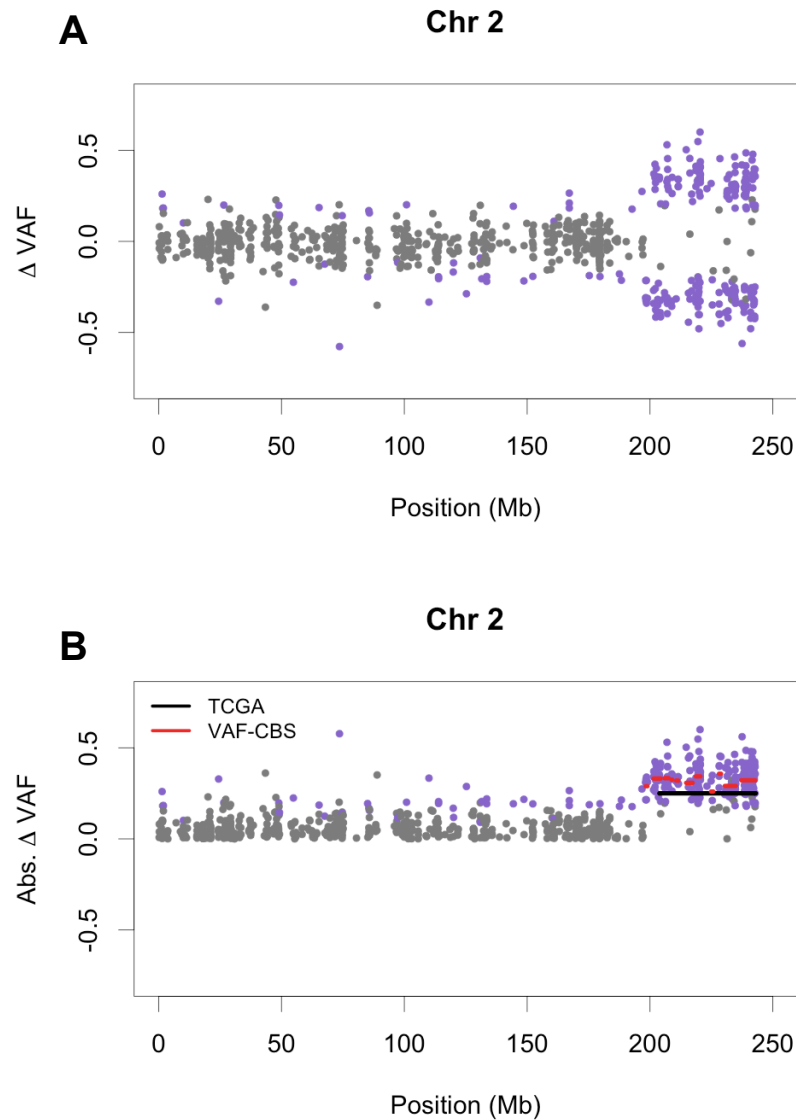


Figure C.2: Example Δ VAF Data. Δ VAF changes in SCNA regions for chromosome 2 of sample TCGA-Y8-A8RZ. (A) Δ VAF for 812 germline heterozygous variants. (B) Absolute Δ VAF for the same variants as (A). SCNA segments identified using TCGA data are shown as black lines, SCNA segments identified using VAF-CBS are shown in red. Color indicates p-value obtained from a Fisher's exact test on tumor and normal read counts: gray: $p \geq 0.05$, purple: $p < 0.05$.

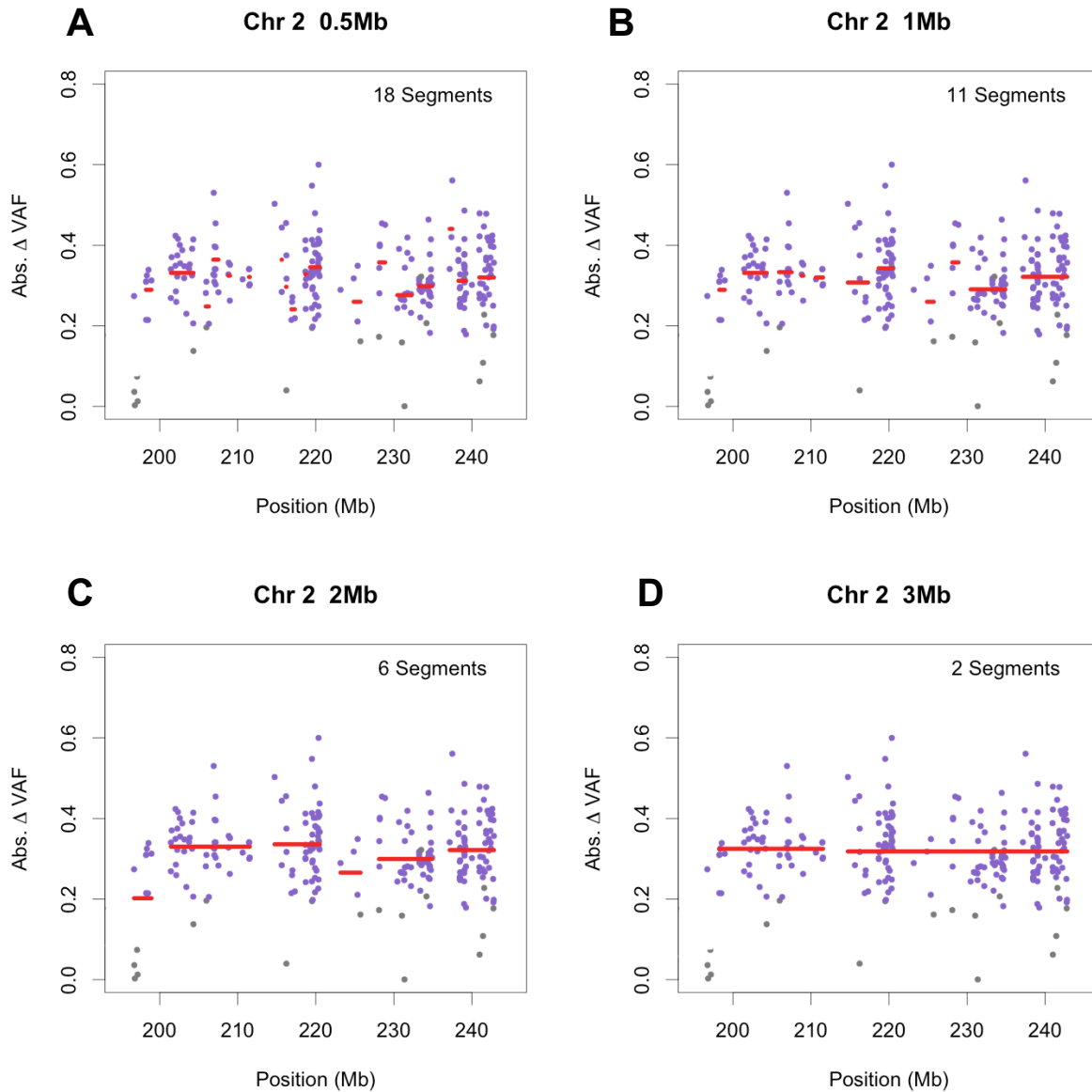


Figure C.3: Example VCF-CBS Data. VAF-CBS segments identified using different smoothing parameters for a SCNA region in chromosome 2 of sample TCGA-Y8-A8RZ. VAF-CBS segments are shown as red lines and the total number of segments in the region is shown in the upper right corner. Color indicates Fisher exact test p-value using tumor and normal read counts: gray: $p \geq 0.05$, purple: $p < 0.05$. Four different smoothing parameters were tested: (A) 0.5 megabases (Mb), (B) 1 Mb, (C) 2 Mb, (D) 3 Mb.

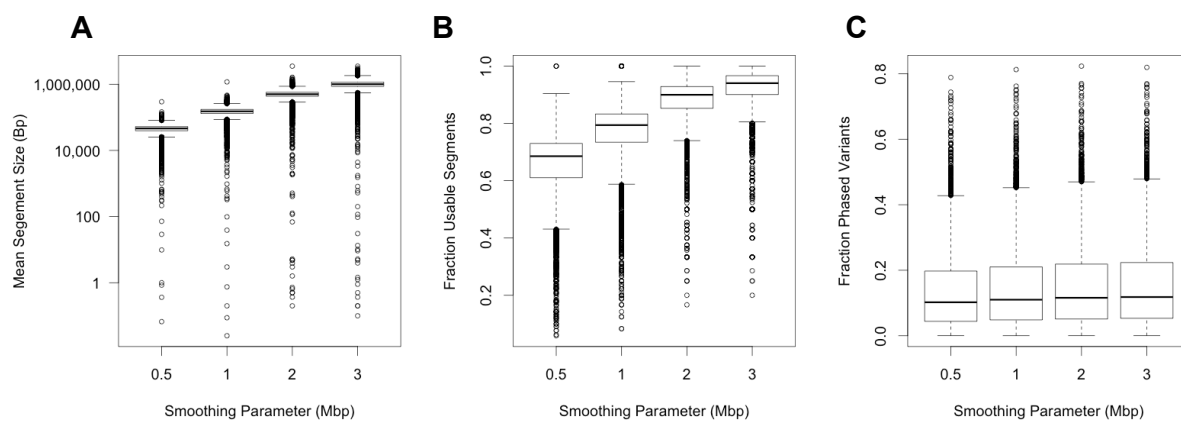


Figure C.4: VAF-CBS Segment Metrics. SCNA segment metrics from $n = 6,158$ samples using four different VAF-CBS smoothing parameters: 0.5 Mb, 1 Mb, 2 Mb, and 3 Mb. (A) Mean size in base pairs of segments identified. (B) Fraction of segments identified that contain more than one heterozygous variant. (C) Fraction of all germline heterozygous variants that can be phased.

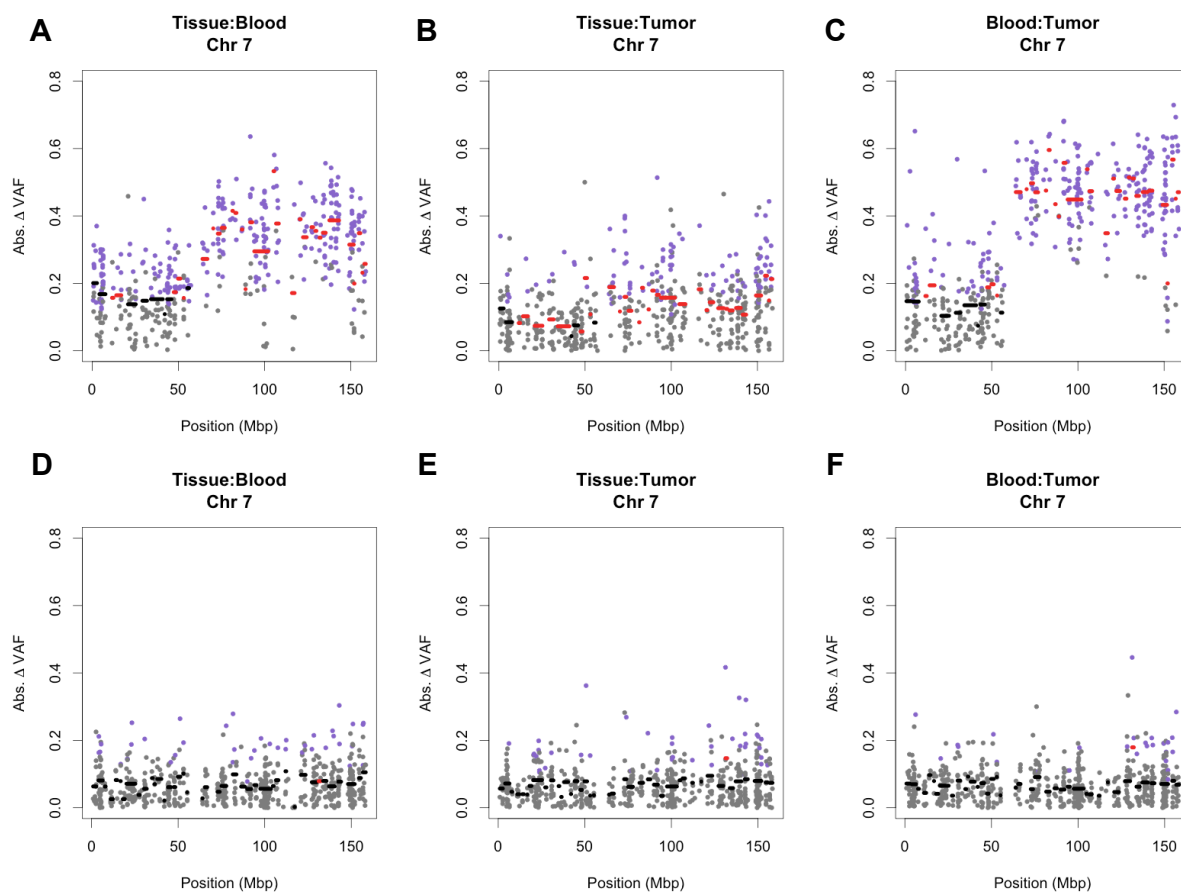


Figure C.5: Δ VAF Data From a Contaminated Sample. Example data illustrating absolute Δ VAF changes in (A) paired normal tissue:normal blood, (B) normal tissue:tumor, and (C) normal blood:tumor for chromosome 7 of sample TCGA-V5-AASX. Example data illustrating absolute Δ VAF changes in (D) paired normal tissue:normal blood, (E) normal tissue:tumor, and (F) normal blood:tumor for chromosome 7 of sample TCGA-Y8-A8RZ. Color indicates p-value obtained from a Fisher's exact test on tumor and normal read counts: gray: $p \geq 0.05$, purple: $p < 0.05$. Segments with an absolute Δ VAF ≥ 0.14 are red.

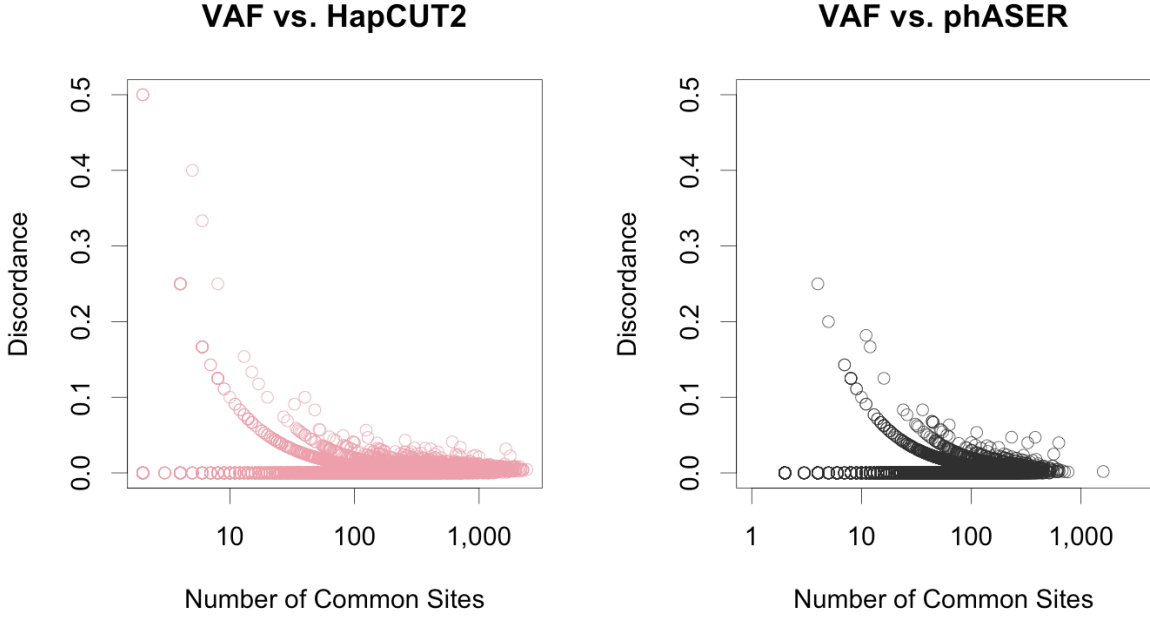


Figure C.6: Discordance Between Phasing Methods. (A) Discordance between VAF phasing and HapCUT2 for $n = 6,180$ samples. Number of common sites are number of sites phased by both methods. (B) Discordance between VAF phasing and phASER for $n = 6,180$ samples.

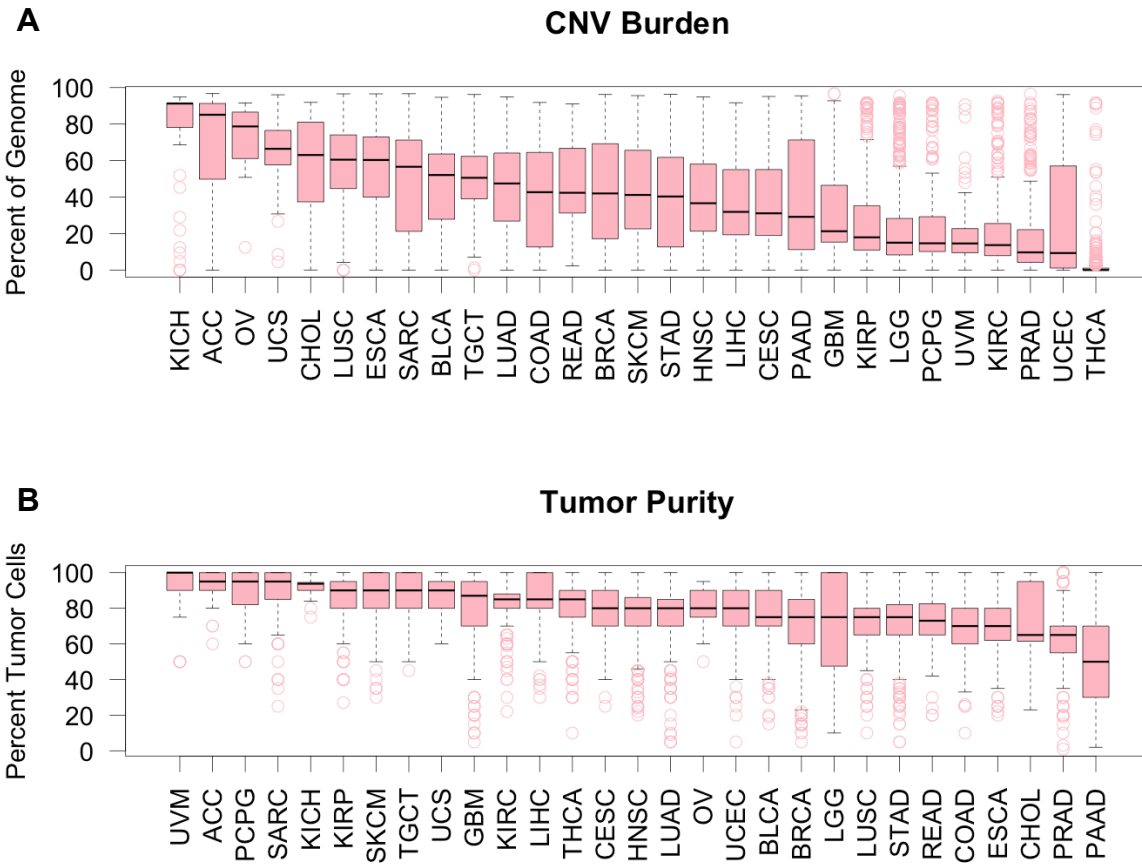


Figure C.7: Sample Metrics that Affect Δ VAF Phasing. (A) Percent of the genome involved in a SCNA event for $n = 8,542$ TCGA samples across 29 cancer types. Fraction of the genome calculated as: total length in base pairs of TCGA SCNA regions (predicted FC > 1.1 or < 0.9) / $3e^{09}$ (B) Percent tumor nuclei for the same samples as (A). Percent tumor nuclei values were obtained from TCGA biospecimen histology slide data. Samples with missing values were imputed to cancer type median.

