DECIPHERING CANCER DEVELOPMENT AND PROGRESSION THROUGH LARGE-SCALE COMPUTATIONAL ANALYSES OF GERMLINE AND SOMATIC GENOMES

A DISSERTATION SUBMITTED TO THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES AND THE PRITZKER SCHOOL OF MEDICINE IN CANDIDACY FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

JASON JAMES PITT

CHICAGO, ILLINOIS
AUGUST 2017
# Table of Contents

**LIST OF FIGURES** ................................................................. v

**LIST OF TABLES** ................................................................. vii

**ACKNOWLEDGMENTS** .............................................................. viii

**ABSTRACT** ........................................................................... x

1 INTRODUCTION ........................................................................ 1

1.1 Cancer — a public health challenge ........................................ 1

1.2 Cancer — a genetic disease .................................................... 3

1.3 Monogenic cancer predisposition ............................................ 7

1.4 Cancer as a genetically complex disease ................................. 11

1.5 Undercutting cancer’s resilience ............................................. 13

1.6 Genetic identification of high risk individuals .......................... 14

1.7 Clinical and genomic features of breast cancer ......................... 16

1.8 Racial/ethnic disparities in breast cancer ................................. 18

2 ROBUST SCALING OF DNA SEQUENCING ANALYSES USING THE MODULAR SWIFTSEQ WORKFLOW ....................................................... 20

2.1 Introduction ........................................................................ 20

2.2 Results ................................................................................ 21

2.2.1 Anatomy of a SwiftSeq run — input gathering ....................... 21

2.2.2 Anatomy of a SwiftSeq run — initiation ............................... 23

2.2.3 Anatomy of a SwiftSeq run — execution ............................. 24

2.2.4 Bioinformatic nuances handles under-the-hood .................... 25

2.2.5 Maximizing performance — parallelization strategies ............ 26

2.2.6 Maximizing performance — efficiency and scalability .......... 28

2.2.7 Facilitating tumor-normal pair analyses ............................... 30

2.2.8 Flexible analyses through a graphical user interface ............... 33

2.2.9 Portability across systems ................................................ 34

2.3 Discussion .......................................................................... 35

3 AGGREGATE ALLELIC BURDEN FOR CANCER RISK GENES ASSOCIATES WITH AGE AT DIAGNOSIS ..................................................... 38

3.1 Introduction ........................................................................ 38

3.2 Results ................................................................................ 39

3.2.1 Allele burden is negatively associated with age at diagnosis .... 39

3.2.2 Orthogonal support from seven control analyses ................ 41

3.2.3 Enrichment analyses of cancer-associated variants and genes 44

3.2.4 Allele burden helps interpret variants of unknown significance 47

3.2.5 High allele burden acts independently of $BRCA1/2$ in breast cancer 51
3.3 Discussion ................................................................. 55
3.4 Supplementary information ........................................... 56
  3.4.1 Supplementary tables ............................................. 56

4 COMBINING COMPUTATIONAL AND FUNCTIONAL ANALYSES TO IDENTIFY NOVEL TWO-HIT TUMOR SUPPRESSOR GENES ................................................................. 58
  4.1 Introduction ........................................................... 58
  4.2 Results ........................................................................ 60
    4.2.1 Two-hit identification strategy .............................. 60
    4.2.2 Quantifying two-hit frequency in known cancer predisposition genes ........................ 62
    4.2.3 Novel two-hit genes pan-cancer ............................... 64
    4.2.4 Sexual dimorphism in two-hit acquisition ................ 65
    4.2.5 Functional analysis of pan-cancer candidates .......... 66
    4.2.6 Characterization of cancer-specific candidates ROBO1 and DBRI .................................... 68
    4.2.7 ROBO1 knockdown represses DNA damage response ................................. 70
  4.3 Discussion ................................................................. 70

5 COMPARISON OF BREAST CANCER MUTATIONAL PATTERNS ACROSS AFRICAN AND EUROPEAN ANCESTRY POPULATIONS ................................................................. 73
  5.1 Introduction ............................................................. 73
  5.2 Results ........................................................................ 74
    5.2.1 Mutational landscape across study populations .... 74
    5.2.2 Mutation signatures across subtypes and driver mutations ........................................ 78
    5.2.3 Mutation signatures across races/ethnicities ......... 83
    5.2.4 The APOBEC-HRD signature balance ................. 89
    5.2.5 Tumor immune microenvironment characterization ................................. 93
  5.3 Discussion ................................................................. 95
  5.4 Supplementary information ........................................... 98
    5.4.1 Supplementary tables ............................................. 98

6 CONCLUSION ...................................................................... 100
  6.1 Current and future endeavors in data-intensive genomics ................................................. 100
  6.2 Implications of age at diagnosis and harmful allele burden ............................................... 102
  6.3 Two-hit genes in cancer risk, development, and progression ........................................... 103
  6.4 Understanding racial/ethnic disparities in breast cancer .................................................. 105

7 MATERIALS AND METHODS .................................................. 107
  7.1 Processing blood germline exomes ................................ 107
  7.2 Allele-specific copy number analysis in tumors ....................................................... 110
  7.3 ClinVar variants and genes ............................................ 110
  7.4 Classifying deleterious variants in exomes .......................................................... 112
  7.5 Age at diagnosis and allele burden associations .................................................... 113
  7.6 Synchronous/bilateral clinical data extraction ..................................................... 114
  7.7 One- versus two-hit assessment ....................................... 114
## List of Figures

2.1 Diagram of a SwiftSeq run. ................................................................. 22
2.2 SwiftSeq parallelization strategies. ................................................... 27
2.3 SwiftSeq processing speed compared to standard pipeline approaches. ... 29
2.4 Naive exome scaling tests with SwiftSeq ........................................... 31
2.5 Comparing optimized and naive exome scaling tests. ......................... 32
2.6 Tumor-normal pair directory structure. ............................................. 33
2.7 Graphical user interface scheme for designing and retrieving workflows. 35

3.1 Samples per cancer type and deleterious allele counts per individual. .... 40
3.2 Increased burden of harmful alleles in cancer risk genes is associated with earlier age at cancer diagnosis. .................................................. 42
3.3 Age at diagnosis by allele burden using the union and intersection of gene sets. ................................................................. 43
3.4 No observed relationship between age at diagnosis and allele burden using non-cancer ClinVar variants. .............................................. 45
3.5 Age at diagnosis against deleterious allele burden exome-wide and within random gene sets. ......................................................... 46
3.6 Age at diagnosis by allele burden in genes significantly somatically mutated in cancer. ................................................................. 47
3.7 Enrichment of deleterious alleles in individuals with cancer. ............... 48
3.8 Age at diagnosis associated with allele burden when high and moderate risk genes are excluded. ......................................................... 49
3.9 Associations remain after excluding alleles with predicted high impact on gene function. ................................................................. 50
3.10 Obese women with uterine/endometrial carcinoma are diagnosed earlier. 51
3.11 *BRCA1/2* carrier status and high allele burden independently associate with earlier breast cancer diagnosis. ........................................ 53
3.12 Highly burdened individuals have earlier breast cancer diagnosis when excluding terminal *BRCA2* variants and *BRCA1/2* carriers. ............... 54

4.1 Workflow for identifying candidate two-hit genes. ............................ 61
4.2 Two-hit enrichment pan- and per-cancer type. .................................... 63
4.3 Top male and female candidates display sexually dimorphism. ............. 66
4.4 Knockdown of candidate two-hit genes induces cancerous phenotypes. ... 67
4.5 *ROBO1* and *DBR1* show cell type-specific phenotypes. .................. 69

5.1 The number of Nigerian samples with each NGS data type. ................. 75
5.2 Landscape of breast cancer in Nigerians compared to Black and White Americans. ................................................................. 76
5.3 Lollipop plots for novel significantly mutated breast cancer genes. ......... 77
5.4 Oncoprint of short mutations and CNAs in Nigerians. ......................... 78
5.5 Tri-nucleotide substitution patterns of nine inferred mutation signatures. 79
5.6 Derived mutation signatures compared to COSMIC mutation signatures and correlation between WES and WGS signature contributions. .... 80
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.7 Correlations between WES and WGS mutation signature contributions.</td>
<td>81</td>
</tr>
<tr>
<td>5.8 Mutation signature contributions across race/ethnicity and subtype.</td>
<td>82</td>
</tr>
<tr>
<td>5.9 Mutation signature contributions between tumors positive and negative for IHC markers.</td>
<td>84</td>
</tr>
<tr>
<td>5.10 Associations between genome-wide oncogenic features and the mutation status of common driver genes.</td>
<td>85</td>
</tr>
<tr>
<td>5.11 The proportion of APOBEC C&gt;T, APOBEC C&gt;G, and aging signatures by race/ethnicity and IHC subtype using WES.</td>
<td>86</td>
</tr>
<tr>
<td>5.12 Mutation signature contributions and structural variant counts partitioned by race/ethnicity and IHC subtype.</td>
<td>87</td>
</tr>
<tr>
<td>5.13 Mutation signature contributions by race/ethnicity using WGS.</td>
<td>88</td>
</tr>
<tr>
<td>5.14 Driver genes associate with APOBEC and HRD signature balance in HR+/HER2-breast cancer.</td>
<td>91</td>
</tr>
<tr>
<td>5.15 Driver genes associate with APOBEC and HRD signature balance across all breast cancer IHC subtypes.</td>
<td>92</td>
</tr>
<tr>
<td>5.16 Gene signatures of immune cell infiltration.</td>
<td>94</td>
</tr>
<tr>
<td>5.17 Pairwise Pearson correlation of immune signatures and potential predictors of response to immunotherapy.</td>
<td>95</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>List of 57 ClinVar Cancer Genes.</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>List of 60 autosomal dominant cancer predisposition genes.</td>
<td>41</td>
</tr>
<tr>
<td>3.3</td>
<td>List of 21 moderate and high risk cancer predisposition genes.</td>
<td>49</td>
</tr>
<tr>
<td>3.4</td>
<td>Demographic and clinical information for individuals from The Cancer Genome Atlas.</td>
<td>56</td>
</tr>
<tr>
<td>3.5</td>
<td>Curated set of cancer-associated ClinVar variants.</td>
<td>56</td>
</tr>
<tr>
<td>3.6</td>
<td>List of 21 moderate and high risk cancer predisposition genes.</td>
<td>49</td>
</tr>
<tr>
<td>3.7</td>
<td>Unadjusted beta and $P$ values for age at diagnosis by allele burden linear models.</td>
<td>57</td>
</tr>
<tr>
<td>3.8</td>
<td>Burden $P$ values from age at diagnosis by allele burden linear models after adjusting for race and cancer type.</td>
<td>57</td>
</tr>
<tr>
<td>3.9</td>
<td>Genic leave-one-out regression analyses for ClinVar cancer genes.</td>
<td>57</td>
</tr>
<tr>
<td>3.10</td>
<td>Genic leave-one-out regression analyses for autosomal dominant cancer predisposition genes.</td>
<td>57</td>
</tr>
<tr>
<td>3.11</td>
<td>Cancer type leave-one-out regression analyses for ClinVar cancer genes.</td>
<td>57</td>
</tr>
<tr>
<td>3.12</td>
<td>Cancer type leave-one-out regression analyses for autosomal dominant cancer predisposition genes.</td>
<td>57</td>
</tr>
<tr>
<td>3.13</td>
<td>Genes associated with cancer phenotypes through genome-wide association studies.</td>
<td>57</td>
</tr>
<tr>
<td>4.1</td>
<td>TCGA cancer types and samples counts used for two-hit analyses.</td>
<td>61</td>
</tr>
<tr>
<td>4.2</td>
<td>Top pan-cancer two-hit genes.</td>
<td>64</td>
</tr>
<tr>
<td>5.1</td>
<td>Summary statistics for WES, WGS, and RNA-seq samples.</td>
<td>98</td>
</tr>
<tr>
<td>5.2</td>
<td>Identifiers of WES samples, their tumor subtype by IHC, and race/ethnicity.</td>
<td>98</td>
</tr>
<tr>
<td>5.3</td>
<td>Identifiers of WGS samples, their tumor subtype by IHC, and race/ethnicity.</td>
<td>98</td>
</tr>
<tr>
<td>5.4</td>
<td>Identifiers of RNA-seq samples, their tumor subtype by PAM50, and race/ethnicity.</td>
<td>98</td>
</tr>
<tr>
<td>5.5</td>
<td>List of 44 driver genes mutated by short variants in breast cancer.</td>
<td>98</td>
</tr>
<tr>
<td>5.6</td>
<td>List of 19 genes recurrently altered by CNAs in breast cancer.</td>
<td>98</td>
</tr>
<tr>
<td>5.7</td>
<td>Summary statistics for WES and WGS samples used for mutation signature analysis.</td>
<td>98</td>
</tr>
<tr>
<td>5.8</td>
<td>Identifiers of samples used for mutation signature analysis, and their sequencing data type.</td>
<td>98</td>
</tr>
<tr>
<td>5.9</td>
<td>Gene sets used for immune signature analyses.</td>
<td>99</td>
</tr>
</tbody>
</table>

1. Note: Additional tables are provided in a supplementary file distributed with this dissertation. The captions for these tables are also provided within each chapter’s Supplementary Information section.
ACKNOWLEDGMENTS

First and foremost, I would like to thank all of my collaborators over the past five years. Tackling these projects would have been impossible without their efforts.

With that being said, I must thank my advisor Kevin White for his fantastic mentorship. He has provided numerous opportunities that have benefited my projects and career. Kevin is incredibly loyal to his graduate students and passionately defends their interests. His insightful comments and criticisms have bolstered each of my projects. Most importantly, he has taught me to think like a rigorous — yet pragmatic — scientist.

The White lab has housed many great individuals during my tenure. Jason Grundstad was always a fantastic coworker. His personality naturally generated an enjoyable work environment. Casey Brown served as a valuable role model and resource throughout my early days in the lab. Chai Bandlamudi, a fellow graduate student, was always a much needed sounding board for problems, both scientific and otherwise. He has likely shaped my projects and thinking in more ways than I will ever know. Also, I appreciate all of the input and functional validation experiments from Mike Bolt and Vineet Dhiman.

I would like to recognize the efforts of my committee members. I will always remember each of your contributions to this arduous process. Specifically, I would like to thank Robert Grossman for introducing me to the wonders of biological data science; Andrey Rzhetsky for always inspiring me to be innovative; and Barbara Stranger for fantastic scientific and career advice. I appreciate all that you have done to help mold my scientific character.

There are four colleagues that deserve special thanks and recognition. Sanjive Qazi, my undergraduate mentor, invested a significant amount of time training me as a quantitatively minded scientist. Repaying his efforts has always been major a driving force for my endeavors. Dominic Fitzgerald was my first summer student and now is a White lab employee. He has made significant contributions to nearly all of my projects, and his work is always reliable and impressive. I am fortunate that we have had so many opportunities to work
and learn together. Lorenzo Pesce has helped me navigate numerous computational and strategic issues since the beginning of my Ph.D studies. In addition to being supportive, he has always been a fountain of knowledge and a great friend. Lastly, Peter Van Loo has been an excellent collaborator and mentor for the past five years. Some of my most memorable and invigorating scientific conversations have occurred with Peter. He is a bright, ardent, and humble young investigator, and I look forward to our continued work together.

For years I have worked with Funmi Olopade and colleagues on the Nigerian breast cancer project. In addition to Funmi, I have really enjoyed collaborating with — and learning from — Toshio Yoshimatsu, Yonglan Zheng, Shengfeng Wang, Dezheng Huo, Jordi Barretina, and Markus Riester. All of their efforts were crucial to the successes of this project. Likewise, I will always be grateful to Stefan Dentro for schooling me in the nuances of cancer life history analyses. Of course, I am severely indebted to the hard-working researchers and generous patients from Nigeria.

I cannot think of a better program administrator than Sue Levison. The level at which she cares and looks out for graduate students is simply incredible. The students and faculty alike are lucky to have her.

I would like to say thank you to Heather Bell, who kept me relaxed, happy, and focused during the final eight months of my doctoral work. She is easily one of the most kind and genuine individuals I have ever met. I have also been lucky to encounter many great graduate students, particularly Michael Turchin and Katie Igartua, who have become even better friends. Lastly, I appreciate my dog Roy whose ridiculous personality makes me laugh each day.
ABSTRACT

Advances in next-generation sequencing (NGS) have propelled genomics into a data-intensive science. Although sufficient hardware resources are necessary for large-scale NGS analyses, robust and scalable software is frequently the more formidable barrier. To address this need, I have been the principle contributor to the development of SwiftSeq, a modular and system-agnostic workflow for end-to-end analysis of NGS data. SwiftSeq offers significant benefits to both small- and large-scale analyses. Parallelization, synchronization, and execution site selection are managed automatically. Tasks are robust to transient software and localized hardware failures, keeping user intervention to a minimum. Analysis jobs can consistently scale to hundreds of nodes and thousands cores. Using a Cray XE6, SwiftSeq can produce annotated germline and somatic genotypes for standard depth whole exomes and genomes in approximately 36 minutes and 11 hours, respectively. SwiftSeq is freely available, and harmonized variant calls representing nearly 10,000 exomes from The Cancer Genome Atlas (TCGA) have been made available to the genomics community through the Bionimbus Protected Data Cloud.

The value of the aforementioned exome dataset is the abundance of unique biological insights it enables. I was interested in using germline cancer genetics to better understand epidemiological phenotypes, particularly age at diagnosis. 5-10% of cancers cases can be attributed to highly penetrant, inherited alleles, which often lead to earlier age at diagnosis. However, the polygenic nature of cancer risk loci and its relationship to age at diagnosis is less understood. Using 8,111 individuals from TCGA representing over 30 cancer types (> 99% solid tumors), I have shown that increased ClinVar and deleterious allele burden within ClinVar cancer risk genes is associated with earlier age at diagnosis. These findings were replicated using a second set of autosomal dominant cancer predisposition genes. Strikingly, high allele burden in breast cancer was an independent predictor of age at diagnosis, and its effect was comparable to mutations in $BRCA1/2$. Overall, greater levels of baseline x
genetic deficiencies likely render individuals more sensitive to somatic events leading to earlier tumorigenesis. Investigating individuals harmful alleles in aggregate could assist in clinical cancer risk assessment.

Combining the aforementioned variation with known mutational mechanisms, I was also able to identify putative cancer genes. The two-hit hypothesis asserts that many cancer risk genes require two-hits (i.e. biallelic loss) in order to promote cancerous phenotypes in cells. In the classical model, the first hit is an inherited deleterious allele, whereas the second is generated through loss-of-heterozygosity (LOH). By jointly analyzing LOH and deleterious, germline variants across 5,146 individuals, I found that the classic tumor suppressors \textit{BRCA1}, \textit{BRCA2}, and \textit{ATM} showed highest, pan-cancer enrichment for two-hit scenarios. Two other genes – \textit{PHLPP2} and \textit{KDEL} – also had a preponderance of two-hits. Performing siRNA knockdowns in multiple cells lines, Mike Bolt showed that reducing \textit{PHLPP2} and \textit{KDEL} expression promotes the cancer-like phenotypes proliferation and migration. Furthermore, malignancy-specific investigations provided strong computational and experimental evidence that \textit{ROBO1} is a novel two-hit gene in breast cancer. Overall, these analyses have shown that integrating germline and somatic genetics can reveal novel cancer genes.

Lastly, I examined how genetic background can affect the somatic mutational landscape. In breast cancer, women of African ancestry are diagnosed younger, have more clinically aggressive disease stage-for-stage, and have higher mortality rates than age-matched women of European or Asian ancestry. Using a combination of exome, genome, and RNA sequencing, Markus Riester and I examined the molecular features of breast cancers across 194 patients from Nigeria and 1,037 patients from the US in TCGA (171 Black, 753 White, 113 other). The mutational landscape and immune signature patterns differed across racial/ethnic populations. Triple Negative (43%) and HER2+ positive (25%) subtypes were enriched in Nigerians whose tumors were characterized by a higher \textit{TP53} mutation rate, increased
structural variation, and greater prevalence of the homologous recombination deficiency signature. *GATA3* mutations were highest in Nigerian hormone receptor positive tumors (25.9%). Higher proportions of APOBEC-mediated substitutions were strongly associated with *PIK3CA* and *CDH1* mutations, which were more prevalent in Whites. Additionally, I identified *PLK2*, *KDM6A*, *GPS2*, and *B2M* as novel significantly mutated genes in breast cancer. These data underscore the importance of genomic research in diverse populations to accelerate progress in precision oncology and reduce global disparities in outcomes.
CHAPTER 1
INTRODUCTION

1.1 Cancer — a public health challenge

Cancer occurs when a cell lineage divides uncontrollably, and this disease can affect almost any tissue in the human body [1]. Not bound by age, cancer is capable of afflicting the young and old alike [2]. Epidemiologically, overall cancer incidence rates have remained relatively stable in the United States (US) for the past two decades. With higher life expectancy fostering population growth, the number of yearly cancer cases is steadily rising. Approximately 1.7 million new diagnoses are expected throughout 2017 [3]. This plight is not exclusive to the US. Compared to 2008, it is estimated that the global cancer burden will nearly double by 2030 [4].

Over time, cancer incidence has shifted for specific populations and anatomical sites. These changes are often influenced by behavioral patterns. Lower tobacco usage amongst males has been a harbinger for diminishing lung cancer rates over the past 30 years. Conversely, lung cancer incidence in women has climbed over the same time period, which coincides with increased use [3]. Throughout Asia — especially the southeast — surging cigarette usage is augmenting lung cancer incidence [5, 6, 7]. Colorectal cancer incidence in young Americans (20 to 29 years of age) has risen sharply since the early 1970s. In contrast to baby boomers, millennials now have a startling two- and four-fold increased risk for colon and rectal cancer, respectively [8]. Unsurprisingly, inflated incidence often precipitates increased mortality [9]. These national and global trends further substantiate the need for public health programs intended to diminish cancer deaths.

Early detection through screening has long been a staple of improved cancer survivorship [10]. A small number of cancer-specific procedures have been developed to this end. For colon cancer, colonoscopies aim to detect pre-malignant polyps, which can then be surgically
removed before tumorigenesis [11]. Mammography, which has been implemented since the latter half of the 20th century, aims to identify breast cancer’s first intimations [12]. The Pap smear has been used to discern early stages of cervical cancer for nearly a century [13]. While experts still debate if screening techniques — particularly mammography — causally decrease mortality, early diagnosis unquestionably affiliates with favorable clinical outcomes [8, 14, 15, 16]. However, these methodologies can also have unintended consequences. The introduction of Prostate Specific Antigen (PSA) testing doubled the prostate cancer incidence rate from 1985 to 1990 [3]. This serves as a cautionary tale of overzealous screening since these additional cases represented asymptomatic disease [17]. Since unnecessary cancer treatment can reduce quality of life, this is non-negligible affair [18]. Epidemiologists and clinicians alike must continually evaluate these approaches to strike a balance between early diagnosis and overtreatment [19].

Across all cancer types, the 5-year survival rate has improved by 20-24% since the 1980s. Yet — despite pervasive screening programs and the development of precision therapies — cancer remains the second leading cause of death in the US today [3]. The Centers for Disease Control estimate that cancer will surpass heart disease to become the primary source of US mortality by 2020 [20]. So while scientific advancements have improved patient survivability, the number of lives continually cut short by cancer is far from acceptable. As a consequence, receiving a cancer diagnosis remains emotionally taxing and instills intense fright [21, 22]. To reduce this strife, researchers must continue leveraging technological capabilities to enhance our knowledge of cancer. As radiation therapy pioneer Marie Curie said, “Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.”
1.2 Cancer — a genetic disease

To effectively combat and prevent a disease, we must understand its genesis. Cancer initiation is catalyzed by the accumulation oncogenic, somatic mutations throughout a cell’s genome [1]. However, the first connection between genetics and cancer was far more coarse. In the early 20th century, Theodor Boveri’s landmark research demonstrated that chromosomes were the courier of hereditary information [23]. Soon after, he astutely predicted that abnormal chromosomal segregation during cell division — namely the acquirement of extra chromosome copies — was a source of oncogenic transformation [24]. Suspicious cytogenetics also prompted one of the next seminal insights into cancer etiology. Peter Nowell and colleagues detected a small, uncharacteristic chromosome in cancerous cells extracted from chronic myelogenous leukemia patients. This irregularity came to be known as the Philadelphia chromosome [25, 26]. Over a decade later, research by Janet Rowley pinpointed this abnormality as a coalescence of q arms from chromosomes 9 and 22 [27, 24]. It was eventually determined that this translocation produces a fusion between BCR and ABL1, and the subsequent protein product engenders constitutive activation of tyrosine kinase signalling [28]. Of course, translocation is not the only mechanism that spurs aberrant gene activity. Nobel prize winning work by Harold Varmus and Michael Bishop demonstrated the transformative ability of retroviruses through genomic insertion of the SRC gene [29, 24]. As time passed, technological advancements permitted researchers to interrogate cancer genomes at finer resolution. After the Human Genome Project was completed in 2003, the Wellcome Trust Sanger Institute launched the Cancer Gene Census — an effort to scan tumor DNA for small, sequence-based alterations [30, 31]. Finally, with after the development of next-generation sequencing (NGS), somatic mutations within individual cancer genomes could be comprehensively cataloged [32, 33].

Today, it is estimated that cells can require as few as three critical mutations, typically genic, to spawn malignancy [34]. Genes that facilitate clonal expansions when mutated are
known as drivers [35]. For most cancers, the sequence of driver mutations and its subsequent phenotypic impact is not known. Given the heterogeneity of driver mutations even within cancer types, it is unlikely this process is deterministic. However, rigorous experimentation in colorectal carcinoma has shown that — most frequently — the first clonal expansion is triggered by an inactivating \(\textit{APC}\) mutation. A subsequent expansion occurs with \(\textit{KRAS}\) activation, and finally a third mutation in one of a handful known cancer genes initiates tumorigenesis [36]. Even though definitive ordering of clonal driver mutations is non-trivial, implicating genes helps establish the steps and pathways that are crucial to oncogenic transformation.

