THE UNIVERSITY OF CHICAGO

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

List of Tables [1]

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

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 through loss-of-heterozygosity (LOH). By jointly analyzing LOH and deleterious, germline variants across 5,146 individuals, I found that the classic tumor suppressors *BRCA1*, *BRCA2*, and *ATM* showed highest, pan-cancer enrichment for two-hit scenarios. Two other genes – *PHLPP2* and *KDELC2* – also had a preponderance of two-hits. Performing siRNA knockdowns in multiple cells lines, Mike Bolt showed that reducing *PHLPP2* and *KDELC2* expression promotes the cancer-like phenotypes proliferation and migration. Furthermore, malignancy-specific investigations provided strong computational and experimental evidence that *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 *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 [3, 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 *APC* mutation. A subsequent expansion occurs with *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 *PIK3CA* [38], V600E in *BRAF* [39], amino acid residues 12 and 13 in *KRAS* [40], and PEST domain mutations in *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 *ERCC2* in bladder cancer [39, 43].

The categorization of some driver genes is more nuanced. In breast cancer, *GATA3* accumulates small insertions and deletions (indels) in its terminal exons that are seemingly loss-of-function [44]. However, *GATA3* mutant tumors have substantially higher *GATA3* expression than those that are wild type (WT), suggesting that these indels promote *GATA3* expression and should be considered activating [45, 46]. Some classifications can even be

tissue specific. *NOTCH1* behaves as an oncogene in acute lymphoblastic leukemia [41] and, paradoxically, like a tumor suppressor in squamous cell carcinomas [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 [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 [56]. In both research and clinical settings, each technique has advantages and drawbacks that have been discussed extensively elsewhere [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 [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) [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 10.6×10^{-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 *BRCA1/2*, 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, *BRCA1* mutation carriers have a striking 65% chance breast cancer will manifest before 70 years of age. Those with mutated copies of *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]. BRCA1 and *BRCA2* mutations confer increased risk for other cancer types as well; both genes are associated with substantially higher ovarian cancer incidence while males harboring *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. 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 ATM's substrates, 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 *BRCA1/2* sequencing to identify possible mutations [198, 199]. If families were negative for mutations in either of these genes, *ATM*, *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 mutational 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; however, 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)

[243]. In fact, they are 40-70% more likely than Whites to present with stage IV breast cancer, regardless of the underlying IHC subtype [244]. Even when adjusting for age, young Black women (< 40 years of age) have a mortality rate double that of young White women [243]. More adverse breast cancer mortality rates persist in older Black women as well [245]. 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 [246]. Not only are these Women beset with aggressive subtypes, but also those with prognostically favorable HR+ tumors still encounter lower survivability [247].

Socioeconomic factors such as income, diet, obesity, lack of exercise, and access to health care are prime candidates to explain these disparities [248]. Weight gain has been shown to explain explain approximately 16% of postmenopausal cancer risk [249]. Conversely, disease frequency drops in individuals who partake in rigorous daily exercise [250]. Studies have suggested that access to breast cancer screening techniques can reduce mortality rates [251, 252, 253, 254]. Although evidence indicates mammography usage can normalize racial/ethnic differences in tumor stage, Black women still present with higher grade tumors [255]. Black women continued to have more aggressive histological features even after adjusting for numerous clinical and socioeconomic factors [256]. So while they undoubtedly contribute, non-biological variables alone cannot sufficiently account for the magnitude of disparities amongst races/ethnicities [249, 257, 258, 259].

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).

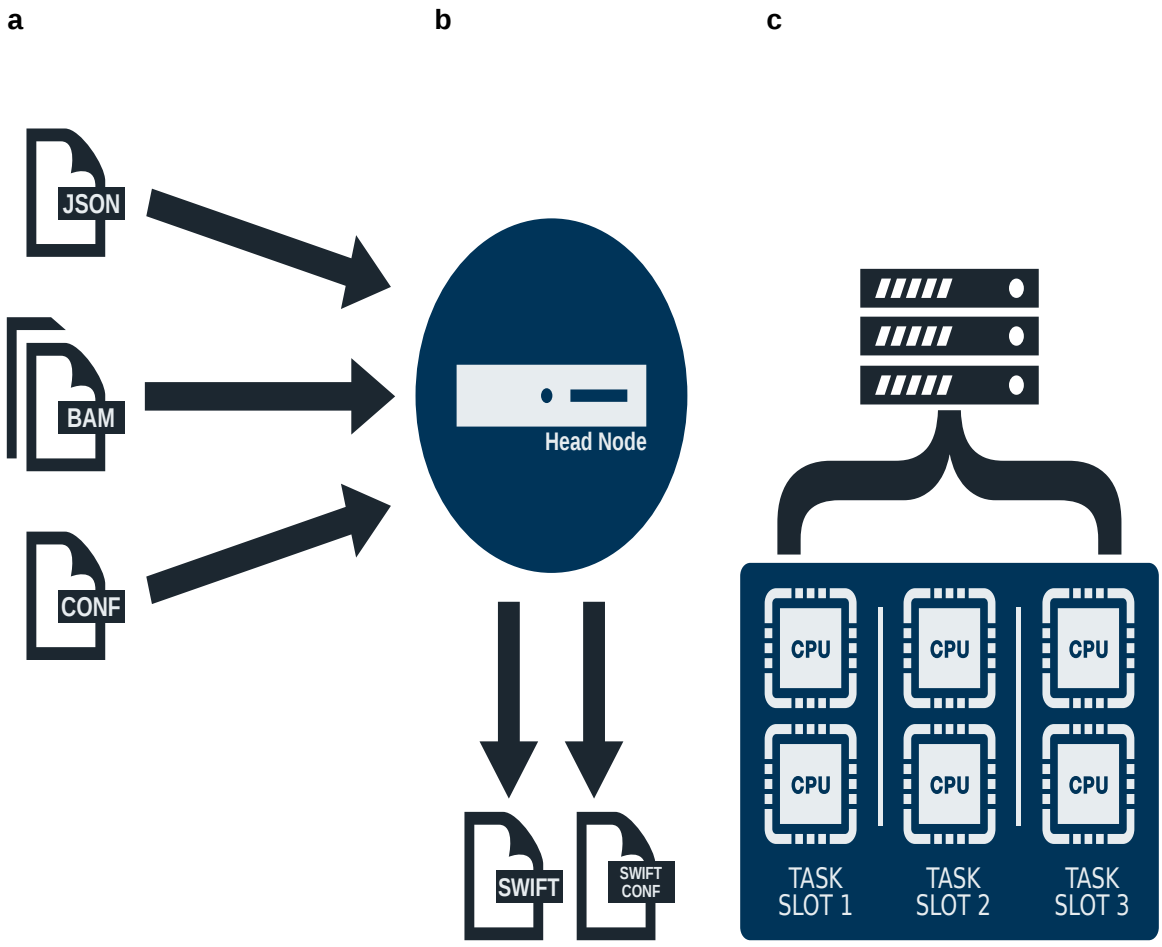

Figure 2.1: **Diagram of a SwiftSeq run.** (a) Input files required for a SwiftSeq run. (b) Initiation step where the Swift script, Swift configuration files, and bash wrappers are generated based on input provided in panel a. (c) The execution step where tasks are distributed to workers that run in 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 bioinfomatics 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 [281]. 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 formated 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-