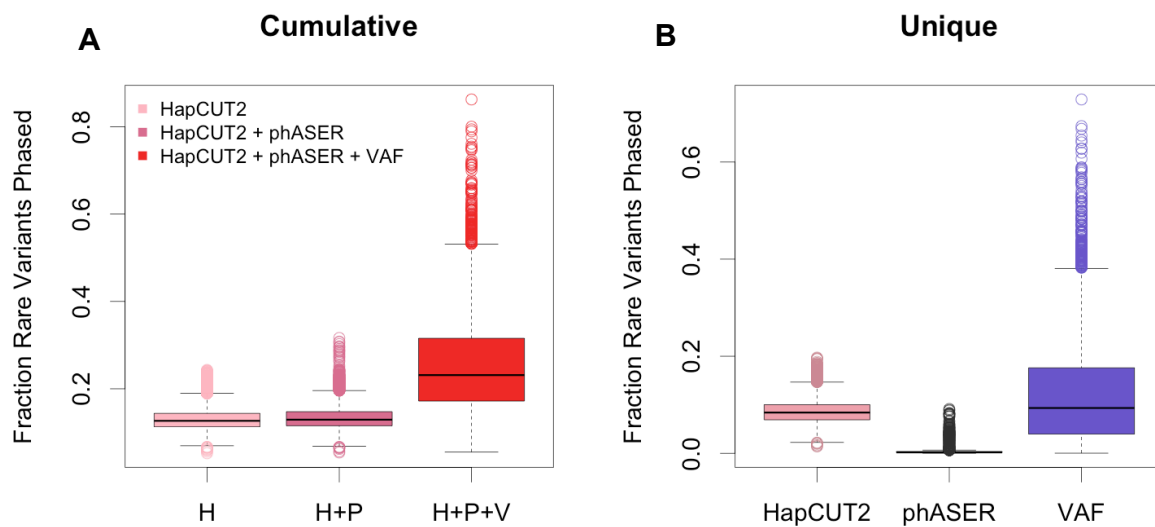


Figure C.8: Phasing Performance on Rare Variants. (A) The fraction of rare (allele frequency ≤ 0.01) germline heterozygous variants phased by HapCUT2 alone, HapCUT2 and phASER, and by HapCUT2, phASER, and VAF in $n = 6,180$ samples. The addition of VAF phasing increased the number of phased variants by 98%. (B) The fraction of rare germline variants phased that are unique to each method.

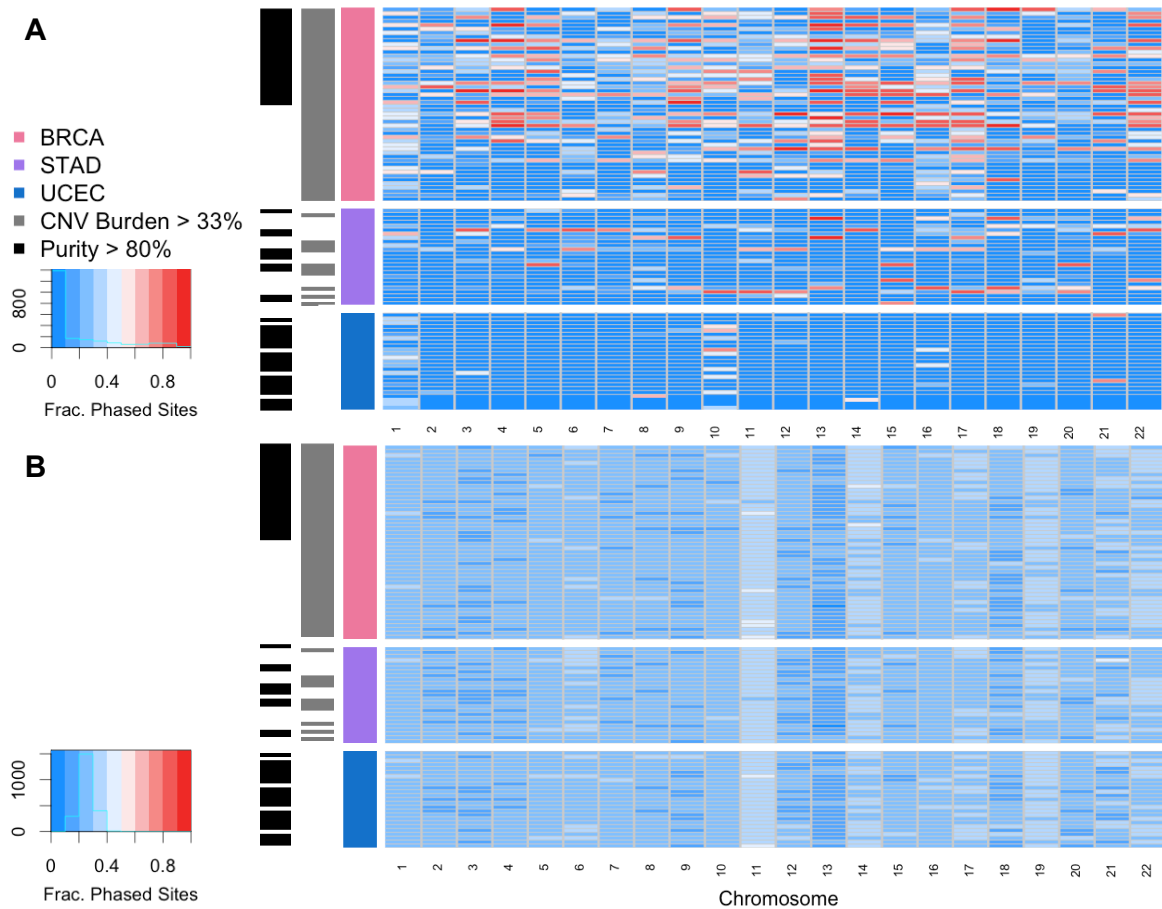


Figure C.9: Fraction of Phased Variants Visualized by Chromosome. (A) Fraction of heterozygous germline variants phased using VAF phasing for $n = 100$ samples. Cancer type is indicated by color: BRCA = Breast Invasive Carcinoma, STAD = Stomach Adenocarcinoma, UCEC = Uterine Corpus Endometrial Carcinoma. Samples with more than 33% of the genome involved in an SCNA are indicated by gray bars. Samples with a histology slide based purity $> 80\%$ are indicated by black bars. (B) Fraction of heterozygous germline variants phased using HapCUT2 for the same samples as (A).

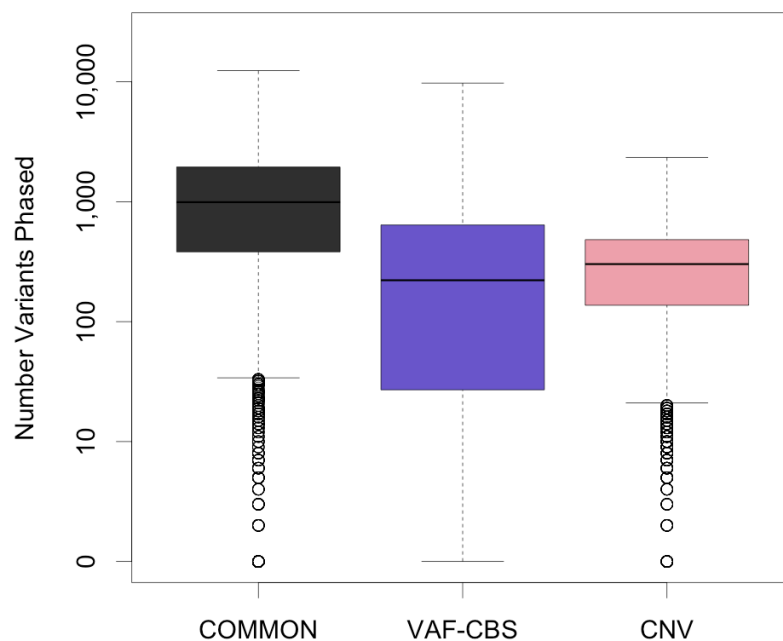


Figure C.10: Comparison Between TCGA and VAF-CBS SCNAs. Comparing number of variants phased using TCGA SCNA segments vs. VAF-CBS segmentation. COMMON = variants phased by both methods, VAF-CBS = variants phased only using VAF-CBS segmentation, CNV = variants phased only using TCGA CNV calls. A median 63% of variants were phased by both methods in $n = 6,180$ samples.

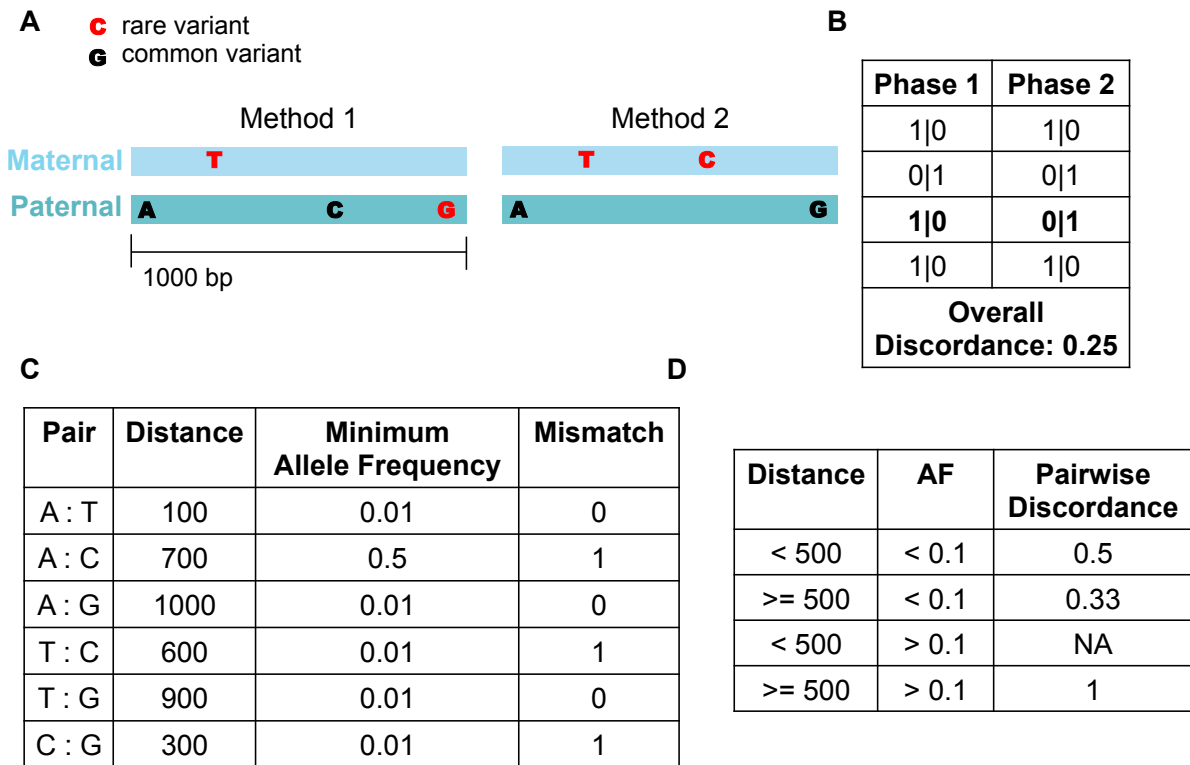


Figure C.11: Method Used to Calculate Pairwise Error. (A) Example of a shared phase segment showing phase calls from two methods. Rare variants (allele frequency ≤ 0.01) are shown in red. (B) The overall discordance was calculated as the fraction of discordant phase calls within a block. (C) Table showing all pairwise phase pairs from the segment in (A). Distance between pairs is calculated as distance in base pairs between the variants. Minimum allele frequency is the smaller allele frequency of the two variants. Error is a binary variable that indicated whether the two variants are in the same orientation. (D) Example showing how pairwise error was binned by distance and allele frequency as in Figure 4.3. For each category the mean error was calculated.

VAF vs. 10X Genomics

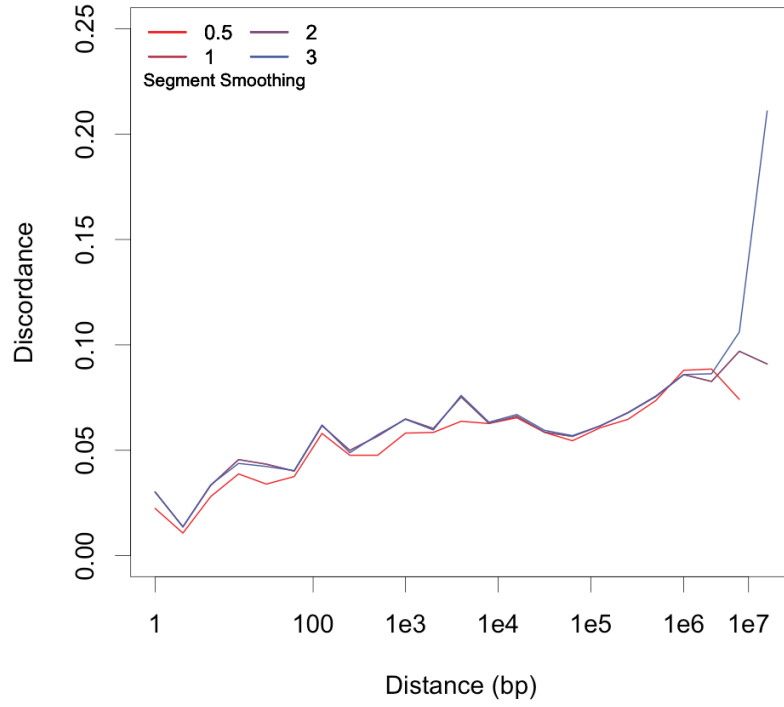


Figure C.12: Discordance Between VAF and 10X Genomics Phasing. Pairwise discordance for the COLO829 cell line as a function of distance and allele frequency. Colors represent different values of the smoothing parameter.

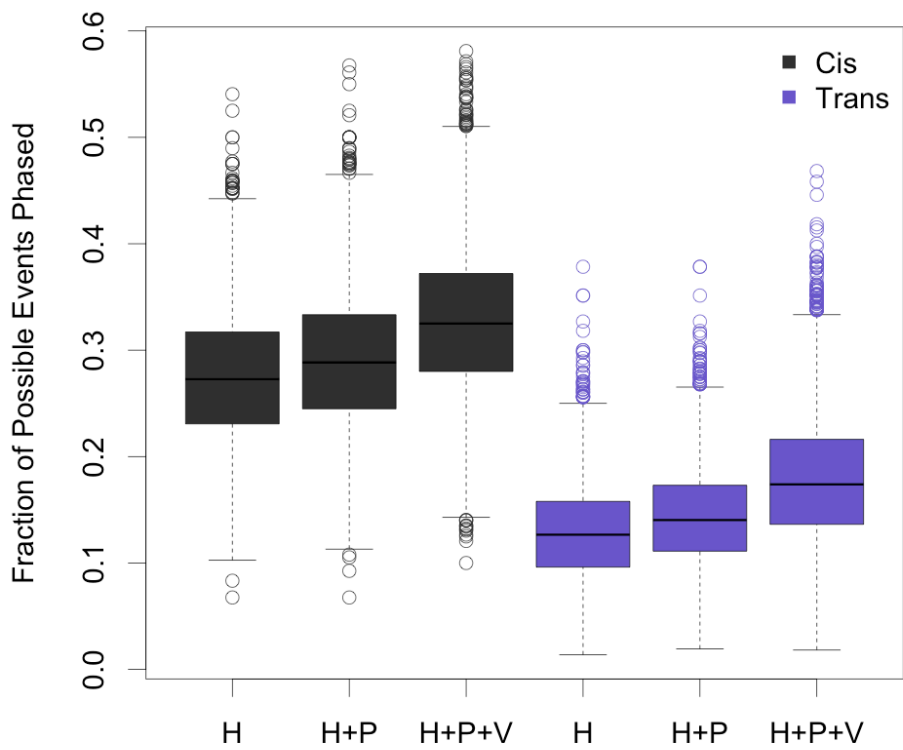


Figure C.13: Fraction of Compound Heterozygosity Events Phased]. Fraction was calculated for each individual as: the number of genes with multiple CADD ≥ 15 variants phased / the number of genes with multiple CADD ≥ 15 germline variants. H = HapCUT2 only, H+P = HapCUT2 and phASER, H+P+V = HapCUT2, phASER, and VAF phasing.