In the simplest of terms, driver genes fall into two categories: oncogenes or tumor suppressors. The former is a gene whose increased activity promotes cancer phenotypes, while inactivation of the latter serves the same function [37]. Oncogenes often acquire activating mutations at “hotspots,” which are specific protein domains or amino acid residues. Examples of oncogenic hotspot mutations include H1047R in \(\textit{PIK3CA}\) [38], V600E in \(\textit{BRAF}\) [39], amino acid residues 12 and 13 in \(\textit{KRAS}\) [40], and PEST domain mutations in \(\textit{NOTCH1}\) [41]. Only a single copy of an oncogene needs to harbor an activating mutation to promote cancer development and progression. Conversely, depending on the tumor suppressor gene, monoallelic (haploinsufficient) or biallelic inactivation is required to confer cells with a selective advantage [42]. Akin to oncogenes, some tumor suppressors are cancer-promoting primarily when disrupted in hotspot regions, as evidenced by \(\textit{ERCC2}\) in bladder cancer [39, 43].

The categorization of some driver genes is more nuanced. In breast cancer, \(\textit{GATA3}\) accumulates small insertions and deletions (indels) in its terminal exons that are seemingly loss-of-function [44]. However, \(\textit{GATA3}\) mutant tumors have substantially higher \(\textit{GATA3}\) expression than those that are wild type (WT), suggesting that these indels promote \(\textit{GATA3}\) expression and should be considered activating [45, 46]. Some classifications can even be
tissue specific. \textit{NOTCH1} behaves as an oncogene in acute lymphoblastic leukemia \cite{41} and, paradoxically, like a tumor suppressor in squamous cell carcinomas \cite{47}. Importantly, functional elements affecting cancer development are not limited to protein-coding genes. Pseudogenes, small RNAs (miRNAs, snoRNA, etc.), and long interspersed non-coding RNAs (LINCs) have all been implicated in oncogenesis, either through mutation or aberrant expression \cite{48, 49, 50, 51, 52, 53, 54, 55}.

Advanced sequencing technology, namely the development of NGS, has significantly enhanced our ability to interrogate cancer genomes. Whole genome sequencing (WGS) — as its name implies — provides a means to query nearly all base pairs in the genome. Whole exome sequencing (WES), which is a less expensive alternative to genome sequencing, uses capture protocol to pull down targeted DNA so only coding regions are assessed \cite{56}. In both research and clinical settings, each technique has advantages and drawbacks that have been discussed extensively elsewhere \cite{57, 58, 59, 60, 61}. Large projects and consortia such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have provided the community with thousands of genomes and over ten thousand exomes from dozens of cancer types. These data are rich with untapped information and will likely contribute to numerous discoveries for years to come.

An array of mutation types are detectable through NGS, providing additional complexity — and intrigue — to an already complicated cancer genome interpretation process. Single nucleotide variants (SNVs) occur when somatic DNA has a native nucleic acid replaced by another. Although seemingly minor, these mutations can result in catastrophic amino acid changes, protein truncations, distorted small RNA binding sites, and promoter inactivation amongst other disruptions \cite{62, 63, 64}. Indels — typically between one and a few dozen nucleotides — can generate all of the aforementioned alterations in addition to garbling protein translation (i.e. frameshift mutations) \cite{65}. Modifications to DNA architecture — also known as structural variants (SVs) — can cause inverted DNA segments, chromosomal
translocations, or the fusing of two distinct transcriptional units [66, 67]. Copy number alterations (CNAs), which are a subset of structural variation, occur when genomic DNA segments are amplified or deleted. Recurrently amplified genomic regions typically carry oncogenes while deleted regions harbor tumor suppressor genes [68]. Epigenetic alterations can also influence cancer phenotypes. Driver genes can be silenced or expressed as a result of DNA hypermethylation and hypomethylation, respectively [69]. Lastly, chromatin architecture, while still underexplored, is suspected of playing a broad role in cancer regulation [70].

Many factors, both intrinsic and extrinsic, can generate the aforementioned DNA lesions [71]. These mutagenic processes are constitutively active in all cells throughout our lifetime. The human genome encodes several pathways evolutionarily derived to reduce and repair DNA damage [72]. These processes are not flawless, and human cells acquire approximately $10.6 \times 10^{-7}$ new somatic mutations per cell division [73]. In fact, over half of the somatic mutations detected in tumors arise before disease onset [74]. Mismatch repair (MMR) and base excision repair (BER) fix defects resulting from single improperly paired and chemically modified nucleotides, respectively [75, 76, 77]. Sunlight-induced pyrimidine dimers are remedied via nucleotide excision repair (NER) [78, 79, 80]. Exogenous radiological insults such as X- and cosmic rays generate DNA double strand breaks, which are corrected by homologous recombination or the more error-prone nonhomologous end joining (NHEJ) [81, 82, 83]. Failing to properly reunite fractured DNA generates indels and SVs [84].

Mutagens, or mutation-inducing agents, are a major contributor to carcinogenesis in some cancer types. Outside of ionizing radiation, mutagens are often chemicals that interfere with DNA structure and/or react with nucleotides to alter their composition [85]. Intercalating agents, which are compounds that insert between DNA base pairs, are one of the most common classes of mutagens. Paradoxically, due to their ability to attenuate DNA replication, intercalating compounds are also used as cancer therapeutics [86]. Cigarette smoke is a well-
known carcinogen that contains benzo[a]pyrene, amongst other alkylating agents [87, 88]. Benzo[a]pyrene is an intercalating polycyclic aromatic hydrocarbon whose genotoxic effect causes a preponderance of C>A transversions [89, 90].

Cigarette smoke isn’t the only entity that imprints a hallmark lesion onto DNA. Exposure to ultraviolet light causes CC>TT dinucleotide substitutions as a consequence of NER eliminating photo-induced pyrimidine-pyrimidine dimers [71]. Endogenous processes can also leave a unique mutagenic footprint. The APOBEC family of cytosine deaminases are highly active in multiple cancer types, leading to excessive C>T and C>G substitutions [91, 92, 93]. However, it is presumed that many important mutagenic processes have not been characterized. Using the trinucleotide context of SNVs and non-negative matrix factorization, researchers at The Wellcome Trust Sanger Institute pioneered an approach that unbiasedly derives mutation signatures active in cancer genomes [94]. With this approach, they were able to identify dozens of unique mutation signatures, with at least one corresponding to each aforementioned mutagenic processes [95, 96]. This study also identified a signature indicative of defective DNA double-strand break repair. Tumors enriched for this marker often lack functional copies of BRCA1/2, which are crucial to the homologous recombination process [97, 98, 99, 100]. Many mutation signatures have been associated with specific cancer types and even some with defective driver genes [95, 101, 102]. Perhaps most importantly, this approach can discern which mutagenic mechanisms contribute to the development and progression of individual malignancies.

1.3 Monogenic cancer predisposition

Of course, cancer effectors are not limited to somatic variation. Germline (inherited) variation has long been known to play a significant role in disease development. One of the first studies proposing a heritable component to cancer occurred in the 1860’s when researchers noted high breast cancer incidence within a single family [103]. Over time, many cases
of familial aggregation were shown to behave in a Mendelian fashion. Some malignancies demonstrated clear autosomal dominance by occurring in multiple successive generations \[104, 105, 106, 107\]. For over a century clinical evidence continued to accumulate. Finally — approximately three decades ago — molecular geneticists began to affirm these findings by mapping cancer predisposition genes \[108\].

Today, it is estimated that mutations in over 100 genes can cause moderate to high risk (greater than 2 fold relative-risk) to various cancer types \[103\]. In the context of cancer predisposition, a genic “mutation” refers to a risk conferring variant, which is often presumed to damage gene function. Two of the most well-recognized cancer predisposition genes are \textit{BRCA}1/2, which were both mapped in 1994. As implied by their names, mutations in either of these genes increase risk to breast cancer. Evidence for the existence of these genes came through familial aggregation and autosomal dominant patterns of inheritance. In females, \textit{BRCA1} mutation carriers have a striking 65% chance breast cancer will manifest before 70 years of age. Those with mutated copies of \textit{BRCA2} have similar prognoses as 45% of these women will also develop breast cancer by age 70 \[109\]. Incidence is not limited to females; males who carry mutated copies of either gene are also more likely to develop breast cancer, albeit with lower lifetime risks \[110, 111\]. \textit{BRCA1} and \textit{BRCA2} mutations confer increased risk for other cancer types as well; both genes are associated with substantially higher ovarian cancer incidence while males harboring \textit{BRCA2} mutations more frequently develop prostate cancer \[109, 112\].

As mentioned previously, these genes play critical and direct roles in the homologous recombination pathway. Unsurprisingly, other genes involved with DNA double-strand break repair also serve as cancer predisposition genes. \textit{ATM} proteins, which are recruited to double strand breaks and activated by the MRN complex, phosphorylate downstream mediators of the DNA damage response \[113\]. Mutations in this gene have been affiliated with a variety of cancer types such as breast, lung and thyroid \[114\]. Two of \textit{ATM’s} substrates, \textit{CHK2} and
p53, are both products of cancer predisposition genes \[115\]. Stomach, breast, and prostate cancers have all been associated with mutations in *CHEK2* \[116, 117\]. Harmful alleles in *TP53* lead to Li Fraumeni syndrome, a condition that causes individuals to develop early — and frequently multiple — cancers \[103, 118\]. Likewise, other DNA double-strand break repair effectors such as *RAD51C, NBN, PALB2, MRE11A, BRIP1*, and *FANCC* have all been implicated in cancer susceptibility \[103\].

Of course, germline deficiencies in other cellular pathways also contribute to cancer development. Mutations in the MMR genes *MSH2, MSH6*, and *MLH1* cause Lynch syndrome and carriers — especially for *MSH2* and *MLH1* — have substantially higher lifetime risks of colorectal, ovarian, and endometrial cancers \[119, 120, 121\]. Monoallelic mutations in *MUTYH* — a BER gene responsible for correcting 8-oxoG alterations due to oxidative damage — confer a 2.9 fold increased risk for colorectal cancer \[122, 123\]. Non-DNA repair pathways, expressly PI3kinase/mTOR, also contain genes that increase cancer risk \[103\]. The catalog of susceptibility-mediating pathways will certainly grow as we enhance our knowledge of cancer genetic architecture.

Since familial transmission was clinically apparent, most of the original cancer predisposition genes (*RB1, BRCA1, BRCA2*, etc.) have an autosomal dominant mode of inheritance. As a matter of fact, the majority cancer predisposition gene pedigrees display this pattern \[103\]. Although less common, some genes behave in an autosomal recessive fashion. *BLM* (Bloom syndrome), *FANCC* (Fanconi anemia), and *WRN* (Werner’s syndrome) mutations classically require biallelic defects to increase cancer risk \[124\]. Although, some reports suggest *BLM* mutations are incompletely penetrant and heterozygous mutations increase risk to colorectal cancer \[125, 126\]. Additionally, occupying the X chromosome, *TGCT1* is sex-linked and recessive, though evidence suggests it is incompletely penetrant \[127, 128\].

While knowing which genes confer cancer risk is useful for disease screening, understanding the means underlying that increased risk helps elucidate disease etiology. One
mechanistic hypothesis is the “two-hit” model. In 1971, Alfred Knudson published an epidemiological study on retinoblastoma, an eye malignancy that often afflicts young children [129]. He first noted that bilateral tumors were more frequent in children with a family history of the disease. Suspicious that biallelic inactivation of a gene (later identified as $RB1$ [108, 130]) triggers disease onset, he constructed mathematical one-hit and two-hit models to determine which was consistent with tumor bilaterality, age at diagnosis, and family history. His assumption was that familial cases inherited a defective copy of $RB1$, thus only the WT allele needed to be mutated somatically to catalyze disease onset. Contrastingly, non-familial cases require somatic inactivation of both $RB1$ copies, which — on average — would require more time as random mutations accumulate. After observing that the youngest patients had hereditary bilateral disease and the oldest had sporadic unilateral disease, Knudson concluded that these clinical discrepancies could only be explained by the two-hit model.

Since Knudson published his landmark findings, many tumor suppressors have been found to adhere to the two-hit model [131, 132]. Biallelic germline mutations in susceptibility genes such as $BRCA2$, $MSH2$, and $MSH6$ can generate more extreme phenotypes like early childhood tumors [133]. Somatic inactivation of a cancer predisposition gene’s WT allele can occur through small mutations (SNVs and indels) or even DNA methylation. However, most commonly, the second event is a copy number deletion causing loss-of-heterozygosity (LOH) [131, 132, 134, 135]. This is the preferred method of biallelic inactivation for $BRCA1/2$, $ATM$, $RB1$, and $NF1$ amongst others [136, 137, 138, 139, 140, 141, 142].

One of Knudson’s key findings was that hereditary retinoblastomas occur earlier than those that are sporadic. This phenomenon isn’t limited to retinoblastomas and is a property of many other cancer types. In familial breast cancer pedigrees, affected individuals are often diagnosed earlier than patients from the general population [143]. In the United Kingdom, individuals were significantly more likely to be diagnosed with lung cancer before 60 years of age if a first-degree relative also had lung cancer [144]. A recent study found that one
out of six early onset colorectal cancer patients had a pathogenic mutation in a known cancer predisposition gene [145]. This pattern becomes even more evident at the genic level. BRCA1/2 mutation carriers are diagnosed with both breast and ovarian cancer far earlier than their WT counterparts [143, 146]. This holds true for familial and unselected populations [147, 148]. A large study of familial sarcomas, TP53, ATM, ATR, and BRCA2 mutation carriers all had earlier age at diagnosis [149]. A multi-cancer study on a European ancestry population also noted significantly younger diagnosis in ATM carriers [150]. Due to the plethora of these associations, early age at diagnosis can be seen as a surrogate for increased heritability or risk.

Of course, the penetrance, relative risk, and lifetime risk are not the same across all cancer predisposition genes [103]. Even mutations within the same cancer predisposition gene have variable penetrance and effects on clinical characteristics [151]. This makes comprehensive risk estimates for each cancer predisposition gene difficult [152]. As data continues to accumulate, multi-factorial models that aggregate both genetic and environmental data will be necessary to estimate personalized risk.

1.4 Cancer as a genetically complex disease

Despite substantial impact on individuals, moderate and high penetrance mutations are estimated to contribute to only 5-10% of malignancies [153]. This alone cannot account for a pan-cancer heritability estimate of 33% derived from hundreds of thousands of monozygotic and dizygotic twins [154]. Accordingly, cancer is considered a complex disease as its manifestation is the result of numerous genetic and environmental factors, and, like with other complex diseases, it exhibits the missing heritability problem [155, 156]. Genome-wide association studies (GWAS) are routinely performed to search for disease-specific loci that contribute to heritability. The concept behind GWAS is simple: take a large sample individuals with (cases) and without (controls) a phenotype of interest; genotype each individual at
loci known to harbor single nucleotide polymorphisms (SNPs); and then determine if cases are enriched — or in some cases depleted — for a given SNP. When significant enrichment persists after multiple testing correction, it is concluded that the SNP is associated with the phenotype of interest [157, 158]. However, in a technical sense, GWAS implicate genomic regions with phenotypes; the associated SNPs are not necessarily causal [159]. Due to linkage disequilibrium, genomically proximal SNPs often co-occur and serve to “tag” one another [160, 161]. Identification of the bona fide causal SNP requires detailed fine-mapping and functional studies [162, 163].

With monogenic disorders, pathogenic alleles are typically rare, and they confer moderate to high disease risk. GWAS are designed to query SNPs that are at common (greater than 0.05) or low (between 0.01 and 0.05) allele frequencies throughout the populations [164]. To be consistent with canonical population genetics theory, we’d expect common SNPs to only have low disease risk contributions. This expectation is the driving force behind the common disease-common variant hypothesis [165, 166, 167, 168]. Indeed, most significant GWAS findings — or hits — have odds ratios less than 1.5 [164]. Notably, due to increasing sample sizes and sophisticated genotype imputation methods, some modern GWAS have the power to test SNPs with allele frequencies < 0.01. Details of these advancements and their implementation are discussed exhaustively elsewhere [169, 170, 171].

GWAS have been utilized extensively for multiple cancer types. As of July 2017, the GWAS Catalog reports 3,708 significant SNP associations for 305 cancer-related traits [172]. The phenotype of interest for many of these studies is not simply the presence of a particular cancer type. Comparing nearly 10,000 BRCA1 mutation carriers with and without breast or ovarian cancer, Couch and colleagues identified loci that modify BRCA1-mediated risk [173]. GWAS have also identified SNPs associated with clinical outcomes in colorectal cancer [174]. As the number of genotyped individuals grows and the quality of clinical phenotypes increases, more genetic associations are sure to follow [164].
The scope of this methodology extends beyond associations; GWAS have no doubt increased our knowledge of fundamental cancer biology. For example, SNPs within the 8q24 locus have been associated with increased risk to breast, colorectal, ovarian, and prostate cancers [175]. Described as a “gene desert,” this region has garnered great interest due to its proximity to the MYC oncogene [176]. Functional studies have shown that one of the identified colorectal cancer risk alleles, rs6983267(G), is disproportionately occupied by the beta-catenin-TCF4 transcription factor complex. Furthermore, when bound, this region forms a 335 kb chromatin loop that directly interacts with the MYC promoter [177]. Additional investigations have determined that looping is tissue-specific, thus elucidating why some cancer types have 8q24 SNP associations not shared by others [178]. Importantly, the majority of cancer GWAS hits do not map to known moderate or high risk loci [103]. Functionally exploring candidates furnishes an opportunity to unveil novel, cancer-relevant genes.

1.5 Undercutting cancer’s resilience

Whether hereditary or sporadic — at its core — cancer is the morbid consequence of the potent evolutionary process. Even within an organism, cells survive through adaptation. Defective DNA repair processes entwined with rampant growth often ensure that tumor cell populations have substantial genetic diversity [179]. Cells with marginally dissimilar genetics are constantly competing with one another where “winners” are those whose progeny thrive within the environment. Each cell within the population has a constant, almost anthropomorphic drive to acquire any feature that promotes viability. In their fervor, this process also produces an increasingly pathogenic cellular community. This persistence is a key contributor to the arduous — and unfortunately sometimes futile — task of eradicating an individual’s cancer [180].

Cancer’s scrappiness makes it evasive to even the most sophisticated, modern treat-
ments. In ovarian cancer, germline BRCA1/2 mutations can be exploited therapeutically by chemically crosslinking DNA using platinum-based drugs, which — without homologous recombination repair — often leads to cell cycle arrest \[181, 182, 183\]. However, in response to treatment, cells whose defective germline allele was somatically reverted to WT can expand \[184\]. Now, presumably better able to repair interstrand DNA crosslinks, malignant cells display chemoresistance \[185, 186, 184, 187\]. Similarly in chronic lymphoblastic leukemia, chemotherapeutic treatment can eliminate the majority of cancerous cells; nonetheless, months to years later some malignancies recur with a previously subclonal driver mutation now ubiquitous \[188, 189\].

In light of the aforementioned findings, it is important to recognize that therapeutic advances did not lead to decades of survival improvements alone. Early detection also played a pivotal role in promoting favorable clinical outcomes. For essentially all cancer types, diagnosing malignancies when they are local and metastasis-free leads to dramatically better patient outcomes \[3\]. Developing enhanced and effective screening techniques could help prevent as many if not more cancer-related deaths than novel therapeutics.

### 1.6 Genetic identification of high risk individuals

One of the tenets of precision medicine is to deliver patients personalized disease risk using a combination of genetic and environmental factors \[190, 191, 192\]. Although the former is a fixed factor and the latter mutable, together they provide multiple avenues to mitigate disease manifestation and morbidity. First, individuals with higher cancer risk may be more vigilant with mammograms, colonoscopies, and other screening techniques \[190, 193\]. Second, for patients with unequivocal, highly penetrant breast cancer mutations, prophylactic mastectomies are a risk-reduction option \[194, 195\]. Third, individuals genetically predisposed to cancer may feel empowered to reduce or eliminate harmful behavioral traits like smoking and sedentarism. This latter point cannot be overstated since an estimated 20-40%
of cancer cases are preventable with environmental modifications [196]. Furthermore, reducing environmental risk is especially important for individuals with strong genetic cancer predisposition. Absolute risk in already susceptible individuals can be exacerbated by even small relative risk increases from other factors [197].

For years patients with familial aggregation of breast cancer have undergone targeted \textit{BRCA1/2} sequencing to identify possible mutations [198, 199]. If families were negative for mutations in either of these genes, \textit{ATM}, \textit{STK11}, or other possible candidates could be queried [200, 201]. As a result of NGS, screening for risk-amplifying mutations has become substantially more comprehensive. Numerous panels have been developed to simultaneously evaluate dozens of known cancer predisposition genes [202, 203, 204, 205]. Given observed and assumed pleiotropy, these panels are often applicable across cancer types. Some cancer centers are even utilizing these panels on seemingly sporadic disease, which has revealed that a surprising number of patients carry putatively pathogenic alleles [206]. As such, these panels are becoming relatively standard practice, as exemplified by a recent Mayo Clinic study that applied a 21 gene panel to over 65,000 breast cancer patients [207]. Of course, risk prediction is not limited to moderate and high penetrance variation. When considered in concert, cancer type-specific GWAS hits are also beneficial to risk stratification. Polygenic risk scores have been derived for breast, prostate, and colorectal cancers amongst others [208, 209, 210, 211, 212]. Leedham and Tomlinson noted a monotonic rise in colorectal cancer risk as the number of GWAS alleles an individual carried increased [213]. Overall, current approaches tend to look at low frequency, moderate/high penetrance alleles in isolation, while highly frequent, low penetrant alleles are assessed as a collection. Few studies have amalgamated these methodologies to explore low frequency, moderate/high penetrance alleles in aggregate.
1.7 Clinical and genomic features of breast cancer

An estimated 252,710 women in the US will be diagnosed with breast cancer throughout 2017, and, in the same time period, the deaths of 40,610 will be attributable to this disease [3]. Unequivocal evidence has shown that breast cancer is not a single disease. Histological classifications — each with distinctive features — indicate the cell type from which the malignancy originated [214]. In standard pathological practice, the expression of two critical hormone signaling regulators — estrogen receptor and progesterone receptor — are assessed immunohistochemically. Along with genomic amplification of HER2, typically determined via fluorescence in situ hybridization or FISH, pathologists assign each breast carcinoma an immunohistochemical (IHC) subtype [215]. Commonly, these are divided into three IHC classes: 1) HR+, which express ER and/or PR, but lack $HER2$ amplification; 2) HER2+, which have HER2 amplification but lack expression of either hormone receptor; and 3) HR-/HER2-, or triple negative, which neither express ER and PR nor possess amplified HER2. Throughout the entire US population, HR+, HER2+, and triple negative IHC subtypes constitute 70, 20, and 10 percent of breast cancer cases, respectively [216].

IHC subtype classifications are critical since they, along with tumor stage, can dictate clinical prognosis and therapeutic options [217]. HR+ patients have the highest rates of disease-free survival, and — due to ER expression — can be treated with Tamoxifen, a drug whose metabolites act as efficous ER inhibitors [218, 219]. The 5- and 10-year survival rates for individuals bearing HER2+ disease are relatively similar to HR+, though this is at least partially attributable to targeted therapies [216]. Drugs, most notably Herceptin, have been engineered to specifically combat HER2+ tumors by inhibiting HER2’s protein product, ERBB2. Receiving FDA-approval in 1998, this was one of the first genomically-targeted cancer therapies. Clinical outcomes — namely likelihood of recurrence and death — are most dire for the triple negative subtype [220]. This subtype is also at the highest risk for deadly brain and lung metastases [221, 222]. These poor outcomes have stimulated
widespread effort to develop therapies that address triple negative-specific molecular features [223, 224, 225]. Concerning heritable forms of the disease, germline BRCA1 mutations typically lead to triple negative tumors, while a majority of BRCA2 carriers are HR+ [226]. Due to conflicting literature, it is not known whether sporadic and BRCA1/2-mediated cancers have divergent survivability [227].

Many studies have sought to understand the mutational and transcriptional characteristics of breast cancer. Microarray analyses have identified expression signatures associated with prognosis, and these signatures have been clinically implemented to determine the necessity of chemotherapeutic treatment for localized, early-stage tumors [228]. Transcriptional patterns have also been used to identify “intrinsic” tumor subtypes that provide further granularity to IHC classification. This approach, termed PAM50 subtyping, has experienced greater adoption in the laboratory than in the clinic [229, 230]. DNA-based arrays have also been utilized to investigate the landscape of CNAs. CCND1 and MYC amplifications are relatively common across all subtypes, while EGFR amplifications are far more prevalent in triple negative tumors [231]. More broadly, multiple studies have concluded that triple negative tumor genomes — especially those with BRCA1 mutations — exhibit substantial structural changes or genomic complexity, which is consistent with defective DNA double-strand break repair [232, 233].

NGS approaches have providing invaluable insight into the somatic mutations and mutational processes that drive breast cancer. In a landmark paper using multi-omics data — including WES from tumor and normal tissue — the TCGA Network identified three genes (PIK3CA, GATA3, and TP53) that were mutated in over 10% of samples. TP53 mutations were preferential to HR- tumors, while the remaining three are observed predominantly in HR+ tumors [234, 235]. PIK3CA mutations often co-occur with those in CDH1 and both typically do not coincide with GATA3 alterations. Further studies tracked this mutual exclusivity to histology, with CDH1 and GATA3 mutations found almost ex-
clusively in lobular and ductal carcinomas, respectively [236]. A small fraction of luminal
tumors, which are typically HR+, had aberrant copies of ESR1 — a gene associated with
therapeutic efficacy [234, 235]. Later, two independent studies conducted by the Broad and
Wellcome Trust Sanger Institutes have identified 40 unique genes significantly mutated in
breast cancer [237, 238].