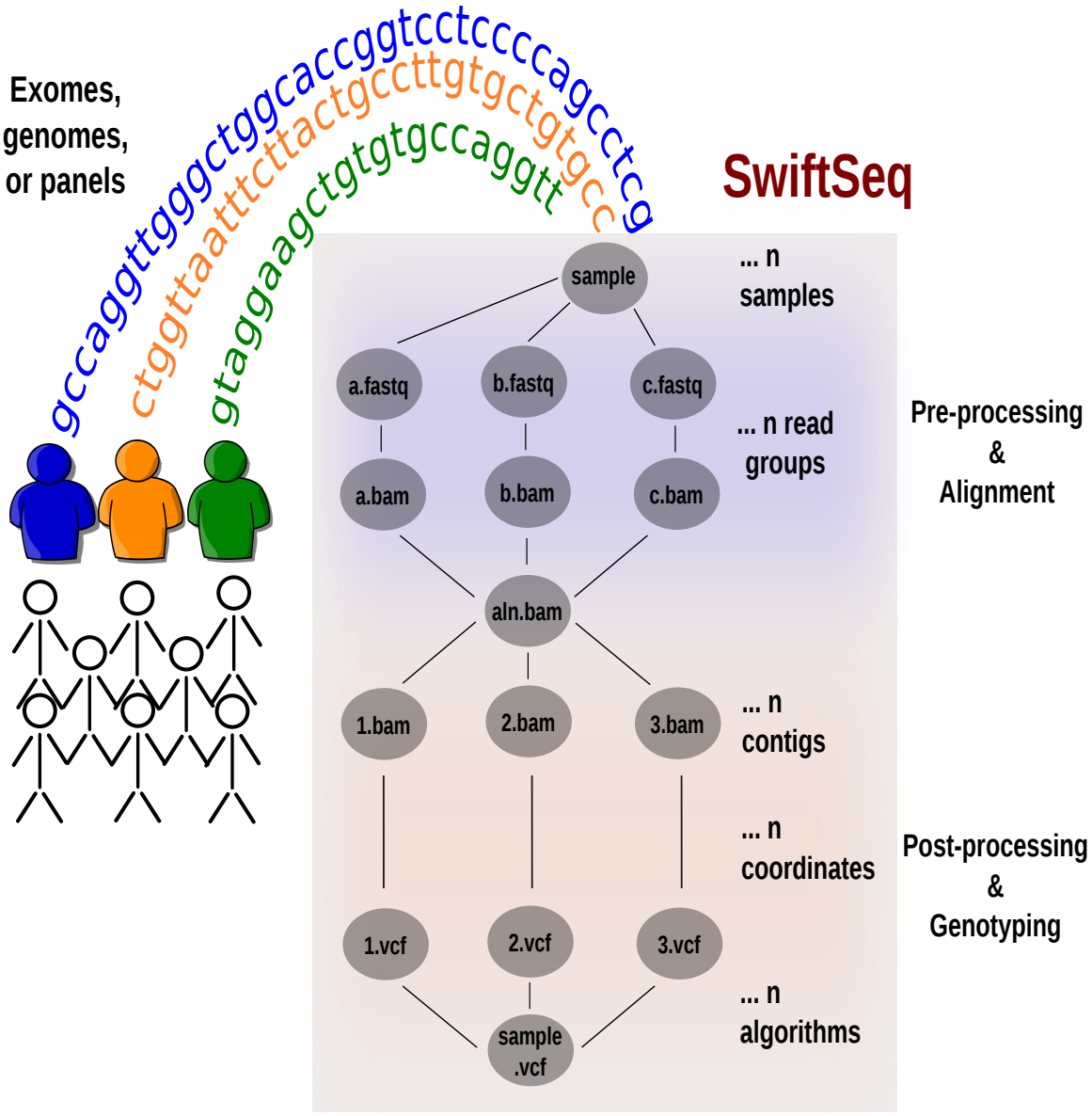

Figure 2.2: **SwiftSeq parallelization strategies.** Five levels of fork-join parallelism (samples, read groups, contigs, genomic coordinates, and algorithms) that are implemented throughout the workflow to maximize resources and decrease runtime.

imately 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 bioinfomatics 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

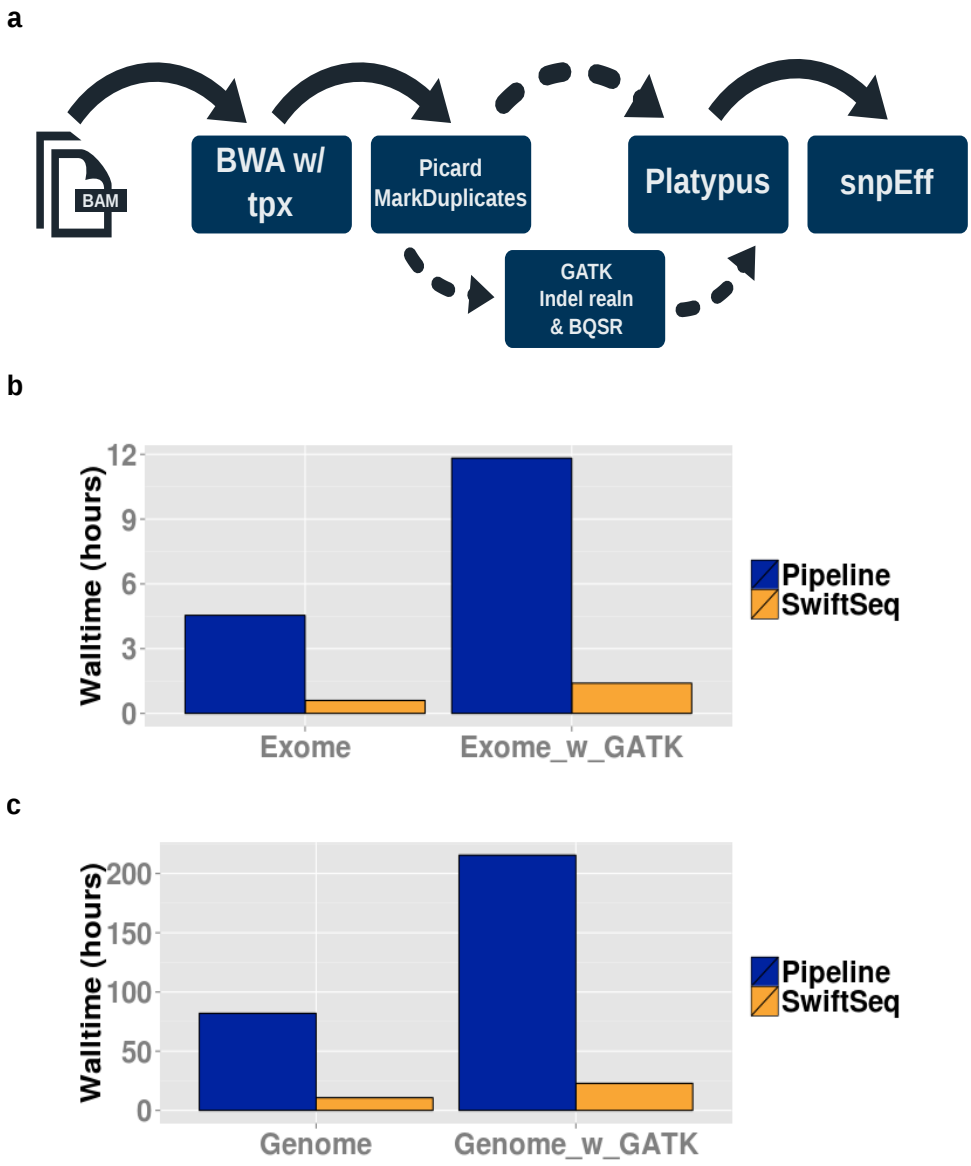

Figure 2.3: **SwiftSeq processing speed compared to standard pipeline approaches.** (a) The software workflow used to compare the two approaches with and without Genome Analysis Toolkit (GATK) bam cleaning steps. Walltime comparisons between SwiftSeq and pipeline approaches using an (b) exome and (c) 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

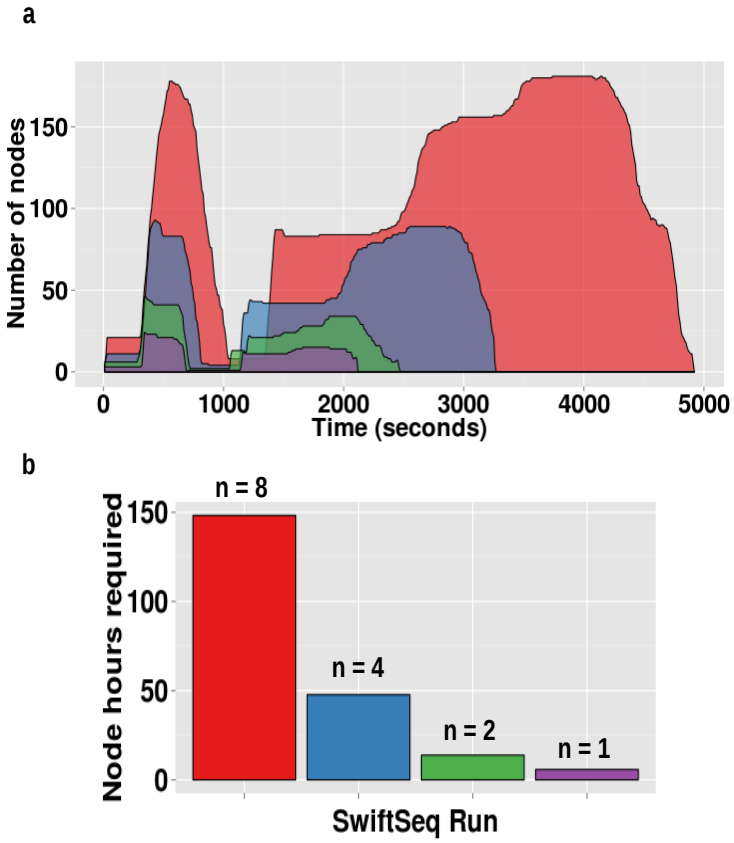

Figure 2.4: **Naive exome scaling tests with SwiftSeq** Using the workflow depicted in Fig. 2.3a without Genome Analysis Toolkit (GATK) bam cleaning steps, we tested the scalability of SwiftSeq by performing runs with 1, 2, 4, and 8 copies of the same exome sample. (a) The number of nodes utilized at 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 Chicagos *Beagle* supercomputer (Cray XE6). Each worker node contained two AMD Opteron 6100 series processors (24 cores total) and 32 GB of RAM.