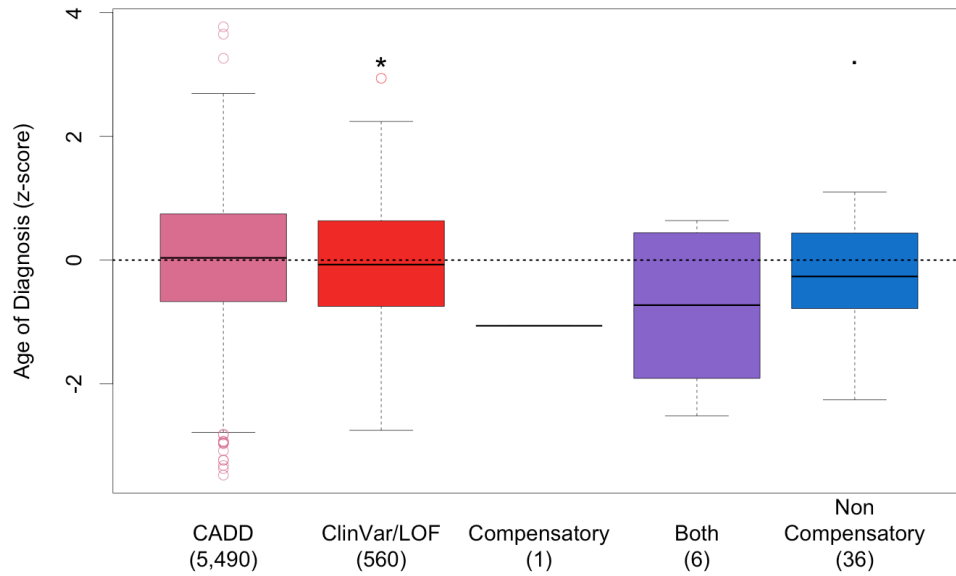


Figure C.14: Association Between Age and Damaging Germline Variants. Age of cancer diagnosis Z-score in $n = 6,093$ TCGA individuals grouped by type of germline alteration in a set of 144 cancer predisposition genes. Groups are the same as Figure 4.4B with the exception of six individuals carrying both a non-compensatory variant set and a ClinVar/LOF variant that are grouped separately (Both). * = $p < 0.05$, . = $p < 0.1$; p-values were determined using a linear model to predict age of diagnosis while accounting for cancer type.

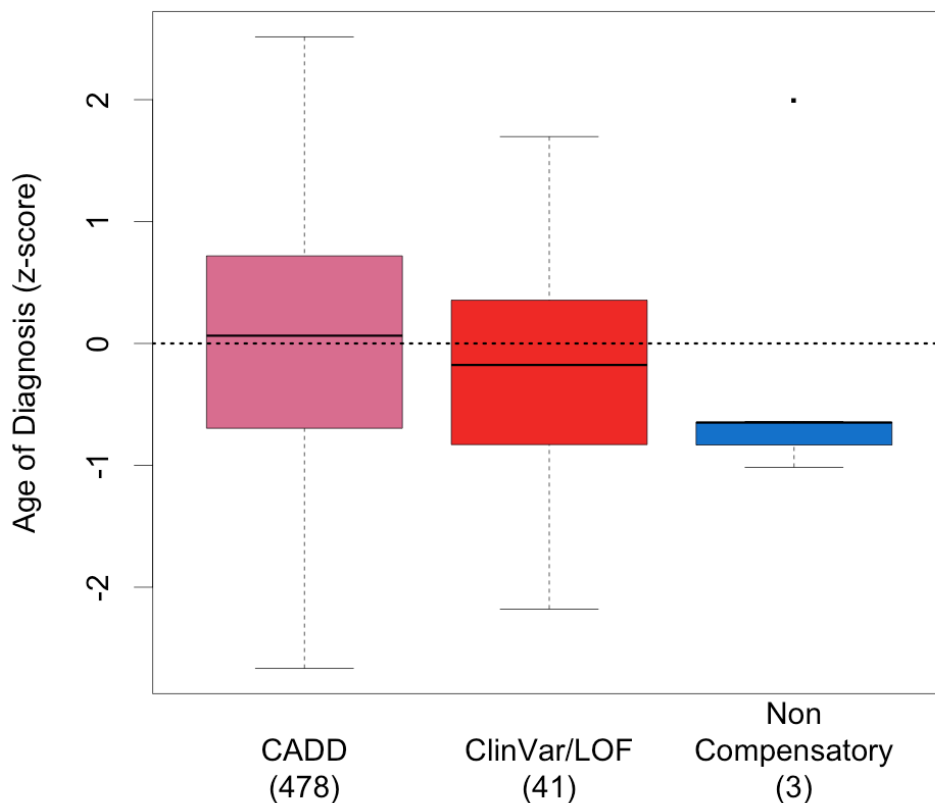


Figure C.15: Association Between Age and *BRCA1/2* Germline Variants. Age of cancer diagnosis Z-score in 6,093 TCGA individuals grouped by type of germline alteration in *BRCA1* or *BRCA2*. (B) Individuals were grouped using HMMvar cis variant scores: CADD = individuals carrying a germline variant with a CADD score ≥ 15 , ClinVar/LOF = individuals carrying a ClinVar pathogenic or LOF germline variant, Non-Compensatory = individuals carrying multiple nonsynonymous variants in a gene predicted to be more deleterious collectively than independently. The number of samples is shown in parentheses. * = $p < 0.05$, . = $p < 0.1$; p-values were determined using a linear model to predict age of diagnosis while accounting for cancer type.

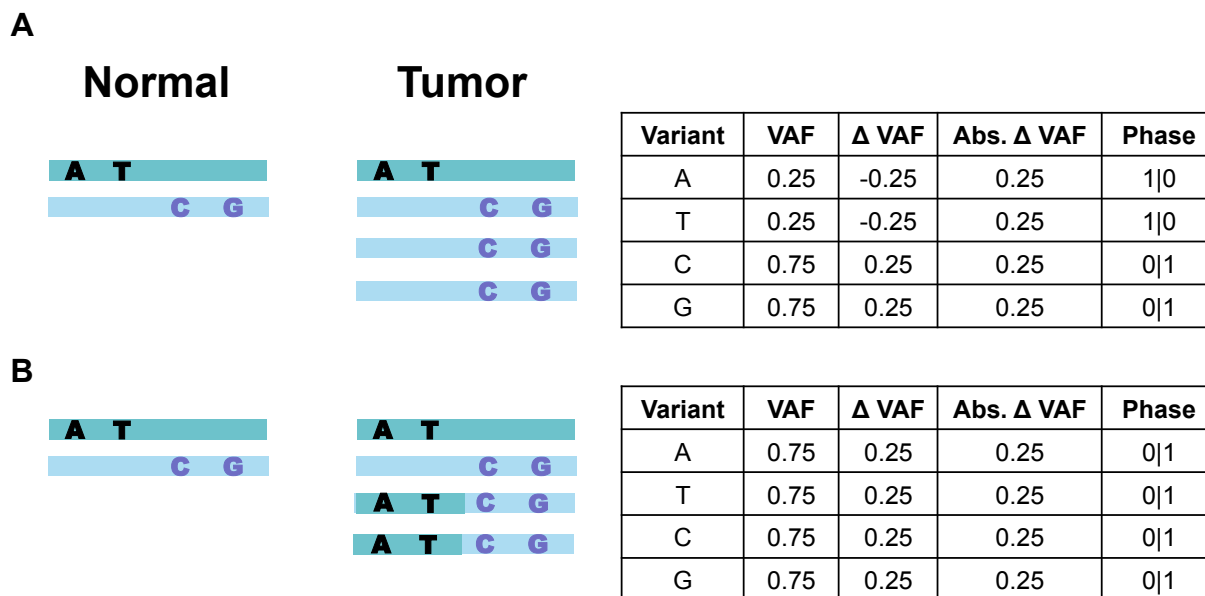


Figure C.16: Assumptions of VAF Phasing Model. VAF phasing assumes that SCNA segments originate from a single homologous chromosome. (A) Under this assumption, Δ VAF can be used to correctly phase variants. (B) Should a SCNA segment result from equal amplification of both homologous chromosomes, using Δ VAF to phase will result in switch errors. In this example, amplification of both homologous chromosomes results in all variants being assigned to the same chromosome incorrectly.

Table C.1: Possible Contaminated Normal Tissue Samples. TCGA barcodes of suspected contaminated normal tissue samples.

Contaminated Normal Sample Barcodes
TCGA-4G-AAZT
TCGA-BH-A0BQ
TCGA-BH-AB28
TCGA-BR-6710
TCGA-G9-6362
TCGA-V5-A7RE
TCGA-V5-AASX

Table C.2: Discordance Between VAF Phasing and Other Methods.
Null model refers to VAF phasing using duplicated samples directly to determine significant VAF-CBS segments, cutoff refers to using a absolute Δ VAF cutoff of 0.14 to determine significant VAF-CBS segments.

Phase Comparison	Smoothing Parameter (Mb)	Region Specific				Cutoff			
		Mean Disc.	Std. Error	Mean Num. Sites	Std. Error	Mean Disc.	Std. Error	Mean Num. Sites	Std. Error
VAF : HapCUT2	5.00e ⁺⁰⁵	0.00321	1.04e ⁻⁰⁴	374.13	4.499	0.00740	2.65e ⁻⁰⁴	388.25	4.638
VAF : HapCUT2	1.00e ⁺⁰⁶	0.00358	9.78e ⁻⁰⁵	386.04	4.596	0.00654	2.64e ⁻⁰⁴	383.62	4.636
VAF : HapCUT2	2.00e ⁺⁰⁶	0.00374	9.76e ⁻⁰⁵	395.21	4.655	0.00517	1.85e ⁻⁰⁴	381.066	4.632
VAF : HapCUT2	3.00e ⁺⁰⁶	0.00400	1.32e ⁻⁰⁴	400.03	4.679	0.00513	2.23e ⁻⁰⁴	379.66	4.629
VAF : phASER	5.00e ⁺⁰⁵	0.00237	1.38e ⁻⁰⁴	91.63	1.201	0.00406	1.53e ⁻⁰⁴	97.367	1.249
VAF : phASER	1.00e ⁺⁰⁶	0.00269	1.55e ⁻⁰⁴	95.38	1.233	0.00395	1.50e ⁻⁰⁴	96.302	1.251
VAF : phASER	2.00e ⁺⁰⁶	0.00289	1.04e ⁻⁰⁴	97.84	1.250	0.00371	1.42e ⁻⁰⁴	95.527	1.250
VAF : phASER	3.00e ⁺⁰⁶	0.00302	1.10e ⁻⁰⁴	98.81	1.257	0.00388	1.62e ⁻⁰⁴	95.223	1.251

Table C.3: Factors That Influence VAF Phasing. Factors that influence number of variants phased per sample by VAF phasing. P-values, β estimates, and standard errors determined using a linear model.

	Beta	Std. Error	t value	P
Purity	0.00665	0.000297	22.339	$2.60e^{-106}$
CNV Burden	0.0122	0.000201	60.676	0.0
Tumor Depth	-0.000305	0.000712	-0.428	0.669
Normal Depth	0.000504	0.000620	0.813	0.416
Depth Ratio	0.000675	0.0504	0.0134	0.989

Table C.4: Discordance Between VAF Phasing with TCGA SCNA and Other Methods. Mean discordance between VAF phasing, HapCUT2 and phASER in $n = 6,180$ samples either using VAF-CBS segments or TCGA SCNA calls.

Segmentation	Phase Comparison	Mean Discordance	Std. Error
VAF-CBS	VAF : HapCUT2	0.0062	$4.92e^{-04}$
VAF-CBS	VAF : phASER	0.0029	$2.37e^{-04}$
TCGA CNV	VAF : HapCUT2	0.0321	$5.89e^{-04}$
TCGA CNV	VAF : phASER	0.0173	$6.10e^{-04}$

Table C.5: Discordance Between VAF and 10X Genomics Phasing.

Smoothing Parameter (Mb)	Fraction Variants Phased	Number Variants Phased	Number of Errors	Discordance
5.00E+05	0.483	17588	656	0.037298158
1.00E+06	0.499	18175	705	0.038789546
2.00E+06	0.498	18161	705	0.038819448
3.00E+06	0.504	18366	822	0.044756615

Table C.6: Features of VAF Phasing Errors. Features of incorrect VAF phasing pairs from COLO829. Features were calculated for all possible pairs of phased heterozygous germline variants from COLO829 (see Methods).

Feature	Correct Pairs Mean	Incorrect Pairs Mean	% Difference (correct - incorrect)	Wilcox p-value
Minimum Read Depth	226.9916	104.1111	-118.0283	$< 2.2e^{-16}$
Segment Size	15844026	16066431	1.384282	$< 2.2e^{-16}$
Segment Abs. Δ VAF	0.1633855	0.1629823	-0.2474177	0.0004492
$\Delta\Delta$ VAF	0.1269109	0.06417217	-97.76631	$< 2.2e^{-16}$
Pair Distance	1208129	1369051	11.75427	$< 2.2e^{-16}$
Δ Allele Frequency	0.2497401	0.2491799	-0.2248064	0.5987
Minimum Allele Frequency	0.2221392	0.2243572	0.9885961	$5.84e^{-09}$

Table C.7: Association Between Age and Compound Heterozygosity. P-values, β estimates, and standard errors determined using a linear model.

	Beta	Std. Error	t value	p
ClinVar/LOF	-1.33	0.565	-2.352	0.0187
Cis	-0.812	1.11	-0.731	0.464
Trans	1.64	1.35	1.212	0.225

Table C.8: Association Between Age and Non-Compensatory Variants. P-values, β estimates, and standard errors determined using a linear model.

	Beta	Std. Error	t value	p
ClinVar/LOF	-1.25	0.5.46	-2.288	0.0221
compensatory	-12.0	12.3	-0.978	0.328
non compensatory	-4.34	1.90	-2.281	0.0225

Table C.9: Association Between Age and Non-Compensatory Variants, ClinVar Samples Removed. P-values, β estimates, and standard errors determined using a linear model.

	Beta	Std. Error	t value	p
ClinVar/LOF	-1.250	0.545	-2.29	0.0220
Compensatory	-12.002	12.268	-0.978	0.327
Both	-8.128	5.013	-1.621	0.105
Non-Compensatory	-3.710	2.054	-1.806	0.071

Table C.10: Association Between Age and Non-Compensatory Variants in *BRCA1/2*. P-values, β estimates, and standard errors determined using a linear model.

	Beta	Std. Error	t value	p
ClinVar/LOF	-1.199	1.975	-0.607	0.544
non compensatory	-12.88	6.99	-1.843	0.065

Table C.11: Germline Non-Compensatory Variants in *BRCA1/2*

Indv	Cancer	Chr	Pos	Ref	Alt	AA	Variant	CADD	ExAC AF	ClinVar	Detail
A	KIRC	17	41246481	T	C	p.Q309R	Miss	9.735	0.04407	B	
A	KIRC	17	41246662	T	C	p.R249G	Miss	16.08	8.24E-06	conflicting	likely benign 1, VUS 2
B	UCS	17	41245071	G	T	p.T779K	Miss	11.61	0.0001813	B	
B	UCS	17	41246662	T	C	p.R249G	Miss	9.735	0.04407	B	
C	HNSC	13	32906729	A	C	p.N372H	Miss	6.228	0.2779	B	
C	HNSC	13	32911278	T	C	p.L929S	Miss	8.566	0.0008861	B	
C	HNSC	13	32911452	A	T	p.N987I	Miss	9.104	0.0008728	B	

Bibliography

- [1] Eric R. Fearon and Bert Vogelstein. A Genetic Model for Colorectal tumorigenesis. *Cell*, 61:759–767, 1990.
- [2] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 340(6127):1546–1558, 2013.
- [3] Siân Jones, Valsamo Anagnostou, Karli Lytle, Sonya Parpart-Li, Monica Nesselbush, David R Riley, Manish Shukla, Bryan Chesnick, Maura Kadan, Eniko Papp, Kevin G Galens, Derek Murphy, Theresa Zhang, Lisa Kann, Mark Sausen, Samuel V Angiuoli, Luis A Diaz, and Victor E Velculescu. Personalized genomic analyses for cancer mutation discovery and interpretation. *Science Translational Medicine*, 7(283), 2015.
- [4] Divyansh Agarwal, Christoph Nowak, Nancy R Zhang, Lajos Pusztai, and Christos Hatzis. Functional germline variants as potential co-oncogenes. *npj Breast Cancer*, 3(1):46, 2017.
- [5] Heather S Feigelson, Katrina A.B. Goddard, Celine Hollombe, Sharna R Tingle, Elizabeth M Gillanders, Leah E Mechanic, and Stefanie A Nelson. Approaches to integrating germline and tumor genomic data in cancer research. *Carcinogenesis*, 35(10):2157–2163, 2014.
- [6] Alfred G Knudson. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.
- [7] C. O. Nordling. A new theory on the cancer-inducing mechanism. *British Journal of Cancer*, 7(1):68–72, mar 1953.
- [8] Manuela Santarosa and Alan Ashworth. Haploinsufficiency for tumour suppressor genes: When you don't need to go all the way. *Biochimica et Biophysica Acta - Reviews on Cancer*, 1654(2):105–122, 2004.
- [9] Steven A. Frank. Genetic predisposition to cancer - Insights from population genetics. *Nature Reviews Genetics*, 5(10):764–772, oct 2004.

- [10] Judy E Garber and Kenneth Offit. Hereditary Cancer Predisposition Syndromes. *J Clin Oncol*, 23(2):276–292, 2005.
- [11] Katherine A Rauhen. The RASopathies. *Annual Review of Genomics and Human Genetics*, 14(1):355–369, 2013.
- [12] Cristian Tomasetti, Lu Li, and Bert Vogelstein. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, 355(6331):1330–1334, 2017.
- [13] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.
- [14] Lorelei A. Mucci, Jacob B. Hjelmborg, Jennifer R. Harris, Kamila Czene, David J. Havelick, Thomas Scheike, Rebecca E. Graff, Klaus Holst, Sören Möller, Robert H. Unger, Christina McIntosh, Elizabeth Nuttall, Ingunn Brandt, Kathryn L. Penney, Mikael Hartman, Peter Kraft, Giovanni Parmigiani, Kaare Christensen, Markku Koskenvuo, Niels V. Holm, Kauko Heikkilä, Eero Pukkala, Axel Skyttthe, Hans Olov Adami, Jaakko Kaprio, Julia Isaeva, Thomas Nilsen, Tellervo Korhonen, and Ulrich Halekoh. Familial risk and heritability of cancer among twins in nordic countries. *JAMA - Journal of the American Medical Association*, 315(1):68–76, jan 2016.
- [15] J. Hall, M. Lee, B Newman, J. Morrow, L. Anderson, B Huey, and M. King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689, dec 1990.
- [16] Paul Lichtenstein, Niels V. Holm, Pia K. Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skyttthe, and Kari Hemminki. Environmental and heritable factors in the causation of cancer analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, 343(2):78–85, 2000.
- [17] Nazneen Rahman. Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–308, 2014.
- [18] Kuan-lin Huang, R. Jay Mashl, Yige Wu, Deborah I. Ritter, Jiayin Wang, Clara Oh, Marta Paczkowska, Sheila Reynolds, Matthew A. Wyczalkowski, Ninad Oak, Adam D. Scott, Michal Krassowski, Andrew D. Cherniack, Kathleen E. Houlahan, Reyka Jayasinghe, Liang-Bo Wang, Daniel Cui Zhou, Di Liu, Song Cao, Young Won Kim, Amanda Koire, Joshua F. McMichael, Vishwanathan Huchtagowder, Tae-Beom Kim, Abigail Hahn, Chen Wang, Michael D. McLellan, Fahd Al-Mulla, Kimberly J. Johnson, Olivier Lichtarge, Paul C. Boutros, Benjamin Raphael, Alexander J. Lazar, Wei Zhang, Michael C. Wendl, Ramaswamy Govindan, Sanjay Jain, David Wheeler, Shashikant Kulkarni, John F. Dipersio, Jüri Reimand, Funda Meric-Bernstam, Ken Chen, Ilya Shmulevich, Sharon E. Plon, Feng Chen, Li Ding, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C.

Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M.

Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavoilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliani, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bublely, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Small-

ridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bitá Esmali, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaut, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Anto-

nio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffrey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, and Armaz Mariamidze. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*, 173(2):355–370.e14, apr 2018.

- [19] Sebastian M. Waszak, Grace Tiao, Bin Zhu, Tobias Rausch, Francesc Muiyas, Bernardo Rodriguez-Martin, Raquel Rabionet, Sergei Yakneen, Georgia Escaramis, Yilong Li, Natalie Saini, Steven A. Roberts, German M. Demidov, Esa Pitkanen, Olivier Delaneau, Jose M. Heredia-Genestar, Joachim Weischenfeldt, Suyash S. Shringarpure, Jieming Chen, Hidewaki Nakagawa, Ludmil B. Alexandrov, Oliver Drechsel, L. Jonathan Dursi, Ayellet V. Segre, Erik Garrison, Serap Erkek, Nina Habermann, Lara Urban, Ekta Khurana, Andy Cafferkey, Shuto Hayashi, Seiya Imoto, Lauri A. Aaltonen, Eva G. Alvarez, Adrian Baez-Ortega, Matthew Bailey, Mattia Bosio, Alicia L. Bruzos, Ivo Buchhalter, Carlos D. Bustamante, Claudia Calabrese, Anthony DiBiase, Mark Gerstein, Aliaksei Z. Holik, Xing Hua, Kuan-lin Huang, Ivica Letunic, Leszek J. Klimczak, Roelof Koster, Sushant Kumar, Mike McLellan, Jay Mashl, Lisa Mirabello, Steven Newhouse, Aparna Prasad, Gunnar Raetsch, Matthias Schlesner, Roland Schwarz, Pramod Sharma, Tal Shmaya, Nikos Sidiropoulos, Lei Song, Hana Susak, Tomas Tanskanen, Marta Tojo, David C. Wedge, Mark Wright, Ying Wu, Kai Ye, Venkata D. Yellapantula, Jorge Zamora, Atul J. Butte, Gad Getz, Jared Simpson, Li Ding, Tomas Marques-Bonet, Arcadi Navarro, Alvis Brazma, Peter Campbell, Stephen J. Chanock, Nilanjan Chatterjee, Oliver Stegle, Reiner Siebert, Stephan Ossowski, Olivier Harismendy, Dmitry A. Gordenin, Jose M. C. Tubio, Francisco M. De La Vega, Douglas F. Easton, Xavier Estivill, Jan Korbel, PCAWG Germline Working Group, and ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Net. Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. *bioRxiv*, page 208330, nov 2017.

- [20] Susanne N. Gröbner, Barbara C. Worst, Joachim Weischenfeldt, Ivo Buchhalter, Kortine Kleinheinz, Vasilisa A. Rudneva, Pascal D. Johann, Gnana Prakash Balasubramanian, Maia Segura-Wang, Sebastian Brabetz, Sebastian Bender, Barbara Hutter, Dominik Sturm, Elke Pfaff, Daniel Hübschmann, Gideon Zipprich, Michael Heinold, Jürgen Eils, Christian Lawerenz, Serap Erkek, Sander Lambo, Sebastian Waszak, Claudia Blattmann, Arndt Borkhardt, Michaela Kuhlen, Angelika Eggert, Simone Fulda, Manfred Gessler, Jenny Wegert, Roland Kappler, Daniel Baumhoer, Stefan Burdach, Renate Kirschner-Schwabe, Udo Kontny, Andreas E. Kulozik, Dietmar Lohmann, Simone Hettmer, Cornelia Eckert, Stefan Bielack, Michaela Nathrath, Charlotte Niemeyer, Günther H. Richter, Johannes Schulte, Reiner Siebert, Frank Westermann, Jan J. Molenaar, Gilles Vassal, Hendrik Witt, Marc Zapatka, Birgit Burkhardt, Christian P. Kratz, Olaf Witt, Cornelis M. Van Tilburg, Christof M. Kramm, Gudrun Fleischhack, Uta Dirksen, Stefan Rutkowski, Michael Frühwald, Katja Von Hoff, Stephan Wolf, Thomas Klingebiel, Ewa Koscielniak, Pablo Landgraf, Jan Koster, Adam C. Resnick, Jinghui Zhang, Yanling Liu, Xin Zhou, Angela J. Waanders, Danny A. Zwiijnenburg, Pichai Raman, Benedikt Brors, Ursula D. Weber, Paul A. Northcott, Kristian W. Pajtler, Marcel Kool, Rosario M. Piro, Jan O. Korbel, Matthias Schlesner, Roland Eils, David T.W. Jones, Peter Lichter, Lukas Chavez, and Stefan M. Pfister. The landscape of genomic alterations across childhood cancers. *Nature*, 555(7696):321–327, feb 2018.
- [21] Jinghui Zhang, Michael F. Walsh, Gang Wu, Michael N. Edmonson, Tanja A. Gruber, John Easton, Dale Hedges, Xiaotu Ma, Xin Zhou, Donald A. Yergeau, Mark R. Wilkinson, Bhavin Vadodaria, Xiang Chen, Rose B. McGee, Stacy Hines-Dowell, Regina Nuccio, Emily Quinn, Sheila A. Shurtleff, Michael Rusch, Aman Patel, Jared B. Becksfort, Shuoguo Wang, Meaghann S. Weaver, Li Ding, Elaine R. Mardis, Richard K. Wilson, Amar Gajjar, David W. Ellison, Alberto S. Pappo, Ching-Hon Pui, Kim E. Nichols, and James R. Downing. Germline Mutations in Predisposition Genes in Pediatric Cancer. *New England Journal of Medicine*, 373(24):2336–2346, 2015.
- [22] Douglas F Easton. How many more breast cancer predisposition genes are there? *Breast cancer research*, 1(1):14–7, 1999.
- [23] Amit Sud, Ben Kinnersley, and Richard S Houlston. Genome-wide association studies of cancer: Current insights and future perspectives. *Nature Reviews Cancer*, 17(11):692–704, 2017.
- [24] Paul D.P. Pharoah, Antonis C Antoniou, Douglas F Easton, and Bruce A.J. Ponder. Polygenes, Risk Prediction, and Targeted Prevention of Breast Cancer. *New England Journal of Medicine*, 358(26):2796–2803, 2008.
- [25] Qiyuan Li, Ji-Heui Seo, Barbara Stranger, Aaron McKenna, Itsik Pe’er, Thomas LaFramboise, Myles Brown, Svitlana Tyekucheva, and Matthew L Freedman. Inte-

- grative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. *Cell*, 152(3):633–641, jan 2013.
- [26] Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan, and Nancy J. Cox. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*, 6(4):e1000888, apr 2010.
- [27] Y.-P. Fu, Indu Kohaar, Nathaniel Rothman, Julie Earl, Jonine D Figueroa, Yuanqing Ye, Núria Malats, Wei Tang, Luyang Liu, Montserrat Garcia-Closas, Brian Muchmore, Nilanjan Chatterjee, Mcanthy Tarway, Manolis Kogevinas, Patricia Porter-Gill, Dalsu Baris, Adam Mumy, Demetrius Albanes, Mark P Purdue, Amy Hutchinson, Alfredo Carrato, A. Tardon, Consol Serra, R. Garcia-Closas, Josep Lloreta, Alison Johnson, Molly Schwenn, Margaret R Karagas, Alan Schned, W Ryan Diver, Susan M Gapstur, Michael J Thun, Jarmo Virtamo, Stephen J Chanock, Joseph F Fraumeni, Debra T Silverman, Xifeng Wu, Francisco X Real, and Ludmila Prokunina-Olsson. Common genetic variants in the PSCA gene influence gene expression and bladder cancer risk. *Proceedings of the National Academy of Sciences*, 109(13):4974–4979, 2012.
- [28] Alan M Pittman, Silvia Naranjo, Emily Webb, Peter Broderick, Esther H Lips, Tom Van Wezel, Hans Morreau, Kate Sullivan, Sarah Fielding, Philip Twiss, Jayaram Vijayakrishnan, Fernando Casares, Mobshra Qureshi, José Luis Gómez-Skarmeta, and Richard S Houlston. The colorectal cancer risk at 18q21 is caused by a novel variant altering SMAD7 expression. *Genome Research*, 19(6):987–993, 2009.
- [29] Franklin W. Huang, Eran Hodis, Mary Jue Xu, Gregory V. Kryukov, Lynda Chin, and Levi A. Garraway. Highly recurrent TERT promoter mutations in human melanoma. *Science*, 339(6122):957–959, 2013.
- [30] Nasim Ahmadiyeh, Mark M Pomerantz, Chiara Grisanzio, Paula Herman, Li Jia, Vanessa Almendro, Housheng Hansen He, Myles Brown, Shirley Liu, Matt Davis, Jennifer L Caswell, Christine A Beckwith, Adam Hills, Laura MacConaill, Gerhard A Coetzee, Meredith M Regan, and Matthew L Freedman. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *PNAS*, 107(21):9742–8746, 2010.
- [31] Cristina Bozzao, Patrizia Lastella, and Alessandro Stella. Anticipation in Lynch Syndrome: Where We Are Where We Go. *Current Genomics*, 12(7):451–465, nov 2011.
- [32] Bente A Talseth-Palmer, Juul T Wijnen, Desma M Grice, and Rodney J Scott. Genetic modifiers of cancer risk in Lynch syndrome: A review. *Familial Cancer*, 12(2):207–216, jun 2013.
- [33] Hany Ariffin, Pierre Hainaut, Anna Puzio-Kuter, Soo Sin Choong, Adelyne Sue Li Chan, Denis Tolkunov, Gunaretnam Rajagopal, Wenfeng Kang, Leon Li Wen Lim,

- Shekhar Krishnan, Kok-Siong Chen, Maria Isabel Achatz, Mawar Karsa, Jannah Shamsani, Arnold J Levine, and Chang S Chan. Whole-genome sequencing analysis of phenotypic heterogeneity and anticipation in Li-Fraumeni cancer predisposition syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, 111(43):15497–501, 2014.
- [34] Antonis C Antoniou, Olga M Sinilnikova, Jacques Simard, Mélanie Léoné, Martine Dumont, Susan L Neuhausen, Jeffery P Struewing, Dominique Stoppa-Lyonnet, Laure Barjhoux, David J Hughes, Isabelle Coupier, Muriel Belotti, Christine Lasset, Valérie Bonadona, Yves-Jean Bignon, Timothy R Rebbeck, Theresa Wagner, Henry T Lynch, Susan M Domchek, Katherine L Nathanson, Judy E Garber, Jeffrey Weitzel, Steven A Narod, Gail Tomlinson, Olufunmilayo I Olopade, Andrew Godwin, Claudine Isaacs, Anna Jakubowska, Jan Lubinski, Jacek Gronwald, Bohdan Górski, Tomasz Byrski, Tomasz Huzarski, Susan Peock, Margaret Cook, Caroline Baynes, Alexandra Murray, Mark Rogers, Peter A Daly, Huw Dorkins, Rita K Schmutzler, Beatrix Versmold, Christoph Engel, Alfons Meindl, Norbert Arnold, Dieter Niederacher, Helmut Deissler, Amanda B Spurdle, Xiaoqing Chen, Nicola Waddell, Nicole Cloonan, Tomas Kirchhoff, Kenneth Offit, Eitan Friedman, Bella Kaufmann, Yael Laitman, Gilli Galore, Gad Rennert, Flavio Lejbkowitz, Leon Raskin, Irene L Andrulis, Eduard Ilyushik, Hilmi Ozcelik, Peter Devilee, Maaïke P.G. Vreeswijk, Mark H Greene, Sheila A Prindiville, Ana Osorio, Javier Benítez, Michal Zikan, Csilla I Szabo, Outi Kilpivaara, Heli Nevanlinna, Ute Hamann, Francine Durocher, Adalgeir Arason, Fergus J Couch, Douglas F Easton, and Georgia Chenevix-Trench. RAD51 135G>C Modifies Breast Cancer Risk among BRCA2 Mutation Carriers: Results from a Combined Analysis of 19 Studies. *The American Journal of Human Genetics*, 81(6):1186–1200, 2007.
- [35] Montserrat Garcia-Closas, Nathaniel Rothman, Jonine D Figueroa, Ludmila Prokunina-Olsson, Summer S Han, Dalsu Baris, Eric J. Jacobs, Nuria Malats, Immaculata De Vivo, Demetrius Albanes, Mark P Purdue, Sapna Sharma, Yi Ping Fu, Manolis Kogevinas, Zhaoming Wang, Wei Tang, Adonina Tardon, Consol Serra, Alfredo Carrato, Reina García-Closas, Josep Lloreta, Alison Johnson, Molly Schwenn, Margaret R Karagas, Alan Schned, Gerald Andriole, Robert Grubb, Amanda Black, Susan M Gapstur, Michael Thun, William Ryan Diver, Stephanie J Weinstein, Jarmo Virtamo, David J Hunter, Neil Caporaso, Maria Teresa Landi, Amy Hutchinson, Laurie Burdett, Kevin B Jacobs, Meredith Yeager, Joseph F Fraumeni, Stephen J Chanock, Debra T Silverman, and Nilanjan Chatterjee. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Research*, 73(7):2211–2220, apr 2013.
- [36] Rosalie Fisher, Stuart Horswell, Andrew Rowan, Maximilian P Salm, Elza C. de Bruin, Sakshi Gulati, Nicholas McGranahan, Mark Stares, Marco Gerlinger, Ignacio Varela, Andrew Crockford, Francesco Favero, Virginie Quidville, Fabrice André, Carolina Navas, Eva Grönroos, David Nicol, Steve Hazell, David Hrouda,

- Tim O'Brien, Nik Matthews, Ben Phillimore, Sharmin Begum, Adam Rabinowitz, Jennifer Biggs, Paul A Bates, Neil Q. McDonald, Gordon Stamp, Bradley Spencer-Dene, James J Hsieh, Jianing Xu, Lisa Pickering, Martin Gore, James Larkin, and Charles Swanton. Development of synchronous VHL syndrome tumors reveals contingencies and constraints to tumor evolution. *Genome biology*, 15(8):433, 2014.
- [37] Amy M. Dworkin, Katie Ridd, Dianne Bautista, Dawn C. Allain, O. Hans Iwenofu, Ritu Roy, Boris C. Bastian, and Amanda Ewart Toland. Germline Variation Controls the Architecture of Somatic Alterations in Tumors. *PLoS Genetics*, 6(9):e1001136, sep 2010.
- [38] Mel F Greaves, Ana Teresa Maia, Joseph L Wiemels, and Anthony M Ford. Leukemia in twins: Lessons in natural history. *Blood*, 102(7):2321–2333, oct 2003.
- [39] Ana Teresa Maia, Anthony M Ford, G. Reza Jalali, Christine J Harrison, G. Malcolm Taylor, Osborn B Eden, and Mel F Greaves. Molecular tracking of leukemogenesis in a triplet pregnancy. *Blood*, 98(2):478–482, 2001.
- [40] G. Cazzaniga, F. W. van Delft, L. Lo Nigro, A. M. Ford, J. Score, I. Iacobucci, E. Mirabile, M. Taj, S. M. Colman, A. Biondi, and M. Greaves. Developmental origins and impact of BCR-ABL1 fusion and IKZF1 deletions in monozygotic twins with Ph+ acute lymphoblastic leukemia. *Blood*, 118(20):5559–5564, nov 2011.
- [41] Bruce A Hamilton and Benjamin D Yu. Modifier genes and the plasticity of genetic networks in mice, 2012.
- [42] Michele Harvey, M J McArthur, Charles A Montgomery, A Bradley, and L a Donehower. Genetic background alters the spectrum of tumors that develop in p53-deficient mice. *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 7(10):938–943, 1993.
- [43] Melina MacPhee, Kenneth P Chepenik, Rebecca A Liddell, Kelly K Nelson, Linda D Siracusa, and Arthur M Buchberg. The secretory phospholipase A2 gene is a candidate for the Mom1 locus, a major modifier of ApcMin-induced intestinal neoplasia. *Cell*, 81(6):957–966, 1995.
- [44] Yuan Zhu, Pritam Ghosh, Patrick Charnay, Dennis K Burns, and Luis F Parada. Neurofibromas in NF1: Schwann cell origin and role of tumor microenvironment. *Science*, 296:920–922, 2002.
- [45] Laurence H Pearl, Amanda C Schierz, Simon E Ward, Bissan Al-Lazikani, and Frances M G Pearl. Therapeutic opportunities within the DNA damage response. *Nature Reviews Cancer*, 15(3):166–180, 2015.
- [46] Stephen P Jackson and Jiri Bartek. The DNA-damage response in human biology and disease. *Nature*, 461(7267):1071–1078, 2009.