Expanding beyond individual, genic lesions, mutation signature analyses have provided
a glimpse of the mutagenic processes shaping breast cancer genomes. Thirteen distinct mu-
tational processes — some with mechanistic explanations — riddle breast cell DNA with
errors [95, 239]. Some of these signatures are characteristic of other cancer types and afflict
few breast tumors. APOBEC and homologous recombination deficiency (HRD) signatures
along with signature 8 — which has unknown etiology — compose much of the mutational
landscape. The aging signature is also influential, though it is omnipresent in cancer [95].
HRD has been proposed as a biomarker that is indicative of efficacious PARP inhibition
therapy [240]. It is possible that other signatures have useful medical applications; how-
ever, the value of mutation signatures transcends possible clinical associations. Determining
signature timing and interactions with driver genes could unmask hidden features of early
breast cancer development. Although a few studies have begun this processes, this space
remains comparatively uncharted [241, 102, 242].

1.8 Racial/ethnic disparities in breast cancer

Breast cancer incidence and outcomes in the US are not the same throughout racial/ethnic
cohorts. Women with African ancestry (Black) experience much high rates of mortality than
those of European ancestry (White). Historically, White women are more likely to receive a
breast cancer diagnosis [3]. This has led to the adage, “White women are more likely to get
breast cancer, but Black women are more likely to die from it.” Black women are also far
more likely to be diagnosed with advanced disease (i.e. with regional or distant metastases)
In fact, they are 40-70% more likely than Whites to present with stage IV breast cancer, regardless of the underlying IHC subtype. Even when adjusting for age, young Black women (< 40 years of age) have a mortality rate double that of young White women. More adverse breast cancer mortality rates persist in older Black women as well. There are also substantial and consequential differences in tumor biology between Black and White cohorts. Namely, triple negativity is far more common in Black individuals. Not only are these Women beset with aggressive subtypes, but also those with prognostically favorable HR+ tumors still encounter lower survivability.

Socioeconomic factors such as income, diet, obesity, lack of exercise, and access to healthcare are prime candidates to explain these disparities. Weight gain has been shown to explain approximately 16% of postmenopausal cancer risk. Conversely, disease frequency drops in individuals who partake in rigorous daily exercise. Studies have suggested that access to breast cancer screening techniques can reduce mortality rates. Although evidence indicates mammography usage can normalize racial/ethnic differences in tumor stage, Black women still present with higher grade tumors. Black women continued to have more aggressive histological features even after adjusting for numerous clinical and socioeconomic factors. So while they undoubtedly contribute, non-biological variables alone cannot sufficiently account for the magnitude of disparities amongst races/ethnicities.
CHAPTER 2
ROBUST SCALING OF DNA SEQUENCING ANALYSES
USING THE MODULAR SWIFTSEQ WORKFLOW

2.1 Introduction

Advancing next-generation sequencing (NGS) techniques and decreasing costs have stimulated the production of unprecedented amounts of data [260]. This technological revolution has propelled genomics into a data-intensive science, and subsequently unveiled new challenges to the field [261]. There are multiple barriers ubiquitous to data-intensive applications [262, 263]. Data size — which for consortium projects such as The Cancer Genome Atlas (TCGA) can exceed petabytes — is unwieldy and requires substantial disk space for storage as well as scratch for active computations. Transferring data between remote sites demands significant bandwidth. Large computational resources, particularly those optimized for heavy input and output (I/O) operations, are crucial to shaping raw data into an informative resource. While necessary, substantial hardware infrastructure is not sufficient. Robust and scalable software is needed to efficiently automate processing and analyses. Some have proposed that inadequate software fosters a greater barrier to scientific discovery than limited hardware [264]. Consequently, researchers are unable to appropriately leverage their own data, let alone the vast datasets available through public repositories.

Typical pipeline approaches (i.e. those implementing map-reduce nave, serial execution) do not provide a sufficient solution as they are inefficient and lack the structure for effective horizontal scaling. In the past few years, there has been a surge of software, both commercial and academic, attempting to better facilitate NGS analyses [265, 266, 267, 268, 269, 270, 271, 272, 273]. However, most — if not all — fail to satisfy all of the following traits: 1) fault tolerance, 2) speed, 3) scalability, 4) efficiency, 5) analytic diversity (e.g. tumor-normal pairs), 6) workflow flexibility, 7) portability, and 8) open source licensing.
To address this deficiency in genomics software, Lorenzo Pesce and I have developed the SwiftSeq workflow. Powered by the Swift scripting language [274], SwiftSeq delivers a sleek and transparent workflow that emphasizes bioinformatics rather than the complexities of parallel programming. This framework offers significant benefits to both small- and large-scale DNA sequencing analyses. Alignment and genotyping of targeted gene panels, whole exome sequencing (WES), and whole genome sequencing (WGS) are all natively supported. While best practice workflows are provided for both germline and tumor-normal pair analyses, SwiftSeq maintains flexibility by allowing users to specify algorithms and parameters through a graphical user interface (GUI). Through effective parallelism, samples can be processed in a fraction of the time required by standard pipeline approaches. Task execution is fully automated, with synchronization and failure detection managed under-the-hood. Importantly, executing SwiftSeq over a single sample is just as easy as thousands.

2.2 Results

2.2.1 Anatomy of a SwiftSeq run — input gathering

Each SwiftSeq run requires three sets of information: the files to be processed; locations of reference files and bioinformatics tools; and the workflow to be executed (Fig. 2.1a).

Large DNA sequencing datasets are often distributed as binary alignment map (bam) files because of their ability preserve metadata and differentiate read groups (i.e. collections of reads that share covariates such as sample, library, sequencer, flow cell lane, etc.) [275, 276, 277, 278]. As such, SwiftSeq requires per sample bam files with properly formatted read groups as input. Consequently, input bams need to be split by read group with Samtools [279] and converted to fastq with BamUtil [280] prior to alignment. Runtime increases resulting from these additional steps are partially mitigated through piping and streaming techniques (discussed below).
Figure 2.1: **Diagram of a SwiftSeq run.** (a) Input files required for a SwiftSeq run. (b) Initiation step where the Swift script, Swift configuration files, and bash wrappers are generated based on input provided in panel a. (c) The execution step where tasks are distributed to workers that run in parallel. The number of tasks packed onto each worker node depends on hardware specifications and software requirements.
Genomics applications often require a standardized genome to use as a reference. Multiple consortia maintain and distribute their own reference genome, each containing subtle — but non-negligible — discrepancies, and reference genome preference can vary from user to user.

Similarly, users often desire specific versions of bioinformatics algorithms for functionality as well as consistency. Some algorithms (e.g. Genome Analysis Toolkit [GATK]) can not be independently distributed due to the terms of use. These nuances make it difficult to bundle SwiftSeq with a comprehensive set of references and software to satisfy all users’ needs. Lorenzo Pesce and Dominic Fitzgerald have constructed a script that will automatically install essential applications; however, users need to ensure certain algorithms, such as GATK, are available on their systems. To make SwiftSeq aware of the location of these algorithms, runs require two configurations files: one that contains paths to reference files; and another provides paths to each bioinformatics tool.

One of SwiftSeq’s core tenets is to provide a flexible framework suitable for customized analyses. Instead of using a monolithic pipeline, analyses are represented as workflows, which are contained in JSON format. Each workflow file contains information regarding analysis type, bioinformatics tools to be used, and desired parameters. During the initiation step, SwiftSeq will parse the JSON file and construct Swift syntax reflective of the desired workflow (discussed below).

### 2.2.2 Anatomy of a SwiftSeq run — initiation

Once the aforementioned inputs are received, SwiftSeq can be partitioned into two processes: run initiation (Fig. 2.1b), which is controlled by a Python back-end; and run execution, underpinned by the parallel scripting language Swift. The initiation step generates files necessary to execute a SwiftSeq run. These include a Swift configuration file, the Swift script, and Bash wrappers for bioinformatics applications.

Broadly, workflows can be described as a set of inputs and outputs stitched together by
computation applications \[282\] \[283\] \[284\]. The Swift script maps data files to applications, where each application is designed to perform a specific bioinformatics task. The properties of these applications (required input, output, and computational resources) are defined in Swift configuration file. Managing each of these applications is a Bash wrapper that executes the desired algorithms (i.e. BWA alignment \[285\]) over input files. Each Bash wrapper has a template that is populated by the workflow and configuration information provided by user. As an additional benefit, these wrappers provide a transparent record of the exact procedures executed over all files throughout the workflow. No file operations — including basic formatting and error checking — are hidden from users.

### 2.2.3 Anatomy of a SwiftSeq run — execution

SwiftSeq has been optimized to run in distributed and parallel environments. Swift implements a master/slave style of computing \[286\] \[287\] \[288\]. SwiftSeq will be executed on the head node, and this process will coordinate and control all tasks (invocation of applications) on worker nodes (Fig. 2.1c). Clusters and high-performance computing machines are often a shared commodity and utilize resource managers to fairly allocate worker nodes \[289\] \[290\]. Using Coasters \[291\], Swift processes control interactions with numerous resource managers. Any parameters required by resource managers (walltime, accounts, etc.) are handled within the Swift configuration file. In 30 second intervals throughout a run, terminal output indicates how many tasks are running, have completed, and have failed — if any.

To leverage distributed/parallel systems to robustly run workflows, the underlying software needs to properly handle workflow patterns \[292\] \[293\]. While there are numerous complex workflow patterns \[294\], NGS tasks typically encounter forks and joins \[295\]. Forked tasks are executed in parallel and are joined once all forked tasks are complete. This creates a “handshake” scenario where task C cannot be executed until tasks A and B are synchronized. Even serial tasks need to be properly managed so task B does not execute prior to
the completion of task A. The deterministic, file-based nature of Swift ensues that workflow patterns are properly handled under-the-hood [274]. Tasks are only executed after all required input files are successfully generated.

Through the combination of Swift’s native framework and customized error checking within Bash wrappers, tasks are robust to transient software and localized hardware failures. SwiftSeq will attempt to re-run any failed task \( x \) times, where \( x \) is defined as a parameter during initiation process. Importantly, this keeps user intervention to a minimum because most of the failures in large scale bioinformatics applications are caused by transient issues in compute node or I/O. Since the underlying bioinformatics applications are written to be modular, the workflow can be halted and restarted. A restart log file records which tasks have completed successfully, while a restart configuration file maintains all crucial information about run initiation and execution. Complex systems are subject to maintenance, power disruptions, and node failures [296, 297], all of which can prematurely terminate workflow execution. “Bad” behavior from other users can also disrupt runs on a shared head node. Therefore, being able to restart a SwiftSeq run with a single, simple command is an advantageous — if not crucial — feature for large-scale genomic analyses.

2.2.4 Bioinformatic nuances handles under-the-hood

Bioinformatics algorithms and NGS file formats contain nuances that are tedious to handle without automation. One of which — read group management — is handled by SwiftSeq under-the-hood at alignment and bam merging steps. Furthermore, single- and paired-end read groups are automatically detected and aligned accordingly. Sanity checks are included at the time of workflow execution to ensure file integrity. Users are warned if input appears truncated or any file lacks properly formatted read groups. These check are necessarily since these issues can easily go unnoticed undermining the integrity of downstream output. Simply basing failure detection on the return status of the various applications cannot produce
reliable results.

2.2.5 Maximizing performance — parallelization strategies

Besides using native multi-threading options included with some bioinformatics algorithms, SwiftSeq employs five levels of fork-join parallelism (sample, read group, contig, genomic coordinate, and algorithm) to maximize compute resources and decrease analysis time (Fig. 2.2). 1) Each sample (bam file) can progress through SwiftSeq independently, barring tumor-normal pairs (discussed below). 2) During alignment, read groups are partitioned and mapped to the genome independently. Once all read groups are aligned, they are merged. 3) The subsequent merged bam is then split into contigs, which are processed (e.g. duplicate marking) concurrently. Processed contigs are then rejoined into a final bam file. 4) Within each contig, variant calling (SNVs and indels) is distributed across genomic coordinates whenever possible. Since SVs can span multiple megabases [298, 299], those variant calling processes are not subdivided by genomic coordinates. 5) Lastly, when multiple variant calling methods are being used they are executed autonomously.

To gauge the effect of these parallelization strategies on overall walltime, I performed alignment and germline variant calling over a standard depth exome (92.6 million reads) and genome (1.11 billion reads). Using state-of-the-art tools (Fig. 2.3a), I compared the SwiftSeq approach to that of a typical serial pipeline using Beagle, a 17,472 core Cray XE6 supercomputer at the University of Chicago. Each Beagle worker node contained two AMD Opteron 6100 series processors (for a total of 24 cores) and 32 GB of RAM (NUMA, as in Non-Uniform Memory Access). The GATK development team recommends two computation-heavy bam cleaning steps — realignment around indels and base quality score recalibration — both of which are properly parallelized in SwiftSeq. Performance was assessed with and without these cleaning procedures. In all four cases, SwiftSeq was substantially faster than pipeline approaches (Fig. 2.3b-c). The genome was aligned, genotyped, and annotated in approx-
Figure 2.2: **SwiftSeq parallelization strategies.** Five levels of fork-join parallelism (samples, read groups, contigs, genomic coordinates, and algorithms) that are implemented throughout the workflow to maximize resources and decrease runtime.
approximately 11 hours, which is even faster than other highly-optimized — and monolithic — parallel approaches [266]. Similarly, the standard depth exome was completed in approximately 36 minutes.

2.2.6 Maximizing performance — efficiency and scalability

Many genomic applications do not have native multi-threading capabilities, and those that do often fail to scale linearly [300]. Inefficient compute node usage wastes valuable resources, which precipitates increased runtimes and monetary expenses. Reconfiguring bioinformatics algorithms to optimize performance was beyond the scope of this project; however, these algorithms were profiled to gauge RAM usage and multi-threading performance. This information helped determine optimal ways to pack multiple tasks on a single compute node (Fig. 2.1c). Accordingly, SwiftSeq aggregates tasks based on algorithm requirements and worker specifications (e.g. the amount of RAM and number of cores). For example, since BWA exhibits linear scaling [300], each alignment task is assigned a personal compute node. Tasks that require more than 8 GB of RAM are assigned to “high memory” pool; so a machine with 32 GB RAM can run four of these tasks concurrently. Of course, optimal packing schemes depend heavily on algorithm, input data, and system architecture. Users can make further, individualized refinements by simply editing the Swift configuration file.

Theoretically — as described above — deconstructing large, sequential tasks into smaller parallel processes should lead to decreased walltime. However, substantial slowdowns were noted with some contig and genomic coordinate levels tasks. SnpEff [301], like many other variant annotation programs, requires a database of sequences and predicted functional effects to be read into RAM at execution time. Executing this process per contig led to increased walltime and wasted resources, presumably caused by I/O bottlenecks (Fig. 2.4a-b). Consequently, contig level annotation was eliminated, and the concurrency of other I/O heavy tasks (e.g. bam file indexing) was throttled. These adjustments led to slightly
Figure 2.3: **SwiftSeq processing speed compared to standard pipeline approaches.**
(a) The software workflow used to compare the two approaches with and without Genome Analysis Toolkit (GATK) bam cleaning steps. Walltime comparisons between SwiftSeq and pipeline approaches using an (b) exome and (c) genome with 92.6 million and 1.11 billion reads, respectively. These comparisons were made on the University of Chicagos **Beagle** supercomputer (Cray XE6). Each worker node contained two AMD Opteron 6100 series processors (24 cores total) and 32 GB of RAM.
increased walltime for a single exome (data not shown); however, scalability improved substantially and resource usage was decreased by as much as 42% (Fig. 2.5a-b).

Named pipes (FIFOs) have also been implemented throughout the workflow, most notably during bam to fastq conversion. Relegating fastq files to FIFO objects allows SwiftSeq to avoid writing unnecessary temporary files, which reduces filesystem I/O. Standard Linux pipes were used whenever possible, so long as it didn’t significantly disrupt workflow modularity. For example, when read group bams are merged and sorted, that output is piped directly into Bamtools [302] and split into contig bams. Together, these modifications have notably reduced I/O constraints, limited disk footprint, and boosted performance.

Overall, the SwiftSeq framework has proven to be scalable. Using *Beagle*, SwiftSeq has consistently utilized 200-300 worker nodes (4,800 – 7,200 cores) concurrently without performance decreases. As many as 671 nodes have been used to effectively manage thousands of simultaneous tasks; although, at this scale I/O limitations that plague data-intensive analyses on parallel filesystems were encountered [303, 304, 305, 306]. Nonetheless, from a practical perspective, SwiftSeq has allowed a single individual to uniformly process nearly 12,000 WES and over 500 WGS samples. Analyses of this magnitude have typically been limited to multi-institutional consortia [307, 308, 309].

### 2.2.7 Facilitating tumor-normal pair analyses

Detecting somatic variants (SNVs, indels, and SVs) within malignant tissue — whether using gene panels, WES, or WGS — is a common procedure in genomics studies [310]. These analyses add an additional layer of computational complexity as bam files representing both malignant and normal tissue (i.e. tumor-normal pairs) must be coordinated at variant calling steps. Through simple directory structure requirements (Fig. 2.6), SwiftSeq is able to identify tumor-normal pairs and efficiently complete somatic variant calling. In cases where multiple tumor and/or normal samples from a single individual are provided, SwiftSeq
Figure 2.4: Naive exome scaling tests with SwiftSeq Using the workflow depicted in Fig. 2.3a without Genome Analysis Toolkit (GATK) bam cleaning steps, we tested the scalability of SwiftSeq by performing runs with 1, 2, 4, and 8 copies of the same exome sample. (a) The number of nodes utilized at any given time during a run. Overall walltime can be inferred from the X-axis. (b) The number of node hours required to complete each run, which was derived from the area under the curves in panel a. These comparisons were made on the University of Chicago's Beagle supercomputer (Cray XE6). Each worker node contained two AMD Opteron 6100 series processors (24 cores total) and 32 GB of RAM.
Figure 2.5: **Comparing optimized and naive exome scaling tests.** Using the workflow depicted in Fig. 2.3a without Genome Analysis Toolkit (GATK) bam cleaning steps, we compared the performance of a naive approach (blue) to that of an optimized approach (orange) after throttling I/O intensive processes. (a) The number of nodes utilized by naive and optimized approaches at any given time during the 8 exome run. Overall walltime can be inferred from the X-axis. (b) Scaling comparison between naive and optimized approaches based on the number of node hours required to process 1, 2, 4, and 8 sample exome runs. These comparisons were made on the University of Chicago’s Beagle supercomputer (Cray XE6). Each worker node contained two AMD Opteron 6100 series processors (24 cores total) and 32 GB of RAM.
Figure 2.6: **Tumor-normal pair directory structure.** Depicted above is the required directory structure for tumor-normal pairs analyses performed by SwiftSeq. Each sample directory must only contain a single bam file. For a given individual, each sample within the tumor directory will be compared to each sample within the normal directory. Consequently, variant calls will be generated for all possible tumor-normal pairs.

will perform all pairwise analyses while avoiding superfluous computation and file generation. This feature accommodates common study designs such as regional tumor sequencing, primary versus metastasis comparisons, and “double normal” (i.e. blood and tumor-adjacent tissue sequenced) analyses. Furthermore, multiple studies have shown that somatic variant calling accuracy can be improved by integrating results from multiple algorithms. As such, SwiftSeq allows users to select numerous callers that will be implemented within the same run.

### 2.2.8 Flexible analyses through a graphical user interface

While SwiftSeq’s modularity provides multiple benefits, one of its primary purposes is to permit workflow flexibility. Bioinformatic algorithms are rapidly changing, which creates a tacit
expiration date for any static workflow. The optimal software for an analysis depends on both
the user’s question and preferences. Enabling users to select their own algorithms and param-
eters offers significant utility. To deliver this functionality, Dominic Fitzgerald and I devel-
oped a graphical user interface accessible (GUI) via the web (https://swiftseq.uchicago.edu)
as a front end to SwiftSeq. The GUI was built using HTML5/CSS3 with a couple JQuery
libraries (JQuery 1.11.3 and JQuery UI 1.11.4) on top of a Django 1.9.1 backend. Peripheral
elements of the website include a front-facing landing page, documentation, and a link to
the SwiftSeq Git repository.

Through the GUI, users are able to specify the run type (germline or tumor-normal pair
processing), which aligner and genotype(s) to use, and the parameters to be passed these
algorithms. Similarly, users can elect to include bam processing steps such as duplicate
removal, realignment around indels, and base quality score recalibration. These workflow
specifications will be written as a downloadable JSON file, which is passed to SwiftSeq at
the time of execution (Fig. 2.7). This approach also promotes reproducibility as the JSON
file can be shared amongst users and systems. For neutral or less experienced users, we
provide a page that serves pre-defined, “best practice” workflows. By and large, allowing
users straightforward and transparent ways to be in charge of numerous analysis details
distinguishes SwiftSeq from other NGS workflows [269, 267, 266, 265, 268].

2.2.9 Portability across systems

Swift, and consequently SwiftSeq, is able to run on a variety of architectures and re-
source managers. SwiftSeq has been successfully deployed on commodity clusters and high-
performance supercomputers, which used Torque [328] and PBS [329] scheduling systems,
respectively. Other popular resource managers such as Slurm [330] and Sun Grid Engine
[331] are natively supported. This framework has also been implemented on personal com-
puters (e.g. multi-core laptops and desktops) and commodity hardware, ranging from low-
end to enterprise-grade computational servers.

In order to maintain scalability, many large genomic projects are moving to cloud environments, both private [278] or commercial [332] [333]. In contrast to typical, shared filesystems, these clouds utilize object storage [334]. With this data management structure, files must be pushed/pulled to virtual machines (workers) before they are subjected to computation. Because of Swift’s file-based design, interacting with object storage is not straightforward. Recently developed software abstracts object-storage and provides users with a shared filesystem interface. This approach provides Swift with a mechanism to see and manage files within object storage. Using Amazon Web Services and their Elastic File System, minimal tests with medium-sized data indicate this is a viable run strategy for SwiftSeq. However, porting applications to the cloud can be notoriously nuanced [335], and slow interconnect could hinder data-intensive genomic tasks [336]. Full scaling tests and subsequent optimizations still need to be performed.

2.3 Discussion

SwiftSeq successfully delivers a 1) fault tolerant, 2) fast, 3) scalable, 4) efficient, 5) analytically diverse, 6) flexible, 8) portable, and 8) open source workflow. Unlike other workflows
that simply claim scalability, I have demonstrated this by completing multiple large-scale projects. With *Beagle* at the University of Chicago, I have harmonized over 10,000 exomes and hundreds of whole genomes, a truly herculean task that would have be impractical — if not impossible — without SwiftSeq. This framework should help empower bioinformaticians and biologists to tackle other seemingly insurmountable NGS projects.

Excessive fork-join usage can hinder performance if system latency exceeds compute time reductions [337]. This phenomenon that has been noted particularly in exascale tests [338]. However, latency-bound tasks typically refer lower-level algorithmic procedures rather than high-level workflow management. Since many genomic tasks, even when parallelized, take minutes to hours to complete, SwiftSeq-mediated latency issues were not noted. Nonetheless, substantial slowdowns were observed when trying to parallelize tasks that repeatedly queried a single file. Preliminary profiling suggests that some algorithms, such as Platypus, may read data inefficiently (data not shown). These issues can go unnoticed by developers since they only manifest with hundreds of concurrent invocations on the same system. So while workflow managers can be optimized to run tasks as efficiently as possible, overcoming some limitations requires altering the underlying bioinformatics algorithms.

One powerful aspect of Swift not yet exploited by SwiftSeq is the “sites” concept. Swift has the capability to run on one system while farming tasks to others. Theoretically, this would allow us to initiate a run on *Beagle* and burst to other resources such as Amazon Web Services or Google Compute Engine. In the case where local resources are occupied, having SwiftSeq distribute tasks to remote sites would be useful, especially if analysis turn-around time is crucial. The benefits of this approach will, of course, be throttled by network speed since data needs to be transferred from the execution site to the remote site. Nonetheless, this feature would help formulate more complex runs that can maximize resource usage.

Despite its merits, like any software stack, SwiftSeq can be improved to make it even more user friendly and desirable. Currently, this framework has only been tested using human and
mouse sequencing data. As long as the organism of interest has a reliable reference genome, SwiftSeq should be capable of performing a full analysis. However, users will need to exert caution when selecting genotyping algorithms and parameters. Platypus, for example, is designed to perform variant calling on diploid genomes [339]. Applying this algorithm to non-diploid genomes may generate unreliable results. Similarly, the genomics community is transitioning to a new reference genome model, HG38. The National Cancer Institute’s Genomic Data Commons has started reprocessing all data from The Cancer Genome Atlas to conform to HG38. This genome build needs fully tested across diverse SwiftSeq workflows to ensure functionality and flexibility remain. Failing to do so would prohibit community adoption and guarantee a short shelf-life.

From a software standpoint, SwiftSeq is a polyglot program where one programming language (e.g. Python) is constructing programs in the syntax of another (e.g. Swift and Bash). While this style isn’t necessarily uncommon [340], it makes the code base harder to maintain. One obvious reason is because it requires developers to have programmatic competency in multiple languages. Recently, the Swift Development Team has created the Python module Parsl, which delivers the power of Swift with the simplicity of Python syntax. A Parsl version of SwiftSeq is currently being tested. This homogenized “Pythonic” framework will not only make SwiftSeq more intuitive, but also will help attract a community of users and developers. In the same vein, the GUI is currently an entirely separate entity with no direct programmatic connection to SwiftSeq. Ideally, the GUI would be responsible for workflow execution and monitoring in addition to design. Due to a variety of factors, GUIs interacting with underlying compute resources are not trivial to generalize across systems. A sensible solution would be to develop a GUI strictly for Amazon Web Services since it already provides a scalable and easily accessible compute environment.
CHAPTER 3
AGGREGATE ALLELIC BURDEN FOR CANCER RISK
GENES ASSOCIATES WITH AGE AT DIAGNOSIS

3.1 Introduction

Cancer is a complex disease with many known environmental and genetic risk factors. 5-10% of cancers cases can be attributed to highly-penetrant, inherited alleles [341], which often lead to earlier age at diagnosis [342, 343]. Large-scale twins studies have estimated pan-cancer heritability at 33% [344], indicating that monogenic approaches cannot explain familial cancer aggregation and phenotypes. The polygenic nature of cancer concerning low to moderate risk loci, and particularly their relationship to age at diagnosis, is less understood. Modeling of common variation has helped estimate overall cancer risk for particular tumor types [345, 346, 347, 348, 349]. However, risk estimators aggregated across many different cancer types, especially those considering lower frequency variation, are not well-established.