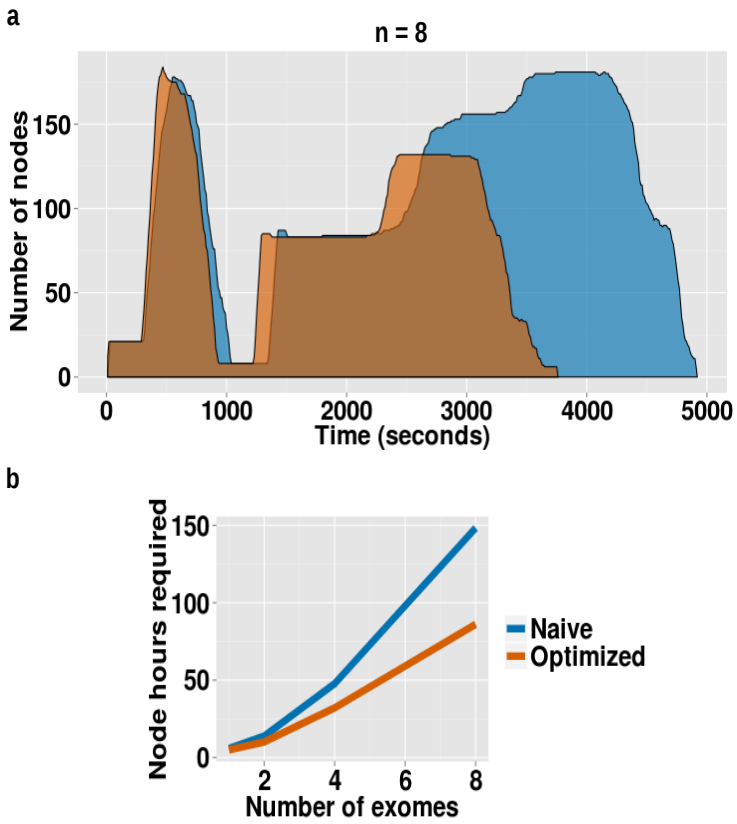

Figure 2.5: **Comparing optimized and naive exome scaling tests.** Using the workflow depicted in Fig. 2.3a without Genome Analysis Toolkit (GATK) bam cleaning steps, we compared the performance of a naive approach (blue) to that of an optimized approach (orange) after throttling I/O intensive processes. (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 processes 1, 2, 4, and 8 sample exome runs.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.

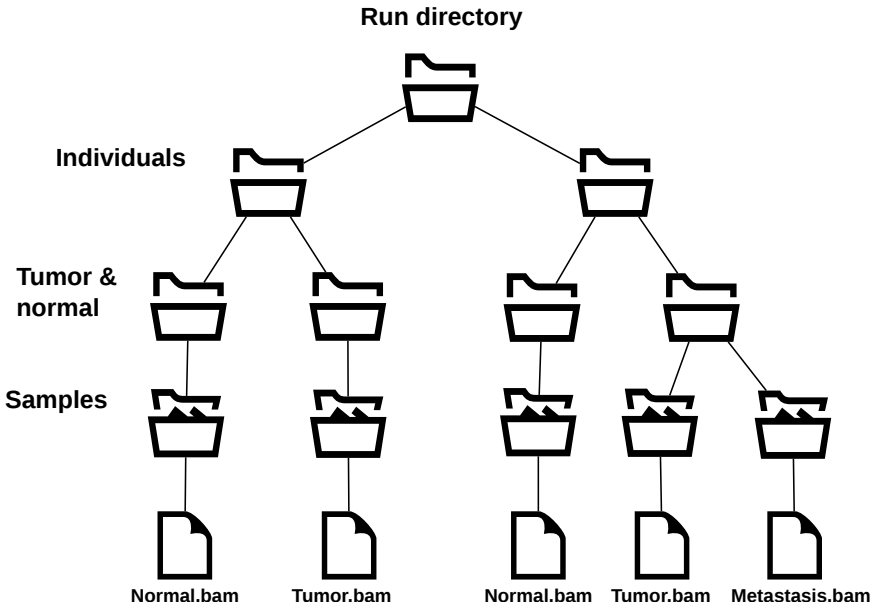

Figure 2.6: **Tumor-normal pair directory structure.** Depicted above is the required directory structure for tumor-normal pairs analyses performed by SwiftSeq. Each sample directory must only contain a single bam file. For a given individual, each sample within the tumor directory will be compared to 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 [310, 311, 312, 313, 314], primary versus metastasis comparisons [315, 316, 317, 318, 319], and "double normal" (i.e. blood and tumor-adjacent tissue sequenced) analyses [320]. Furthermore, multiple studies have shown that somatic variant calling accuracy can be improved by integrating results from multiple algorithms [321, 322, 323, 324]. As such, SwiftSeq allows users to select numerous callers [325, 326, 327] 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 parameters offers significant utility. To deliver this functionality, Dominic Fitzgerald and I developed 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 genotyper(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 resource 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 computers (e.g. multi-core laptops and desktops) and commodity hardware, ranging from low-

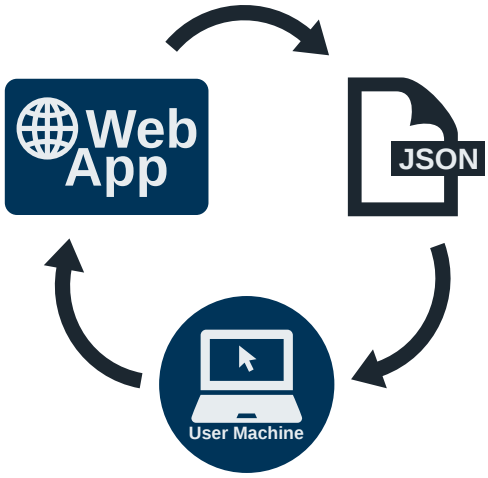

Figure 2.7: **Graphical user interface scheme for designing and retrieving workflows.**

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 filesytems, 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-