- [47] Isidro Cortes-Ciriano, Sejoon Lee, Woong Yang Park, Tae Min Kim, and Peter J Park. A molecular portrait of microsatellite instability across multiple cancers. *Nature Communications*, 8, 2017.
- [48] Ronald J Hause, Colin C Pritchard, Jay Shendure, and Stephen J Salipante. Classification and characterization of microsatellite instability across 18 cancer types. *Nature Medicine*, 22(11):1342–1350, oct 2016.
- [49] Christopher J Lord and Alan Ashworth. BRCAness revisited. *Nature Reviews Cancer*, 16(2):110–120, 2016.
- [50] Marianne Berwick and Paolo Vineis. Markers of DNA repair and susceptibility to cancer in humans: an epidemiologic review. *Journal of the National Cancer Institute*, 92(11):874–897, jun 2000.
- [51] Harvey W Mohrenweiser, David M Wilson, and Irene M Jones. Challenges and complexities in estimating both the functional impact and the disease risk associated with the extensive genetic variation in human DNA repair genes. *Mutation research*, 526(1-2):93–125, 2003.
- [52] R S Campeljoh, S Hodgson, N Carter, B S Kato, and T D Spector. Heritability of DNA-damage-induced apoptosis and its relationship with age in lymphocytes from female twins. *British Journal of Cancer*, 95(4):520–524, aug 2006.
- [53] Paul Finnon, Naomi Robertson, Sylwia Dziwura, Claudine Ravy, Wei Zhang, Liz Ainsbury, Jaakko Kaprio, Christophe Badie, and Simon Bouzer. Evidence for significant heritability of apoptotic and cell cycle responses to ionising radiation. *Hum Genet*, 123:485–493, 2008.
- [54] Harald Surowy, Antje Rinckleb, Manuel Luedeke, Madeleine Stuber, Anna Wecker, Dominic Varga, Christiane Maier, Josef Hoegel, and Walther Vogel. Heritability of baseline and induced micronucleus frequencies. *Mutagenesis*, 26(1):111–117, 2011.
- [55] Jacqueline Cloos, Eline J.C. Nieuwenhuis, Dorret I. Boomsma, Dirk J. Kuik, Marianne L.T. Van Der Sterre, Fré Arwert, Gordon B. Snow, and Boudewijn J.M. Braakhuis. Inherited susceptibility to bleomycin-induced chromatid breaks in cultured peripheral blood lymphocytes. *Journal of the National Cancer Institute*, 91(13):1125–1130, jul 1999.
- [56] Sara Gutiérrez-Enríquez, Marie Fernet, Thilo Dork, Michael Bremer, Anthony Lauge, Dominique Stoppa-Lyonnet, Norman Moullan, Sandra Angèle, and Janet Hall. Functional Consequences of ATM Sequence Variants for Chromosomal Radiosensitivity. *Genes Chromosomes and Cancer*, 40(2):109–119, 2004.
- [57] M R Spitz, J J Fueger, S Halabi, S P Schantz, D Sample, and T C Hsu. Mutagen sensitivity in upper aerodigestive tract cancer: a case-control analysis. *Cancer epidemiology, biomarkers & prevention*, 2(4):329–33, jul 1993.

- [58] X Wu, J Gu, Y Patt, M Hassan, M R Spitz, R P Beasley, and L Y Hwang. Mutagen sensitivity as a susceptibility marker for human hepatocellular carcinoma. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 7(7):567–70, jul 1998.
- [59] N. Rajeswari, Y. R. Ahuja, U. Malini, S. Chandrashekar, N. Balakrishna, K. V. Rao, and Ashok Khar. Risk assessment in first degree female relatives of breast cancer patients using the alkaline Comet assay. *Carcinogenesis*, 21(4):557–561, 2000.
- [60] Isabel García-Cao, Marta García-Cao, Juan Martín-Caballero, Luis M. Criado, Peter Klatt, Juana M. Flores, Jean Claude Weill, María A. Blasco, and Manuel Serrano. 'Super p53' mice exhibit enhanced DNA damage response, are tumor resistant and age normally. *EMBO Journal*, 21(22):6225–6235, 2002.
- [61] Ludmil B. Alexandrov, Serena Nik-Zainal, David C. Wedge, Samuel A.J.R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, Niccolò Bolli, Ake Borg, Anne Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P. Butler, Carlos Caldas, Helen R. Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A. Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T.W. Jones, David Jonas, Stian Knappskog, Marcel Koo, Sunil R. Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C. Munshi, Hiromi Nakamura, Paul A. Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V. Pearson, Xose S. Puente, Keiran Raine, Manasa Ramakrishna, Andrea L. Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N. Schumacher, Paul N. Span, Jon W. Teague, Yasushi Totoki, Andrew N.J. Tutt, Rafael Valdés-Mas, Marit M. Van Buuren, Laura Van 'T Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R. Yates, Jessica Zucman-Rossi, P. Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M. Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M. Pfister, Peter J. Campbell, and Michael R. Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013.
- [62] Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A Stebbings, Andrew Menzies, Sancha Martin, Kenric Leung, Lina Chen, Catherine Leroy, Manasa Ramakrishna, Richard Rance, King Wai Lau, Laura J Mudie, Ignacio Varela, David J McBride, Graham R Bignell, Susanna L Cooke, Adam Shlien, John Gamble, Ian Whitmore, Mark Maddison, Patrick S Tarpey, Helen R Davies, Elli Papaemmanuil, Philip J Stephens, Stuart McLaren, Adam P Butler, Jon W Teague, Göran Jönsson, Judy E Garber, Daniel Silver, Penelope Miron, Aquila Fatima, Sandrine Boyault, Anita Langerod, Andrew Tutt, John W.M. Martens, Samuel A.J.R. Aparicio, Åke Borg, Anne Vincent Salomon, Gilles Thomas, Anne Lise Borresen-Dale, Andrea L Richardson, Michael S Neuberger, P Andrew

- Futreal, Peter J Campbell, and Michael R Stratton. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, may 2012.
- [63] Scott D. Brown, Lisa A. Raeburn, and Robert A. Holt. Profiling tissue-resident T cell repertoires by RNA sequencing. *Genome Medicine*, 7(1):125, dec 2015.
- [64] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, may 2015.
- [65] Charles Lu, Mingchao Xie, Michael C. Wendl, Jiayin Wang, Michael D. McLellan, Mark D.M. Leiserson, Kuan Lin Huang, Matthew A. Wyczalkowski, Reyka Jayasinghe, Tapahsama Banerjee, Jie Ning, Piyush Tripathi, Qunyuan Zhang, Beifang Niu, Kai Ye, Heather K. Schmidt, Robert S. Fulton, Joshua F. McMichael, Prag Batra, Cyriac Kandoth, Maheetha Bharadwaj, Daniel C. Koboldt, Christopher A. Miller, Krishna L. Kanchi, James M. Eldred, David E. Larson, John S. Welch, Ming You, Bradley A. Ozenberger, Ramaswamy Govindan, Matthew J. Walter, Matthew J. Ellis, Elaine R. Mardis, Timothy A. Graubert, John F. Dipersio, Timothy J. Ley, Richard K. Wilson, Paul J. Goodfellow, Benjamin J. Raphael, Feng Chen, Kimberly J. Johnson, Jeffrey D. Parvin, and Li Ding. Patterns and functional implications of rare germline variants across 12 cancer types. *Nature Communications*, 6:10086, dec 2015.
- [66] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B Brinkman, Sandro Morganello, Miriam R Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung Min Ahn, Sandrine Boyault, Jane E Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A Foekens, Moritz Gerstung, Gerrit K.J. Hooijer, Se Jin Jang, David R Jones, Hyung Yong Kim, Tari A King, Savitri Krishnamurthy, Hee Jin Lee, Jeong Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O’Meara, Iris Pauporté, Xavier Pivot, Colin A Purdie, Keiran Raine, Kamna Ramakrishnan, F. Germán Rodríguez-González, Gilles Romieu, Anieta M Sieuwerts, Peter T Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G. Van Den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura Van t. Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T. Ueno, Christos Sotiriou, Alain Viari, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Ewan Birney, Hendrik G Stunnenberg, Marc J Van De Vijver, John W.M. Martens, Anne Lise Børresen-Dale, Andrea L Richardson, Gu Kong, Gilles Thomas, and Michael R Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

- [67] Nadeem Riaz, Pedro Bledua, Raymond S. Lim, Ronglai Shen, Daniel S. Higginson, Nils Weinhold, Larry Norton, Britta Weigelt, Simon N. Powell, and Jorge S. Reis-Filho. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nature Communications*, 8(1):857, dec 2017.
- [68] Nicola Waddell, Marina Pajic, Ann Marie Patch, David K. Chang, Karin S. Kassahn, Peter Bailey, Amber L. Johns, David Miller, Katia Nones, Kelly Quek, Michael C.J. Quinn, Alan J. Robertson, Muhammad Z.H. Fadlullah, Tim J.C. Bruxner, Angelika N. Christ, Ivon Harliwong, Senel Idrisoglu, Suzanne Manning, Craig Nourse, Ehsan Nourbakhsh, Shivangi Wani, Peter J. Wilson, Emma Markham, Nicole Cloonan, Matthew J. Anderson, J. Lynn Fink, Oliver Holmes, Stephen H. Kazakoff, Conrad Leonard, Felicity Newell, Barsha Poudel, Sarah Song, Darrin Taylor, Nick Waddell, Scott Wood, Qinying Xu, Jianmin Wu, Mark Pinese, Mark J. Cowley, Hong C. Lee, Marc D. Jones, Adnan M. Nagrial, Jeremy Humphris, Lorraine A. Chantrill, Venessa Chin, Angela M. Steinmann, Amanda Mawson, Emily S. Humphrey, Emily K. Colvin, Angela Chou, Christopher J. Scarlett, Andreia V. Pinho, Marc Giry-Laterriere, Ilse Rooman, Jaswinder S. Samra, James G. Kench, Jessica A. Pettitt, Neil D. Merrett, Christopher Toon, Krishna Epari, Nam Q. Nguyen, Andrew Barbour, Nikolajs Zeps, Nigel B. Jamieson, Janet S. Graham, Simone P. Niclou, Rolf Bjerkgvig, Robert Grützmann, Daniela Aust, Ralph H. Hruban, Anirban Maitra, Christine A. Iacobuzio-Donahue, Christopher L. Wolfgang, Richard A. Morgan, Rita T. Lawlor, Vincenzo Corbo, Claudio Bassi, Massimo Falconi, Giuseppe Zamboni, Giampaolo Tortora, Margaret A. Tempero, Anthony J. Gill, James R. Eshleman, Christian Pilarsky, Aldo Scarpa, Elizabeth A. Musgrove, John V. Pearson, Andrew V. Biankin, and Sean M. Grimmond. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, 518(7540):495–501, feb 2015.
- [69] Paz Polak, Jaegil Kim, Lior Z Braunstein, Rosa Karlic, Nicholas J Haradhavala, Grace Tiao, Daniel Rosebrock, Dimitri Livitz, Kirsten Kübler, Kent W Mouw, Atanas Kamburov, Yosef E Maruvka, Ignaty Leshchiner, Eric S Lander, Todd R Golub, Aviad Zick, Alexandre Orthwein, Michael S Lawrence, Rajbir N Batra, Carlos Caldas, Daniel A Haber, Peter W Laird, Hui Shen, Leif W Ellisen, Alan D D’Andrea, Stephen J Chanock, William D Foulkes, and Gad Getz. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*, 49(10):1476–1486, aug 2017.
- [70] Jenni Nikkilä, Ann Christin Parpys, Katri Pylkäs, Muthiah Bose, Yanying Huo, Kerstin Borgmann, Katrin Rapakko, Pentti Nieminen, Bing Xia, Helmut Pospiech, and Robert Winqvist. Heterozygous mutations in PALB2 cause DNA replication and damage response defects. *Nature Communications*, 4:2578, oct 2013.
- [71] Mary I Coolbaugh-Murphy, Jing Ping Xu, Louis S. Ramagli, Brian C Ramagli, Barry W Brown, Patrick M Lynch, Stanley R Hamilton, Marsha L Frazier, and Michael J Siciliano. Microsatellite instability in the peripheral blood leukocytes of HNPCC patients. *Human Mutation*, 31(3):317–324, mar 2010.

- [72] Henry T. Lynch and Thomas Smyrk. Hereditary nonpolyposis colorectal cancer (Lynch syndrome). An updated review. *Cancer*, 78(6):1149–1167, sep 1996.
- [73] Candace D Middlebrooks, A Rouf Banday, Konichi Matsuda, Krizia Ivana Udquim, Olusegun O Onabajo, Ashley Paquin, Jonine D Figueroa, Bin Zhu, Stella Koutros, Michiaki Kubo, Taro Shuin, Neal D Freedman, Manolis Kogevinas, Nuria Malats, Stephen J Chanock, Montserrat Garcia-Closas, Debra T Silverman, Nathaniel Rothman, and Ludmila Prokunina-Olsson. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nature Genetics*, 48(11):1330–1338, sep 2016.
- [74] Rachel Marty, Saghar Kaabinejadian, David Rossell, Michael J Slifker, Joris van de Haar, Hatice Billur Engin, Nicola de Prisco, Trey Ideker, William H Hildebrand, Joan Font-Burgada, and Hannah Carter. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell*, 171(6):1272–1283.e15, 2017.
- [75] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, mar 2011.
- [76] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, James S. Ware, andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A. Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N. Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I. Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M. Peloso, Ryan Poplin, Manuel A. Rivas, Valentin Ruano-Rubio, Samuel A. Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T. Tusie-Luna, Ben Weisburd, Hong Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C. Florez, Stacey B. Gabriel, Gad Getz, Stephen J. Glatt, Christina M. Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, shaun M Purcell, Danish Saleheen, Jeremiah M. Scharf, Pamela Sklar, Patrick F. Sullivan, Jaakko Tuomilehto, Ming T. Tsuang, Hugh C. Watkins, James G. Wilson, Mark J Daly, and Daniel G. MacArthur. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.
- [77] Hannah Carter, Rachel Marty, Matan Hofree, Andrew M Gross, James Jensen, Kathleen M Fisch, Xingyu Wu, Christopher Deboever, Eric L Van Nostrand, Yan Song, Emily Wheeler, Jason F Kreisberg, Scott M Lippman, Gene W Yeo, J Silvio Gutkind, and Trey Ideker. Interaction landscape of inherited polymorphisms with somatic events in cancer. *Cancer Discovery*, 7(4):410–423, apr 2017.

- [78] Roberto Puzone and Ulrich Pfeffer. SNP variants at the MAP3K1/SETD9 locus 5q11.2 associate with somatic PIK3CA variants in breast cancers. *European Journal of Human Genetics*, 25(3):384–387, mar 2017.
- [79] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1118, 2013.
- [80] Duane A Compton. Mechanisms of aneuploidy. *Current Opinion in Cell Biology*, 23(1):109–113, feb 2011.
- [81] M. A. Didraga, E. H. Van Beers, S. A. Joosse, K. I.M. Brandwijk, R. A. Oldenburg, L. F.A. Wessels, F. B.L. Hogervorst, M. J. Ligtenberg, N. Hoogerbrugge, S. Verhoef, P. Devilee, and P. M. Nederlof. A non-BRCA1/2 hereditary breast cancer sub-group defined by aCGH profiling of genetically related patients. *Breast Cancer Research and Treatment*, 2011.
- [82] Tobias Rausch, David T W Jones, Marc Zapatka, Adrian M Stü Tz, Thomas Zichner, Joachim Weischenfeldt, Natalie Jä, Marc Remke, David Shih, Paul A Northcott, Elke Pfaff, Jelena Tica, Qi Wang, Luca Massimi, Hendrik Witt, Sebastian Bender, Sabrina Pleier, Huriye Cin, Cynthia Hawkins, Christian Beck, Andreas Von Deimling, Volkmar Hans, Benedikt Brors, Roland Eils, Wolfram Scheurlen, Jonathon Blake, Vladimir Benes, Andreas E Kulozik, Olaf Witt, Dianna Martin, Cindy Zhang, Rinat Porat, Diana M Merino, Jonathan Wasserman, Nada Jabado, Adam Fontebasso, Lars Bullinger, Jan Koster, Jan J Molenaar, Rogier Versteeg, Marcel Kool, Uri Tabori, David Malkin, Andrey Korshunov, Michael D Taylor, Peter Lichter, Stefan M Pfister, and Jan O Korbel. Genome Sequencing of Pediatric Medulloblastoma Links Catastrophic DNA Rearrangements with TP53 Mutations. *Cell*, 148:59–71, 2012.
- [83] Megan P Hitchins, Robert W Rapkins, Chau To Kwok, Sameer Srivastava, Justin J.L. Wong, Levon M Khachigian, Patsie Polly, Jack Goldblatt, and Robyn L Ward. Dominantly Inherited Constitutional Epigenetic Silencing of MLH1 in a Cancer-Affected Family Is Linked to a Single Nucleotide Variant within the 5’UTR. *Cancer Cell*, 20(2):200–213, 2011.
- [84] Shuji Ogino, Aditi Hazra, Gregory J Tranah, Gregory J Kirkner, Takako Kawasaki, Katsuhiko Nosho, Mutsuko Ohnishi, Yuko Suemoto, Jeffrey A Meyerhardt, David J Hunter, and Charles S Fuchs. MGMT germline polymorphism is associated with somatic MGMT promoter methylation and gene silencing in colorectal cancer. *Carcinogenesis*, 28(9):1985–1990, 2007.
- [85] Aditi Hazra, Charles S Fuchs, Takako Kawasaki, Gregory J Kirkner, David J Hunter, and Shuji Ogino. Germline polymorphisms in the one-carbon metabolism pathway and DNA methylation in colorectal cancer. *Cancer Causes and Control*, 21(3):331–345, 2010.