Previous sequencing studies have shown that individuals carry on average approximately one hundred rare, loss-of-function, protein-coding alleles [350]. Yet, aggregating a sufficient number of sequenced cases and controls to perform an adequately powered rare variant association study has remained a challenge [351]. This challenge provides incentive for alternative study designs to identify clinically-relevant, genetic contributions to cancer susceptibility. Across a large cohort of heterogeneous malignancies — using age at diagnosis as a surrogate for risk — I associate earlier diagnosis with increased harmful allele burden in cancer predisposition genes. These results promote an avenue to explore, interpret, and potentially manage variants from cancer gene sequencing.
3.2 Results

3.2.1 Allele burden is negatively associated with age at diagnosis

Using the SwiftSeq workflow, I uniformly processed and genotyped the blood germline exomes of 8,111 unique individuals, which represented 31 cancer types from The Cancer Genome Atlas (Fig. 3.1a and Supplementary Table 3.4). For each of these individuals, I categorized rare/low-frequency cancer-associated (ClinVar) and deleterious variants (Fig. 3.1b and Supplementary Table 3.5) [352]. Here, “deleterious” refers to loss-of-function variants (stop gained, frameshift, splice donor, etc.) and missense variants predicted to disrupt normal protein function; it does not necessarily mean the variant is disease causative.

I hypothesized that younger age at diagnosis may be indicative of multiple underlying genetic vulnerabilities, particularly rare/low-frequency variants within known cancer risk genes. Across two distinct gene sets — ClinVar cancer genes (CCGs) (n = 57) and autosomal dominant cancer predisposition genes [353] (ADGs) (n = 60) (Tables 3.1 and 3.2) — I tested if increasing burden of known and putatively harmful alleles is associated with earlier age at diagnosis. I regressed age at diagnosis against the number of ClinVar, deleterious, and ClinVar/deleterious alleles within CCGs and ADGs jointly across all cancer types. The burden of ClinVar (P = 4.5 x 10^{-4}), deleterious (P = 1.2 x 10^{-3}), and ClinVar/deleterious (P = 4.2 x 10^{-4}) alleles in CCGs were all negatively associated with age at diagnosis (Fig. 3.2a-c). On average, each additional ClinVar, deleterious, and ClinVar/deleterious allele decreased age at diagnosis by 0.91 (95% confidence interval (CI) = 0.37 1.45), 0.64 (95% CI = 0.23 1.06), and 0.59 (95% CI = 0.24 0.94) years, respectively. Within ADGs, each deleterious allele contributed to 1.31 years earlier age at diagnosis (95% CI = 0.70 1.93; P = 1.6 x 10^{-5}), and each ClinVar/deleterious allele lead to 1.12 years earlier age at diagnosis (95% CI = 0.60 1.63; P = 1.6 x 10^{-5}) (Fig. 3.2d-e). These associations were recapitulated using both the union (n = 87) and intersection (n = 30) of CCGs and ADGs (Fig 3.3a-
Figure 3.1: **Samples per cancer type and deleterious allele counts per individual.**
(a) The number of individuals representing each of the 31 cancer types included from TCGA. 
(b) Mapped reads versus the number of called deleterious, autosomal alleles. Each individual is colored by self-reported race.
Table 3.1: **List of 57 ClinVar Cancer Genes.**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Gene</th>
<th>Gene</th>
<th>Gene</th>
<th>Gene</th>
<th>Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALDH2</td>
<td>CDKN2A</td>
<td>MITF</td>
<td>PTEN</td>
<td>SMAD4</td>
<td></td>
</tr>
<tr>
<td>APC</td>
<td>CHEK2</td>
<td>MLH1</td>
<td>RAD50</td>
<td>STK11</td>
<td></td>
</tr>
<tr>
<td>ATM</td>
<td>CYP17A1</td>
<td>MLH3</td>
<td>RAD51</td>
<td>TGFB2</td>
<td></td>
</tr>
<tr>
<td>ATR</td>
<td>ELAC2</td>
<td>MRE11A</td>
<td>RAD51C</td>
<td>TMEM127</td>
<td></td>
</tr>
<tr>
<td>BARD1</td>
<td>EPHB2</td>
<td>MSH2</td>
<td>RAD51D</td>
<td>TP53</td>
<td></td>
</tr>
<tr>
<td>BLM</td>
<td>FAM175A</td>
<td>MSH6</td>
<td>RET</td>
<td>TSC1</td>
<td></td>
</tr>
<tr>
<td>BMPR1A</td>
<td>FANCC</td>
<td>MUTYH</td>
<td>RNASEL</td>
<td>TSC2</td>
<td></td>
</tr>
<tr>
<td>BRCA1</td>
<td>FH</td>
<td>NBN</td>
<td>SDHA</td>
<td>VHL</td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td>FLCN</td>
<td>NF1</td>
<td>SDHAF2</td>
<td>XRCC2</td>
<td></td>
</tr>
<tr>
<td>BRIP1</td>
<td>MAX</td>
<td>PALB2</td>
<td>SDHB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDH1</td>
<td>MEN1</td>
<td>PLA2G2A</td>
<td>SDHC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDK4</td>
<td>MET</td>
<td>PMS2</td>
<td>SDHD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.2: **List of 60 autosomal dominant cancer predisposition genes.**

To further ensure robustness of these findings, I employed numerous orthogonal control analyses. Using ClinVar, I observed no association between age at diagnosis and the burden of 1) non-pathogenic cancer alleles and 2) pathogenic non-cancer alleles (Fig. 3.4a-b).
Figure 3.2: Increased burden of harmful alleles in cancer risk genes is associated with earlier age at cancer diagnosis. Using 8,111 individuals from 31 cancer types, we regressed age of diagnosis (mean and 95% confidence interval) against the burden of (a) cancer-associated (ClinVar), (b) deleterious, and (c) ClinVar/deleterious alleles in ClinVar cancer genes (CCGs). This relationship was also assessed using the burden of (d) deleterious and (e) ClinVar/deleterious alleles in autosomal dominant cancer predisposition genes (ADGs). Slope ($\beta_{\text{Burden}}$) estimates and $P$ values are provided. The number of individuals included in each allele burden group is shown. For plotting purposes, the 5 allele category in panel d and the 5 and 6 category in panel e were excluded since each contained only a single individual.
Figure 3.3: Age at diagnosis by allele burden using the union and intersection of gene sets. Using the union of ClinVar cancer genes (CCGs) and autosomal dominant cancer predisposition genes (ADGs) (n = 87), age of diagnosis (mean and 95% confidence interval) was regressed against the burden of (a) deleterious and (b) combined cancer-associated (ClinVar)/deleterious alleles. Similarly, using the CCG and ADG intersection (n = 30), we assessed the relationship between (c) deleterious and (d) ClinVar/deleterious alleles and age at diagnosis. Slope ($\beta_{\text{Burden}}$) estimates and one-sided $P$ values are provided. For plotting purposes, the 6 allele category in panel b was excluded since it contained only a single individual.
I also regressed the burden of deleterious alleles across all genes against age at diagnosis. While I found a slight negative association ($\beta_{\text{Burden}} = -0.0076$; 95% CI = -0.015 0.0007; $P = 0.015$), this dissipated after adjusting for self-reported race ($P = 0.19$) (Fig. 3.5a).

4) Compared to random gene set simulations, the observed associations between deleterious allele burden and age at diagnosis remained unlikely to occur by chance (Fig. 3.5b-c). 5) Genes significantly enriched for somatic mutations [354] (oncogenes and predisposition genes excluded) lacked evidence for a negative association between diagnosis age and allele burden (Fig. 3.6 and Supplementary Table 3.8). 6) Leave-one-out analyses demonstrated that the aforementioned associations were not dependent on any single gene (Supplementary Tables 3.9 and 3.10) or 7) cancer type (Supplementary Tables 3.11 and 3.12). Taken together, these analyses strongly support the interpretation that increased harmful allele burden across disease-related loci contributes to younger age at cancer diagnosis.

3.2.3 Enrichment analyses of cancer-associated variants and genes

Together these results indicated that CCG and ADG allele burden may confer cancer susceptibility; however, its utility as a surrogate for risk remained unclear. Using ExAC (version 0.3) [355], I evaluated if deleterious CCG and ADG alleles were more prevalent in individuals with cancer by comparing allele counts (AC) and the number of alleles called (AN) between TCGA and non-TCGA individuals with European ancestry. To determine enrichment, empirical odds ratios were compared to null distributions constructed by randomly sampling 10,000 equally sized sets of genes and variants. The cancer cohort was enriched for deleterious variation in both CCGs and ADGs when considering either gene- ($P = 0.011$; $P = 0.039$) or variant-based ($P = 0.044$; $P = 0.073$) background distributions (Fig. 3.7a-b).

As expected, gene ontology and KEGG pathway enrichment analyses found common cancer terms (i.e. “double-strand break repair” and “negative regulation of cell proliferation”) and pathways (i.e. “Pathways in cancer” and “PI3K-Akt signaling pathway”) over-
Figure 3.4: No observed relationship between age at diagnosis and allele burden using non-cancer ClinVar variants. Boxplots depicting age at diagnosis as a function of allele burden using ClinVar variants (a) not asserted as [ Likely] pathogenic (i.e. labeled only Benign, Likely benign, Other, not provided, etc.) in a cancer phenotype(s) and (b) [ Likely] pathogenic in non-cancer phenotypes. Slope ($\beta_{\text{Burden}}$) and allele burden $P$ value are reported in each panel.
Figure 3.5: **Age at diagnosis against deleterious allele burden exome-wide and within random gene sets.** (a) Age at diagnosis against all deleterious, autosomal alleles exome-wide with unadjusted and race-adjusted $P$ values reported. The estimate and fit line are not adjusted for race. Hotter colors represent a greater density of data points. (b) The distribution of $-\log P$ values after 5,000 simulations regressing age at diagnosis against deleterious allele burden using random sets of 57 and (c) 60 genes. The empirical $-\log P$ value for ClinVar cancer genes (CCGs) and autosomal dominant cancer predisposition genes (ADGs) are represented as a vertical dashed line in panels b and c, respectively. $P$ values in the aforementioned panels indicate the fraction of simulated $P$ values that were lower than the empirical $P$ value.
Figure 3.6: **Age at diagnosis by allele burden in genes significantly somatically mutated in cancer.** Regressions were performed using deleterious allele burden across significantly somatically mutated in cancer (n = 185). Genes were excluded (n = 75) if they were a ClinVar cancer gene (CCG), an autosomal dominant cancer predisposition gene (ADG), or deemed an oncogene by the Cancer Gene Census. Slope ($\beta_{\text{Burden}}$) and allele burden P value are reported. Represented in both CCGs and ADGs. Taking the union of CCGs and ADGs, I found that 9 genes overlapped genome-wide association study (GWAS) genes (n = 362) (Supplementary Table 3.13). The number of shared genes were more than expected by chance ($P = 3.5 \times 10^{-5}$, Fisher’s Exact). While there was significant enrichment, it’s notable that the majority of these cancer predisposition genes do not overlap genes mapped to cancer GWAS hits.

### 3.2.4 Allele burden helps interpret variants of unknown significance

Cancer risk evaluation via clinical sequencing is plagued with difficulties due to variants of unknown significance (VUS) [356], which can be found disproportionately in lower or uncharacterized risk genes. To determine if allele burden can assist interpretation of VUS, I repeated the regression analysis while disregarding ClinVar/deleterious alleles in clinically actionable moderate/high risk genes curated by Slavin and colleagues [357] (Table 3.3). The negative effects of CCG and ADG allele burden on age at diagnosis remained ($\beta_{\text{Burden}} = 0.0054$, $P = 0.52$).
Figure 3.7: **Enrichment of deleterious alleles in individuals with cancer.** Odds ratios (ORs) were calculated by comparing AC and AN between cancer (TCGA) and non-cancer cohorts using data from ExAC. Distributions were generated using 10,000 random sets of genes (ClinVar cancer genes [CCG] n = 57; autosomal dominant cancer predisposition genes [ADG] n = 60) and deleterious variants (CCG n = 4,420; ADG n = 3,278). The vertical dashed line represents the empirical OR calculated using (a) CCGs, (b) ADGs, and the deleterious variants they harbor. *P* values in the aforementioned panels indicate the fraction of simulated ORs that were higher than the respective empirical ORs.
Table 3.3: **List of 21 moderate and high risk cancer predisposition genes.** These genes were curated by Slavin and colleagues [357].

<table>
<thead>
<tr>
<th>APC</th>
<th>BRCA1</th>
<th>CDH1</th>
<th>MLH1</th>
<th>MUTYH</th>
<th>PTEN</th>
<th>SMAD4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATM</td>
<td>BRCA2</td>
<td>CHEK2</td>
<td>MSH2</td>
<td>PALB2</td>
<td>RAD51C</td>
<td>STK11</td>
</tr>
<tr>
<td>BMPR1A</td>
<td>BRIP1</td>
<td>EPCAM</td>
<td>MSH6</td>
<td>PMS2</td>
<td>RAD51D</td>
<td>TP53</td>
</tr>
</tbody>
</table>

Figure 3.8: **Age at diagnosis associated with allele burden when high and moderate risk genes are excluded.** Age at diagnosis against cancer-associated (ClinVar) and deleterious allele burden in (a) ClinVar cancer genes (CCGs) and (b) autosomal dominant cancer predisposition genes (ADGs) while excluding genes known to have moderate or high effects on risk. Each category is represented by the mean and 95% confidence interval. Slope ($\beta_{\text{burden}}$) and $P$ value for burden are reported.

-0.79; 95% CI = -1.14 - 0.26; $P = 2.6 \times 10^{-4}$ and $\beta_{\text{burden}} = -1.59$; 95% CI = -1.14 - 0.26; $P = 3.2 \times 10^{-6}$, respectively) (Fig. 3.8a-b). Additionally — after loss-of-function variants were removed — the burden of missense variants was still associated with earlier age at diagnosis, albeit more weakly (Fig. 3.9a-d). These results demonstrate that well-characterized susceptibility genes and loss-of-function variants are not solely responsible for the association signal and subsequently suggest that collective interpretation of variants of lesser known significance could provide information when ascertaining cancer risk.
Figure 3.9: Associations remain after excluding alleles with predicted high impact on gene function. Characterization of the relationship between age at diagnosis and allele burden after removing variants deemed to have a HIGH (i.e. presumed loss-of-function) impact (e.g. stop gained, frameshift, splice donor, etc.) by Variant Effect Predictor. Regression using non-HIGH impact deleterious alleles in (a) ClinVar Cancer Genes (CCGs) and (b) autosomal dominant cancer predisposition genes (ADGs). Similarly, within CCGs, age at diagnosis by the burden of (c) cancer-associated (ClinVar) alleles and (d) the combination of ClinVar and deleterious alleles. The number of individuals in each allele burden category is shown. Slope (β\text{Burden}) and burden P value are reported within each panel. Mean and 95% confidence interval are provided for each allele burden group.
Figure 3.10: Obese women with uterine/endometrial carcinoma are diagnosed earlier. Individuals were partitioned into six groups based on BMI (kg/m\(^2\)), and age at diagnosis was subsequently summarized via boxplots. Statistical comparisons were made between obese (BMI ≥ 30) and non-obese (BMI < 30) individuals. The \(P\) value (shown) was calculated using a one-sided Welch two-sample T-test.

3.2.5 High allele burden acts independently of BRCA1/2 in breast cancer

Using the union of CCGs and ADGs, I explored the relative relationships of allele burden (ClinVar/deleterious alleles) and canonical risk factors/phenotypes with age at diagnosis. Patients with synchronous/bilateral tumors had earlier age at diagnosis (\(P = 1.0 \times 10^{-3}\)) and greater allele burden (\(P = 6.4 \times 10^{-3}\)). Higher body mass index (BMI) confers susceptibility to multiple cancer types [358], including a remarkable 60% increased risk per 5 kg/m\(^2\) in uterine/endometrial carcinoma [359, 358]. Within this cancer type, individuals classified as obese (BMI ≥ 30 kg/m\(^2\)) were diagnosed 3.73 (95% CI = 1.72 – 5.73) years earlier than non-obese counterparts (\(P = 1.5 \times 10^{-4}\)) (Fig. 3.10). This difference is on par with the 2.84 (95% CI = 0.11 – 5.58; \(P = 0.021\)) years earlier age at diagnosis seen in individuals with high allele burden (four or more alleles) across all cancer types.

51
Since breast was the most prevalent cancer type across the dataset, I wanted to determine the relative effects of high allele burden, BRCA1, and BRCA2 on age at diagnosis. Using a linear regression framework, high burden ($\beta_{\text{HighBurden}} = -7.64$, 95% CI = -14.0 to -1.26; $P = 9.5 \times 10^{-3}$), BRCA1 status ($\beta_{\text{BRCA1}} = -6.63$, 95% CI = -11.4 to -1.82; $P = 3.5 \times 10^{-3}$), and BRCA2 status ($\beta_{\text{BRCA2}} = -4.80$, 95% CI = -8.29 to -1.31; $P = 3.6 \times 10^{-3}$) were all independently associated with earlier age at diagnosis (Fig. 3.11), even after adjusting for self-reported race. High burden displayed an even stronger effect than BRCA1 alleles. The majority (78.6%) of these BRCA1 loci had LOH that retained the harmful allele, supporting their classification as risk variants. Even after removing potentially non-functional BRCA2 terminal variants and BRCA1/2 carriers entirely, high burden still had a significant effect on age at diagnosis (Fig. 3.12a-b). Finally, I noted two individuals who carried harmful BRCA2 alleles and had overall greater burden. One harbored mutant copies of BRCA2, ATM, and MUTYH and the other BRCA2, ATM, RAD51C, and MSH6. Both of these women were diagnosed with breast cancer at 26 years, which was the earliest across the dataset.
Figure 3.11: *BRCA1/2* carrier status and high allele burden independently associate with earlier breast cancer diagnosis. Violin plots representing the age at diagnosis distribution for all female breast cancers, *BRCA1* carriers, *BRCA2* carriers, and individuals with four or more cancer-associated (ClinVar) and deleterious alleles across the union of ClinVar cancer genes (CCGs) and autosomal dominant cancer predisposition genes (ADGs) (n = 87). Boxplots within each violin depict the mean the standard deviation. The sample size, mean, and adjusted *P* value for each category is shown.
Figure 3.12: Highly burdened individuals have earlier breast cancer diagnosis when excluding terminal BRCA2 variants and BRCA1/2 carriers. (a) Violin plots representing the age at diagnosis distribution for all female breast cancers, BRCA1 carriers, BRCA2 carriers, and individuals with four or more cancer-associated (ClinVar) and deleterious alleles in ClinVar cancer genes (CCGs) and autosomal dominant cancer predisposition genes (ADGs). Within this panel, all potential terminal variants in BRCA2 have been removed. (b) The same analysis as panel a except with all BRCA1/2 carriers were excluded. Boxplots within each violin depict the mean the standard deviation. Sample sizes, means, and $P$ values are shown.
3.3 Discussion

Overall, we’ve highlighted the polygenic nature of cancer risk by demonstrating — across multiple sets of predisposition genes — that the burden of rare/low-frequency deleterious and cancer-associated ClinVar alleles associates with increasingly earlier age at diagnosis. While a recent study [360] associated harmful allele burden with age at diagnosis in sarcoma, to my knowledge, this is the first report showing extensive evidence of this association across a large, heterogeneous cohort. Specifically in breast cancer, we’ve shown that the effect of high burden on age at diagnosis is as strong as harmful BRCA1 alleles. I propose that greater levels of baseline genetic vulnerability renders individuals more sensitive to somatic mutation insults, which subsequently manifests in earlier oncogenesis. Evaluating individuals’ putative and bona fide predisposition alleles in aggregate may elucidate disease etiology and assist in cancer risk assessment. Further large-scale studies are required to assess the value of this approach, particularly with clinical screening.

Use of computational methods such as CADD can be scrutinized since predicted harmful variants are sometimes biologically benign. CADD scores have been frequently utilized as a “deleterious” metric for population [361, 362], disease [363, 364, 365, 366, 367, 368], and clinical [369, 370, 371] genetics studies. It has been touted by the InSIGHT Group for it’s ability to identify pathogenic variants as well as prioritize variants of unknown significance [369]. Guidelines published by the American College of Medical Genetics (ACMG) cites CADD and states in silico methods can be used to support claims of pathogenicity [372]. NIH’s Centers for Mendelian Genomics developed a joint protocol that uses CADD to support clinical interpretation of highly-penetrant alleles [373].

A CADD score of > 20 was used to define deleterious variants in another mutational burden study highlighting sex-bias in neurodevelopmental disorders [361]. Additionally, while elucidating genetic underpinnings schizophrenia and developmental disorders, the UK 10K Consortium considered any missense variant with a CADD score >= 15 as damaging [374].
This same “deleterious” or “damaging” criterion of CADD $\geq 15$ was utilized multiple other studies as well [371, 375, 376]. Throughout the analyses, I considered missense variants with a CADD score $\geq 25$ as deleterious, a threshold that is more conservative than other similar studies. So while CADD scores do not prove impact on protein function or clinical deleteriousness, they are accepted and state-of-the-art metrics to assess relative pathogenicity of variants.

In breast cancers, I demonstrated that BRCA2 alleles are associated with earlier age at diagnosis, as one would expect. However, it is possible that some non-pathogenic terminal BRCA2 variants (i.e. lack an autosomal dominant pattern of inheritance) may have been included as ClinVar or deleterious variants. I showed that removing potential terminal variants did not alter conclusions. K3326X — the most frequent terminal stop gain variant in BRCA2 (accounts for approximately 93% [13 of 14] of the terminal BRCA2 variants in breast cancer) — is often considered clinically benign [377]. Nonetheless, this allele is still enriched in familial and sporadic cancers from a variety of anatomical sites [378, 379, 380]. Additionally, the remaining BRCA2 terminal variant had LOH affecting the opposite allele (data not shown). The inclusion of ambiguous alleles such as these was intentional since their effects could manifest in a polygenic context.

### 3.4 Supplementary information

#### 3.4.1 Supplementary tables

Table 3.4: **Demographic and clinical information for individuals from The Cancer Genome Atlas.** (See accompanying supplementary file) Information from 8,210 individuals is depicted.

Table 3.5: **Curated set of cancer-associated ClinVar variants.** (See accompanying supplementary file).
Table 3.6: Unadjusted beta and $P$ values for age at diagnosis by allele burden linear models. (See accompanying supplementary file).

Table 3.7: Burden $P$ values from age at diagnosis by allele burden linear models after adjusting for race and cancer type. (See accompanying supplementary file).

Table 3.8: Significantly mutated genes with and without oncogenes. (See accompanying supplementary file). Significantly mutated genes were defined by Lawrence and colleagues [381]. Oncogenes were curated from the Cancer Gene Census.

Table 3.9: Genic leave-one-out regression analyses for ClinVar cancer genes. (See accompanying supplementary file).

Table 3.10: Genic leave-one-out regression analyses for autosomal dominant cancer predisposition genes. (See accompanying supplementary file).

Table 3.11: Cancer type leave-one-out regression analyses for ClinVar cancer genes. (See accompanying supplementary file).

Table 3.12: Cancer type leave-one-out regression analyses for autosomal dominant cancer predisposition genes. (See accompanying supplementary file).

Table 3.13: Genes associated with cancer phenotypes through genome-wide association studies. (See accompanying supplementary file).
CHAPTER 4

COMBINING COMPUTATIONAL AND FUNCTIONAL ANALYSES TO IDENTIFY NOVEL TWO-HIT TUMOR SUPPRESSOR GENES

4.1 Introduction

After thorough mathematical characterization of a series of hereditary and sporadic retinoblastomas, Alfred Knudson proposed his classic two-hit hypothesis [382]. It postulates that many tumor suppressor genes require biallelic inactivation — or “two-hits” — to facilitate a cancer promoting phenotype. The first gene hit is an inherited harmful allele, while the second hit is acquired somatically and disrupts the remaining wild type (WT) allele. This inactivation can occur genetically through mutations or epigenetically via DNA methylation, though the latter is less common [383, 384, 385, 386, 387]. Most frequently the second hit occurs through loss-of-heterozygosity (LOH), where a deletion removes the WT allele [388]. This mechanism not only reveals how predisposition genes can functionally contribute to the oncogenic transformation cells, but also why heritable malignancies are often entwined with earlier age at diagnosis [389, 390, 382].

Many susceptibility genes follow the two-hit model. Inactivating germline mutations in TP53 lead to Li Fraumeni syndrome, an autosomal dominant disorder that can cause multiple cancers throughout an individual’s lifetime. Childhood cancers are common in this syndrome, underlining the highly penetrant constitution of TP53 mutations [391]. Harmful variants in other tumor suppressor genes such as BRCA1 and BRCA2 induce cancer, typically after reproductive years [392, 393, 394, 395]. These two genes are critical in DNA double-strand break repair, and women who carry defective copies have substantially increased risk for breast and ovarian cancers [396, 397, 398]. In fact, many cancer risk genes serve to mitigate DNA lesions [396]. ATM, which produces a serine/threonine kinase that is
an upstream regulator of cell cycle and DNA damage repair, is also a biallelically inactivated cancer predisposition gene [399]. Inherited deficiencies in this gene are linked to increased risk of multiple solid tumors and blood cancers [400]. Many moderate to high penetrance predisposition genes exhibit some level of cancer type pleiotropy [396, 401]. However, the full extent of risk gene pleiotropy has been insufficiently explored.

Even with a plethora of familial, linkage, and genome-wide association studies, many sources of cancer’s heritability remain “missing” [402, 403]. Both empirical and theoretical population genetics studies indicate that fitness reducing alleles (i.e. those that are disease causative) should be rare across the population (allele frequency < 0.01) [405, 406, 407]. Nonetheless, the presence of loss-of-function alleles within a given individual is not rare. The 1,000 Genomes Project estimates that, on average, each individual carries over 100 non-functional gene copies [407]. Most of these alleles are assumed to be neutral, that is they do not reduce fitness [406]. Separating neutral loss-of-function alleles from those that contribute to disease risk has garnered substantial interest from the community. One way to identify new risk genes and alleles is through rare variant association studies, though conducting such studies comes with numerous obstacles, mainly the required number of cases and controls [408, 409].