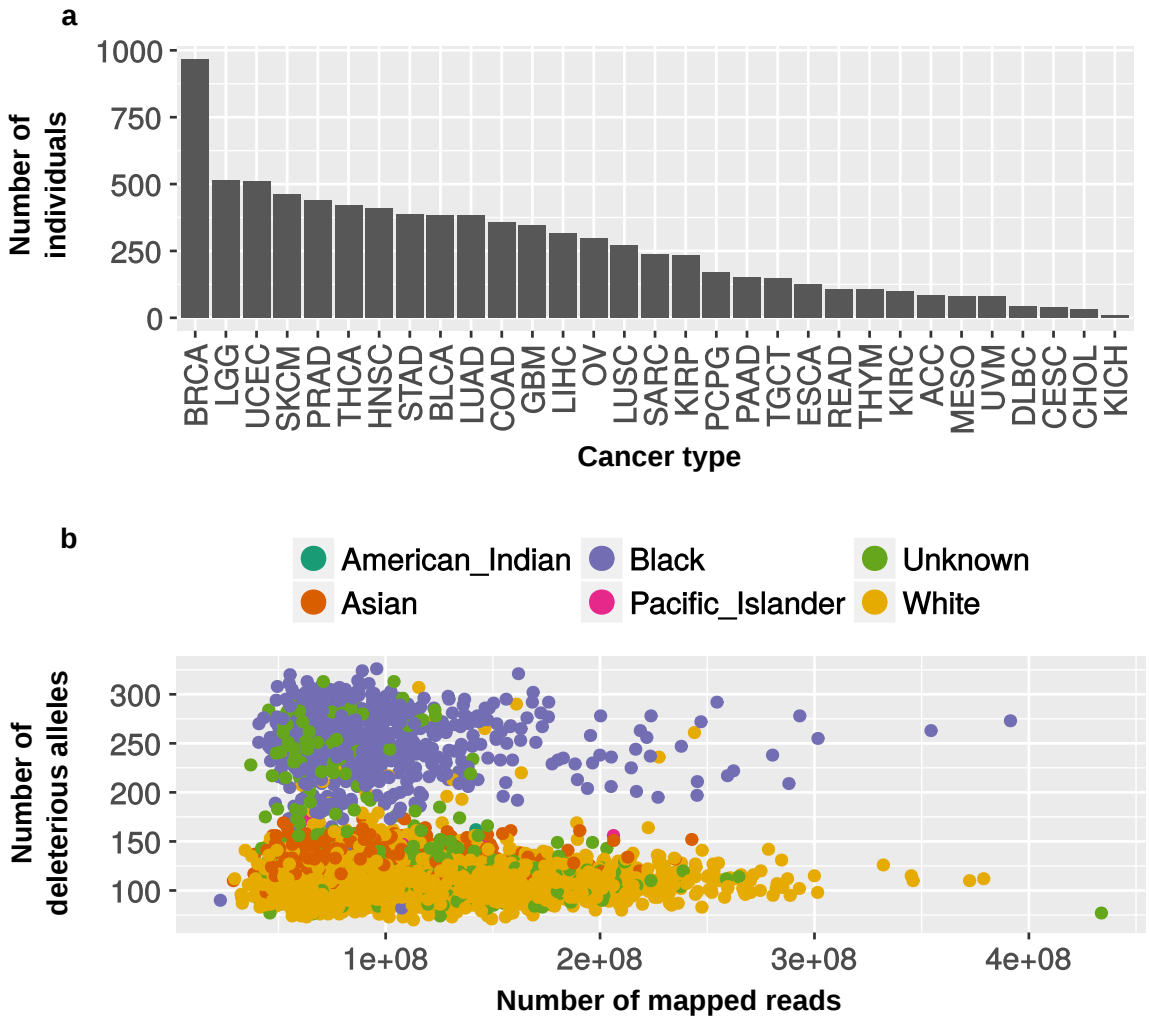

Figure 3.1: **Samples per cancer type and deleterious allele counts per individual.** (a) The number of individuals representing each of the 31 cancer types included from TCGA. (b) Mapped reads versus the number of called deleterious, autosomal alleles. Each individual is colored by self-reported race.

ALDH2	*CDKN2A*	*MITF*	*PTEN*	*SMAD4*
APC	*CHEK2*	*MLH1*	*RAD50*	*STK11*
ATM	*CYP17A1*	*MLH3*	*RAD51*	*TGFBR2*
ATR	*ELAC2*	*MRE11A*	*RAD51C*	*TMEM127*
BARD1	*EPHB2*	*MSH2*	*RAD51D*	*TP53*
BLM	*FAM175A*	*MSH6*	*RET*	*TSC1*
BMPR1A	*FANCC*	*MUTYH*	*RNASEL*	*TSC2*
BRCA1	*FH*	*NBN*	*SDHA*	*VHL*
BRCA2	*FLCN*	*NF1*	*SDHAF2*	*XRCC2*
BRIP1	*MAX*	*PALB2*	*SDHB*	
CDH1	*MEN1*	*PLA2G2A*	*SDHC*	
CDK4	*MET*	*PMS2*	*SDHD*	

Table 3.1: **List of 57 ClinVar Cancer Genes.**

ALK	*CDKN2A*	*MSH6*	*RB1*	*SMARCB1*	*CBL*
APC	*CEBPA*	*NF1*	*RET*	*STK11*	*HRAS*
BAP1	*DICER1*	*NF2*	*RUNX1*	*SUFU*	*KRAS*
BMPR1A	*EPCAM*	*PALB2*	*SDHA*	*TMEM127*	*MAP2K1*
BRCA1	*FH*	*PAX5*	*SDHAF2*	*TP53*	*MAP2K2*
BRCA2	*GATA2*	*PHOX2B*	*SDHB*	*TSC1*	*NRAS*
CDC73	*MAX*	*PMS2*	*SDHC*	*TSC2*	*PTPN11*
CDH1	*MEN1*	*PRKAR1A*	*SDHD*	*VHL*	*RAF1*
CDK4	*MLH1*	*PTCH1*	*SMAD4*	*WT1*	*SHOC2*
CDKN1C	*MSH2*	*PTEN*	*SMARCA4*	*BRAF*	*SOS1*

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

d). For all combinations of variant and gene sets (Figs. 3.2a-e and 3.3a-d), the significant effect of allele burden remained after adjusting for self-reported race and cancer type, both independently and simultaneously (Supplementary Tables 3.6 and 3.7).

3.2.2 Orthogonal support from seven control analyses

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). 3)

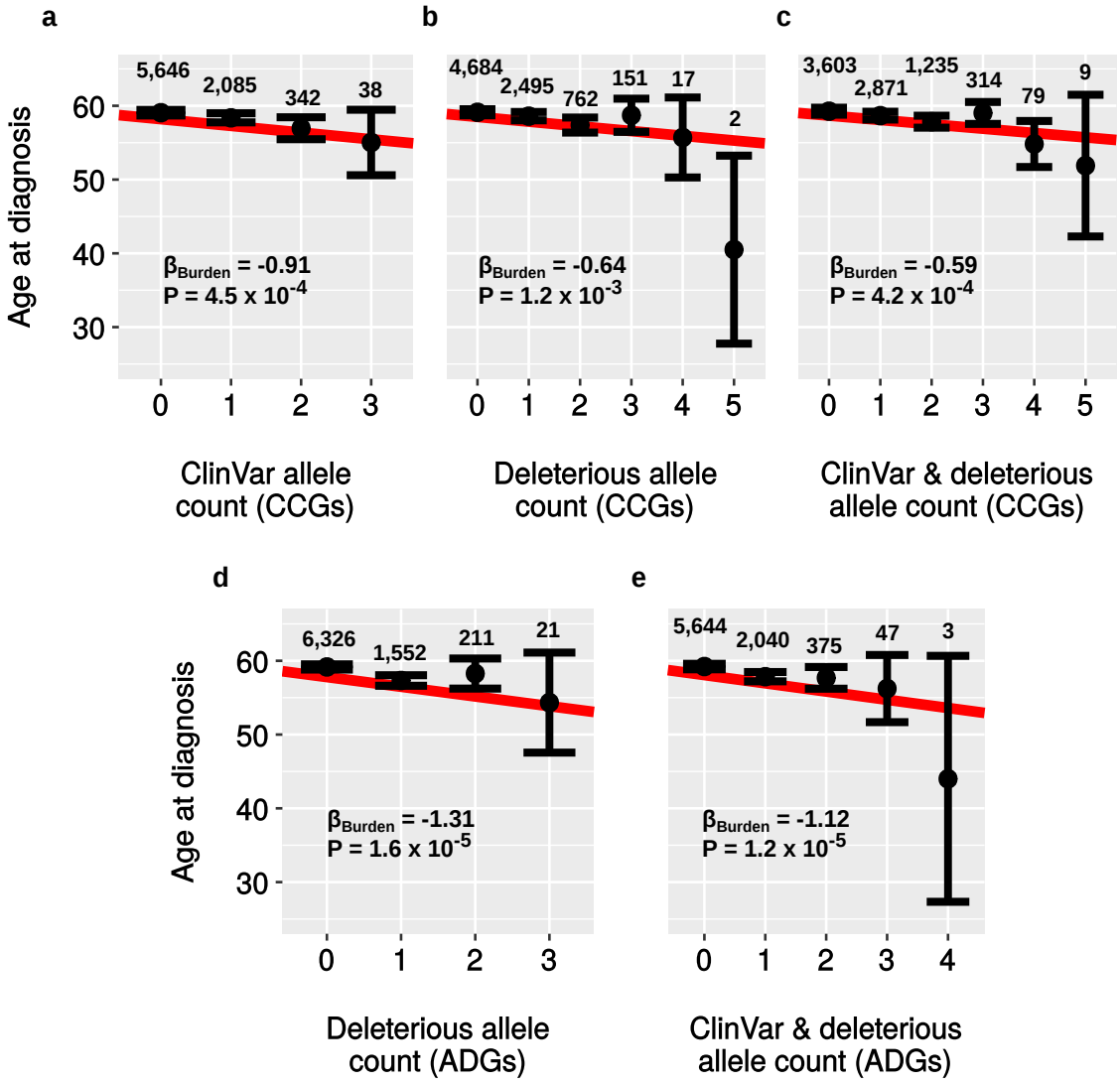

Figure 3.2: **Increased burden of harmful alleles in cancer risk genes is associated with earlier age at cancer diagnosis.** Using 8,111 individuals from 31 cancer types, we regressed age of diagnosis (mean and 95% confidence interval) against the burden of (a) cancer-associated (ClinVar), (b) deleterious, 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 (β_{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.