- [86] David Capper, David T.W. Jones, Martin Sill, Volker Hovestadt, Daniel Schrimpf, Dominik Sturm, Christian Koelsche, Felix Sahm, Lukas Chavez, David E Reuss, Annekathrin Kratz, Annika K Wefers, Kristin Huang, Kristian W Pajtler, Leonille Schweizer, Damian Stichel, Adriana Olar, Nils W Engel, Kerstin Lindenberg, Patrick N Harter, Anne K Braczynski, Karl H Plate, Hildegard Dohmen, Boyan K Garvalov, Roland Coras, Annett Hölsken, Ekkehard Hewer, Melanie Bewerunge-Hudler, Matthias Schick, Roger Fischer, Rudi Beschorner, Jens Schittenhelm, Ori Staszewski, Khalida Wani, Pascale Varlet, Melanie Pages, Petra Temming, Dietmar Lohmann, Florian Selt, Hendrik Witt, Till Milde, Olaf Witt, Eleonora Aronica, Felice Giangaspero, Elisabeth Rushing, Wolfram Scheurlen, Christoph Geisenberger, Fausto J Rodriguez, Albert Becker, Matthias Preusser, Christine Haberler, Rolf Bjerkgvig, Jane Cryan, Michael Farrell, Martina Deckert, Jürgen Hench, Stephan Frank, Jonathan Serrano, Kasthuri Kannan, Aristotelis Tsirogos, Wolfgang Brück, Silvia Hofer, Stefanie Brehmer, Marcel Seiz-Rosenhagen, Daniel Hänggi, Volkmar Hans, Stephanie Rozsnoki, Jordan R Hansford, Patricia Kohlhof, Bjarne W Kristensen, Matt Lechner, Beatriz Lopes, Christian Mawrin, Ralf Ketter, Andreas Kulozik, Ziad Khatib, Frank Heppner, Arend Koch, Anne Jouvét, Catherine Keohane, Helmut Mühleisen, Wolf Mueller, Ute Pohl, Marco Prinz, Axel Benner, Marc Zapatka, Nicholas G Gottardo, Pablo Hernáiz Driever, Christof M Kramm, Hermann L Müller, Stefan Rutkowski, Katja Von Hoff, Michael C Frühwald, Astrid Gnekow, Gudrun Fleischhack, Stephan Tippelt, Gabriele Calaminus, Camelia Maria Monoranu, Arie Perry, Chris Jones, Thomas S Jacques, Bernhard Radlwimmer, Marco Gessi, Torsten Pietsch, Johannes Schramm, Gabriele Schackert, Manfred Westphal, Guido Reifenberger, Pieter Wesseling, Michael Weller, Vincent Peter Collins, Ingmar Blümcke, Martin Bendszus, Jürgen Debus, Annie Huang, Nada Jabado, Paul A Northcott, Werner Paulus, Amar Gajjar, Giles W Robinson, Michael D Taylor, Zane Jaunmuktane, Marina Ryzhova, Michael Platten, Andreas Unterberg, Wolfgang Wick, Matthias A Karajannis, Michel Mittelbronn, Till Acker, Christian Hartmann, Kenneth Aldape, Ulrich Schüller, Rolf Buslei, Peter Lichter, Marcel Kool, Christel Herold-Mende, David W Ellison, Martin Hasselblatt, Matija Snuderl, Sebastian Brandner, Andrey Korshunov, Andreas Von Deimling, and Stefan M Pfister. DNA methylation-based classification of central nervous system tumours. *Nature*, 555(7697):469–474, 2018.
- [87] Philip S Bernard, Joel S Parker, Michael Mullins, Maggie C.U. Cheung, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Inge J Stijleman, Juan Palazzo, J. S. Matron, Andrew B Nobel, Elaine Mardis, Torsten O Nielsen, Matthew J Ellis, and Charles M Perou. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, mar 2009.
- [88] Roel G.W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D. Wilkerson, C. Ryan Miller, Li Ding, Todd Golub, Jill P. Mesirov, Gabriele Alexe, Michael Lawrence, Michael O’Kelly, Pablo Tamayo, Barbara A.

- Weir, Stacey Gabriel, Wendy Winckler, Supriya Gupta, Lakshmi Jakkula, Heidi S. Feiler, J. Graeme Hodgson, C. David James, Jann N. Sarkaria, Cameron Brennan, Ari Kahn, Paul T. Spellman, Richard K. Wilson, Terence P. Speed, Joe W. Gray, Matthew Meyerson, Gad Getz, Charles M. Perou, and D. Neil Hayes. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, jan 2010.
- [89] Martin J Larsen, Mads Thomassen, Qihua Tan, Anne-Vibeke Lænkholm, Martin Bak, Kristina P Sørensen, Mette Klarskov Andersen, Torben A Kruse, and Anne-Marie Gerdes. RNA profiling reveals familial aggregation of molecular subtypes in non-BRCA1/2 breast cancer families. *BMC Medical Genomics*, 7(1):9, dec 2014.
- [90] Andrea Sottoriva, Inmaculada Spiteri, Sara G M Piccirillo, Anestis Touloumis, V Peter Collins, John C Marioni, Christina Curtis, Colin Watts, Simon Tavaré, and S Tavaré. Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc Natl Acad Sci U S A*, 110(10):4009–4014, 2013.
- [91] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R Santos, Mahrokh Nohadani, Aron C Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P. Andrew Futreal, and Charles Swanton. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, 366(10):883–892, 2012.
- [92] Jing Gong, Shufang Mei, Chunjie Liu, Yu Xiang, Youqiong Ye, Zhao Zhang, Jing Feng, Renyan Liu, Lixia Diao, An Yuan Guo, Xiaoping Miao, and Leng Han. Pan-cancerQTL: Systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Research*, 46(D1):D971–D976, 2018.
- [93] Halit Ongen, Claus L Andersen, Jesper B Bramsen, Bodil Oster, Mads H Rasmussen, Pedro G Ferreira, Juan Sandoval, Enrique Vidal, Nicola Whiffin, Alexandra Planchon, Ismael Padioleau, Deborah Bielser, Luciana Romano, Ian Tomlinson, Richard S Houlston, Manel Esteller, Torben F Orntoft, and Emmanouil T Dermitzakis. Putative cis-regulatory drivers in colorectal cancer. *Nature*, 512(1):87–90, 2014.
- [94] Nasim Mavaddat, Daniel Barrowdale, Irene L Andrulis, Susan M Domchek, Diana Eccles, Heli Nevanlinna, Susan J Ramus, Amanda Spurdle, Mark Robson, Mark Sherman, Anna Marie Mulligan, Fergus J Couch, Christoph Engel, Lesley McGuffog, Sue Healey, Olga M Sinilnikova, Melissa C Southey, Mary Beth Terry, David Goldgar, Frances O’Malley, Esther M John, Ramunas Janavicius, Laima Tihomirova, Thomas V.O. Hansen, Finn C Nielsen, Ana Osorio, Alexandra Stavropoulou, Javier

Benítez, Siranoush Manoukian, Bernard Peissel, Monica Barile, Sara Volorio, Barbara Pasini, Riccardo Dolcetti, Anna Laura Putignano, Laura Ottini, Paolo Radice, Ute Hamann, Muhammad U Rashid, Frans B Hogervorst, Mieke Kriege, Rob B. Van Der Luijt, Susan Peock, Debra Frost, D Gareth Evans, Carole Brewer, Lisa Walker, Mark T Rogers, Lucy E Side, Catherine Houghton, Jo Ellen Weaver, Andrew K Godwin, Rita K Schmutzler, Barbara Wappenschmidt, Alfons Meindl, Karin Kast, Norbert Arnold, Dieter Niederacher, Christian Sutter, Helmut Deissler, Doroteha Gadzicki, Sabine Preisler-Adams, Raymonda Varon-Mateeva, Ines Schönbuchner, Heidrun Gevensleben, Dominique Stoppa-Lyonnet, Muriel Belotti, Laure Barjhoux, Claudine Isaacs, Beth N Peshkin, Trinidad Caldes, Miguel De Al Hoya, Carmen Cañadas, Tuomas Heikkinen, Päivi Heikkilä, Kristiina Aittomäki, Ignacio Blanco, Conxi Lazaro, Joan Brunet, Bjarni A Agnarsson, Adalgeir Arason, Rosa B Barkardottir, Martine Dumont, Jacques Simard, Marco Montagna, Simona Agata, Emma D’Andrea, Max Yan, Stephen Fox, Timothy R Rebbeck, Wendy Rubinstein, Nadine Tung, Judy E Garber, Xianshu Wang, Zachary Fredericksen, Vernon S Pankratz, Noralane M Lindor, Csilla Szabo, Kenneth Offit, Rita Sakr, Mia M Gaudet, Christian F Singer, Muy Kheng Tea, Christine Rappaport, Phuong L Mai, Mark H Greene, Anna Sokolenko, Evgeny Imyanitov, Amanda Ewart Toland, Leigha Senter, Kevin Sweet, Mads Thomassen, Anne Marie Gerdes, Torben Kruse, Maria Caligo, Paolo Aretini, Johanna Rantala, Anna Von Wachenfeld, Karin Henriksson, Linda Steele, Susan L Neuhausen, Robert Nussbaum, Mary Beattie, Kunle Odunsi, Lara Sucheston, Simon A Gayther, Kate Nathanson, Jenny Gross, Christine Walsh, Beth Karlan, Georgia Chenevix-Trench, Douglas F Easton, and Antonis C Antoniou. Pathology of breast and ovarian cancers among BRCA1 and BRCA2 mutation carriers: Results from the consortium of investigators of modifiers of BRCA1/2 (CIMBA). *Cancer Epidemiology Biomarkers and Prevention*, 21(1):134–147, jan 2012.

- [95] Yi Kan Wang, Ali Bashashati, Michael S Anglesio, Dawn R Cochrane, Diljot S Grewal, Gavin Ha, Andrew McPherson, Hugo M Horlings, Janine Senz, Leah M Prentice, Anthony N Karnezis, Daniel Lai, Mohamed R Aniba, Allen W Zhang, Karey Shumansky, Celia Siu, Adrian Wan, Melissa K McConechy, Hector Li-Chang, Alicia Tone, Diane Provencher, Manon De Ladurantaye, Hubert Fleury, Aikou Okamoto, Satoshi Yanagida, Nozomu Yanaihara, Misato Saito, Andrew J Mungall, Richard Moore, Marco A Marra, C. Blake Gilks, Anne Marie Mes-Masson, Jessica N McAlpine, Samuel Aparicio, David G Huntsman, and Sohrab P Shah. Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nature Genetics*, 49(6):856–864, 2017.
- [96] Maria N. Timofeeva, Rayjean J. Hung, Thorunn Rafnar, David C. Christiani, John K. Field, Heike Bickeböller, Angela Risch, James D. McKay, Yufei Wang, Juncheng Dai, Valerie Gaborieau, John Mclaughlin, Darren Brenner, Steven A. Narod, Neil E. Caporaso, Demetrius Albanes, Michael Thun, Timothy Eisen, H. Erich Wichmann, Albert Rosenberger, Younghun Han, Wei Chen, Dakai Zhu, Margaret Spitz, Xifeng Wu, Mala Pande, Yang Zhao, David Zaridze, Neonilia

- Szeszenia-dabrowska, Jolanta Lissowska, Peter Rudnai, Eleonora Fabianova, Dana Mates, Vladimir Bencko, Lenka Foretova, Vladimir Janout, Hans E. Krokan, Maiken Elvestad Gabrielsen, Frank Skorpen, Lars Vatten, Inger NjØlstad, Chu Chen, Gary Goodman, Mark Lathrop, Simone Benhamou, Tõnu Vooder, Kristjan Valk, Mari Nelis, Andres Metspalu, Olaide Raji, Ying Chen, John Gosney, Triantafillos Liloglou, Thomas Muley, Hendrik Dienemann, Gudmar Thorleifsson, Hongbing Shen, Kari Stefansson, Paul Brennan, Christopher I. Amos, Richard Houlston, and Maria Teresa Landi. Influence of common genetic variation on lung cancer risk: Meta-analysis of 14,900 cases and 29,485 controls. *Human Molecular Genetics*, 21(22):4980–4995, nov 2012.
- [97] Robert D Schreiber, Lloyd J Old, and Mark J Smyth. Integrating Immunity’s Roles in Cancer Suppression and Promotion. *Science*, 331(6024):1565–1570, 2014.
- [98] Dirk C Strauss and J. Meirion Thomas. Transmission of donor melanoma by organ transplantation. *The Lancet Oncology*, 11(8):790–796, 2010.
- [99] Vesteinn Thorsson, David L. Gibbs, Scott D. Brown, Denise Wolf, Dante S. Bortone, Tai-Hsien Ou Yang, Eduard Porta-Pardo, Galen F. Gao, Christopher L. Plaisier, James A. Eddy, Elad Ziv, Aedin C. Culhane, Evan O. Paull, I.K. Ashok Sivakumar, Andrew J. Gentles, Raunaq Malhotra, Farshad Farshidfar, Antonio Colaprico, Joel S. Parker, Lisle E. Mose, Nam Sy Vo, Jianfang Liu, Yuexin Liu, Janet Rader, Varsha Dhankani, Sheila M. Reynolds, Reanne Bowlby, Andrea Califano, Andrew D. Cherniack, Dimitris Anastassiou, Davide Bedognetti, Arvind Rao, Ken Chen, Alexander Krasnitz, Hai Hu, Tathiane M. Malta, Houtan Noushmehr, Chandra Sekhar Pdamallu, Susan Bullman, Akinyemi I. Ojesina, Andrew Lamb, Wanding Zhou, Hui Shen, Toni K. Choueiri, John N. Weinstein, Justin Guinney, Joel Saltz, Robert A. Holt, Charles E. Rabkin, Alexander J. Lazar, Jonathan S. Serody, Elizabeth G. Demicco, Mary L. Disis, Benjamin G. Vincent, Llya Shmulevich, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips,

Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J.M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Morris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C.S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavoilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hage-

dorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bublely, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bitá Esmali, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Onathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine, Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jen-

nifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter, Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffrey Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Had-

- dad, Jeffrey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, and Armaz Mariamidze. The Immune Landscape of Cancer. *Immunity*, 48(0):1–19, apr 2018.
- [100] Daniel Kogan, Alexander Grabner, Christopher Yanucil, and Christian Faul. Stat3-enhancing germline mutations contribute to tumor-extrinsic immune evasion. *The Journal of Clinical Investigation*, 128(5):1867–1872, 5 2018.
- [101] Vijay K Ulaganathan, Bianca Sperl, Ulf R Rapp, and Axel Ullrich. Germline variant FGFR4 p.G388R exposes a membrane-proximal STAT3 binding site. *Nature*, 528(7583):570–574, 2015.
- [102] Shumei Kato, Aaron Goodman, Vighnesh Walavalkar, Donald A Barkauskas, Andrew Sharabi, and Razelle Kurzrock. Hyperprogressors after immunotherapy: Analysis of genomic alterations associated with accelerated growth rate. *Clinical Cancer Research*, 23(15):4242–4250, 2017.
- [103] Paula J Hurley, Debasish Sundi, Brian Shinder, Brian W Simons, Robert M Hughes, Rebecca M Miller, Benjamin Benzoni, Sheila F Faraj, George J Netto, Ismael A Vergara, Nicholas Erho, Elai Davicioni, R Jeffrey Karnes, Guifang Yan, Charles Ewing, Sarah D. Isaacs, David M Berman, Jennifer R Rider, Kristina M Jordahl, Lorelei A Mucci, Jessie Huang, Steven S An, Ben H Park, William B Isaacs, Luigi Marchionni, Ashley E Ross, and Edward M Schaeffer. Germline variants in asporin vary by race, modulate the tumor microenvironment, and are differentially associated with metastatic prostate cancer. *Clinical Cancer Research*, 22(2):448–458, jan 2016.
- [104] M Streit, L Riccardi, P Velasco, L F Brown, T Hawighorst, P Bornstein, M Detmar, Jack Lawler, Mark Kieran, Amish Shah, and Raghu Kalluri. Thrombospondin-2: a potent endogenous inhibitor of tumor growth and angiogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(26):14888–93, dec 1999.
- [105] Latha Kadalayil, Sofia Khan, Heli Nevanlinna, Peter A Fasching, Fergus J Couch, John L Hopper, Jianjun Liu, Tom Maishman, Lorraine Durcan, Sue Gerty, Carl Blomqvist, Brigitte Rack, Wolfgang Janni, Andrew Collins, Diana Eccles, and William Tapper. Germline variation in ADAMTSL1 is associated with prognosis following breast cancer treatment in young women. *Nature Communications*, 8(1), 2017.
- [106] Howard L. McLeod. Cancer pharmacogenomics: Early promise, but concerted effort needed. *Science*, 340(6127):1563–1566, 2013.

- [107] William J Irvin, Christine M Walko, Karen E Weck, Joseph G Ibrahim, Wing K Chiu, E Claire Dees, Susan G Moore, Oludamilola A Olajide, Mark L Graham, Sean T Canale, Rachel E Raab, Steven W Corso, Jeffrey M Peppercorn, Steven M Anderson, Kenneth J Friedman, Evan T Ogburn, Zeruesenay Desta, David A Flockhart, Howard L. McLeod, James P Evans, and Lisa A Carey. Genotype-guided tamoxifen dosing increases active metabolite exposure in women with reduced CYP2D6 metabolism: A multicenter study. *Journal of Clinical Oncology*, 29(24):3232–3239, 2011.
- [108] Eric J Peters, Alison Motsinger-Reif, Tammy M Havener, Lorraine Everitt, Nicholas E Hardison, Venita G Watson, Michael Wagner, Kristy L Richards, Mike A Province, and Howard L McLeod. Pharmacogenomic characterization of US FDA-approved cytotoxic drugs. *Pharmacogenomics*, 12(10):1407–1415, 2011.
- [109] Kathryn P Pennington, Tom Walsh, Maria I Harrell, Ming K Lee, Christopher C Pennil, Mara H Rendi, Anne Thornton, Barbara M Norquist, Silvia Casadei, Alexander S Nord, Kathy J Agnew, Colin C Pritchard, Sheena Scroggins, Rochelle L Garcia, Mary Claire King, and Elizabeth M Swisher. Germline and somatic mutations in homologous recombination genes predict platinum response and survival in ovarian, fallopian tube, and peritoneal carcinomas. *Clinical Cancer Research*, 20(3):764–775, feb 2014.
- [110] Elizabeth M. Swisher, Wataru Sakai, Beth Y. Karlan, Kaitlyn Wurz, Nicole Urban, and Toshiyasu Taniguchi. Secondary BRCA1 mutations in BRCA1-mutated ovarian carcinomas with platinum resistance. *Cancer Research*, 68(8):2581–2586, aug 2008.
- [111] John Burn, John C Mathers, and D. Tim Bishop. Chemoprevention in Lynch syndrome. *Familial Cancer*, 12(4):707–718, 2013.
- [112] Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J Topol, and Nicholas J Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223, 2011.
- [113] Henry T. Lynch, Carrie L. Snyder, Trudy G. Shaw, Christopher D. Heinen, and Megan P. Hitchins. Milestones of Lynch syndrome: 1895-2015. *Nature Reviews Cancer*, 15(3):181–194, 2015.
- [114] Rebecca J. Allred and Wade Samowitz. Constitutional mismatch repair-deficiency syndrome. *Pathology Case Reviews*, 18(2):79–80, may 2013.
- [115] Mingming Liu, Layne T. Watson, and Liqing Zhang. Predicting the combined effect of multiple genetic variants. *Human genomics*, 9(1):18, dec 2015.
- [116] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology*, 1A(1A):68–77, 2015.