Since cancer is characterized by the accumulation of oncogenic mutations, it provides us with clues regarding its own etiology. Through genomic inquiry, the genetic disruptions each malignancy acquired can be determined. These data are typically used to identify somatic disease contributors. However, cancer development is an evolutionary process that begins in normal cells, and germline variation sets the stage on which somatic mutation acts. Leveraging known mutational phenomena — such as biallelic inactivation through LOH — could improve our understanding of germline disease contributions.

Here, by integrating somatic and germline data, I examined Knudson’s two-hit hypothesis across a large collection of malignancies from numerous cancer types. With this approach,
I explored the 1) prevalence of two-hits within known cancer predisposition genes; 2) evidence for risk gene pleiotropy; 3) sexual dimorphism in two-hit frequency; and 4) candidate risk/tumor suppressive genes. Since these classical two-hit events are relatively rare, these analyses could be statistically underpowered, particularly for candidate gene discovery. To supplement computational findings, functional genomics assays in multiple cells lines were performed. This approach subsequently demonstrated that multiple candidates, when depleted, promote oncogenic phenotypes.

**4.2 Results**

4.2.1 Two-hit identification strategy

Using the blood germline exomes and SNP array-based copy number alterations from 5,146 individuals representing 25 cancer types (Table 4.1), I classified two-hit cases and subsequent two-hit enrichment using a straightforward and intuitive approach. For each gene of interest, individuals that carried a deleterious germline allele and have somatic LOH at that locus were identified. For each of these individuals — using the tumor exome data — I calculated the variant allele fraction (VAF) of the deleterious germline allele. Those with a VAF > 0.5 were deemed “two-hit” since evidence suggested the WT allele was lost — causing biallelic inactivation. Those with a VAF < 0.5 were considered “one-hit” as the deleterious allele was likely lost as the result of LOH. Importantly, these one-hit cases still have a functional copy of the gene. Then the numbers of one- and two-hit cases for each gene were aggregated across the cohort. If biallelic loss did not increase cell fitness, one- and two-hit events would be expected to occur at the same frequency. However, if biallelic loss promotes a more oncogenic phenotype, there would be an excess of two-hit cases. Statistical enrichment for two-hit cases was calculated with a one-way binomial test (Fig. 4.1).
Table 4.1: TCGA cancer types and samples counts used for two-hit analyses.

<table>
<thead>
<tr>
<th>Type</th>
<th>n</th>
<th>Type</th>
<th>n</th>
<th>Type</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>79</td>
<td>HNSC</td>
<td>313</td>
<td>PAAD</td>
<td>68</td>
</tr>
<tr>
<td>BLCA</td>
<td>200</td>
<td>KICH</td>
<td>9</td>
<td>PRAD</td>
<td>242</td>
</tr>
<tr>
<td>BRCA</td>
<td>883</td>
<td>KIRC</td>
<td>72</td>
<td>READ</td>
<td>102</td>
</tr>
<tr>
<td>CESC</td>
<td>37</td>
<td>LGG</td>
<td>365</td>
<td>SARC</td>
<td>114</td>
</tr>
<tr>
<td>COAD</td>
<td>247</td>
<td>LIHC</td>
<td>110</td>
<td>SKCM</td>
<td>78</td>
</tr>
<tr>
<td>DLBC</td>
<td>23</td>
<td>LUAD</td>
<td>360</td>
<td>STAD</td>
<td>266</td>
</tr>
<tr>
<td>ESCA</td>
<td>49</td>
<td>LUSC</td>
<td>261</td>
<td>THCA</td>
<td>409</td>
</tr>
<tr>
<td>GBM</td>
<td>314</td>
<td>OV</td>
<td>287</td>
<td>UCEC</td>
<td>258</td>
</tr>
</tbody>
</table>

Figure 4.1: Workflow for identifying candidate two-hit genes.
4.2.2 Quantifying two-hit frequency in known cancer predisposition genes

First, a set of 114 cancer predisposition genes of varying penetrance and modes of inheritance were assessed [396]. The top three two-hit genes were BRCA1 (54 of 59 two-hit; \( P = 6.7 \times 10^{-14}; \ Q = 2.5 \times 10^{-10} \)), BRCA2 (66 of 98; \( P = 8.5 \times 10^{-6}; \ Q = 0.011 \)), and ATM (60 of 85; \( P = 2.0 \times 10^{-6}; \ Q = 3.8 \times 10^{-3} \)), all of which were significantly enriched for two-hits after multiple testing correction (Fig. 4.2a-b and Table 4.2). Signal from BRCA1 was almost entirely attributable to breast and ovarian cancer types (Fig. 4.2c). Only five two-hits in BRCA1 were from other cancer types; three of these occurred in uterine/endometrial carcinoma, which has been linked to BRCA1-mediated susceptibility [410]. Contrastingly, two-hits in BRCA2 were more heterogeneous occurring in 15 distinct cancer types (Fig. 4.2d), most notably glioblastoma multiforme (n = 5; 4 of 5 two-hit), suggesting cell type-promiscuous tumor suppressive activity. Overall, 37.9% (n = 25) two-hit cases occurred in cancer types other than breast and ovarian, many of which (e.g., glioblastoma, liver carcinoma, head and neck carcinoma) have little or no evidence of BRCA2-mediated risk [411]. ATM exhibited a similar pattern with two-hits in 16 different cancer types. These occurred most frequently in stomach (6 of 8) and lung (6 of 7) adenocarcinomas as well as breast cancer (24 of 35), all of which have been associated with ATM-mediated susceptibility [412, 413, 414].

I found that only eight of the 114 cancer predisposition genes (BRCA1, BRCA2, ATM, WRN, COL7A1, CHEK2, NF1, and RAD51D) had pan-cancer \( P \) values < 0.1 [396]. Collectively, the entire set of predisposition genes was implicated in 483 one-hit and 589 two-hit cases, representing notable enrichment (\( P = 1.25 \times 10^{-10} \)), although this effect was reduced when BRCA1, BRCA2, and ATM were excluded (\( P = 0.011 \)). 44 of these genes did not contain a single two-hit case and only 40 genes contained more than two. These findings potentially highlight the varying penetrance and etiology of predisposition genes, especially when considering heterogeneous cancer types.
Figure 4.2: Two-hit enrichment pan- and per-cancer type. (a) Plotting the highest per-cancer $P$ value (-log10) against the pan-cancer $P$ value (-log10) for each gene. Each gene is colored by cancer type where it is the most enriched. If a gene’s greatest enrichment occurred in two or more cancer types, it is colored grey for multiple. (b) Closer view of genes from panel a with candidate two-hit genes highlighted. Waterfall plots depicting the variant allele fraction (VAF) for all one- and two-hit cases in (c) $BRCA1$, (d) $BRCA2$, (e) $GPR126$, (f) $ACO1$, (g) $PHLPP2$, and (h) $KDELRC2$. Black, horizontal lines indicate a VAF of 0.5. Cases with a VAF below and above this threshold are considered one-hit and two-hit, respectively. Each case is colored by its cancer type.
<table>
<thead>
<tr>
<th>Gene</th>
<th>One-hit count</th>
<th>Two-hit count</th>
<th>Total</th>
<th>One-way textit P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>5</td>
<td>54</td>
<td>59</td>
<td>9.5 x 10^{-12}</td>
</tr>
<tr>
<td>ATM</td>
<td>25</td>
<td>60</td>
<td>85</td>
<td>9.3 x 10^{-5}</td>
</tr>
<tr>
<td>BRCA2</td>
<td>32</td>
<td>66</td>
<td>98</td>
<td>3.8 x 10^{-4}</td>
</tr>
<tr>
<td>GPR126</td>
<td>23</td>
<td>47</td>
<td>70</td>
<td>2.8 x 10^{-3}</td>
</tr>
<tr>
<td>ACO1</td>
<td>9</td>
<td>25</td>
<td>34</td>
<td>4.5 x 10^{-3}</td>
</tr>
<tr>
<td>KDELC2</td>
<td>28</td>
<td>52</td>
<td>80</td>
<td>4.8 x 10^{-3}</td>
</tr>
<tr>
<td>PHLPP2</td>
<td>1</td>
<td>10</td>
<td>11</td>
<td>5.9 x 10^{-3}</td>
</tr>
<tr>
<td>DBR1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>7.8 x 10^{-3}</td>
</tr>
</tbody>
</table>

Table 4.2: Top pan-cancer two-hit genes.

4.2.3 Novel two-hit genes pan-cancer

In addition to known susceptibility genes, multiple candidates showed a preponderance of two-hit scenarios, although none were significant after multiple testing correction (Fig. 4.2b and Table 4-2). Adhesion G protein-coupled receptor G6 (ADGRG6), also known as GPR126 (47 of 70 two-hit; $P = 1.8 \times 10^{-4}; Q = 0.17$) had an excess of two-hit cases. Its highest rates of two-hits came from head and neck cancers (5 of 6) and lower grade glioma (5 of 5) (Fig. 4.2e). Another candidate, ACO1 is a highly conserved aconitase that catalyzes the conversion of citrate to isocitrate in the tricarboxylic acid cycle [415]. It displayed two-hits across a variety of cancer types (25 of 34; $P = 7.8 \times 10^{-4}; Q = 0.49$), though these scenarios were particularly prevalent in head and neck cancer (Fig. 4.2f). Interestingly, 18 of the 25 two-hit cases involved the glycine to arginine missense variant rs34630459 (ExAC AF = 0.0078; CADD = 35.0). PHLPP2, a member of the PI3K-Akt signaling pathway, encodes a phosphatase responsible for dephosphorylating AKT1, which subsequently initiates apoptosis [416]. Across multiple cancers, 91% of individuals with a PHLPP2 deleterious allele and LOH lost the WT allele (10 of 11; $P = 2.3 \times 10^{-3}; Q = 0.88$) (Fig. 4.2g). PHLPP2 has exhibited tumor suppressive activity in multiple cancer types [417, 418, 419, 420]; however, behavior consistent with the Knudson’s two-hit hypothesis has not yet been reported. Lastly, two-hits in the relatively unexplored gene KDELC2 (52 of 80; $P = 2.9 \times 10^{-4}; Q = 0.22$) occurred
primarily in breast cancer (Fig. 4.2h).

### 4.2.4 Sexual dimorphism in two-hit acquisition

Genetic architecture of complex diseases can vary between sexes [421, 422]. Additionally, recent reports have shown that males and females have different patterns of somatic mutation and gene expression across malignancies [423, 424]. I wanted to determine if the most prominent two-hit genes within each sex are depleted in the other. For this analysis, predominantly sex-specific cancer types (BRCA, CESC, OV, PRAD, and UCEC) were removed. From the remaining cancer types, I compiled the top 10 pan-cancer two-hit genes for each sex and performed Fisher’s exact tests to determine if two-hits in any gene was biased towards males or females. Notably, *ATM* and *GPR126* displayed two-hit signal in both males and females, providing no evidence of sex specificity (Fig. 4.3). Relatively understudied genes *LEPRE1* \( (P = 0.031) \) and *ZNF488* \( (P = 0.011) \) were significantly enriched for two-hits in females. *MYO15A* \( (P = 0.043) \), a gene implicated in congenital hearing loss [425, 426, 427], was also more prevalent in females. Reciprocally, two-hits in *SGSM3* \( (P = 0.024) \), a G protein signalling modulator, were almost exclusive to males. Variants proximal to this gene have been implicated in breast [428, 429] and hepatocellular carcinoma [430] as well as mammographic density [431]. Most notably, *BRCA2* two-hit cases were enriched in females \( (P = 0.013) \) (Fig. 4.3). While carriers are most at risk for female cancers such as breast and ovarian, *BRCA2* mutations also confer risk to a variety of sex-independent malignancies such as pancreatic cancer and acute myeloid leukemia [432, 433]. These results suggest that biallelic *BRCA2* loss, at least with respect to two-hit mediated susceptibility, may be more prominent in females.
Figure 4.3: **Top male and female candidates display sexually dimorphism.** The union of the top ten two-hit candidates from both males and females. Gender enriched $P$ values were calculated using a Fishers exact test which compared one- and two-hit counts in males versus females. Circle size is proportional to gender enriched $P$ values with smaller $P$ values depicted by larger circles. Genes showing evidence of two-hit sexual dimorphism ($P < 0.05$) are colored yellow.

### 4.2.5 Functional analysis of pan-cancer candidates

Since the intersection of relatively rare events was being explored, the ability to make definitive conclusions with purely statistical framework was limited. As such, Mike Bolt and I leveraged functional assays to gather orthogonal support for our computational findings. As uncontrolled cell growth is a key step in oncogenesis, we wanted to determine if depletion of two-hit candidates increased proliferation. MCF10A (normal breast epithelial) and MRC-5 (normal lung fibroblast) cells were subjected to a 96 hour siRNA knockdown ($\text{GPR126, ACO1, PHLPP2, and KDELC2}$), and cells were subsequently measured for proliferative capacity. Specifically, in MCF10A cells, siRNA to $\text{KDEL C2}$ ($P = 1.8 \times 10^{-3}$, Welch two-sample T-test) and $\text{PHLPP2}$ ($P = 1.3 \times 10^{-10}$) caused significant increases in proliferation (Fig. 4.4a). siRNA to $\text{PHLPP2}$ caused a borderline significant proliferation increase in MRC-5 cells ($P = 0.063$). Cell type independent effects are consistent with $\text{PHLPP2}$’s
Figure 4.4: **Knockdown of candidate two-hit genes induces cancerous phenotypes.** (a) Proliferation and (b) scratch assays performed in MCF10A (left) and MRC-5 (right) cell lines after KDELC2 and PHLPP2 knockdown with siRNA. The mean the standard deviation are provided with each bar. Results that were significantly different (Welch two-sample T-test) from control siRNA (siCon) are indicated within each panel.

Another key oncogenic step is the acquisition of a migratory phenotype. To determine if any two-hit candidates affected cellular migration, Mike Bolt performed scratch assays in MCF10A and MRC-5 cells again after siRNA treatment. Knockdown of PHLPP2 in MCF10A ($P = 2.1 \times 10^{-4}$, Welch two-sample T-test) as well as both PHLPP2 and KDELC2 in MRC-5 ($P = 8.5 \times 10^{-3}$ and $P = 0.011$, respectively) lead to decreases in scratch area, which is indicative of increased migration (Fig. 4.4b). Together these results promote a role for these two-hit candidates — particularly PHLPP2 — in oncogenesis across cell types.

KDELC2 showed an interesting phenotype as it was a two-hit in breast and lung cancers, but it showed different oncogenic properties when knocked down in those cell types. More
specifically, proliferation and migration increase were observed in MCF10A and MRC-5 cells, respectively (Fig. 4.4a-b). Furthermore, closer inspection revealed that KDELC2 is proximal to ATM (approximately 0.1 Mb apart). Interestingly, nearly all (92.8%) KDELC2 two-hit cases co-occurred with two-hits in ATM, strongly suggesting that some putatively harmful alleles in these genes share haplotypes. In light of these functional results, evidence indicates that KDELC2 loss may play a legitimate role in tumorigenesis and not simply hitchhike with ATM driver events.

4.2.6 Characterization of cancer-specific candidates ROBO1 and DBR1

Throughout these analyses I discovered a variety of two-hit candidates that were observed in some cancer types and not others. I wanted to determine whether these findings could be extrapolated to phenotypic changes in cell lines from the same cancer type and chose to test this on ROBO1 and DBR1. ROBO1 was a top cancer-type specific candidate with two-hits mainly observed in breast cancer, while two-hit DBR1 is primarily seen in lung, with no occurrences in breast cancer (Fig. 4.5a). In order to determine if loss of these genes conferred cell-type specific phenotypes, Mike Bolt performed proliferation and scratch assays after siRNA knockdown in MCF10A and MRC-5 cell lines. For both assays, ROBO1 knockdown only triggered oncogenic properties in MCF10A cells, while siRNA targeting DBR1 generated the same phenotypes only in MRC-5. Conversely, DBR1 depletion led to decreased proliferative ability in MCF10A cells while ROBO1 has insignificant effects on MRC-5 cells (Fig. 4.5b-c). Even though data were limited, Cox-proportional hazards models suggest that having a ROBO1 two-hit in breast cancer (hazards ratio = 4.4; P = 0.11) or DBR1 two-hit in lung cancer (hazards ratio = 5.5; P = 0.01) confers worse prognosis (overall survival) than not having a two-hit. These results provide computational and functional evidence that ROBO1 and DBR1 act as cancer-specific, two-hit genes. Patient survival data also indicate that both genes may have clinical relevance.
Figure 4.5: **ROBO1** and **DBR1** show cell type-specific phenotypes. (a) Waterfall plots depicting the variant allele fraction (VAF) for all one- and two-hit cases in **ROBO1** (left) and **DBR1** (right). Black, horizontal bars indicate a VAF of 0.5. Cases with a VAF below and above this threshold are considered one-hit and two-hit, respectively. Each case is colored by its cancer type. (b) Proliferation and (c) scratch assays performed in MCF10A (left) and MRC-5 (right) cell lines after **ROBO1** and **DBR1** knockdown with siRNA. The former cell line is derived from normal breast epithelium and the latter from normal lung fibroblasts. Each individual data point (red) is presented with error bars representing one standard deviation above and below the mean. Results that were significantly different (Welch two-sample T-test) from control siRNA (siCon) are indicated within each panel.
4.2.7 ROBO1 knockdown represses DNA damage response

Many tumor suppressor genes are involved in DNA repair pathways [396]. In breast cancer, I noted that individuals harboring two-hits in ROBO1 had increased copy number segmentation compared to the rest of the cohort \((P < 0.05,\) Mann-Whitney U), implying greater genomic complexity. Consequently, Mike Bolt and I wanted to determine if ROBO1 loss could affect DNA damage repair, both basally and in response to cisplatin. In order to measure the effect of gene candidate knockdown on DNA damage, fluorescence microscopy analysis were performed using an antibody that targets yH2AX, the histone modification that marks DNA damage within the genome. Using MCF10A cells, we observed a 14.5\% increase in yH2AX/DAPI signal with knockdown of ROBO1 \((P < 1.0 \times 10^{-7},\) ANOVA with Tukey HSD). This effect was enhanced to 50.8\% when comparing siROBO1 plus cisplatin to control siRNA plus cisplatin \((P < 1.0 \times 10^{-7},\) ANOVA with Tukey HSD). These findings indicate that ROBO1 deficient cells have increased basal DNA damage and, most of all, have reduced capacity to fix DNA damage caused by cisplatin. To our knowledge, this is the first study implicating ROBO1 with DNA damage.

4.3 Discussion

Expanding the computational approach to include experimental validations afforded the ability to determine the oncogenic characteristics of rarer two-hit genes (e.g. DBR1) in a variety of cell lines. Consequently, the pro-proliferative and pro-migratory effects of DBR1 knockdown in MRC-5 lung fibroblasts became an alluring finding. DBR1 breaks down RNA lariats formed during mRNA splicing to allow the discarded introns to be further enzymatically broken down into single bases [436, 437]. Until recently, there have been no studies implicating DBR1 in cancer. Han and colleagues found that downregulation of DBR1 led to oncogenic transformation and defects in RNA processing (e.g. exon skipping) [438]. Interestingly, they also used lung cells when identifying DBR1’s tumor suppressive activity.
Finally, the fact that two-hits in *DBR1* all involved the same variant (rs36061810) make it an interesting candidate to interrogate using CRISPR-mediated allelic recombination [439].

The true standout from both computational and experimental analyses was *ROBO1*. *ROBO1* is a cell surface receptor for SLIT1 and SLIT2 and plays roles in neural development and cellular migration [440, 441, 442]. The findings of increased proliferation, migration, and decreased DNA damage response in *ROBO1* knockdown cells provide novel functional insight into the receptor. Previously, high *ROBO1* levels in breast cancer have been demonstrated to correlate with better outcome [443] while low levels correlate with poor prognosis and brain-specific metastasis [444]. Further, while *ROBO1* has not been reported as a two-hit gene, it has been shown to be hypermethylated in some breast cancer samples where the other allele is deleted, creating a de facto two-hit state [445]. These findings also suggest a previously unknown role for *ROBO1* signaling in the DNA repair pathway, both basally and in reaction to cisplatin treatment. Since many two-hit cancer predisposition genes affect DNA repair, this gives further credence to *ROBO1*’s potential as a risk gene. These finding are also supported by a report that germline *ROBO1* deletions segregate with affected individuals in families with hereditary cancer [446].

The observed sexually dimorphic effects in *BRCA2* could be related to hormone signaling. Estrogen receptor alpha binds to the promoter of *BRCA2* and can subsequently mediate transcription [447]. Additionally, harmful *BRCA2* allele carriers frequently develop estrogen receptor positive breast tumors [448] and exogenous estrogens exacerbate their breast cancer risk [449]. Given the combination of epidemiological and molecular evidence, it is plausible that an interaction exists between *BRCA2* and estrogen with respect to tumorigenesis.

Even though these analyses emphasized two-hit genes from a risk perspective, it is possible that candidate genes play no role in susceptibility. Biallelic inactivation may only confer a selective advantage for an already oncogenic cell population. This would suggest that harmful alleles in these genes do not increase risk, though they will alter the evolutionary
trajectory once cancer manifests. Such an assertion is certainly plausible, and it is supported by previous findings. Multiple studies have identified inherited variants associated with therapeutic response and outcome \([450, 451, 452, 453]\), indicating they function as a modifier for disease severity. Further studies — both computational and experimental — should be conducted to interrogate the etiology of biallelic inactivation in the candidate genes, especially \(ROBO1\).

This study was intentionally designed to identify genes that require biallelic inactivation. Numerous studies have shown that many tumor suppressor genes, including predisposition genes, are haploinsufficient and do not adhere to Knudson’s two-hit model \([389, 454]\). This is exemplified by the recent discovery of \(HAPB2\) as a dominant-negative predisposition gene in nonmedullary thyroid carcinoma \([455]\). These analyses also did not detect enrichment in known two-hit tumor suppressor genes \(TP53\) and \(PTEN\). This could be due to a number of factors. \(Bona fide\) harmful alleles in these genes are highly penetrant, leading to Li Fraumeni \([456, 457, 458]\) and Cowden syndromes \([459]\), respectively. These syndromes are relatively rare \([460, 461]\), and thus we’d expect to see only a small proportion of randomly sampled individuals with cancer harboring mutations in these genes. It’s also possible that TCGA sample collection procedures were biased against individuals diagnosed with these syndromes. Importantly, identifying two-hit tumor suppressor genes with this approach depends on a variety of technical factors such as variant calling accuracy, sufficient tumor exome coverage, and deleterious allele classification. Any negative results must be interpreted with caution.
CHAPTER 5

COMPARISON OF BREAST CANCER MUTATIONAL PATTERNS ACROSS AFRICAN AND EUROPEAN ANCESTRY POPULATIONS

5.1 Introduction

Breast cancer is a heterogeneous disease comprised of distinct subtypes. Global burden of the disease and severity also vary widely across populations, with women of African ancestry being diagnosed at a younger age, having more clinically aggressive disease and advanced stage, and having higher mortality rates than age-matched women of European or Asian ancestry [462, 463, 464, 465]. Molecular and genetic characteristics strongly influence prognosis and treatment, with human epidermal growth factor receptor 2 (HER2) amplification and hormone receptor (HR; estrogen receptor [ER] and progesterone receptor [PR]) expression being the best examples.

Recent large sequencing studies, for instance the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), have refined our knowledge of the genomic landscape and pathogenesis of breast cancer; have provided insight into tumor evolution and mechanisms of drug resistance; and have laid a pathway to deployment of precision therapeutics [466, 467, 468, 469, 470, 471, 472, 473, 474, 475]. Moreover, these large public datasets have enhanced our understanding on the divergent mutation accretion processes; most notably in breast cancer, studies have shown high APOBEC-related mutagenesis especially in HER2+ tumors [476], whereas BRCA1/2 mutations are strongly associated with signatures depicting DNA repair deficiency [477].

The cases used to elucidate the genetic basis of breast cancer have been overwhelmingly from women with European ancestry, which reiterates the need for data from underrepresented ethnicities [478, 479, 480]. This paucity of data from African countries potentially
widens the knowledge gap that contributes to disparities in breast cancer outcomes. To get a comprehensive understanding of the genetic architecture of breast cancer in West Africa, the founder population of a large proportion of women throughout the African Diaspora, researchers from Novartis and the University of Chicago conducted whole-genome sequencing (WGS), whole-exome sequencing (WES) and transcriptome sequencing (RNA-seq) of 194 tumors from Nigerian patients. With this data, Markus Riester and I performed a comparative analysis with Black patients of African ancestry and White patients of European ancestry in TCGA. To the best of my knowledge, combined with the Black patients in TCGA, this is the largest breast cancer genomics study on African ancestry individuals to date.

5.2 Results

5.2.1 Mutational landscape across study populations

The Nigerian cohort is comprised of 194 breast cancer patients: 40 with WGS data, 129 with WES data and 103 with RNA-seq data (Fig. 5.1). Of the 1,097 TCGA breast cancer patients with either WES (n = 1,035) or WGS (n = 84), 1,030 were assigned without ambiguity to three ancestry race groups, and the other 67 had mixed racial background. DNA sequencing data from all samples was uniformly processed using the SwiftSeq workflow. Patient numbers and characteristics are provided in Tables 5.1, 5.2, 5.3, and 5.4.

Congruous with previous studies including Surveillance Epidemiology End Results (SEER) dataset [463, 481], a strong enrichment of HR- (ER- and PR-)/HER2- (43% in Nigerian vs. 33% in Black and 13% in White) and HR-/HER2+ (25% vs. 6% and 2%) subtypes was observed in African ancestry individuals (Fig. 5.2a). PAM50 subtyping revealed a similar enrichment of Basal (32% vs. 35% and 15%) and HER2 enriched (29.1% vs. 8.8% and 5.2%) in Nigerians (Fig. 5.2b).