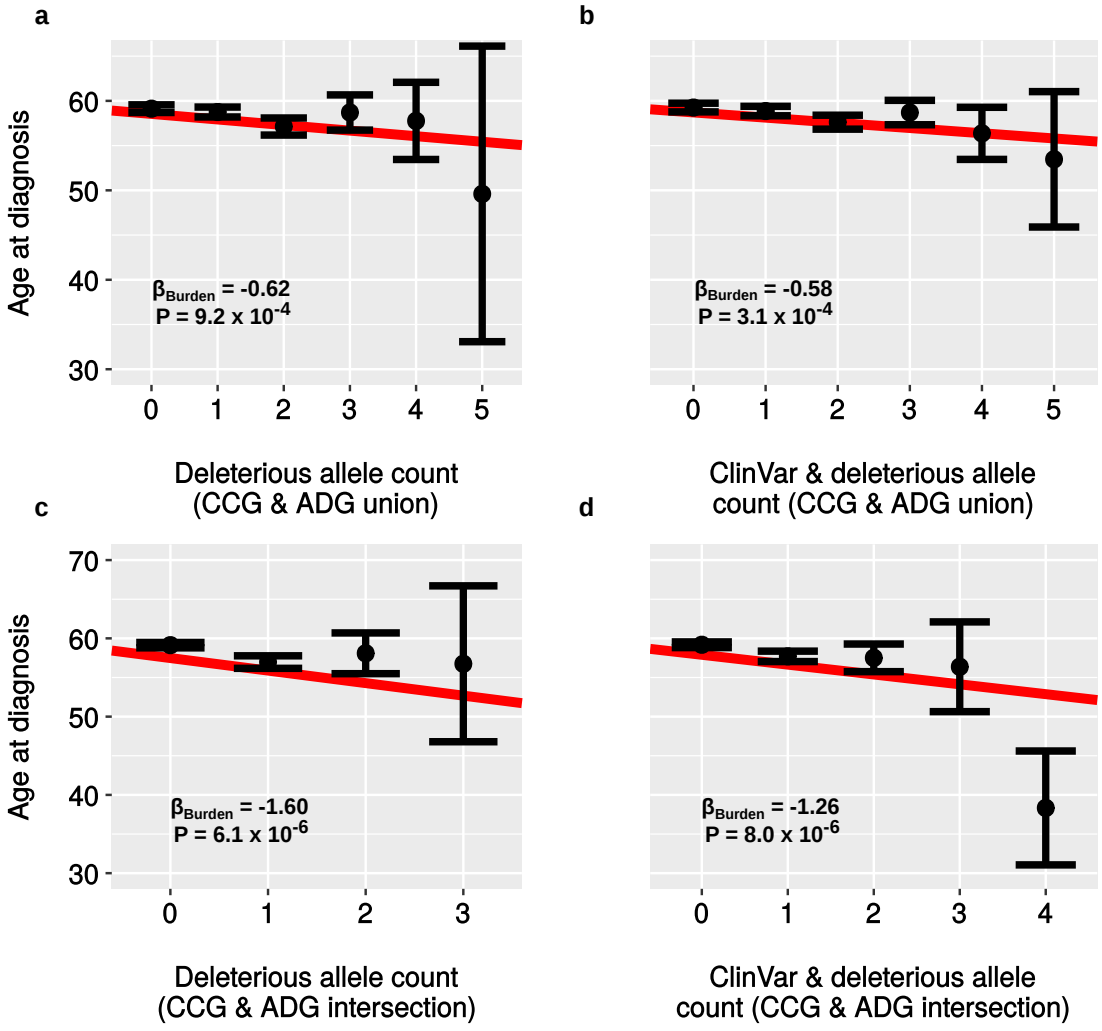

Figure 3.3: **Age at diagnosis by allele burden using the union and intersection of gene sets.** Using the union of ClinVar cancer genes (CCGs) and autosomal dominant cancer predisposition genes (ADGs) (n = 87), age of diagnosis (mean and 95% confidence interval) was regressed against the burden of (a) 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 (β_{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 (β_{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-

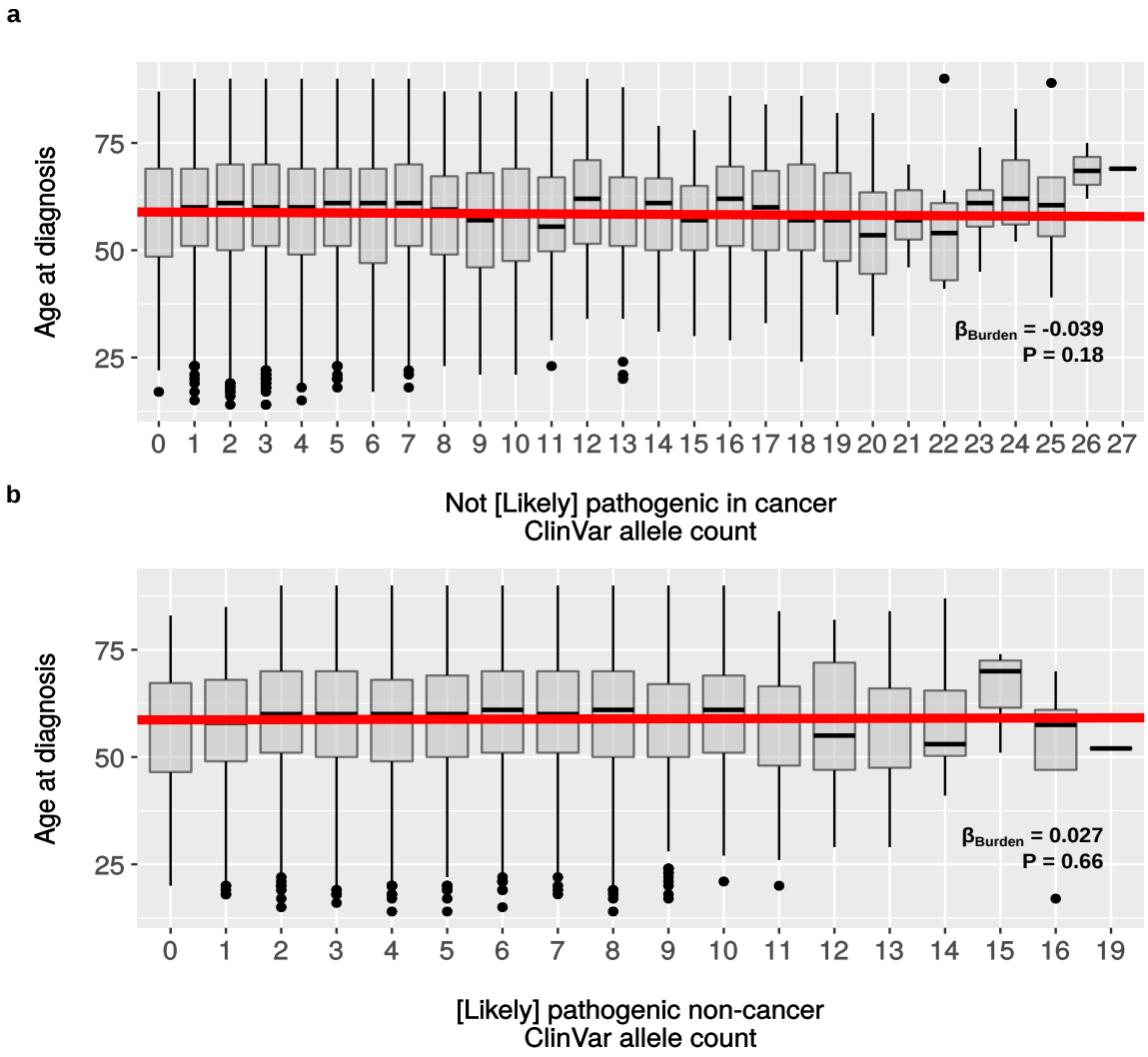

Figure 3.4: **No observed relationship between age at diagnosis and allele burden using non-cancer ClinVar variants.** Boxplots depicting age at diagnosis as a function of allele burden using ClinVar variants (a) not asserted as [Likely] pathogenic (i.e. labeled only Benign, Likely benign, Other, not provided, etc.) in a cancer phenotype(s) and (b) [Likely] pathogenic in non-cancer phenotypes. Slope (β_{Burden}) and allele burden P value are reported in each panel.