- [117] Alexandra R. Buckley, Kristopher A. Standish, Kunal Bhutani, Trey Ideker, Roger S. Lasken, Hannah Carter, Olivier Harismendy, and Nicholas J. Schork. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC Genomics*, 18(1):458, dec 2017.
- [118] Adam Shlien, Brittany B Campbell, Richard de Borja, Ludmil B Alexandrov, Daniele Merico, David Wedge, Peter Van Loo, Patrick S Tarpey, Paul Coupland, Sam Behjati, Aaron Pollett, Tatiana Lipman, Abolfazl Heidari, Shriya Deshmukh, ama Avitzur, Bettina Meier, Moritz Gerstung, Ye Hong, Diana M Merino, Manasa Ramakrishna, Marc Remke, Roland Arnold, Gagan B Panigrahi, Neha P Thakkar, Karl P Hodel, Erin E Henninger, A Yasemin Göksenin, Doua Bakry, George S Charames, Harriet Druker, Jordan Lerner-Ellis, Matthew Mistry, Rina Dvir, Ronald Grant, Ronit Elhasid, Roula Farah, Glenn P Taylor, Paul C Nathan, Sarah Alexander, Shay Ben-Shachar, Simon C Ling, Steven Gallinger, Shlomi Constantini, Peter Dirks, Annie Huang, Stephen W Scherer, Richard G Grundy, Carol Durno, Melyssa Aronson, Anton Gartner, M Stephen Meyn, Michael D Taylor, Zachary F Pursell, Christopher E Pearson, David Malkin, P Andrew Futreal, Michael R Stratton, Eric Bouffet, Cynthia Hawkins, Peter J Campbell, and Uri Tabori. Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermuted cancers. *Nature Genetics*, 47(3), 2015.
- [119] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–20, oct 2013.
- [120] Helen Cavanagh and Katherine M A Rogers. The role of BRCA1 and BRCA2 mutations in prostate, pancreatic and stomach cancers. *Hereditary cancer in clinical practice*, 13(1):16, 2015.
- [121] Brennan Decker, Danielle M. Karyadi, Brian W. Davis, Eric Karlins, Lori S. Tillmans, Janet L. Stanford, Stephen N. Thibodeau, and Elaine A. Ostrander. Biallelic BRCA2 Mutations Shape the Somatic Mutational Landscape of Aggressive Prostate Tumors. *The American Journal of Human Genetics*, 98(5):818–829, may 2016.
- [122] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, oct 2010.
- [123] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly.

A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–498, apr 2011.

- [124] Geraldine A. Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From fastq data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 43(1):11.10.1–11.10.33, 2013.
- [125] Christopher Wilks, Melissa S. Cline, Erich Weiler, Mark Diehkans, Brian Craft, Christy Martin, Daniel Murphy, Howdy Pierce, John Black, Donovan Nelson, Brian Litzinger, Thomas Hatton, Lori Maltbie, Michael Ainsworth, Patrick Allen, Linda Rosewood, Elizabeth Mitchell, Bradley Smith, Jim Warner, John Groboske, Haifang Telc, Daniel Wilson, Brian Sanford, Hannes Schmidt, David Haussler, and Daniel Maltbie. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database : the journal of biological databases and curation*, 2014(0), sep 2014.
- [126] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje Van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, 22(9):1760–1774, sep 2012.
- [127] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, dec 2015.
- [128] John W. Belmont, Paul Hardenbol, Thomas D. Willis, Fuli Yu, Huanming Yang, Lan Yang Ch’Ang, Wei Huang, Bin Liu, Yan Shen, Paul Kwong Hang Tam, Lap Chee Tsui, Mary Miu Yee Waye, Jeffrey Tze Fei Wong, Changqing Zeng, Qingrun Zhang, Mark S. Chee, Luana M. Galver, Semyon Kruglyak, Sarah S. Murray, Arnold R. Oliphant, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Michael S. Phillips, Andrei Verner, Shenghui Duan, Denise L. Lind, Raymond D. Miller, John Rice, Nancy L. Saccone, Patricia Taillon-Miller, Ming Xiao, Akihiro Sekine, Koki Sorimachi, Yoichi Tanaka, Tatsuhiko Tsunoda, Eiji Yoshino, David R. Bentley, Sarah Hunt, Don Powell, Houcan Zhang, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R. Macer, Eiko Suda, Charles Rotimi, Clement A. Adebamowo,

Toyin Aniagwu, Patricia A. Marshall, Olayemi Matthew, Chibuzor Nkwodimmah, Charmaine D.M. Royal, Mark F. Leppert, Missy Dixon, Fiona Cunningham, Ardavan Kanani, Gudmundur A. Thorisson, Peter E. Chen, David J. Cutler, Carl S. Kashuk, Peter Donnelly, Jonathan Marchini, Gilean A.T. McVean, Simon R. Myers, Lon R. Cardon, Andrew Morris, Bruce S. Weir, James C. Mullikin, Michael Feolo, Mark J. Daly, Renzong Qiu, Alastair Kent, Georgia M. Dunston, Kazuto Kato, Norio Niikawa, Jessica Watkin, Richard A. Gibbs, Erica Sodergren, George M. Weinstock, Richard K. Wilson, Lucinda L. Fulton, Jane Rogers, Bruce W. Birren, Hua Han, Hongguang Wang, Martin Godbout, John C. Wallenburg, Paul L'Archevêque, Guy Bellemare, Kazuo Todani, Takashi Fujita, Satoshi Tanaka, Arthur L. Holden, Francis S. Collins, Lisa D. Brooks, Jean E. McEwen, Mark S. Guyer, Elke Jordan, Jane L. Peterson, Jack Spiegel, Lawrence M. Sung, Lynn F. Zacharia, Karen Kennedy, Michael G. Dunn, Richard Seabrook, Mark Shillito, Barbara Skene, John G. Stewart, David L. Valle, Ellen Wright Clayton, Lynn B. Jorde, Aravinda Chakravarti, Mildred K. Cho, Troy Duster, Morris W. Foster, Maria Jasperse, Bartha M. Knoppers, Pui Yan Kwok, Julio Licinio, Jeffrey C. Long, Pilar Ossorio, Vivian Ota Wang, Charles N. Rotimi, Patricia Spallone, Sharon F. Terry, Eric S. Lander, Eric H. Lai, Deborah A. Nickerson, Gonçalo R. Abecasis, David Altshuler, Michael Boehnke, Panos Deloukas, Julie A. Douglas, Stacey B. Gabriel, Richard R. Hudson, Thomas J. Hudson, Leonid Kruglyak, Yusuke Nakamura, Robert L. Nussbaum, Stephen F. Schaffner, Stephen T. Sherry, Lincoln D. Stein, and Toshihiro Tanaka. The international HapMap project. *Nature*, 426(6968):789–796, dec 2003.

- [129] Friedrich Leisch. A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis*, 51(2):526–544, 2006.
- [130] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R.S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1):122, dec 2016.
- [131] Kai Wang, Mingyao Li, and Hakon Hakonarson. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, sep 2010.
- [132] D. Karolchik. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(90001):493D–496, jan 2004.
- [133] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufu, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad

- Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–D745, jan 2016.
- [134] Robert L. Grossman, Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12):1109–1112, sep 2016.
- [135] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.
- [136] Broad Institute. Picard tools, 2018. Accessed: 2016/09/27; version 1.131.
- [137] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, oct 1990.
- [138] Roger S Lasken and Timothy B Stockwell. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnology*, 7(1):19, apr 2007.
- [139] Daniel C. Koboldt, Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, and Richard K. Wilson. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, mar 2012.
- [140] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011.
- [141] Angelo Canty and B. D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2017. R package version 1.3-20.
- [142] Brandi L. Cantarel, Yunping Lei, Daniel Weaver, Huiping Zhu, Andrew Farrell, Graeme Benstead-Hume, Justin Reese, and Richard H. Finnell. Analysis of archived residual newborn screening blood spots after whole genome amplification. *BMC Genomics*, 16(1):602, dec 2015.
- [143] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M. Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5(1):17875, nov 2015.

- [144] Mohammad Shabbir Hasan, Xiaowei Wu, and Liqing Zhang. Performance evaluation of indel calling tools using real short-read data. *Human genomics*, 9(1):20, aug 2015.
- [145] Steve Laurie, Marcos Fernandez-Callejo, Santiago Marco-Sola, Jean Remi Trotta, Jordi Camps, Alejandro Chacón, Antonio Espinosa, Marta Gut, Ivo Gut, Simon Heath, and Sergi Beltran. From Wet-Lab to Variations: Concordance and Speed of Bioinformatics Pipelines for Whole Genome and Whole Exome Sequencing. *Human Mutation*, 37(12):1263–1271, dec 2016.
- [146] Ryan E. Mills, W. Stephen Pittard, Julienne M. Mullaney, Umar Farooq, Todd H. Creasy, Anup A. Mahurkar, David M. Kemeza, Daniel S. Strassler, Chris P. Ponting, Caleb Webber, and Scott E. Devine. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Research*, 21(6):830–839, jun 2011.
- [147] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M.-M. Suner, T. Hunt, I. H. A. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, C. Tyler-Smith, and Chris Tyler-Smith. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science*, 335(6070):823–828, feb 2012.
- [148] Han Fang, Yiyang Wu, Giuseppe Narzisi, Jason A. O’Rawe, Laura T Jimenez Barrón, Julie Rosenbaum, Michael Ronemus, Ivan Iossifov, Michael C Schatz, and Gholson J Lyon. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Medicine*, 6(10):89, dec 2014.
- [149] Krishna L Kanchi, Kimberly J Johnson, Charles Lu, Michael D. McLellan, Mark D.M. Leiserson, Michael C Wendl, Qunyuan Zhang, Daniel C Koboldt, Mingchao Xie, Cyriac Kandoth, Joshua F. McMichael, Matthew A Wyczalkowski, David E Larson, Heather K Schmidt, Christopher A Miller, Robert S Fulton, Paul T Spellman, Elaine R Mardis, Todd E Druley, Timothy A Graubert, Paul J Goodfellow, Benjamin J Raphael, Richard K Wilson, and Li Ding. Integrated analysis of germline and somatic variants in ovarian cancer. *Nature Communications*, 5, 2014.
- [150] T. Walsh, S. Casadei, M. K. Lee, C. C. Pennil, A. S. Nord, A. M. Thornton, W. Roeb, K. J. Agnew, S. M. Stray, A. Wickramanayake, B. Norquist, K. P. Pennington, R. L. Garcia, M.-C. King, and E. M. Swisher. Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. *Proceedings of the National Academy of Sciences*, 108(44):18032–18037, nov 2011.

- [151] Amit R Indap, Regina Cole, Christina L Runge, Gabor T Marth, and Michael Olivier. Variant discovery in targeted resequencing using whole genome amplified DNA. *BMC Genomics*, 14(1):468, jul 2013.
- [152] NCI Genomic Data Commons. Mutect2 insertion artifacts, 2016. Accessed: 2016/12/08.
- [153] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219, mar 2013.
- [154] Kristopher A. Standish, Tristan M. Carland, Glenn K. Lockwood, Wayne Pfeiffer, Mahidhar Tatineni, C Chris Huang, Sarah Lamberth, Yauheniya Cherkas, Carrie Brodmerkel, Ed Jaeger, Lance Smith, Gunaretnam Rajagopal, Mark E. Curran, and Nicholas J. Schork. Group-based variant calling leveraging next-generation super-computing for large-scale whole-genome sequencing studies. *BMC Bioinformatics*, 16(1):304, dec 2015.
- [155] Alex A.T. Bui and John Darrell Van Horn. Envisioning the future of ‘big data’ biomedicine. *Journal of Biomedical Informatics*, 69:115–117, may 2017.
- [156] Xiuzhen Huang, Steven F Jennings, Barry Bruce, Alison Buchan, Liming Cai, Pengyin Chen, Carole L Cramer, Weihua Guan, Uwe KK Hilgert, Hongmei Jiang, Zenglu Li, Gail McClure, Donald F McMullen, Bindu Nanduri, Andy Perkins, Bhanu Rekepalli, Saeed Salem, Jennifer Specker, Karl Walker, Donald Wunsch, Donghai Xiong, Shuzhong Zhang, Yu Zhang, Zhongming Zhao, and Jason H Moore. Big data - A 21st century science Maginot Line? No-boundary thinking: Shifting from the big data paradigm. *BioData Mining*, 8(1):7, jun 2015.
- [157] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saaboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, Susan Sullivan, Jason Bridge, Harold Magazine, John Syron, Johnelle Fleming, Laura Siminoff, Heather Traino, Maghboeba Mosavel, Laura Barker, Scott Jewell, Dan Rohrer, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Eric Hudson, Kristin Feenstra, Leslie Sobin, James Robb, Phillip Branton, Greg Korzeniewski, Charles Shive, David Tabor, Liqun Qi, Kevin Groch, Sreenath Nampally, Steve Buia, Angela Zimmerman, Anna Smith, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, Mark Cosentino, Norma Diaz-Mayoral, Mary Kennedy, Theresa Engel, Penelope Williams, Kenyon Erickson, Kristin Ardlie, Wendy Winckler, Gad Getz, David DeLuca, Daniel MacArthur, Manolis Kellis, Alexander Thomson, Taylor Young, Ellen Gelfand, Molly Donovan, Yan Meng, George Grant, Deborah Mash, Yvonne Marcus, Margaret Basile, Jun Liu, Jun Zhu, Zhidong Tu, Nancy J Cox, Dan L Nicolae, Eric R Gamazon, Hae Kyung Im, Anuar

- Konkashbaev, Jonathan Pritchard, Matthew Stevens, Timothée Flutre, Xiaoquan Wen, Emmanouil T Dermitzakis, Tuuli Lappalainen, Roderic Guigo, Jean Monlong, Michael Sammeth, Daphne Koller, Alexis Battle, Sara Mostafavi, Mark McCarthy, Manual Rivas, Julian Maller, Ivan Rusyn, Andrew Nobel, Fred Wright, Andrey Shabalina, Mike Feolo, Nataliya Sharopova, Anne Sturcke, Justin Paschal, James M Anderson, Elizabeth L Wilder, Leslie K Derr, Eric D Green, Jeffery P Struewing, Gary Temple, Simona Volpi, Joy T Boyer, Elizabeth J Thomson, Mark S Guyer, Cathy Ng, Assya Abdallah, Deborah Colantuoni, Thomas R Insel, Susan E Koester, A Roger Little, Patrick K Bender, Thomas Lehner, Yin Yao, Carolyn C Compton, Jimmie B Vaught, Sherilyn Sawyer, Nicole C Lockhart, Joanne Demchok, and Helen F Moore. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, may 2013.
- [158] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, 42(Database issue):D980–5, jan 2014.
- [159] Liacine Bouaoun, Dmitriy Sonkin, Maude Ardin, Monica Hollstein, Graham Byrnes, Jiri Zavadil, and Magali Olivier. TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Human Mutation*, 37(9):865–876, sep 2016.
- [160] Martin Kircher, Daniela M Witten, Preti Jain, Brian J. O’roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014.
- [161] Robert D. Finn, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A. Salazar, John Tate, and Alex Bateman. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, jan 2016.
- [162] Daniel R. Zerbino, Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G. Izuogu, Sophie H. Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R. Laird, Ilias Lavidas, Zhicheng Liu, Jane E. Loveland, Thomas Maurer, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N. Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E. Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J. Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M. Staines, Stephen J.

- Trevanion, Bronwen L. Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, nov 2018.
- [163] Alexey Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*, page 060012, jun 2016.
- [164] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, oct 2005.
- [165] Rachel Rosenthal, Nicholas McGranahan, Javier Herrero, Barry S. Taylor, and Charles Swanton. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, dec 2016.
- [166] Simon A. Forbes, David Beare, Harry Boutselakis, Sally Bamford, Nidhi Bindal, John Tate, Charlotte G. Cole, Sari Ward, Elisabeth Dawson, Laura Ponting, Raymond Stefancsik, Bhavana Harsha, Chai Yin Kok, Mingming Jia, Harry Jubb, Zbyslaw Sondka, Sam Thompson, Tisham De, and Peter J. Campbell. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, jan 2017.
- [167] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2):175–191, may 2007.
- [168] Lauren Fishbein, Ignaty Leshchiner, Vonn Walter, Ludmila Danilova, A. Gordon Robertson, Amy R. Johnson, Tara M. Lichtenberg, Bradley A. Murray, Hans K. Ghayee, Tobias Else, Shiyun Ling, Stuart R. Jefferys, Aguirre A. de Cubas, Brandon Wenz, Esther Korpershoek, Antonio L. Amelio, Liza Makowski, W. Kimryn Rathmell, Anne Paule Gimenez-Roqueplo, Thomas J. Giordano, Sylvia L. Asa, Arthur S. Tischler, Rehan Akbani, Adrian Ally, Laurence Amar, Antonio L. Amelio, Harindra Arachchi, Sylvia L. Asa, Richard J. Auchus, J. Todd Auman, Robert Baertsch, Miruna Balasundaram, Saianand Balu, Detlef K. Bartsch, Eric Baudin, Thomas Bauer, Allison Beaver, Christopher Benz, Rameen Beroukhim, Felix Beuschlein, Tom Bodenheimer, Lori Boice, Jay Bowen, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Suzie Carter, Clarissa A. Cassol, Andrew D. Cherniack, Lynda Chin, Juok Cho, Eric Chuah, Sudha Chudamani, Leslie Cope, Daniel Crain, Erin Curley, Ludmila Danilova, Aguirre A. de Cubas, Ronald R. de Krijger, John A. Demchok, Timo Deutschbein, Noreen Dhalla, David Dimmock, Winand N.M. Dinjens, Tobias Else, Charis Eng, Jennifer Eschbacher, Martin Fassnacht, Ina Felau, Michael Feldman, Martin L. Ferguson, Ian Fiddes, Lauren Fishbein, Scott Frazer, Stacey B. Gabriel, Johanna Gardner, Julie M. Gastier-Foster, Nils Gehlenborg, Mark Gerken, Gad

Getz, Jennifer Geurts, Hans K. Ghayee, Anne Paule Gimenez-Roqueplo, Thomas J. Giordano, Mary Goldman, Kiley Graim, Manaswi Gupta, David Haan, Stefanie Hahner, Constanze Hantel, David Haussler, D. Neil Hayes, David I. Heiman, Katherine A. Hoadley, Robert A. Holt, Alan P. Hoyle, Mei Huang, Bryan Hunt, Carolyn M. Hutter, Stuart R. Jefferys, Amy R. Johnson, Steven J.M. Jones, Corbin D. Jones, Katayoon Kasaian, Electron Kebebew, Jaegil Kim, Patrick Kimes, Theo Knijnenburg, Esther Korpershoek, Eric Lander, Michael S. Lawrence, Ronald Lechan, Darlene Lee, Kristen M. Leraas, Antonio Lerario, Ignaty Leshchiner, Tara M. Lichtenberg, Pei Lin, Shiyun Ling, Jia Liu, Virginia A. LiVolsi, Laxmi Lolla, Yair Lotan, Yiling Lu, Yussanne Ma, Nicole Maison, Liza Makowski, David Mallery, Massimo Mannelli, Jessica Marquard, Marco A. Marra, Thomas Matthew, Michael Mayo, Tchao Méatchi, Shaowu Meng, Maria J. Merino, Ozgur Mete, Matthew Meyerson, Piotr A. Mieczkowski, Gordon B. Mills, Richard A. Moore, Olena Morozova, Scott Morris, Lisle E. Mose, Andrew J. Mungall, Bradley A. Murray, Rashi Naresh, Katherine L. Nathanson, Yulia Newton, Sam Ng, Ying Ni, Michael S. Noble, Fiemu Nwariaku, Karel Pacak, Joel S. Parker, Evan Paul, Robert Penny, Charles M. Perou, Amy H. Perou, Todd Pihl, James Powers, Jennifer Rabaglia, Amie Radenbaugh, Nilsa C. Ramirez, Arjun Rao, W. Kimryn Rathmell, Anna Riester, Jeffrey Roach, A. Gordon Robertson, Sara Sadeghi, Gordon Saksena, Sofie Salama, Charles Saller, George Sandusky, Silviu Sbiera, Jacqueline E. Schein, Steven E. Schumacher, Candace Shelton, Troy Shelton, Margi Sheth, Yan Shi, Juliann Shih, Ilya Shmulevich, Janae V. Simons, Payal Sipahimalani, Tara Skelly, Heidi J. Sofia, Artem Sokolov, Matthew G. Soloway, Carrie Sougnez, Josh Stuart, Charlie Sun, Teresa Swatloski, Angela Tam, Donghui Tan, Roy Tarnuzzer, Katherine Tarvin, Nina Thiessen, Leigh B. Thorne, Henri J. Timmers, Arthur S. Tischler, Kane Tse, Vlado Uzunangelov, Anouk van Berkel, Umadevi Veluvolu, Ales Vicha, Doug Voet, Jens Waldmann, Vonn Walter, Yunhu Wan, Zhining Wang, Tracy S. Wang, Joellen Weaver, John N. Weinstein, Dirk Weismann, Brandon Wenz, Matthew D. Wilkerson, Lisa Wise, Tina Wong, Christopher Wong, Ye Wu, Liming Yang, Tomas Zelinka, Jean C. Zenklusen, Jia-shan (Julia) Zhang, Wei Zhang, Jingchun Zhu, Franck Zinzindohoué, Erik Zmuda, Karel Pacak, Katherine L. Nathanson, and Matthew D. Wilkerson. Comprehensive Molecular Characterization of Pheochromocytoma and Paraganglioma. *Cancer Cell*, 31(2):181–193, feb 2017.