Across all 1,164 individuals with uniformly processed WES data, I identified 25 genes
that were significantly mutated above background ($Q < 0.05$, MutSigCV). Four of these genes ($PLK2$, $KDM6A$, $GPS2$, and $B2M$) had little or no previous evidence of harboring mutations that drive breast carcinogenesis (Fig. 5.3a-d) [482]. Notably, mutations in $PLK2$ ($P = 0.048$) and $KDM6A$ ($P = 0.06$) were enriched within HER2+ individuals. Combined with previously reported significantly mutated genes in breast cancer [474, 381], this resulted in 44 driver genes (Fig. 5.2c and Table 5.5).

Consistent with the aggressive subtype composition in Nigerians, there was an enrichment of $TP53$ alterations (62% vs. 45% and 29%, BH $P < 1.0 \times 10^{-4}$, Fisher’s exact) as well as a lower rate of $PIK3CA$ mutations (17% vs. 20% and 36%, BH $P < 1.0 \times 10^{-4}$) among these 44 breast cancer drivers (Fig. 5.2c). Combined $BRCA1$ germline and somatic variants were also enriched in the Nigerian cohort (11.6% vs. 7.0% and 4.0%, BH $P = 0.03$). $CDH1$ mutation was rare in Nigerians (0.8% vs 6.4% and 16.2%, BH $P < 1.0 \times 10^{-4}$), whereas $GATA3$ alterations were more common in Nigerians (17.1% vs. 10.0% and 9.5%, BH $P = 0.24$). When comparing recurrently gained or lost regions as identified by GISTIC2, all high confidence peaks identified in the Nigerian cohort had corresponding peaks within 10 Mb in the combined TCGA cohort. In line with IHC and PAM50, the $ERBB2$ locus (17q12) was enriched in Nigerians (amplified in 24% vs. 12% and 10%, BH $P = 2.0 \times 10^{-3}$), as was its
Figure 5.2: Landscape of breast cancer in Nigerians compared to Black and White Americans. (a) Incidence of IHC subtypes in the Nigerian, the Black and White cohorts from TCGA, and in the SEER database. (b) Incidence of PAM50 subtypes in Nigerians, Blacks and Whites. (c) Comparison of the frequencies of short variants (SNVs and indels) in 44 breast cancer drivers in all cohorts. (d) Alteration frequencies of 19 genes recurrently affected by CNAs (homozygous deletions and amplifications). (e) Comparison of key breast cancer drivers stratified by IHC subtype. Both short variants and copy number events are included. * $P < 0.05$; ** $P < 0.001$; *** $P < 0.0001$ (Fishers exact with $P$ values adjusted via the Benjamini-Hochberg method).
Figure 5.3: **Lollipop plots for novel significantly mutated breast cancer genes.** Protein-altering SNVs and indels for (a) *PLK2*, (b) *KDM6A*, (c) *GPS2*, and (d) *B2M*. The start position of a deletion in *B2M* falls outside of the first exon; however, that deletion is represented in panel d since it spans part of the first exon.
### Oncoprint of short mutations and CNAs in Nigerians

Recurrently mutated genes (Tables 5.5 and 5.6) that were altered least 3% of Nigerians are shown in Figure 5.4. A wide neighboring peak at 17q23.1 (TBX2 locus, BH $P = 0.1$) (Fig. 5.2d).

Within IHC subtypes, significantly mutated genes and copy number peaks (Table 5.6) generally displayed similar proportions across ethnicities, suggesting that most mutation frequency differences reflect subtype differences across ethnicities (Fig. 5.2e). Within the HR+/HER2- subtype, however, there were more $TP53$ and $GATA3$ mutations, and fewer $PIK3CA$ and $CDH1$ mutations in Nigerians, compared to TCGA Blacks and Whites (all $P < 0.05$) (Fig. 5.4). This suggests that HR+/HER2- breast cancers in Nigerian women have genomic lesions consistent with more aggressive disease.

### 5.2.2 Mutation signatures across subtypes and driver mutations

I next extracted breast cancer mutational signatures in the 122 WGS and 500 WES samples from all cohorts harboring 100 or more mutations (Tables 5.7 and 5.8). Previously identified
breast cancer signatures closely matched signatures A (APOBEC C>T), B (APOBEC C>G), C (Aging), H (Signature 8) and I (Homologous recombination deficiency [HRD]) (Fig. 5.5 and Fig. 5.6). Given their high correlation between exomes and genomes (Fig. 5.7), I examined these five signatures for subsequent analyses. Combined, they explain the vast majority of mutations regardless of race/ethnicity (Fig. 5.8a) or subtype (Fig. 5.8b).

Increased contributions from APOBEC C>T ($P = 3.5 \times 10^{-9}$, Mann-Whitney U [MWU]) and APOBEC C>G ($P = 0.044$) signatures were observed in HR+ tumors compared to HR- tumors, which is consistent with previous findings [483, 484]. Conversely, the HRD signature was substantially more active in HR- tumors ($P = 2.2 \times 10^{-15}$) (Fig. 5.9a-d). Consistent with previous work [476], HER2+ tumors had the highest contributions from APOBEC C>T and C>G signatures ($P = 1.6 \times 10^{-8}$ and $P = 9.1 \times 10^{-4}$, respectively) (Fig. 5.8b and
Figure 5.6: Derived mutation signatures compared to COSMiC mutation signatures and correlation between WES and WGS signature contributions. Kullback-Leibler divergence was calculated pairwise for derived and COSMiC mutation signatures. For each derived signature, the smallest divergence value — which indicates the most similar COSMiC signature — is denoted by a pink circle.
Figure 5.7: Correlations between WES and WGS mutation signature contributions. Scatterplots of nine mutation signature contributions between WES and WGS for 59 individuals. Spearman correlation was calculated for each signature with Rho depicted on each plot. The derived signature is denoted above each plot.
Figure 5.8: **Mutation signature contributions across race/ethnicity and subtype.**  
(a) The contribution (proportion) of mutation signatures (Signatures D, E, F, and G are combined into "Other") within each individual. Individuals are partitioned by race/ethnicity and ordered by APOBEC C>T signature contributions (high to low). The number of individuals representing each cohort is shown. (b) Mekko plot of the proportional contributions of mutation signatures across IHC subtypes.
Similarly, I recapitulated the known aging signature associations and confirmed higher HRD contributions in individuals harboring deleterious germline or somatic BRCA1/2 mutations ($P = 6.2 \times 10^{-7}$) [476, 477]. TP53 mutations were associated with higher HRD contributions (BH $P = 2.1 \times 10^{-13}$, MWU), higher missense mutation burden (BH $P = 6.5 \times 10^{-45}$), and increased copy number segmentation (BH $P = 2.0 \times 10^{-43}$) (Fig. 5.10a). In contrast, CDH1 or PIK3CA mutations — which frequently co-occur ($P = 3.8 \times 10^{-8}$) — were associated with lower HRD contributions (CDH1 BH $P = 5.2 \times 10^{-11}$; PIK3CA BH $P = 2.1 \times 10^{-17}$) in addition to higher contributions from APOBEC C>T (BH $P = 3.0 \times 10^{-9}$; BH $P = 3.8 \times 10^{-17}$) and C>G (BH $P = 1.7 \times 10^{-4}$; BH $P = 2.1 \times 10^{-6}$) (Fig. 5.10a). Importantly, these significant associations persisted even when considering only HR+/HER2- tumors (Fig. 5.10b). These findings suggest a consistent interplay between driver mutations and the relative activity of mutational processes.

### 5.2.3 Mutation signatures across races/ethnicities

When partitioned by IHC subtypes, the APOBEC C>T signature displayed differences by race/ethnicity in HR+/HER2- with Nigerian ($P = 0.02$) and Black cohorts ($P = 0.05$) having lower APOBEC C>T contributions compared to Whites. In the HR-/HER2- subtype, Nigerians had slightly increased APOBEC C>G signature relative to the Black ($P=0.06$) and White ($P = 6.8 \times 10^{-3}$) cohorts (Fig. 5.11a-b). Signature 8 demonstrated substantial contribution differences between cohorts. This effect was the most pronounced in HR-/HER2- tumors, where Nigerians and Blacks ($P = 4.4 \times 10^{-6}$), Nigerians and Whites ($P = 4.6 \times 10^{-12}$), as well as Blacks and Whites ($P = 0.023$) were significantly different from one another (Fig. 5.12a). Notably, Whites presented with remarkably higher signature 8 in HR-/HER2- (mean = 20.6%) compared to HR+/HER2- (mean = 12.2%) tumors ($P = 3.4 \times 10^{-7}$), which was recapitulated using WGS data ($P = 6.9 \times 10^{-3}$) (Fig. 5.13a-b). These subtype differences were not observed for either Nigerians or Blacks.
Figure 5.9: Mutation signature contributions between tumors positive and negative for IHC markers. Boxplots represent (a) APOBEC C>T, (b) APOBEC C>G, (c) aging, and (d) HRD signatures partitioned by ER status. Similarly, contributions from (e) APOBEC C>T and (f) APOBEC C>G signatures between HER2 positive and negative tumors. $P$ values shown were calculated via Mann-Whitney U.
Figure 5.10: **Associations between genome-wide oncogenic features and the mutation status of common driver genes.** Dot plot depicting the relationships between mutation status in TP53, PIK3CA, CDH1, and GATA3 and mutation signatures (APOBEC C>T, APOBEC C>G, aging, HRD, and signature 8), missense mutation burden, and copy number (CN) segments (a) across all IHC subtypes and (b) within HR+/HER2-. Comparisons between mutation status and genomic features were performed with Mann-Whitney U and \( P \) values were corrected for multiple testing (Benjamini-Hochberg method). Circle size is proportional to the magnitude of the -log\(_{10}\) BH \( P \) value (i.e. lower BH \( P \) values have larger circles). If mutation status associated with a significant increase or decrease of a genomic feature, the corresponding circle is colored red or blue, respectively. Non-significant (N.S.) comparisons are colored black.
Figure 5.11: The proportion of APOBEC $C>T$, APOBEC $C>G$, and aging signatures by race/ethnicity and IHC subtype using WES. The proportion of APOBEC $C>T$, APOBEC $C>G$, and aging signatures by race/ethnicity and IHC subtype using WES. Differences in (a) APOBEC $C>T$, (b) APOBEC $C>G$, and (c) aging signatures contributions by race/ethnicity within each IHC subtype were assessed using Kruskal-Wallis tests with post hoc comparisons made via Dunn’s test. * $P$ values < 0.01; + $P$ values $\leq$ 0.05.
Figure 5.12: Mutation signature contributions and structural variant counts partitioned by race/ethnicity and IHC subtype. Mutation signature contributions from (a) HRD and (b) signature 8 subdivided by race/ethnicity and IHC subtype. (c) Boxplots representing the number of SVs identified across WGS samples partitioned by race/ethnicity and IHC subtype. Asterisks denote significant differences ($P < 0.05$) between groups using Kruskal-Wallis tests followed by post hoc comparisons with Dunn’s test.
Figure 5.13: **Mutation signature contributions by race/ethnicity using WGS.** Boxplots of WGS signature 8 contributions for (a) HR+/HER2- and (b) HR-/HER2- malignancies. (c) The proportion of HRD signature within HR+/HER2- malignancies. Racial/ethnic differences across subtypes were assessed using Kruskal-Wallis tests followed by post hoc comparisons with Dunn's test. Within a race/ethnicity, tests across HR+/HER2- and HR-/HER2- (i.e. White in panels a and b) were performed with a Mann-Whitney U. *P* values < 0.05 are provided.
Strikingly, HR+/HER2- Nigerian tumors had higher HRD signature contributions compared to both Black ($P = 1.8 \times 10^{-4}$) and White ($P = 1.6 \times 10^{-4}$) cohorts (Fig. 5.12b). This finding was confirmed using data from WGS (Fig. 5.13c). Structural variants (SVs) are more prevalent in malignancies with HRD defects such as ovarian and basal-like breast cancers. In this same set of genomes, Nigerians had more SVs than both Black ($P = 0.03$) and White cohorts ($P = 2.8 \times 10^{-4}$). Like with the HRD signature, SVs counts in HR+/HER2- Nigerians (551 SVs per genome) were reminiscent of HR-/HER2- (approximately 626 SVs per genome) (Fig. 5.12c). Differences between Nigerians and Whites in HRD signature and SVs (both $P < 2.0 \times 10^{-3}$) extended to HER2+ cases as well (Fig. 5.12b-c). Taken together, multiple lines of evidence suggest that HR+/HER2- Nigerians have increased HRD and genomic complexity compared to the Black and White cohorts. Furthermore, genome data suggests a potentially more granular stratification by African ancestry.

I postulated that increased HRD in HR+/HER2- Nigerians may be due to an increased $TP53$ mutation rate as well as decreased rates of $PIK3CA$ and $CDH1$. Using multivariate modeling, I investigated the effect of race/ethnicity on HRD adjusting for age and missense burden as well as mutation status in $TP53$, $BRCA1/2$, $PIK3CA$, and $CDH1$. While many of these factors have significant, independent effects, they cannot entirely account for the racial/ethnic HRD disparities seen across HR+/HER2- tumors ($P < 0.05$).

### 5.2.4 The APOBEC-HRD signature balance

Numerous threads of evidence suggest a possible interplay between APOBEC and HRD signature contributions, particularly in HR+/HER2- breast cancers: 1) I identified racial/ethnic mutation rate differences in $TP53$, $CDH1$, and $PIK3CA$; 2) I found associations between these mutations and mutation signatures (Fig. 5.10a); and 3) consistent with differential mutation status, HRD activity was increased in Nigerians while APOBEC C$>$T displayed
reduced activity in Nigerians and Blacks compared to Whites (Fig. 5.11a). Furthermore, within this subtype, HRD had a notable negative correlation with both APOBEC C>T (Rho = -0.56; \( P < 1.0 \times 10^{-4} \), permutation test) and APOBEC C>G (Rho = -0.30; \( P < 1.0 \times 10^{-4} \)). Integrating these findings, I postulated that a balance of APOBEC and HRD signature contributions exists and can be discriminated — if not dictated — by mutations in \( \text{TP53, PIK3CA, CDH1, and BRCA1/2} \) (germline and somatic). For each tumor, APOBEC C>T and C>G contributions were combined and plotted them against that of HRD (Fig. 5.14a). Tumors were partitioned based on the presence of \( \text{CDH1 or PIK3CA mutations} \) (“\( \text{CDH1/PIK3CA} \)”), \( \text{TP53 or BRCA1/2 mutations} \) (“\( \text{TP53/BRCA1/BRCA2} \)”), mutations from both aforementioned categories (“Both”), or mutations in neither of the aforementioned categories (“Neither”). APOBEC contributions were significantly higher in \( \text{CDH1/PIK3CA} \) compared to the \( \text{TP53/BRCA1/BRCA2} \) \( (P = 1.8 \times 10^{-6}, \text{Dunn’s test}) \) and Neither \( (P = 7.2 \times 10^{-9}) \) groups. Tumors harboring mutations from both groups (“Both”) had lower APOBEC contributions than \( \text{CDH1/PIK3CA} \) \( (P = 0.11) \), yet higher than \( \text{TP53/BRCA1/BRCA2} \) \( (P = 5.0 \times 10^{-3}) \) (Fig. 5.14a). In contrast, \( \text{TP53/BRCA1/BRCA2} \) had significantly higher HRD contributions than all other groups \( (P \text{ CDH1/PIK3CA} = 9.9 \times 10^{-15}; \text{Both} = 1.6 \times 10^{-4}; \text{Neither} = 2.1 \times 10^{-4}) \), while \( \text{CDH1/PIK3CA} \) had significantly lower contributions than all other groups \( (P \text{ Both} = 3.5 \times 10^{-3}; \text{Neither} = 1.2 \times 10^{-3}) \) (Fig. 5.14c). These findings were similar when considering all samples simultaneously (Fig. 5.15a-c).

The signature patterns for Neither most closely resembled those of \( \text{TP53/BRCA1/BRCA2} \) (Figure 5-14 a-c), suggesting there may be other mechanisms, such as \( \text{BRCA1/2 methylation} \), that promotes increased HRD activity. When looking at the proportion of these mutational groups across HR+/HER2- samples (including those without signature estimates), the groups with the lowest APOBEC and highest HRD — \( \text{TP53/BRCA1/BRCA2} \) and Neither — encompassed 70.3% of Nigerians and 66.3% of Blacks but only 47.7% of Whites \( (P= 1.2 \times 10^{-3}, \text{Chi-squared test}) \) (Fig. 5.14d). This suggests that individuals with African ancestry
Figure 5.14: Driver genes associate with APOBEC and HRD signature balance in HR+/HER2- breast cancer. (a) For each malignancy, the proportion of APOBEC signatures (sum of APOBEC C>T and C>G) by the proportion of HRD is shown. Each patient is colored based on harboring a CDH1 or PIK3CA mutation (pink), a TP53 or BRCA1/2 (including germline) mutation (blue), mutations from both aforementioned categories (yellow), or mutations in neither of the aforementioned categories (grey). These values are decomposed into violin plots for (b) APOBEC and (c) HRD signatures, respectively. Horizontal black bars represent the median contribution proportion for each group. Between group comparisons were made using a Kruskal-Wallis test followed by Dunn’s test. (d) The proportion of HR+/HER2- individuals falling into each mutational group by race/ethnicity (n White = 465; n Black = 80; n Nigerian = 27). This also includes samples for which mutation signatures were not estimated. ** indicates groups that were significantly different ($P < 0.05$) from all three other categories.
Figure 5.15: **Driver genes associate with APOBEC and HRD signature balance across all breast cancer IHC subtypes.** (a) For each malignancy, the proportion of APOBEC signatures (sum of APOBEC C>T and C>G) by the proportion of HRD is shown. Each patient is colored based on harboring a CDH1 or PIK3CA mutation (pink), a TP53 or BRCA1/2 (including germline) mutation (blue), mutations from both aforementioned categories (yellow), or mutations in neither of the aforementioned categories (grey). These values are decomposed into violin plots for (b) APOBEC and (c) HRD signatures, respectively. Horizontal black bars represent the median contribution proportion for each group. Between group comparisons were made using a Kruskal-Wallis test followed by Dunn’s test. ** indicates groups that are significantly different ($P < 0.05$) from all three other categories.
are more likely to fall within mutational groups associated with increased HRD and lower APOBEC contributions. Consistent with this assertion, the HR+/HER2- Black cohort had greater copy number segmentation ($P = 0.022$, MWU), more structural variation ($P = 0.028$, Dunn’s test), and increased HRD in WGS ($P = 0.015$) compared to Whites (Fig. 5.12 and Fig. 5.13). Throughout African ancestry tumors, prevalent aggressive and limited favorable molecular features help explain racial/ethnic mortality disparities within the HR+/HER2-subtype [487].

5.2.5 Tumor immune microenvironment characterization

Given the enrichment of Triple Negative and HER2+ breast cancer in Nigerians, the fact that these subtypes usually present with higher levels of tumor-infiltrating lymphocytes (TILs) [488], and the relevance of these groups for checkpoint inhibition, Markus Riester and Artur Veloso investigated gene expression signatures related to immune cell infiltration (Fig. 5.16a and Table 5.9). Most immune signatures displayed statistically significant differences across PAM50 subtypes (B-cell, Cytotoxic cell, Fibroblast, IFN, Type I Interferon and Proliferation, all $P < 1.0 \times 10^{-4}$, ANOVA). Racial differences adjusted for PAM50 subtype, however, were modest (Fig. 5.16b). The Cytotoxic cell signature ($P = 4.0 \times 10^{-3}$) was lower in Nigerians in all subtypes but Basal, whereas the Fibroblast signature ($P = 0.01$) was consistently highest in Nigerians. Type I Interferon signature scores ($P = 0.01$) were enriched in Luminal subtypes for both Nigerians and Blacks, which potentially indicate that tumors from these racial groups would respond better to chemotherapy or immunotherapy [489]. Lastly, macrophage infiltration in Nigerians was highest in the Basal subtype, similar to what has been reported in some studies, including one in a small subset of Nigerian patients [490 491].

These gene signatures were tested for association with potential predictors of response to immunotherapy. In addition to mutation burden and chromosome instability (CIN), APOBEC ($C>T$ and $C>G$ combined) and HRD mutation signatures were considered in-
Figure 5.16: Gene signatures of immune cell infiltration. (a) Heatmap visualizing gene signature activation in all 1,040 patients with RNA-seq data from the combined Nigerian, Black and White cohort. High signature scores (red) indicate high overall expression of genes from a given signature, whereas low values (blue) indicate low expression. (b) Distribution of signature scores across PAM50 subtypes and ethnicities. * $P < 0.05$; ** $P < 0.001$; *** $P < 0.0001$ (all adjusted using the Benjamini-Hochberg method).
Figure 5.17: **Pairwise Pearson correlation of immune signatures and potential predictors of response to immunotherapy.** Potential predictors include APOBEC (C>T and C>G combined) and HRD signatures, chromosome instability (CIN), and mutation burden. The Nigerian data is shown in the leftmost panel and the combined Black and White cohorts in the rightmost. * P < 0.05; ** P < 0.001; *** P < 0.0001 (all adjusted using the Benjamini-Hochberg method).

Dependent mutational processes capable of generating putative neoantigens [492, 493, 494]. APOBEC contributions were positively correlated with mutation burden (Rho = 0.35, BH P < 1.0 x 10^{-4}), and, consistent with recent reports, APOBEC contributions were associated with increased T-cell infiltration (Rho = 0.25, BH P < 1.0 x 10^{-4}) [492]. Conversely, CIN positively correlated with mutation burden (Rho = 0.28, BH P < 1.0 x 10^{-4}) yet negatively correlated with T-cell infiltration (Rho = -0.08, BH P < 0.01) [493]. The same trends were observed in the Nigerian and TCGA cohorts separately with similar effect sizes (Fig. 5.17), although, in the former, most were not significant after multiple testing correction — potentially due to the smaller sample size (Fig. 5.16a).

5.3 Discussion

To date, this study is the largest genomic analysis of breast cancer among Black patients of African ancestry. The triple-negative subtype was found to be more frequent in Nigeri-
ans, which supports a previous IHC-based observation that increased HR- or triple-negative breast cancer is more common in Black women in Africa [463]. Recently ER expression was demonstrated to be a heritable trait in breast cancer [495], suggesting that genetically-influenced basal expression levels may contribute to subtype differentiation in breast cancer.

Including Nigerian samples along with TCGA allowed us to identify \textit{PLK2} and \textit{KDM6A} as novel significantly mutated genes in breast cancer, both with a propensity to be altered in HER2+ tumors. \textit{PLK2} is a cell cycle regulator and presumed tumor suppressor, while \textit{KDM6A} is a chromatin modifier frequently mutated in other cancer types (e.g. pancreatic, esophageal, and bladder) [496, 497, 498]. The former is a TP53 protein target [499] and proposed tumor suppressor gene [500] that is epigenetically silenced in various malignancies [501]. It plays key roles in mTOR signalling [502], and its loss mediates sensitivity to paclitaxel [499] and platinum [501] treatment in cell lines. It was also found to be one of the frequent outlier kinases in pancreatic cancer [503]. \textit{KDM6A}, a transcription-inducing H3K27me3 demethylase, is inactivated in large fractions of epigenetically-modified malignancies such as urothelial carcinoma [504]. It is also frequently mutated in squamous pancreatic tumors [496] and metaplastic breast cancers [505], though the latter study had a small sample size (3 of 23 samples mutated). Inactivation has not been limited to small lesions as exon-disrupting SVs were found in cervical cancer samples [506]. In female cancers, it has been reported that biallelic inactivation is common for some tumor-suppressor genes including \textit{KDM6A} [507]. In addition, \textit{KDM6A} might be required for a luminal-to-basal phenotypic switch [508]. Depletion of \textit{GPS2} has been shown to promote cell proliferation in MCF-7 breast cancer cell lines [509]. \textit{B2M} inactivation was recently reported to be a recurrent event in lung cancer and potentially affects response to anti-PD-1/anti-PD-L1 therapies [510]. For samples with additional WGS data, all mutations within novel significantly mutated genes were validated. Further studies are necessary to help characterize the role for these genes in breast cancer.
The mutational landscape and signature patterns differed across racial/ethnic populations. In particular, Nigerian patients had more TP53 and GATA3 mutations than African Americans, whereas both African ancestry groups were higher than Whites. The frequencies of prognostically favorable PIK3CA and CDH1 mutations were lower in individuals of African ancestry than in Whites. Even when restricting to HR+/HER2- breast cancer, tumors from Nigerian women were characterized by canonically aggressive molecular features, such as TP53 mutations, increased structural variation, and higher contributions from the homologous recombination deficiency mutational signature. Along with higher rates of HR negativity and HER2 positivity, aggressive HR+ tumors provide biological insight to why breast cancers in Africa are often fatal [511]. These findings also suggest Nigerians could benefit from FDA-approved genomically-tailored treatments such as HER2-targeted therapy in HER2+ and chemotherapy — specifically PARP inhibitors [512, 513, 514, 515, 516, 517] — in homologous recombination deficient tumors. For HR+ tumors, only a small subset of Nigerian women would benefit from Tamoxifen therapy alone, yet this is commonly prescribed without pathologic confirmation of HR status. Notably, outside of HER2+, these findings also suggest that underrepresentation of individuals of African ancestry in sequencing studies will further worsen disparities and access to genomic therapies currently undergoing clinical trials [518]. This reinforces the need to include diverse populations when identifying and pursuing novel therapeutic targets [478].

It is possible that genetic and environmental factors not only drive subtype differentiation, but also dictate evolutionary dynamics of a tumor. This latter assertion could help explain the observed mutational differences between racial/ethnic groups, which has also been noted comparing Black and White Americans with colorectal cancer [519]. Similarly, strong associations between driver mutations and mutation signature contributions (e.g. PIK3CA and APOBEC signatures) pose a causality dilemma suited for further biological and epidemiological investigations. Overall, these results justify future studies integrating
germline and somatic genetics as well as environmental factors in order to better understand and reduce breast cancer outcomes disparities throughout the African diaspora.

5.4 Supplementary information

5.4.1 Supplementary tables

Table 5.1: Summary statistics for WES, WGS, and RNA-seq samples. (See accompanying supplementary file).