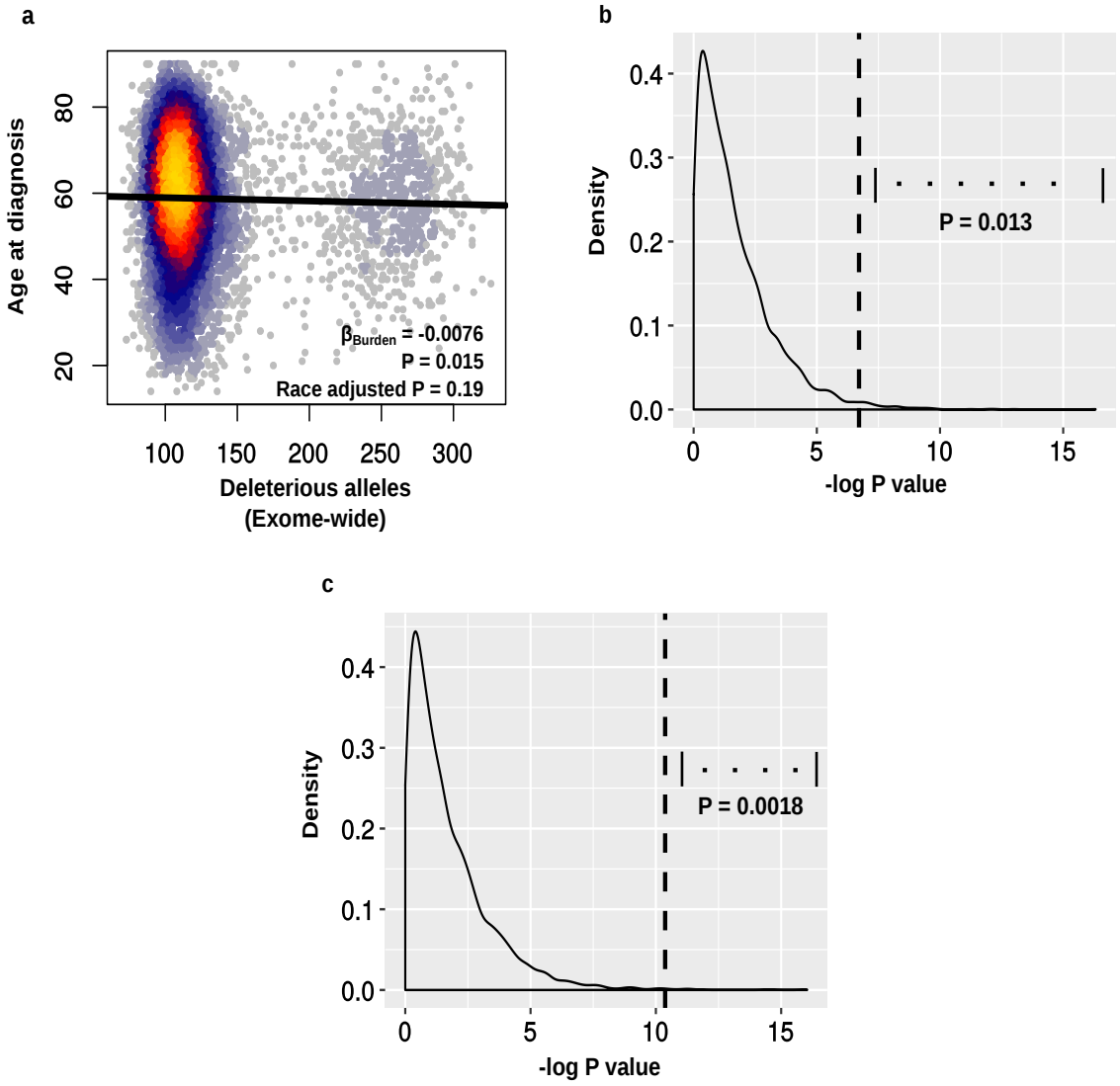

Figure 3.5: **Age at diagnosis against deleterious allele burden exome-wide and within random gene sets.** (a) Age at diagnosis against all deleterious, autosomal alleles exome-wide with unadjusted and race-adjusted P values reported. The estimate and fit line are not adjusted for race. Hotter colors 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 lines 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.

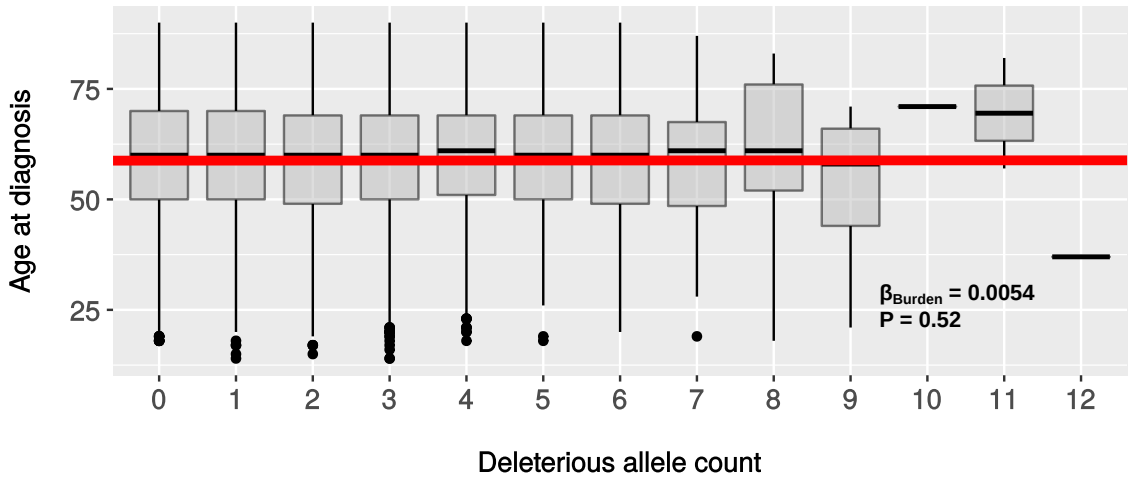

Figure 3.6: **Age at diagnosis by allele burden in genes significantly somatically mutated in cancer.** Regressions were performed using deleterious allele burden across significantly somatically mutated in cancer (n = 185). Genes were excluded (n = 75) if they were a ClinVar cancer gene (CCG), an autosomal dominant cancer predisposition gene (ADG), or deemed an oncogene by the Cancer Gene Census. Slope (β_{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 x 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 (β_{Burden} =

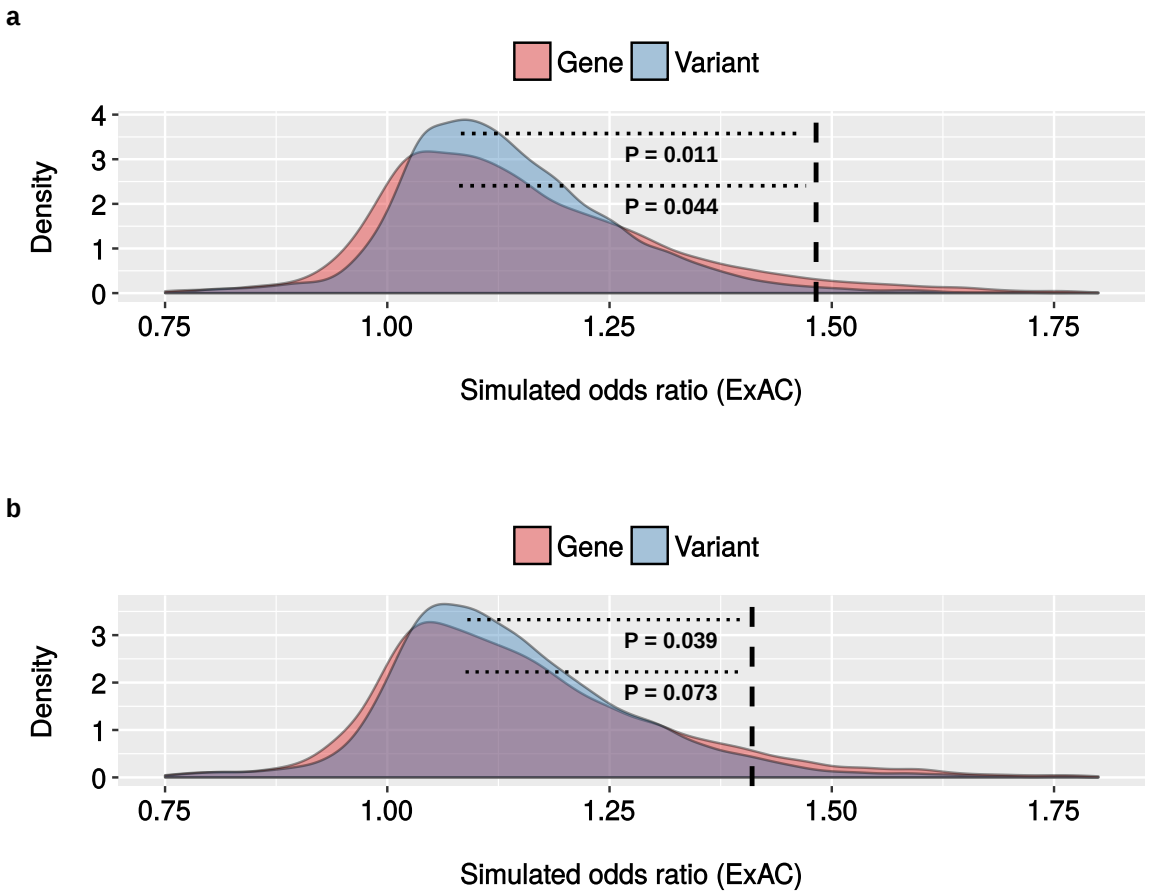

Figure 3.7: **Enrichment of deleterious alleles in individuals with cancer.** Odds ratios (ORs) were calculated by comparing AC and AN between cancer (TCGA) and non-cancer cohorts using data from ExAC. Distributions were generated using 10,000 random sets of genes (ClinVar cancer genes [CCG] n = 57; 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.