- [169] Nelly Burnichon, Jean-Jacques Brière, Rossella Libé, Laure Vescovo, Julie Rivière, Frédérique Tissier, Elodie Jouanno, Xavier Jeunemaitre, Paule Bénit, Alexander Tzagoloff, Pierre Rustin, Jérôme Bertherat, Judith Favier, and Anne-Paule Gimenez-Roqueplo. SDHA is a tumor suppressor gene causing paraganglioma. *Human Molecular Genetics*, 19(15):3011–3020, aug 2010.
- [170] Bo Liu, Nicholas C. Nicolaides, Sanford Markowitz, James K.V. Willson, Ramon E. Parsons, Jin Jen, Nickolas Papadopolous, Päivi Peltomäki, Albert de la Chapelle, Stanley R. Hamilton, Kenneth W. Kinzler, and Bert Vogelstein. Mismatch repair

- gene defects in sporadic colorectal cancers with microsatellite instability. *Nature Genetics*, 9(1):48–55, jan 1995.
- [171] Scott D McCulloch and Thomas A Kunkel. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research*, 18(1):148–161, jan 2008.
- [172] Cyriac Kandoth, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F. McMichael, Matthew A. Wyczalkowski, Mark D.M. Leiserson, Christopher A. Miller, John S. Welch, Matthew J. Walter, Michael C. Wendl, Timothy J. Ley, Richard K. Wilson, Benjamin J. Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, oct 2013.
- [173] Julie M Cunningham, Eric R Christensen, David J Tester, Cheong Yong Kim, Patrick C Roche, Lawrence J Burgart, and Stephen N Thibodeau. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Research*, 58(15):3455–3460, aug 1998.
- [174] M Esteller, M Toyota, M Sanchez-Cespedes, G Capella, M A Peinado, D N Watkins, J P Issa, D Sidransky, S B Baylin, and J G Herman. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is associated with G to A mutations in K-ras in colorectal tumorigenesis. *Cancer research*, 60(9):2368–71, may 2000.
- [175] Akira Motegi, Raman Sood, Helen Moinova, Sanford D Markowitz, Pu Paul Liu, and Kyungjae Myung. Human SHPRH suppresses genomic instability through proliferating cell nuclear antigen polyubiquitination. *Journal of Cell Biology*, 175(5):703–708, dec 2006.
- [176] Mev Dominguez-Valentin, Christina Therkildsen, Srinivas Veerla, Mats Jönsson, Inge Bernstein, Ake Borg, and Mef Nilbert. Distinct gene expression signatures in lynch syndrome and familial colorectal cancer type x. *PloS one*, 8(8):e71755, 2013.
- [177] Sharon R Browning and Brian L Browning. Haplotype phasing: existing methods and new developments. *Nature Publishing Group*, 12, 2011.
- [178] Peter Van Loo, Silje H Nordgard, O. C. Lingjaerde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M Perou, A.-L. Borresen-Dale, and Vessela N Kristensen. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, sep 2010.
- [179] Fatima Zare, Michelle Dow, Nicholas Monteleone, Abdelrahman Hosny, and Sheida Nabavi. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics*, 18(1):286, may 2017.

- [180] Ninad Dewal, Yang Hu, Matthew L Freedman, Thomas LaFramboise, and Itsik Pe'Er. Calling amplified haplotypes in next generation tumor sequence data. *Genome Research*, 22(2):362–374, feb 2012.
- [181] Katharina Wimmer, Christian P Kratz, Hans F.A. Vasen, Olivier Caron, Chrys-telle Colas, Natacha Entz-Werle, Anne Marie Gerdes, Yael Goldberg, Denisa Ilen-cikova, Martine Muleris, Alex Duval, Noémie Lavoine, Clara Ruiz-Ponte, Irene Slave, Brigit Burkhardt, and Laurence Brugieres. Diagnostic criteria for constitutional mis-match repair deficiency syndrome: Suggestions of the European consortium 'Care for CMMRD' (C4CMMRD). *Journal of Medical Genetics*, 51(6):355–365, 2014.
- [182] Stephane E Castel, Pejman Mohammadi, Wendy K Chung, Yufeng Shen, and Tuuli Lappalainen. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nature Communications*, 7, 2016.
- [183] Olivier Delaneau, Jonathan Marchini, Gil A. McVean, Peter Donnelly, Gerton Lunter, Jonathan L. Marchini, Simon Myers, Anjali Gupta-Hinch, Zamin Iqbal, Iain Mathieson, Andy Rimmer, Dionysia K. Xifara, Angeliki Kerasidou, Claire Churchhouse, Olivier Delaneau, David M. Altshuler, Stacey B. Gabriel, Eric S. Lander, Namrata Gupta, Mark J. Daly, Mark A. DePristo, Eric Banks, Gaurav Bhatia, Mauricio O. Carneiro, Guillermo del Angel, Giulio Genovese, Robert E. Handsaker, Chris Hart, Steven A. McCarroll, James C. Nemes, Ryan E. Poplin, Stephen F. Schaffner, Khalid Shakir, Pardis C. Sabeti, Sharon R. Grossman, Shervin Tabrizi, Ridhi Tariya, Heng Li, David Reich, Richard M. Durbin, Matthew E. Hurles, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Qasim Ayub, Yuan Chen, Alison J. Coffey, Vincenza Colonna, Ni Huang, Luke Jostins, Aylwyn Scally, Klaudia Walter, Yali Xue, Yujun Zhang, Ben Blackburne, Sarah J. Lind-say, Zemin Ning, Adam Frankish, Jennifer Harrow, Chris Tyler-Smith, Gonalo R. Abecasis, Hyun Min Kang, Paul Anderson, Tom Blackwell, Fabio Busonero, Chris-tian Fuchsberger, Goo Jun, Andrea Maschio, Eleonora Porcu, Carlo Sidore, Adrian Tan, Mary Kate Trost, David R. Bentley, Russell Grocock, Sean Humphray, Ter-ena James, Zoya Kingsbury, Markus Bauer, R. Keira Cheetham, Tony Cox, Michael Eberle, Lisa Murray, Richard Shaw, Aravinda Chakravarti, Andrew G. Clark, Alon Keinan, Juan L. Rodriguez-Flores, Francisco M. De La Vega, Jeremiah Degen-hardt, Evan E. Eichler, Paul Flicek, Laura Clarke, Rasko Leinonen, Richard E. Smith, Xiangqun Zheng-Bradley, Kathryn Beal, Fiona Cunningham, Javier Herrero, William M. McLaren, Graham R. S. Ritchie, Jonathan Barker, Gavin Kelman, Eu-gene Kulesha, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Ian Streeter, Iliana Toneva, Richard A. Gibbs, Huyen Dinh, Christie Kovar, Sandra Lee, Lora Lewis, Donna Muzny, Jeff Reid, Min Wang, Fuli Yu, Matthew Bainbridge, Danny Challis, Uday S. Evani, James Lu, Uma Nagaswamy, Aniko Sabo, Yi Wang, Jin Yu, Gerald Fowler, Walker Hale, Divya Kalra, Eric D. Green, Bartha M. Knop-pers, Jan O. Korbel, Tobias Rausch, Adrian M. Sttz, Charles Lee, Lauren Grif-

fin, Chih-Heng Hsieh, Ryan E. Mills, Marcin von Grotthuss, Chengsheng Zhang, Xinghua Shi, Hans Lehrach, Ralf Sudbrak, Vyacheslav S. Amstislavskiy, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie-Laure Yaspo, Marie-Laure Sudbrak, Ralf Herwig, Elaine R. Mardis, Richard K. Wilson, Lucinda Fulton, Robert Fulton, George M. Weinstock, Asif Chinwalla, Li Ding, David Dooling, Daniel C. Koboldt, Michael D. McLellan, John W. Wallis, Michael C. Wendl, Qunyuan Zhang, Gabor T. Marth, Erik P. Garrison, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Deborah A. Nickerson, Can Alkan, Fereydoun Hormozdiari, Arthur Ko, Peter H. Sudmant, Jeanette P. Schmidt, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Stephen T. Sherry, Chunlin Xiao, Deanna Church, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Jun Wang, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Guoqing Li, Jingxiang Li, Yingrui Li, Xiao Liu, Yao Lu, Xuedi Ma, Shuaishuai Tai, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Ye Yin, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Lachlan J.M. Coin, Lin Fang, Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Bingqiang Wang, Yinlong Xie, Chen Ye, Chang Yu, Hancheng Zheng, Hongmei Zhu, Hongyu Cai, Hongzhi Cao, Yeyang Su, Zhongming Tian, Huanming Yang, Ling Yang, Jiayong Zhu, Zhiming Cai, Jian Wang, Marcus W. Albrecht, Tatiana A. Borodina, Adam Auton, Seungtae C. Yoon, Jayon Lihm, Vladimir Makarov, Hanjun Jin, Wook Kim, Ki Cheol Kim, Srikanth Gottipati, Danielle Jones, David N. Cooper, Edward V. Ball, Peter D. Stenson, Bret Barnes, Scott Kahn, Kai Ye, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Mark D. Shriver, Carlos D. Bustamante, Simon Gravel, Eimear E. Kenny, Jeffrey M. Kidd, Phil Lacroute, Brian K. Maples, Andres Moreno-Estrada, Fouad Zakharia, Brenna Henn, Karla Sandoval, Jake K. Byrnes, Eran Halperin, Yael Baran, David W. Craig, Alexis Christoforides, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Nils Homer, Kevin Squire, Jonathan Sebat, Vineet Bafna, Kenny Ye, Esteban G. Burchard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman, W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Tuuli Lappalainen, Scott E. Devine, Xinyue Liu, Ankit Maroo, Luke J. Tallon, Jeffrey A. Rosenfeld, Leslie P. Michelson, Andrea Angius, Francesco Cucca, Serena Sanna, Abigail Biggam, Chris Jones, Fred Reinier, Yun Li, Robert Lyons, David Schlessinger, Philip Awadalla, Alan Hodgkinson, Taras K. Oleksyk, Juan C. Martinez-Cruzado, Yunxin Fu, Xiaoming Liu, Momi Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Brian L. Browning, Iman Hajirasouliha, Ken Chen, Cornelis A. Albers, Mark B. Gerstein, Alexej Abyzov, Jieming Chen, Yao Fu, Lukas Habegger, Arif O. Harmanci, Xinmeng Jasmine Mu, Cristina Sisú, Suganthi Balasubramanian, Mike Jin, Ekta Khurana, Declan Clarke, Jacob J. Michaelson, Chris OSullivan, Kathleen C. Barnes, Neda Gharani, Lor-

- raine H. Toji, Norman Gerry, Jane S. Kaye, Alastair Kent, Rasika Mathias, Pilar N. Ossorio, Michael Parker, Charles N. Rotimi, Charmaine D. Royal, Sarah Tishkoff, Marc Via, Walter Bodmer, Gabriel Bedoya, Gao Yang, Chu Jia You, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Nicholas C. Clegg, Mark S. Guyer, Jane L. Peterson, Audrey Duncanson, Michael Dunn, Leena Peltonenz, and Leena Peltonenz. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nature Communications*, 5:3934, jun 2014.
- [184] Peter Edge, Vineet Bafna, and Vikas Bansal. HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research*, 27(5):801–812, may 2017.
- [185] Yulia Mostovoy, Michal Levy-Sakin, Jessica Lam, Ernest T Lam, Alex R Hastie, Patrick Marks, Joyce Lee, Catherine Chu, Chin Lin, Željko Džakula, Han Cao, Stephen A Schlebusch, Kristina Giorda, Michael Schnall-Levin, Jeffrey D Wall, and Pui-Yan Kwok. A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, 13(7):587–590, jul 2016.
- [186] Adam B. Olshen, E. S. Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, oct 2004.
- [187] Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC bioinformatics*, 14 Suppl 1(Suppl 11):S1, sep 2013.
- [188] Victor Guryev, Bart M. G. Smits, Jose van de Belt, Mark Verheul, Norbert Hubner, and Edwin Cuppen. Haplotype Block Structure Is Conserved across Mammals. *PLoS Genetics*, 2(7):e121, 2006.
- [189] David W Craig, Sara Nasser, Richard Corbett, Simon K Chan, Lisa Murray, Christophe Legendre, Waibhav Tembe, Jonathan Adkins, Nancy Kim, Shukmei Wong, Angela Baker, Daniel Enriquez, Stephanie Pond, Erin Pleasance, Andrew J Mungall, Richard A Moore, Timothy McDaniel, Yussanne Ma, Steven J.M. Jones, Marco A Marra, John D Carpten, and Winnie S Liang. A somatic reference standard for cancer genome sequencing. *Scientific Reports*, 6:24607, apr 2016.
- [190] Douglas F Easton, Amie M Deffenbaugh, Dmitry Pruss, Cynthia Frye, Richard J Wenstrup, Kristina Allen-Brady, Sean V Tavtigian, Alvaro N.A. Monteiro, Edwin S Iversen, Fergus J Couch, and David E Goldgar. A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the BRCA1 and BRCA2 Breast Cancer Predisposition Genes. *The American Journal of Human Genetics*, 81(5):873–883, 2007.

- [191] Thomas LaFramboise, Barbara A. Weir, Xiaojun Zhao, Rameen Beroukhim, Cheng Li, David Harrington, William R. Sellers, and Matthew Meyerson. Allele-specific amplification in cancer revealed by SNP array analysis. *PLoS Computational Biology*, 1(6):0507–0517, 2005.
- [192] C Andrew Stewart, Roger Horton, Richard J.N. Allock, Jennifer L Ashurst, Alexey M Atrazhev, Penny Coggill, Ian Dunham, Simon Forbes, Karen Halls, Joanna M.M. Howson, Sean J Humphray, Sarah Hunt, Andrew J Mungall, Kazutoyo Osoegawa, Sophie Palmer, Anne N Roberts, Jane Rogers, Sarah Sims, Yu Wang, Laurens G Wilming, John F. Elliot, Pieter J de Jong, Stephen Sawcer, John A Todd, John Trowsdale, and Stephan Beck. Complete MHC haplotype sequencing for common disease gene mapping. *Genome Research*, 14(6):1176–1187, jun 2004.
- [193] Kazuyoshi Hosomichi, Timothy A Jinam, Shigeki Mitsunaga, Hirofumi Nakaoka, and Ituro Inoue. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. *BMC Genomics*, 14(1):355, may 2013.
- [194] Diego Chowell, Luc G T Morris, Claud M Grigg, Jeffrey K Weber, Robert M Samstein, Vladimir Makarov, Fengshen Kuo, Sviatoslav M Kendall, David Requena, Nadeem Riaz, Benjamin Greenbaum, James Carroll, Edward Garon, David M Hyman, Ahmet Zehir, David Solit, Michael Berger, Ruhong Zhou, Naiyer A Rizvi, and Timothy A Chan. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science (New York, N.Y.)*, 359(6375):582–587, dec 2018.
- [195] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, apr 2010.
- [196] Prateek Kumar, Steven Henikoff, and Pauline C Ng. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, 4(7):1073–1081, jul 2009.
- [197] M. Kelly and C. Semsarian. Multiple Mutations in Genetic Cardiovascular Disease: A Marker of Disease Severity? *Circulation: Cardiovascular Genetics*, 2(2):182–190, apr 2009.
- [198] Lea M Starita, David L Young, Muhtadi Islam, Jacob O Kitzman, Justin Gullingsrud, Ronald J Hause, Douglas M Fowler, Jeffrey D Parvin, Jay Shendure, and Stanley Fields. Massively parallel functional analysis of BRCA1 RING domain variants. *Genetics*, 200(2):413–422, jun 2015.
- [199] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, Dorret I Boomsma, Fred A Wright, Patrick F Sullivan, Elina Nikkola, Marcus Alvarez, Mete Civelek, Aldons J Lusis,

Terho Lehtimäki, Emma Raitoharju, Mika Kähönen, Ilkka Seppälä, Olli T Raitakari, Johanna Kuusisto, Markku Laakso, Alkes L Price, Päivi Pajukanta, and Bogdan Pasaniuc. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, mar 2016.

- [200] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, jan 2011.