Table 5.2: Identifiers of WES samples, their tumor subtype by IHC, and race/ethnicity. (See accompanying supplementary file).

Table 5.3: Identifiers of WGS samples, their tumor subtype by IHC, and race/ethnicity. (See accompanying supplementary file).

Table 5.4: Identifiers of RNA-seq samples, their tumor subtype by PAM50, and race/ethnicity. (See accompanying supplementary file).

Table 5.5: List of 44 driver genes mutated by short variants in breast cancer. (See accompanying supplementary file).

Table 5.6: List of 19 genes recurrently altered by CNAs in breast cancer. (See accompanying supplementary file).

Table 5.7: Summary statistics for WES and WGS samples used for mutation signature analysis. (See accompanying supplementary file).

Table 5.8: Identifiers of samples used for mutation signature analysis, and their sequencing data type. (See accompanying supplementary file).
Table 5.9: Gene sets used for immune signature analyses. (See accompanying supplementary file).
CHAPTER 6
CONCLUSION

6.1 Current and future endeavors in data-intensive genomics

SwiftSeq was developed in order to effectively leverage germline exome data from TCGA. The most important aspects of SwiftSeq that facilitated these analyses were parallelism, scalability, efficiency, and automation. The latter proved crucial in order to curtail person-hours. Prior to SwiftSeq’s development, I performed a number of RNA and DNA sequencing analyses by semi-manual job setup and submission. In addition to transient failures due to system mishaps, there was the complication of human error. Manually rummaging through thousands of log files to identify failed tasks was — and is — an untenable endeavor. Having software that facilitates execution, failure detection, and retries was necessary to ensure analysis integrity.

However, automation does not resolve all obstacles entwined with large-scale NGS data processing. A naive — or pipeline — approach wastes resources. Workflow parallelism coupled with informed task allocation save resources through multiple mechanisms. First, tasks are structured to minimize the number of spare CPU cycles. Second, task modularity makes failures far less detrimental. If a failure occurs when a pipeline is 75% complete, it must be re-run in its entirety. SwiftSeq only requires failed tasks to be repeated. Consequently, the former approach unnecessarily repeats successfully completed tasks while the latter avoids this behavior. Modularity also assists, but does not guarantee, scalability. As demonstrated with SwiftSeq, even sophisticated workflows can squander resources if not properly tuned. Similarly, workflow management systems are only as effective as the algorithms they invoke. There are a finite number of changes that can be made to how data is managed to improve scalability. If an algorithm does not effectively manage I/O, RAM, or exhibits limited multithreading, there is little a workflow manager can do to alleviate this issue. This stresses
the utmost importance of continued NGS algorithm development.

In the era of NGS, there are three overlapping epochs: 1) data generation, 2) data processing, and 3) data synthesis. Today, WES and WGS are becoming centralized. Genomics hubs such as the Garvan Institute, New York Genome Center, and many non-academic entities offer sequencing as a service. Outsourcing data processing is a logical next step. At the time of development, SwiftSeq was necessary to leverage all of the TCGA germline data. Reliable genomics software that contained integral features were nonexistent. The monetary opportunities from widespread NGS use in research and its anticipated clinical applications has attracted data science professionals. Even within academia, NCI funded projects such as the Genomic Data Commons and Cancer Cloud Pilots are attempting to centralize much of the large-scale NGS processing, especially for consortia projects like TCGA and ICGC. That is not to say that workflow development is no longer required. The Global Alliance for Genomics and Health is currently trying to understand the best ways to process data and disseminate workflows across heterogeneous systems. Regardless, unless coupled with truly significant advances in hardware or software, future development of NGS data processing workflows is best performed through a community-wide effort. Above all, this paradigm is beneficial to scientific productivity as centralization promotes higher quality data while allowing researchers to focus on discoveries.

“Cutting edge” research, by definition, is a moving target. Going forward, bioinformaticians should embrace the third epoch, which is using software, hardware, and algorithmic advances to generate efficient analytical systems. This includes algorithms to help expedite analysis interpretation as well as methods that extract novel information from heavily utilized data. Data structures that are amenable to fast queries and calculations are becoming increasingly important, especially given the NGS data generation rate [520, 521]. The primary objective of the Global Alliance for Genomics and Health is to make genomic data flow freely amongst hospitals and institutions, which has enormous implications for research and
healthcare. Voice-activated, personal assistant software can be leveraged to parse, analyze, and report genomics data with a single sentence. Cancer genomics, like any other discipline, can benefit from early adoption of revolutionary technologies. The development of these and related applications are certain to help stimulate biomedical breakthroughs.

6.2 Implications of age at diagnosis and harmful allele burden

The fact that moderate to high penetrance mutations lead to earlier cancer onset has long been known. This is one of the core findings from Knudson’s seminole retinoblastoma study [522]. The analyses presented here are the first to demonstrate that higher harmful allele burden in cancer predisposition genes associates with earlier age at diagnosis across heterogeneous malignancies. As a result, I suggest that “genomic age” plays an important role in age at diagnosis. It has often been said that everyone who lives long enough will develop cancer [523]. As somatic mutations accumulate within cells, the affected genomes gradually become less robust and are susceptible to addition lesions. This provides a mechanistic explanation as to why cancer disproportionately affects the elderly [524]. Having multiple germline mutations in critical DNA repair genes likely fast-forwards the time to cancer onset since fewer somatic driver mutations are required for tumorigenesis. Accordingly, it is as if these genomes were “older” and more fragile from the outset. Of course, not all of these germline alleles have equivalent effects. As DNA sequencing continues and repositories grow, modelling interactions between deficiencies in predisposition genes becomes a more viable strategy. By looking at extreme phenotypes in breast cancer, I identified two women who were diagnosed very young — 26 years of age. Each harbored mutations in the well-known breast cancer susceptibility genes BRCA2 and ATM [525]. Each woman also had one or more mutations in other DNA damage repair genes. Continued analyses of individuals with substantially younger diagnosis may reveal sets of genes that are abnormally detrimental when mutated in unison.
Many targeted gene panels have oligos designed to pull-down many — if not all — of the ClinVar cancer genes and autosomal dominant predisposition genes used throughout this study. This presents an opportunity to replicate my findings in a much larger cohort. Couch and colleagues recently performed gene panel sequencing on over 60,000 breast cancer patients \[526\]. Although this panel contained only 21 genes, it exemplifies the near-term scalability of this approach. Furthermore, gene panels could be combined with common genotypes derived from SNP arrays. Jointly modelling all GWAS susceptibility loci can produce polygenic risk scores (PRS), and this technique has been applied to a variety of cancer types \[527\]. Combining common variant PRS with the burden of harmful, rare/low-frequency alleles is a logical next step for cancer risk prediction \[528\]. Overall, for the majority of the population, harmful allele burden may not provide actionable information. However, in breast cancer, the effect of high allele burden was striking. Those with increased breast cancer risk have estrogen therapy and mastectomy as prophylactic treatment options. My observations, together with these clinical implications, justify rigorous follow-up studies to ascertain the effect of harmful allele burden in breast cancer.

### 6.3 Two-hit genes in cancer risk, development, and progression

Through a multi-omics approach, I was able to confirm known and identify novel two-hit tumor suppressor genes. Although some of these genes (e.g. \(PHLPP2\)) are proposed tumor suppressors, the germline variants they harbor have not been implicated in tumor development. Rare variant association studies could determine if mutations in these genes confer risk. A few years ago, studies like this would seem impractical since harmonized variant calls from a large cohort would be required. The Exome Aggregation Consortium (ExAC) was specifically formed to collate and uniformly genotype WES data housed within dbGaP. At present, ExAC version 0.4 consists of an unprecedented 100,304 samples, many of which are from cancer stricken individuals \[529\]. Once sufficiently mined, datasets like these will
contribute substantially to our understanding of cancer genetic architecture. This includes any roles my two-hit genes play in tumor initiation and progression. Lastly, this two-hit detection approach may even have clinical relevance. Variants of unknown significance plague clinical sequencing panels [530]. Even variants occurring in genes with well-characterized risk can be difficult to interpret [531]. Two-hit status could be used as supporting evidence for pathogenic variant classification [532]. Patients and clinicians are forced to make difficult decisions based on genetic testing results; therefore, every reliable piece of evidence should be considered.

ROBO1 was also identified as putative tumor suppressor gene in breast cancer. Preliminary data suggests that ROBO1 two-hits may preferentially occur in ER- tumors. To ascertain the merits of this finding, MCF-7 cells — derived from an ER+ breast cancer — were transfected with siRNA targeting ROBO1 and then treated with estrogen to assess common ER target activation. Decreased gene expression for ER target genes GREB1 and TFF1 was observed both basally and after estrogen treatment. Furthermore — using proliferation assays under the same conditions — estrogen-induced proliferation of MCF-7 cells was negated by ROBO1 knockdown. Subsequent experiments will attempt to repeat and potentially expand upon these results. If these findings hold, two-hit bias for ER- tumors may be explained by ROBO1 loss negatively affecting estrogen signaling pathways. Since ER-breast cancer has greater genomic complexity than its ER+ counterpart, these findings are also consistent with ROBO1 knockdown increasing DNA damage. It must be stressed that further experimentation is necessary before making conclusions about ROBO1 deficiency and ER status. Nonetheless, it does provide an intriguing angle for continued inquiry.

Beyond disease susceptibility, the role of germline variation in the evolutionary history of malignancies has largely been ignored. Mounting evidence — including two-hit genes — suggests that inherited gene deficiencies and genetic background can influence somatic alterations [533] [532]. By amalgamating germline and somatic genetics, future explorations
will provide a comprehensive picture of the genomic vulnerabilities exploited during cancer initiation. This space invites questions that have large implications for cancer development and progression. For example, when ancestral cells undergo malignant, clonal expansion, what genomic lesions – including germline – are present? Aggregating this information across samples will elucidate the minimal pathway disruptions required for tumorigenesis.

6.4 Understanding racial/ethnic disparities in breast cancer

Women of African ancestry (Black) are much more likely to develop aggressive triple negative breast cancer than those of European ancestry (White). Disparate biology, environmental factors, and the combination thereof are all assumed to contribute [534]. When investigating all IHC subtypes, White women were much more likely to have features — such as PIK3CA mutations — consistent with less aggressive disease [535, 536]. On the other hand, Black individuals were more likely to have triple negative-like characteristics such as TP53 mutations and higher contributions from the HRD mutation signature [537, 538]. It is possible that Black individuals develop more HRD-enriched tumors due to genetic factors. BRCA1 deficiencies are strongly associated with both HRD and the triple negative subtype [539]. Due to conflicting findings, it is unclear if harmful germline variants in BRCA1 are more prevalent in Black women, so additional studies are required [540, 541]. Importantly, only a small subset of breast cancer risk alleles identified in White women replicate in Black cohorts [542]. Moving forward, comprehensive integration of germline and somatic data could help uncover sources of breast cancer disparities.

Although there were clear relationships between somatic PIK3CA mutations and increased APOBEC contributions, it remains unclear if these mutations are a cause or consequence of APOBEC activity. Henderson and colleagues postulated that PIK3CA mutations could arise directly from APOBEC-mediated mutagenesis [543]. They noted that two hotspot amino acid (AA) substitutions — E542K and E545K — occur as a result of TC>[T/G]W
substitutions, which are APOBEC-related. They subsequently showed that head and neck carcinomas positive for human papillomavirus had high TC>[T/G]W mutational burden, and these malignancies were enriched for E542K/E545K PIK3CA substitutions. Within the combined TCGA and Nigerian breast cancer cohort, preliminary analyses indicate that PIK3CA mutated HR+ carcinomas harbor an excess of E542K/E545K AA substitutions compared to their HR- counterparts (P = 0.052, Fisher’s exact). APOBEC signatures (C>T and C>G) were significantly more operative in samples with E542K/E545K substitutions compared to both PIK3CA WT tumors (BH P = 5.2 x 10^{-14}, Dunn’s test) and those with other (non-E542K/E545K) PIK3CA mutations (BH P = 9.2 x 10^{-6}). These results, combined with previous findings, point to a possible causality dilemma between PIK3CA and APOBEC activity. With respect to racial/ethnic differences, a germline APOBEC deletion that is common in Asian populations confers increased breast cancer risk and APOBEC signature activity [544]. Additionally, the APOBEC3B enzyme dictates APOBEC-mediated hypermutation, and its expression could be influenced by genetic and environmental factors [545]. Further genomic and molecular characterization of this PIK3CA-APOBEC association could increase our understanding of subtype incidence disparities.
CHAPTER 7
MATERIALS AND METHODS

7.1 Processing blood germline exomes

Blood germline exomes from The Cancer Genome Atlas (TCGA) were downloaded in bam format from CGhub (https://browser.cghub.ucsc.edu/) as approved by the Database of Genotypes and Phenotypes (dbGaP). Informed consent for all patients was acquired originally by TCGA. Files were assessed for their integrity using MD5 sums. All exomes were processed using the SwiftSeq workflow framework. SwiftSeq, which uses the Swift parallel scripting language [546] to run and manage tasks, was written to provide scalable DNA sequence analyses on clouds and high-performance computing machines. Within SwiftSeq, each file was split by read group using Samtools (v1.2) [547]. Read group bams were converted to single or paired end fastq data using bamUtil (v1.0.13). Each read group was aligned independently to GRCh37 (GATK data bundle version 2.8) with BWA-MEM (v0.7.12) and coordinate sorted with Novosort (v1.00.01). Aligned, sorted, read group bams were consolidated via Novosort and split into contig bams with bamUtil. Each contig bam had duplicate reads removed using Picard Tools (v1.119) Mark Duplicates utility. The subsequent bams were single-sample genotyped in 10 million bp windows using Platypus (v0.7.9.1) [548]. Variants, both single nucleotide variants and indels, were filtered based on seven distinct and empirically validated criteria. Notably, in a recent study using a gold standard set of indels, Platypus outperformed other popular tools in nearly all indel calling categories [549]. Exonic region variants (as defined by Broad.human.exome.b37.interval_list from GATK data bundle version 2.8) were annotated with Variant Effect Predictor (VEP) (v79) [550]. The number of mapped reads was determined with Samtools flagstat. The exact commands and parameters invoked at each step are as follows:

Single-end read group extraction and alignment
Paired-end read group extraction and alignment

```
samtools view -b -r READ_GROUP_NAME INPUT_BAM 2>> LOG_FILE |
/path/to/bamutil/bin/bam bam2FastQ --in -.bam --noeof
--firstOut /dev/null --secondOut /dev/null --unpairedOut FASTQ
2>> LOG_FILE

bwa mem -M -t 32 -R "READ_GROUP" /path/to/human_g1k_v37.fasta FASTQ 2>>
LOG_FILE | samtools view -b - 2>> LOG_FILE | novosort --threads 8
--ram 31000M --tmpcompression 6 --tmpdir TMP_DIR --output OUTPUT_BAM
--index - 2>> LOG_FILE
```

Read group merge and contig split

```
novosort --threads 32 --ram 62000M --tmpcompression 6 --tmpdir TMP_DIR
INPUT_BAMS 2>> LOG_FILE | /path/to/bamutil/bin/bam splitChromosome --in
-.bam --out *.CONTIG. --noeof 2>> LOG_FILE
```

Duplicate removal
java -XX:+UseParallelGC -XX:ParallelGCThreads=1 -Xmx5166m -jar MarkDuplicates.jar INPUT=INPUT_BAM OUTPUT=OUTPUT_BAM METRICS_FILE=METRICS TMP_DIR=TMP_DIR REMOVE_DUPLICATES=true >> LOG_FILE 2>&1

Genotyping
python Platypus.py callVariants --nCPU 2 --output OUTPUT_VCF --refFile /path/to/human_g1k_v37.fasta --regions COORDINATES --bamFiles INPUT_BAM >> LOG_FILE 2>&1

Sort vcf
perl sortByRef.pl INPUT_VCF /path/to/human_g1k_v37.fasta.fai 2>> LOG_FILE >> OUTPUT_VCF

Annotation
perl variant_effect_predictor.pl -i INPUT_VCF -o OUTPUT_VCF --dir_cache=/path/to/.vep --port 3337 --offline --symbol --vcf --plugin ExAC,/path/to/ExAC.r0.3.sites.vep.vcf.gz --plugin CADD,/path/to/whole_genome_SNVs.tsv.gz --fork 30

Merge contig bams
novosort --threads 32 --ram 62000M --tmpcompression 6 --tmpdir TMP_DIR --output OUTPUT_BAM --index INPUT_BAMS >> LOG_FILE 2>&1

Alignment metrics
samtools flagstat INPUT_BAM 2>> LOG_FILE > OUT_METRICS
Of the 9,451 bams processed, 8,268 belonged to a unique individual. For individuals with multiple bams, the sample with the greatest number of mapped reads was used downstream.

7.2 Allele-specific copy number analysis in tumors

Copy number changes across TCGA breast cancer tumors were called using the ASCAT algorithm \[551\]. Initially, Affymetrix SNP 6.0 CEL files provided by TCGA Data Portal, for both malignant and normal tissue, were processed using PennCNV \[552\] to obtain logR and BAF data. Since samples profiled via SNP arrays are prone to wave artifacts, the logR was subsequently corrected for GC content. Copy number profiles for all tumor samples were inferred using the ASCAT computational framework (version 2.4.2) from the BAF and corrected LogR data. For racial/ethnic comparison analyses, only copy number profiles for samples that had matching exome data were used for analysis.

7.3 ClinVar variants and genes

The August 4th, 2015 ClinVar \[553\] vcf was downloaded. Any variant record containing the case-insensitive string “cancer” within the “CLNDBN” field was extracted. This captured desired annotations such as “*cancer_susceptibility”, “Familial_cancer*”, “cancer*familial”, and “Hereditary_cancer-predisposing_syndrome”. A small fraction of the remaining vcf records (109 of 16,339) whose “CLNDBN” field did not contain the words “susceptibility”, “familial”, or “heritary” were removed. The results were annotated using VEP (details above) and all non-coding, non-autosomal, and synonymous variants were discarded. At a given locus, each alternative allele was considered a distinct variant. If only a single pathogenicity assertion was present for multiple alleles, that assertion was assumed to represent both alleles. If an allele had multiple submitters, if any submitter labeled it as “Pathogenic” or “Likely Pathogenic” it was considered as such. If mapping alleles to a
pathogenicity was ambiguous (e.g. three alternative alleles present with only two pathogenicity assertions) the locus was discarded to avoid misclassification.

From this list of annotated cancer alleles, any gene that had “Pathogenic” or “Likely Pathogenic” assertion with the “CLNSIG” field for any phenotype was carried downstream. I chose to include any phenotype since manual inspection noted some variants were labeled “Pathogenic” for a cancer predisposing disease (e.g. Gardner Syndrome), yet labeled “Uncertain significance” for the more general “Hereditary_cancer-predisposing_syndrome” (e.g. rs137854567). Similarly, cancer-associated alleles not explicitly labeled as “Pathogenic” or “Likely pathogenic” for cancer were retained if considered “Unknown significance”, “Not provided”, or “Other” since they are pathogenic in other contexts. Since these would likely be returned as variants of unknown significance during clinical cancer screening, it was important that they were included. Non-silent variants (excluding those affecting splice donor/acceptor sites) and variants not falling within a CCDS region (release 17) were removed. Any variant that passed the filters outlined above, had an Exome Aggregation Consortium (ExAC) frequency < 0.05, and had a cohort allele frequency < 0.05 was considered a “cancer-associated ClinVar variant” for during analyses. This set of variants fell within 57 genes (Table 3.1), which were subsequently referred to as ClinVar Cancer Genes (CCGs). Pathogenic and likely pathogenic variants within non-cancer phenotypes were determined using the same methodology above, except any cancer-associated variant was removed. Similarly, cancer-associated variants that lacked any pathogenic or likely pathogenic assertion (i.e. annotated as ‘Benign’, ‘Likely benign’, ‘Uncertain significance’, ‘not provided’, etc.) and met filtering criteria above (Allele frequency < 0.05, within CCDS regions, non-silent, etc.) were considered non-pathogenic.
7.4 Classifying deleterious variants in exomes

To expand beyond ClinVar, I included CCG variants likely to be deleterious. Variants were classified as deleterious if they had an ExAC allele frequency < 0.05, an allele frequency < 0.05 using the calls from Platypus (across all unique individuals), and if they met either of the following criteria: 1) predicted to have a “HIGH” functional impact (i.e. frameshift, stop gain, splice donor/acceptor, etc.) based on VEP annotation; a 2) a missense variant identified with Combined Annotation Dependent Depletion (CADD) score ≥ 25. These criteria were also used to classify deleterious variants across all genes. For analyses presented in Chapter V, splice donor/acceptor variants were not included as deleterious. All complex variants (multi-nucleotide substitutions and multiple variants on the same haplotype) were broken into allelic primitives and annotated with VEP. Only a very small fraction of these variants passed deleterious variant criteria (data not shown), and no complex variants contained cancer-associated ClinVar alleles. Therefore, due to their minimal contribution and interpretation difficulties, complex variants were removed from further analysis unless predicted to have a “HIGH” functional impact. Notably, based on their annotations, variants could be considered both cancer-associated via ClinVar and deleterious. While deleterious variants were identified on sex chromosomes, they were not included in subsequent analyses. A very small fraction (0.0007%) of deleterious variants had ≥ 2x more homozygous alternative calls than heterozygous calls and were removed. 55 individuals were removed for having a low number of deleterious loci (< 70). Three individuals that had a substantial amount of deleterious variation (> 5 standard deviations from the mean) were removed, leaving 8,210 unique samples for analysis, 8,111 with reported age at diagnosis.
7.5 Age at diagnosis and allele burden associations

Clinical data for were downloaded from the TCGA Data Matrix, and age of onset and self-reported race were extracted for each individual. Associations between age of onset and cancer-associated (ClinVar), deleterious and ClinVar/deleterious allele burdens were determined using the following linear regression models (lm() function in R), where i represents an individual:

Unadjusted:

\[ y_i = \beta_0 + \beta_{\text{burden}} X_{\text{burden}_i} + \epsilon \]

Race adjusted:

\[ y_i = \beta_0 + \beta_{\text{burden}} X_{\text{burden}_i} + \beta_{\text{race}} X_{\text{race}_i} + \epsilon \]

Cancer type adjusted:

\[ y_i = \beta_0 + \beta_{\text{burden}} X_{\text{burden}_i} + \beta_{\text{type}} X_{\text{type}_i} + \epsilon \]

Race and cancer type adjusted:

\[ y_i = \beta_0 + \beta_{\text{burden}} X_{\text{burden}_i} + \beta_{\text{race}} X_{\text{race}_i} + \beta_{\text{type}} X_{\text{type}_i} + \epsilon \]

For models containing race and cancer type, each was represented by a distinct term. The 95% confidence interval was reported for all regression coefficient estimates. The following two models were used when jointly assessing the effects of high burden (Four or more ClinVar/deleterious alleles in the union of CCGs and ADGs), \( BRCA1 \) status, and \( BRCA2 \) status on age at diagnosis in breast cancer:

Unadjusted:

\[ y_i = \beta_0 + \beta_{BRCA1} X_{BRCA1_i} + \beta_{BRCA2} X_{BRCA2_i} + \beta_{\text{HighBurden}} X_{\text{HighBurden}_i} + \epsilon \]
Race adjusted:

\[ y_i = \beta_0 + \beta_{BRCA1} \times BRCA1_i + \beta_{BRCA2} \times BRCA2_i + \beta_{HighBurden} \times HighBurden_i + \beta_{race} \times race_i + \epsilon \]

High burden, BRCA1 status, and BRCA2 status were coded as binary categorical variables.

### 7.6 Synchronous/bilateral clinical data extraction

All TCGA clinical data were downloaded from Firehose (https://gdac.broadinstitute.org/). Across all samples,

- “patient.tumor_samples.tumor_sample.other_dx”
- “patient.tumor_samples.tumor_sample-2.other_dx”

fields were extracted. Possible values for these fields were: 1) “yes, history of prior malignancy”, 2) “yes, history of synchronous/bilateral malignancy”, 3) “both history of synchronous/bilateral and prior malignancy”, 4) “no” and 5) “NA”. Any sample labeled “yes, history of synchronous/bilateral malignancy” or “both history of synchronous/bilateral and prior malignancy” was considered to be a synchronous/bilateral malignancy while “yes, history of prior malignancy” and “no” were not. “NA” was treated as missing data. Overall, of the 8,210 samples, 194 had synchronous/bilateral malignancy, 6,645 did not, and 1,371 lacked an assertion for this field.

### 7.7 One- versus two-hit assessment

In samples containing a ClinVar/deleterious variant of interest, that genomic region was assessed for LOH. For any sample that displayed LOH, the allele fraction of the ClinVar/deleterious variant was extracted from the corresponding tumor exome using pysam (https://github.com/pysam-developers/pysam). Consequently, any sample with LOH and a
variant allele fraction > 50% was considered to have biallelic (two) hits (i.e. the WT allele is lost and the harmful allele retained).

7.8 Compiling high and moderate risk genes

The list (Table 3.3) of high and moderate risk genes was borrowed from previous work by Slavin and colleagues [556]. Scanning the literature, they compiled genes that had generalized risk ratios > 2 in either breast, ovarian, or colorectal cancer. This was not intended to be a comprehensive set of all high and moderate risk genes. However, since it contained well-characterized risk genes from three prominent and highly studied cancer types, it provided a reasonable approximation. Notably, this set of genes overlaps substantially with another set of “high” and “moderate” risk genes that have been implemented for pan-cancer clinical evaluation [557].