APC	*BRCA1*	*CDH1*	*MLH1*	*MUTYH*	*PTEN*	*SMAD4*
ATM	*BRCA2*	*CHEK2*	*MSH2*	*PALB2*	*RAD51C*	*STK11*
BMPR1A	*BRIP1*	*EPCAM*	*MSH6*	*PMS2*	*RAD51D*	*TP53*

Table 3.3: **List of 21 moderate and high risk cancer predisposition genes.** These genes were curated by Slavin and colleagues [357].

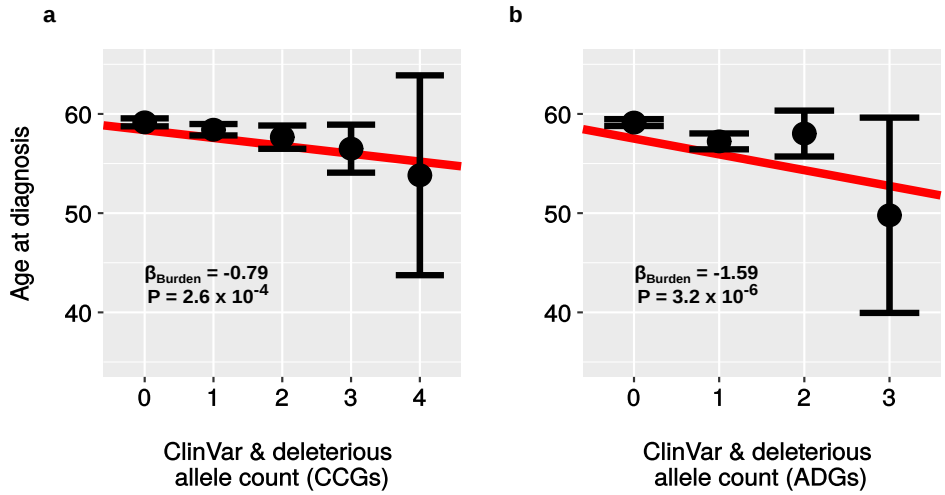

Figure 3.8: **Age at diagnosis associated with allele burden when high and moderate risk genes are excluded.** Age at diagnosis against cancer-associated (ClinVar) and deleterious allele burden in (a) ClinVar cancer genes (CCGs) and (b) autosomal dominant cancer predisposition genes (ADGs) while excluding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

-0.79; 95% CI = -1.14 -0.26; P = 2.6 x 10^{-4} and β_{Burden} = -1.59; 95% CI = -1.14 -0.26; P = 3.2 x 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.

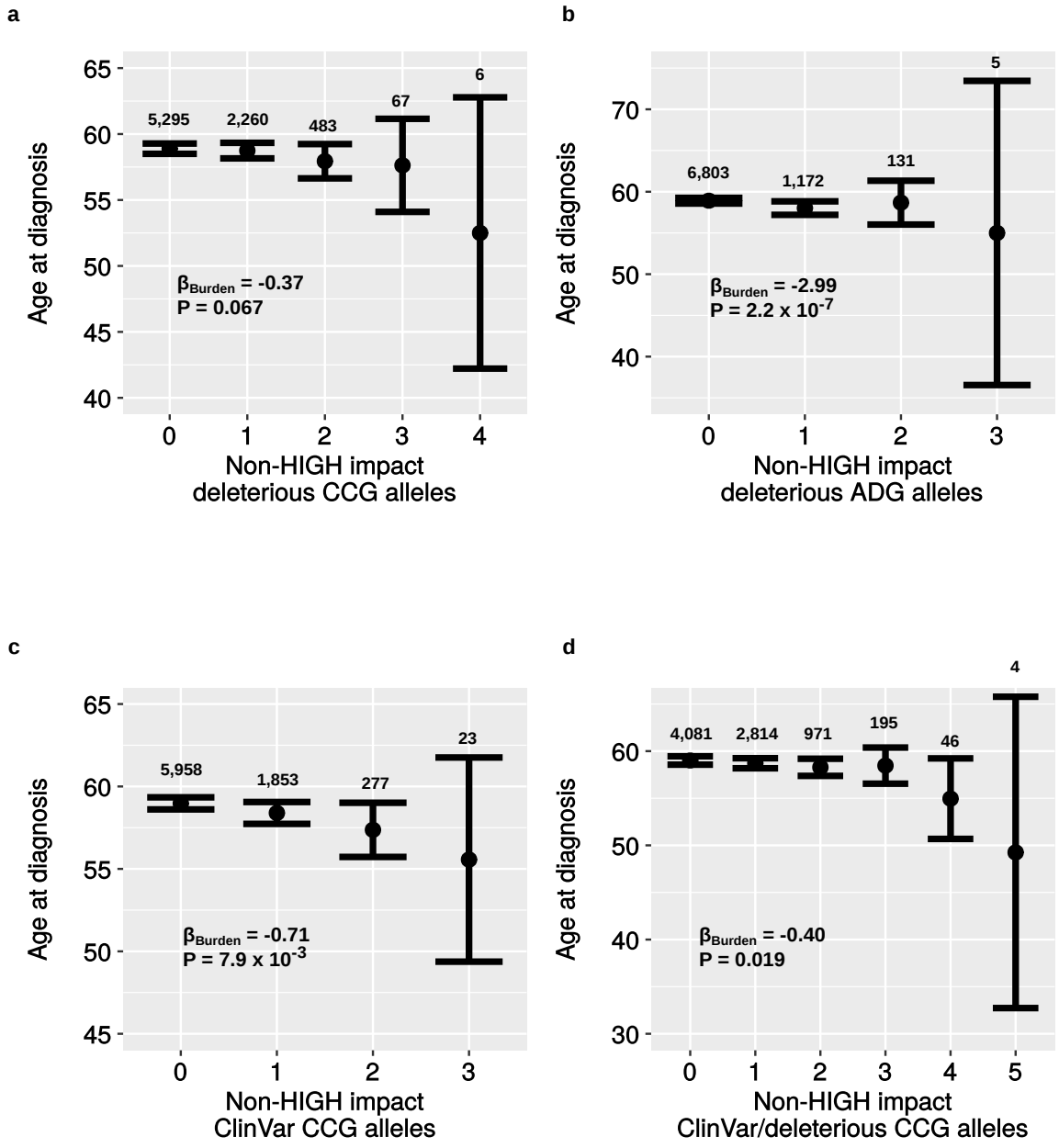

Figure 3.9: **Associations remain after excluding alleles with predicted high impact on gene function.** Characterization of the relationship between age at diagnosis and allele burden after removing variants deemed to have a HIGH (i.e. presumed loss-of-function) impact (e.g. stop gained, frameshift, 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_{Burden}) and burden P value are reported within each panel. Mean and 95% confidence interval are provided for each allele burden group.

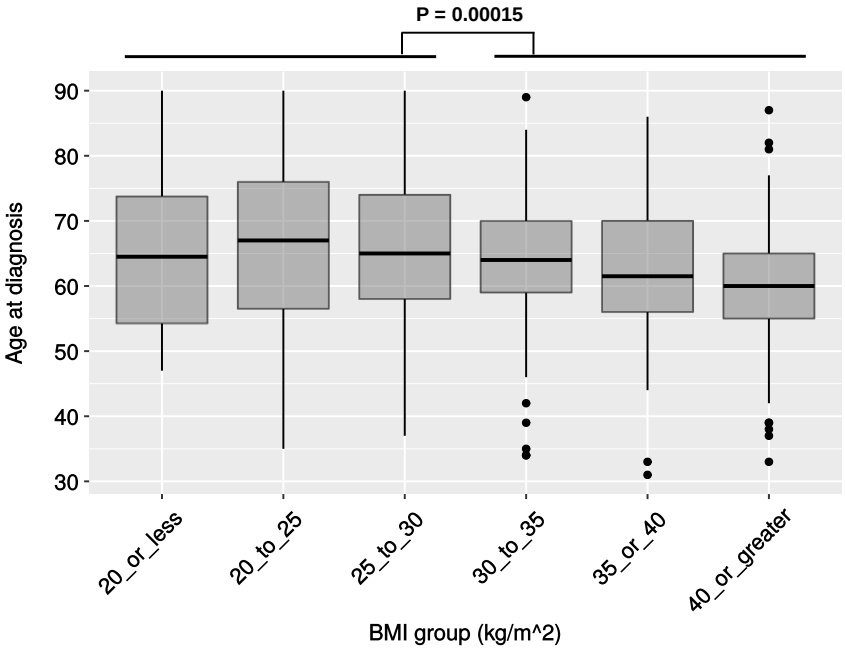

Figure 3.10: **Obese women with uterine/endometrial carcinoma are diagnosed earlier.** Individuals were partitioned into six groups based on BMI (kg/m^2), and age at diagnosis was subsequently summarized via boxplots. Statistical comparisons were made between obese (BMI $\geq$ 30) and non-obese (BMI $<$ 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 ($P = 1.0$ x 10^{-3}) and greater allele burden ($P = 6.4$ x 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 ($P = 1.5$ x 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.