7.9 ExAC allele counts

Two formulations of the ExAC database (Version 0.3) were downloaded. One, referred to as “Total”, contained variation called across all individuals, including those from TCGA. Another, “Non-TCGA”, contained only variant calls from samples that are not in TCGA. Overall, this resulted in 7,601 TCGA and 53,105 Non-TCGA individuals. I re-annotated both databases and called deleterious variants using the same criteria and methods outlined above. 798,334 deleterious, autosomal variants were identified in ExAC with 4,420 and 3,278 falling within CCGs and ADGs, respectively. Individual-specific ExAC calls were not available; however, allele counts (AC) and allele number (AN) were available for both database versions. Using Finnish and non-Finnish Europeans, the AC and AN for all deleterious variants were obtained from Total and Non-TCGA. For each allele, TCGA AC and AN were inferred through the following calculations:
\[ AC_{TCGA_i} = AC_{Total_i} - AC_{Non-TCGA_i} \]

\[ AN_{TCGA_i} = AN_{Total_i} - AN_{Non-TCGA_i} \]

There were instances of incongruent data between Total and Non-TCGA databases. 1) A small fraction of variants were only seen in Total and not Non-TCGA. This implies the variant was only seen within the TCGA cohort. Therefore, Non-TCGA AC and AN values could not be directly obtained for above calculation. In these instances, AC and AN were determined by the following:

\[ AC_{TCGA_i} = AC_{Total_i} \]

\[ AN_{TCGA_i} = \max(AN_{TCGA}) \]

\[ AN_{Non-TCGA_i} = AN_{Total_i} - \max(AN_{TCGA}) \]

This made the conservative assumption that nearly all TCGA samples were callable at this locus, which made it less likely to reject the null hypothesis. 2) Some variants were present in Non-TCGA, but yet were not present in Total. In these cases, the variant was discarded.

### 7.10 ExAC simulations

Using the AN and AC values from above, two independent sets of 10,000 simulations were performed. First, random variants (CCG n = 4,420; ADG n = 3,278) were selected from the 847,303 deleterious alleles identified in ExAC. Second, gene sets were randomly selected (CCG n = 57; ADG n = 60) and aggregated all deleterious variants harbored by those genes. Enrichment was determined via odds ratios (ORs). For alleles of interest (where i represents an allele), each empirical and simulated ORs was calculated by:
\[ Odds_{TCGA} = \sum_i AC_{TCGA_i}/\left(\sum_i AN_{TCGA_i} \sum_i AC_{TCGA_i}\right) \]
\[ Odds_{Non-TCGA} = \sum_i AC_{Non-TCGA_i}/\left(\sum_i AN_{Non-TCGA_i} \sum_i AC_{Non-TCGA_i}\right) \]
\[ OR = Odds_{TCGA}/Odds_{Non-TCGA} \]

Note that the AC and AN counts were summed prior to calculating the OR. Therefore, each OR is calculated using the total AC and AN across all deleterious variants, which effectively assesses if deleterious alleles from a set of variants are enriched in TCGA samples compared to non-TCGA. \( P \) values were calculated by comparing the distribution of simulated ORs to the empirical ORs.

### 7.11 Gene ontology and pathway enrichment

Gene lists of interest were assessed using the STRING database [558]. Gene ontology (biological process, molecular function, and cellular component) and KEGG pathway were assessed using STRING’s native framework. A false discovery rate \( < 0.05 \) was considered significant.

### 7.12 Extracting genes from genome-wide association study hits

The entire catalog (https://www.ebi.ac.uk/gwas/) of GWAS associations (gwas_catalog_v1.0-associations_e88_r2017-04-24.tsv) was downloaded. Any SNP association that had “cancer”, “tumor”, or “carcinoma” in the “DISEASE/TRAIT” field was extracted. Any trait not associated with risk (e.g. outcome, drug response, etc.) was removed. For any remaining association that was genome-wide significant (\( P \leq 5 \times 10^{-8} \)), all genes present in the “MAPPED_GENE” field were extracted. 362 unique genes were present in the CCDS database (Release 17).
7.13 Significance testing for age at diagnosis

All linear regression $P$ values throughout the study are one-sided unless specified otherwise. Age at diagnosis comparisons between two groups were performed using Welch two-sample T-tests (one-sided). Allele burden comparisons between two groups were performed with Wilcoxon rank sum tests (one-sided). The 95% confidence interval was reported for all regression coefficient estimates. When being summarized (e.g. Figure 3.2), age at diagnosis was reported using the mean and corresponding 95% confidence interval. All statistical calculations were performed in R (version 3.3.2).

7.14 Cell culture and RNA interference

MCF-7 (HTB-22), MRC-5 (CCL-171), and MCF10A (CRL-10317) cells were acquired from ATCC. MCF-7 and MRC-5 cells were grown in DMEM with 10% FBS, sodium pyruvate, glutamax, and antibiotic/antimycotic. MCF10A cells were grown in DMEM/F12 with 10% horse serum, EGF, hydrocortisone, insulin, and cholera toxin. siRNAs were ordered from ThermoFisher and transfected into cells using Lipofectamine RNAiMax. Experimental media for MCF-7 cells as well as proliferation assay incubation media for MCF10A and MRC-5 was phenol-red free DMEM + 5% charcoal-stripped PBS, sodium pyruvate glutamax, and antibiotic/antimycotic.

7.15 Quantitative PCR

Following 96hr siRNA knockdown, RNA was extracted from cells using Trizol (ThermoFisher) and Directzol RNA Miniprep Kit (Zymo Research). RNA was converted to cDNA using M-MulV reverse transcriptase (NEB). qPCR was performed using iTaq Universal SYBR Green Supermix (BioRad) on a StepOne Plus qPCR machine. Primers for qPCR were designed using Primer3 (http://bioinfo.ut.ee/ primer3-0.4.0/ primer3/) and ordered from IDT.
7.16 Proliferation assays

After 96hr siRNA knockdown, MCF10A or MRC-5 cell proliferation was quantified using the Vybrant MTT Cell Proliferation Assay kit (ThermoFisher). Briefly, cells were incubated with MTT in incubation media for 3 hours, followed by addition of SDS-HCl solution for 3 hours. The resulting mixture was pipetted and readings were taken on a NanoDrop at 570nm.

7.17 Scratch assays

MCF10A or MRC-5 cells were transfected with siRNA in a 24-well plate for 48hrs. After 48hrs a scratch was made down the center of each well with a 1ml pipette tip. Cells were then allowed to grow for 48hrs in fresh media. Cells were then fixed in 4% paraformaldehyde in PBS +calcium/magnesium for 20min. Cells were then stained with 1% crystal violet in 2% ethanol (Sigma-Aldrich) for 30min. Cells were then washed with PBS 3X until only blue cells remained. Three images were taken along the scratch in each well with a Zeiss Axiovert 40 light microscope.

Scratch area for each image was quantified using ImageJ. Briefly, color images were changed to 8-bit gray images, then edges found and images sharpened. Images were thresholded to produce a black as white images and fill holes was applied, this created a black and white image with the scratch in white and the cell layer in black. After inverting the lookup table (making the scratch area black), analyze particles was used to determine the area of the scratch.

7.18 DNA damage response assays

MCF10A or MRC-5 cells were plated in a 96-well optical glass-bottomed plate in siRNA transfection mix. After 48hrs cells were treated with 1ug/ml cisplatin in growth media for
Following cisplatin incubation, the cisplatin was removed and growth media was added for 24 hrs. Cells were then fixed with 4% paraformaldehyde in PBS + calcium/magnesium for 20 min, quenched in 100 mM NH4Cl for 10 minutes, then fixed in 0.5% Triton-X 100 for 15 min. Cells were then blocked in 5% powdered milk in TBS-T (blotto) for 1 hr. A rabbit primary antibody to γH2AX (Bethyl Laboratories A300-081A) and blotto were incubated on the cells overnight at 4°C. Following washes, cells were incubated with Donkey Anti-Rabbit 488 (ABCAM ab150073) for 1 hr at room temperature. Cells were then fixed with 4% paraformaldehyde in PBS, then quenched in NH4Cl, as stated previously. Finally, cells were incubated with DAPI for 2 minutes. Plates were imaged and images were analyzed for DAPI and γH2AX staining within single cells. DNA damage was calculated as the γH2AX signal divided by the DAPI signal to correct for cells that have gone through S phase and have the potential to display twice as much γH2AX.

7.19 Patient recruitment, biospecimen collection, and pathological assessment

This study was embedded within the Nigerian Breast Cancer Study and approved by the Institutional Review Board of all participating institutions. Patient ascertainment and details of the study have been previously published [559, 560, 561, 562]. In collaboration with Novartis, study was extended to LASUTH. A grand total of 493 subjects were recruited (284 from UCH and 109 from LASUTH) between February 2013 and September 2015. Each patient gave written, informed consent prior to participation of the study. 27 mastectomy tissues were preserved in RNAlater. Six biopsy cores and peripheral blood were collected from each patient. Two biopsy cores were used for routine formalin fixation for clinical diagnosis, and the remaining four cores were preserved in PAXgene Tissue containers (QIAGEN, CA) for subsequent genomic material extraction. Complete pathology assessment was done central by study pathologists. Tumor burden was assessed based on cellularity, histology.
type, and morphological quality of tissue using TCGA best practices and only tissues containing 60% or more tumor cellularity were used for WGS. For WES, tissues containing 30% or more tumor cellularity were used. IHC on ER, PR and HER2 were performed centrally in Nigeria and further reviewed in the US. Cases with discordant results were reviewed by the study pathologists. IHC scoring variables for Allred scoring algorithm were captured according to the 2013 ASCO/CAP standard reporting guidelines. Briefly, for ER and PR testing, immunoreactive tumor cells < 1% was recorded as negative, and those with 1% were reported positive. All the positive ER and PR cases were graded in percentages stained cells, and further scored in line with the Allred scoring system. Percentage of tumor staining for HER2 test were also reported along with a score of 0 and 1+ as negative, 2+ as equivocal, and 3+ as positive case. In addition, genomic copy number calls of HER2 and chromosome 17 ploidy were used as alternative to HER2 fluorescent in situ hybridization (FISH) test.

7.20 Sample selection and genomic material extraction

Breast tumors were selected for sequencing following the TCGA guidelines [563]. Tumor samples containing > 60% tumor cellularity were selected for DNA extraction using PAXgene Tissue DNA kit (QIAGEN, CA). Gentra Puregene Blood Kit (QIAGEN, CA) was used to extract genomic DNA from blood. Extracted DNA were quality controlled for its purity, quantity, integrity. Identity of the extracted DNA were tested using AmpFlSTR Identifiler PCR Amplification Kit (Thermo Fisher Scientific). Samples that match > 80% of the short tandem repeat (STR) profiles between tumor and germline DNA were considered authentic. RNA was extracted from PAXgene fixed tissues using the PAXgene Tissue RNA kit (QIAGEN, CA). RNA integrity (RIN) was determined for all samples by the RIN score given by the TapeStation (Agilent) read out. RNA samples that had RIN scores of 4 and above were included in downstream sequencing analysis.
7.21 Next-generation sequencing data generation

WES and RNA-seq were carried out at the Novartis Next Generation Diagnostics facility. Exome enrichment was performed on libraries (prepared by Illumina TruSeq Nano DNA Library Prep Kit) passing QC using Agilent SureSelect XT Human All Exon V4 baits and SureSelectXT capture enrichment reagents. Passing captured libraries are combined in equimolar pools with other captured libraries of compatible adapter barcodes. These pools were normalized with concentration and were sequenced on the Illumina HiSeq 2500 sequencer. Tumor samples had an average coverage depth of 139x (63-265x), normals 52x (19-205x). WGS was performed at the University of Chicago High-throughput Genome Analysis Core (HGAC) and at the New York Genome Center (NYGC). Libraries were prepared using the Illumina Truseq DNA PCR-free Library Preparation Kit. Libraries were sequenced on an Illumina HiSeq 2000 sequencer at HGAC using 2 x 100bp paired end format, and HiSeq X sequencer (v2.5 chemistry) at NYGC using 2 x 150 bp cycles. Mean coverage depth tumor was at 98.5x and normal was at 34.2x. For RNA-seq, total RNA were constructed into poly-A selected Illumina-compatible cDNA libraries using the Illumina TruSeq RNA Sample Prep kit. Passing cDNA libraries were combined in equimolar pools with other libraries of compatible adapter barcodes and later sequenced on the Illumina HiSeq 2500 sequencer. Average number of mapped reads per sample was 97 million (ranging from 36 to 232 million).

7.22 Tumor-normal pair DNA sequence alignment

For both exomes and genomes, reads were aligned to GRCh37 from GATK data bundle version 2.8 using BWA-MEM (v0.7.12). Duplicate reads were removed using PicardTools MarkDuplicates (v1.119). Using a custom Fluidigm SNP panel, it was confirmed that whole-exome BAM files matched the library DNA, to identify sample swaps in the sequencing lab or bioinformatics pipelines.
7.23 Calling somatic single nucleotide variants

SNVs were called using both MuTect (v1.1.7) \[564\] and Strelka (v1.0.13) \[565\] with default parameters except Strelka’s depth filter was not used for exomes (isSkipDepthFilters = 1). Variants were called on the entirety of the genome in order to detect and retain any high-quality off-target calls. Any variant call that did not meet ‘PASS’ criteria for either algorithm was discarded. For a given tumor-normal pair, only SNVs called by both MuTect and Strelka were retained. Furthermore, using 1,088 blood germline exomes (959 TCGA BRCA; 129 Nigerian), a panel of normal samples was constructed. For a given normal sample, a site needed to be covered by a minimum of 10 reads to be included. Any SNV that was supported by 5\% or more of reads (MAPQ 20; Base quality 20) in two or more samples was removed. SNVs were later annotated with Oncotator \[566\], and those ones that meet the criteria (“COSMIC\_n\_overlapping\_mutation > 1” AND “1000gp3\_AF ≤ 0.005” AND “ExAC\_AF ≤ 0.005”) were considered likely to be somatic and were retained. This panel of normal process was also repeated for genomes (normal sample n = 124). All subsequent SNV calls were annotated by Variant Effect Predictor (VEP) (v79) \[567\].

7.24 Calling somatic insertions and deletions

Small indels were called using Scalpel (v0.5.3) \[568, 569\] in ‘somatic’ mode. Variants were only called in known genic regions as defined by Broad.human.exome.b37.interval.bed from the GATK data bundle version 2.8. To minimize the number of false positive calls, the ‘twopass’ option was employed. Default Scalpel filters were implemented which required a minimum alternative allele count of four in the tumor, no alternative allele present in the normal, and a minimum tumor variant allele frequency of 5\%. Additionally, indel calls located in repetitive genomic regions (via DustMasker \[570\]) or found in the 1000 Genomes Project Phase 3 release \[571\] were removed. Finally, a pseudo panel of normals was generated
by aggregating all putative indel calls that failed Scalpel filters due to ‘HighVafNormal’ or ‘HighAltCountNormal’. Any indel that failed in two or more samples was filtered. The remaining calls were annotated using VEP.

7.25 Calling copy number alterations in exomes

Allele-specific copy number in whole-exome data was called using PureCN 1.5.45 [572]. Alternative purity and ploidy solutions were considered. Genes were called amplified if the median exon copy number was 6 or higher for focal gains (< 3 Mb), or 7 or higher for non-focal gains. Genes with median exon copy number of 0 were called lost. Non-focal amplifications of tumor suppressor genes [573] were excluded. Since Affymetrix Genome-Wide Human SNP Array 6.0 data was available for the TCGA cohort, copy number calling was performed using ASCAT. Amplifications and deletions were called exactly as in the exome data. GISTIC 2.0.22 [574] was used to identify significantly gained or lost genomic regions in the Nigerian cohort. TCGA GISTIC2 results were obtained from the BROAD FireBrowse portal (doi:10.7908/C1NP23RQ). Chromosomal instability (CIN) was defined as the fraction of the genome with copy number alteration.

The required coverages of targeted exons and 200kb off-target bins were calculated and GC-bias corrected using PureCN. To maximize the number of heterozygous SNPs informative for allele-specific copy number estimation, all variants in the 50 base pair flanking regions of targets were included. Position-specific mapping bias estimates of known germline SNPs were obtained by providing PureCN a variant call format (VCF) file containing variants present in five or more of the normal samples. Since accurate copy number calling is notoriously difficult in low purity exome data, the PureCN 1.7.16 [572] calls were compared with two other recently published tools, FACETS 0.5.6 [575] and Sequenza 2.1.2 [576]. Cases for which the PureCN estimates of tumor ploidy or purity differed by 0.5 or 0.1, respectively, from the median of the estimates from the 3 tools were manually curated. PureCN further
flags samples for manual curation and all flagged samples were curated. All tools were run as recommended in the corresponding documentations. In total, 3.9% of samples showed discordant purity estimates. PureCN is the only tool that includes somatic point mutations in the purity estimation and in half of the discordant samples, PureCN correctly identified a very low tumor contribution as revealed by low allelic fractions of somatic SNVs. Ploidy was as expected more discordant, with 7.8% of samples showing a difference in ploidy of 0.5 or higher. Discordant samples had low purity (average 37% versus 47% for concordant samples, two-sided t-test $P = 0.05$). For 2% of samples the ploidy estimate was changed. The remaining samples were either of too low purity or quality to confidently call ploidy (i.e. genome doubling yes versus no) or the PureCN estimates were more plausible.

### 7.26 Calling structural variants in genomes

SVs (deletions, duplications, and inversions) were called with both Delly (v0.6.1) and Lumpy Express (v0.2.10). A panel of normal samples was constructed by taking all Delly SVs calls made in at least one (n = 124) normal sample, regardless of ‘PASS’ or ‘LowQual’ in the FILTER field. Any SV found within the panel of normals was removed from the analysis. All Delly SVs passing the aforementioned filters were queried within the matched Lumpy calls. Delly SVs corroborated by a Lumpy call (same SV type and breakpoints within 500 bp [up or downstream]) were retained. These consensus SVs were filtered if a breakpoint (from either Delly or Lumpy) fell within a repetitive genomic region according to DustMasker. Lastly, inversions were required to have split read evidence (at least one read) from both Delly and Lumpy.
7.27 Estimating genetic ancestry of the study population

The ancestry of breast cancer patients from TCGA was estimated using principal component analysis as practiced by TCGA Analysis Working Group [579]. According to the estimated proportion of ancestry, patients were grouped into genomic Black (≥ 50% African ancestry), genomic White (≥ 90% European ancestry), and genomic Asian (≥ 90% Asian ancestry). All Nigerian patients were assumed to be 100% African with little to no admixture with other populations [580].

7.28 Significantly mutated genes

To detect significantly mutated genes MutSigCV (v1.4) was used [581, 582]. SNV and indel VCFs from 1,164 individuals were annotated with Oncotator [566] using the oncotator_v1_ds_Jan262014 database. MutSigCV was then invoked with default parameters on the Oncotator generated MAF file. To reduce common false positives, only a single non-silent indel within a given gene per sample was allowed. Finally, for any gene to be called significantly mutated, it needed to have more than two individuals harboring non-silent mutations across the entire dataset.

7.29 Mutation signatures in exomes and genomes

The Bioconductor package SomaticSignatures [583] was used to estimate somatic mutational signatures. The ability to reliably call mutation signatures depends on sufficient numbers of mutations. To this point, all high-quality exome SNVs were used, regardless if they are coding or non-coding. Any sample containing at least 100 SNVs was included for downstream assessment. Additionally, in order to stimulate more accurate signature estimates, 122 WGS tumor-normal pairs were also included in addition to 500 WES pairs. To account for variable mutation counts across samples, SomaticSignatures was used to normalize the mutation
matrix prior to performing non-negative matrix factorization. Nine signatures were estimated
(Figure 5.5) since that was consistent with the number of signatures identified previously in
breast cancer, and nine signatures explained approximately 99% of variance when using 122
genomes alone. Using matrix algebra on the resulting exposure and mutation matrices, the
relative contributions of the nine signatures on each sample were calculated. “Contributions”
represent the proportion of mutations assigned to given mutation signature within each
tumor. This distinction is important as high APOBEC contributions, for example, do not
necessarily imply APOBEC hypermutation.

59 individuals had both somatic WES and WGS data. For all nine signatures, the
correlation of contributions between exomes and genomes were examined. Signatures A
(Rho = 0.87), B (0.92), C (0.69), H (0.86), and I (0.75) all exhibited strong correlation (Rho
≈ 0.7) (Figure 5.7). APOBEC C>T (Signature A) and C>G (Signature B) contributions were
highly correlated (Rho = 0.65; \( P < 2.2 \times 10^{-16} \)). Contributions from the aging signature
(Signature C) were also positively correlated with age at diagnosis (Rho = 0.18; \( P = 7.3 \times 10^{-5} \)).

7.30 Comparison with reported mutation signatures

The mutation signature matrices were compared with the 30 previously reported signatures
downloaded from the Catalog of Somatic Mutation in Cancer (COSMiC), containing previ-
ously identified 30 signatures operative across a variety of cancer types. Kullback-Leibler
Divergence \(^584\) was used to compare derived signatures to those from COSMiC. With this
approach, the most representative COSMiC mutation signature for each derived signature
was identified. Signatures A (COSMiC signature 2; APOBEC C>T), B (COSMiC signature
13; APOBEC C>G), C (COSMiC signature 1; Aging), H (COSMiC signature 3; homologous
recombination deficiency [HRD]), and I (COSMiC signature 8; Unknown etiology) all closely
matched to signatures known to be operative in breast cancer (Figure 5.6).
7.31 Mutation signature correlation permutations

For each individual, two signature contribution values were randomly selected without replacement. These values were assigned to dummy signature 1 and dummy signature 2, respectively. Dummy signatures were subsequently correlated using Spearman’s method. This process was repeated 10,000 times to construct a null Rho distribution. Permutated $P$ values were calculated by comparing empirical Rho values to the null distribution.

7.32 RNA-seq analysis, PAM50 classification, and immune signatures

Gene expression measurements were uniformly calculated using Omicsoft ArraySuite software [585] for Nigerian and TCGA samples. The RNA sequencing reads passing quality control were aligned to the Human B37 genome. Read counts for the UCSC gene models were calculated by the software. The gene counts were upper quartile normalized with the edgeR Bioconductor/R package [586] and batch normalized using ComBat as implemented in the sva package [587]. Transcripts per million (TPM) expression values were calculated based on the normalized counts. PAM50 classification was carried out using the pbcM package [588] using the “robust” parameter. To characterize the immune and stromal microenvironment of these tumors, the expression of several pre-specified sets of immune and stromal cell gene expression markers were assessed. Gene signature scores were calculated using the GSVA R/Bioconductor package [588, 589].

7.33 Testing for associations amongst PAM50 subtypes and race

ANOVA and median regression with bootstrapping standard errors was used to test for association of immune signature scores with PAM50 subtypes and race. The median regression is robust to the skewness of the distribution in immune signatures. A model was fitted
with interaction terms for race and subtypes and none of the interactions were significant. Therefore, the analysis using all subjects provided more power and could be justified over subgroup analyses. In addition, PCA analysis showed that subtype, not race was the most significant source of variance.

7.34 GISTIC analysis

GISTIC2 was run with parameters -ta 0.3 -td -0.3 -conf 0.9 -broad 1 -brlen 0.5. 0.3 approximately matched the average log-ratio standard deviation in normal samples. Any GISTIC peaks for which the majority of exons displayed a high variance in tumor versus normal coverage log-ratios in the pool of normal samples were excluded. The affected deletion peaks in large repetitive regions were 2q32, 4q13.3, 7q21.11, 10p11.22 and Xq22.3. Furthermore, only homozygous deletions were counted (i.e. single copy losses were not included in Figure 5.2).
References


135


143


Kwong Wai Choy, Chi Pui Pang, Christopher B O Yu, Hing Lok Wong, Joan S K Ng, Dorothy S P Fan, Kwok Wai Lo, Joshua T Y Chai, Jianhua Wang, Weiling Fu, and Dennis S C Lam. Loss of heterozygosity and mutations are the major mechanisms of RB1 gene inactivation in chinese with sporadic retinoblastoma. *Hum. Mutat.*, 20(5):408, November 2002.


J L Hopper, M C Southey, G S Dite, D J Jolley, G G Giles, M R McCredie, D F Easton, and D J Venter. Population-based estimate of the average age-specific cumulative risk of breast cancer for a defined set of protein-truncating mutations in BRCA1 and


[174] Jingyao Dai, Jian Gu, Maosheng Huang, Cathy Eng, E Scott Kopetz, Lee M Ellis,


[226] Nasim Mavaddat, Daniel Barrowdale, Irene L Andrulis, Susan M Domchek, Diana Eccles, Heli Nevanlinna, Susan J Ramus, Amanda Spurdle, Mark Robson, Mark...


[229] Brett Wallden, James Storhoff, Torsten Nielsen, Naem Dowidar, Carl Schaper, Sean Ferree, Shuzhen Liu, Samuel Leung, Gary Geiss, Jacqueline Snider, Tammi Vickery, Sherri R Davies, Elaine R Mardis, Michael Gnant, Ivana Sestak, Matthew J Ellis,


169


Jessica X Chong, Kati J Buckingham, Shalini N Jhangiani, Corinne Boehm, Nara Sobreira, Joshua D Smith, Tanya M Harrell, Margaret J McMillin, Wojciech Wiszniewski, Tomasz Gambin, Zeynep H Coban Akdemir, Kimberly Doheny, Alan F Scott, Dimitri Avramopoulos, Aravinda Chakravarti, Julie Hoover-Fong, Debra Mathews, P Dane

179


182


[420] Xin Li, Payton D Stevens, Jianyu Liu, Haihua Yang, Wei Wang, Chi Wang, Zheng Zeng, Micheal D Schmidt, Mike Yang, Eun Y Lee, and Tianyan Gao. PHLPP is a negative regulator of RAF1, which reduces colorectal cancer cell motility and prevents tumor progression in mice. *Gastroenterology*, 146(5):1301–12.e1–10, May 2014.


Sohrab P Shah, Andrew Roth, Rodrigo Goya, Arusha Oluomi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarrui Ding, Kane Tse, Gholamreza Haffari, Ali Bashashati, Leah M Prentice, Jaswinder Khattra, Angela Burleigh, Damian Yap, Virginie Bernard, Andrew McPherson, Karey Shumansky, Anamaria Crisan, Ryan Giuliany, Alirea


[486] Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Siewu-


[530] C Sue Richards, Sherri Bale, Daniel B Bellissimo, Soma Das, Wayne W Grody, Mad-


[554] 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,


[560] D Huo, C A Adebamowo, T O Ogundiran, E E Akang, O Campbell, A Adenipekun, S Cummings, J Fackenthal, F Ademuyiwa, H Ahsan, and O I Olopade. Parity and


206


Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D