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 ($\beta_{HighBurden}$ = -7.64, 95% CI = -14.0 -1.26; P = 9.5 x 10^{-3}), *BRCA1* status (β_{BRCA1} = -6.63, 95% CI = -11.4 -1.82; P = 3.5 x 10^{-3}), and *BRCA2* status (β_{BRCA2} = -4.80, 95% CI = -8.29 -1.31; P = 3.6 x 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.

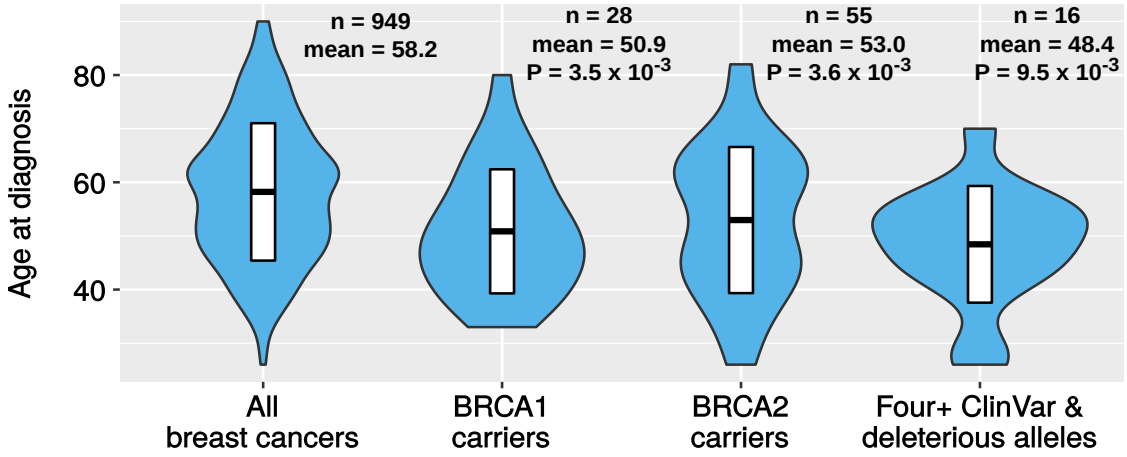

Figure 3.11: ***BRCA1/2* carrier status and high allele burden independently associate with earlier breast cancer diagnosis.** Violin plots representing the age at diagnosis distribution for all female breast cancers, *BRCA1* carriers, *BRCA2* carriers, and individuals with four or more cancer-associated (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.

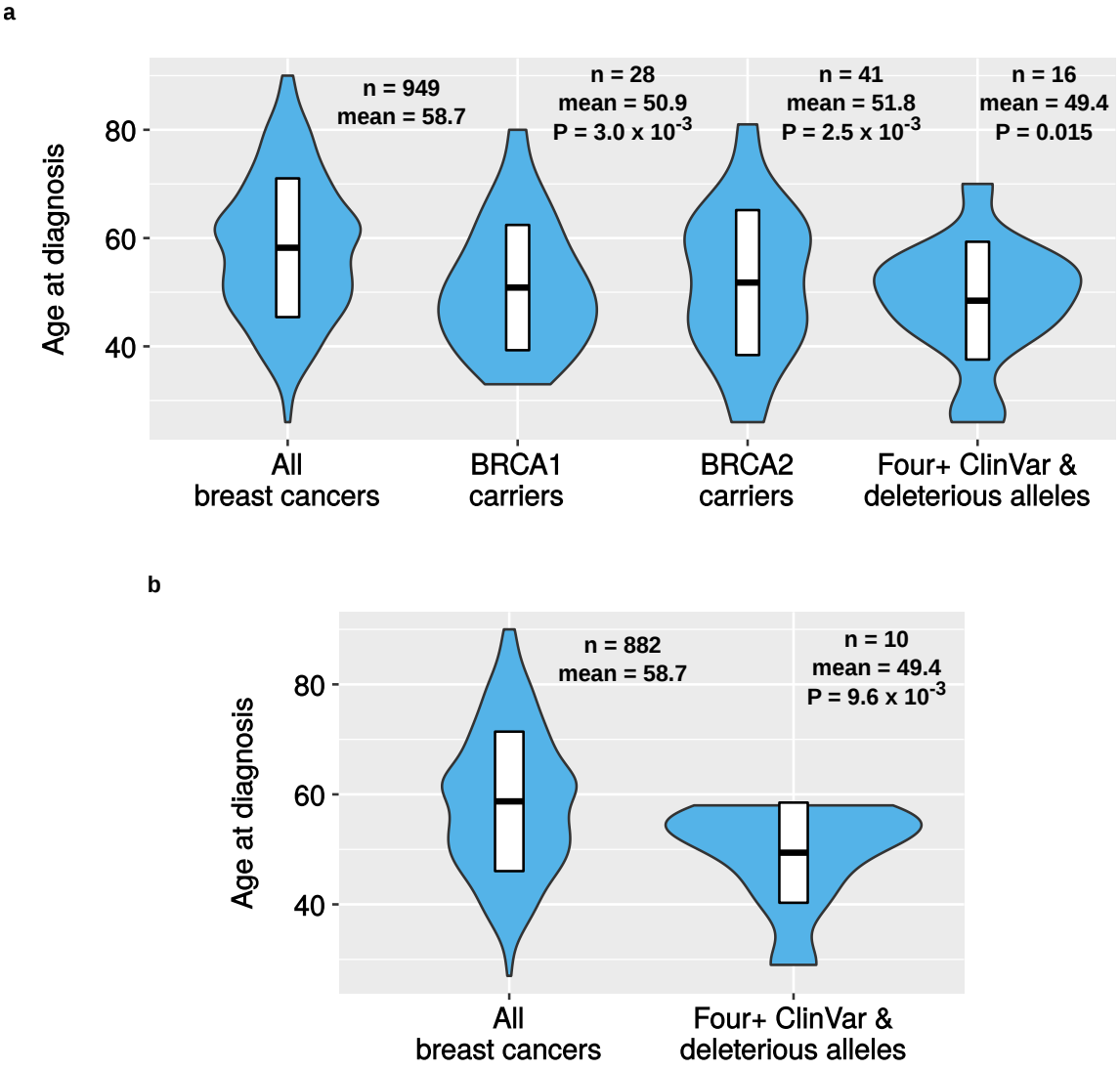

Figure 3.12: **Highly burdened individuals have earlier breast cancer diagnosis when excluding terminal *BRCA2* variants and *BRCA1/2* carriers.** (a) Violin plots representing the age at diagnosis distribution for all female breast cancers, *BRCA1* carriers, *BRCA2* carriers, and individuals with four 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 >= 15 was utilized multiple other studies as well [371, 375, 376]. Throughout the analyses, I considered missense variants with a CADD score >= 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, 404]. 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).

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

Table 4.1: **TCGA cancer types and samples counts used for two-hit analyses.**

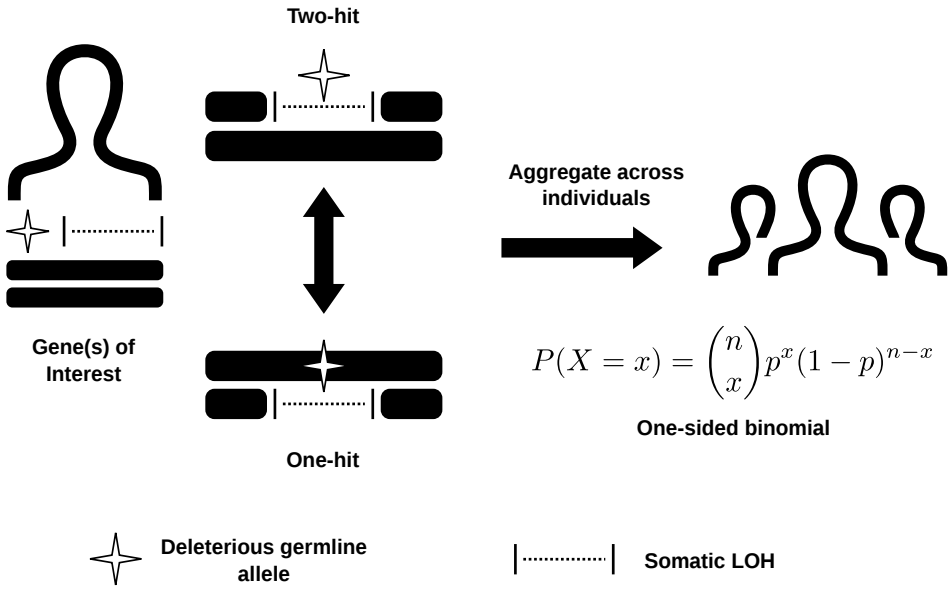

Figure 4.1: **Workflow for identifying candidate two-hit genes.**

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.

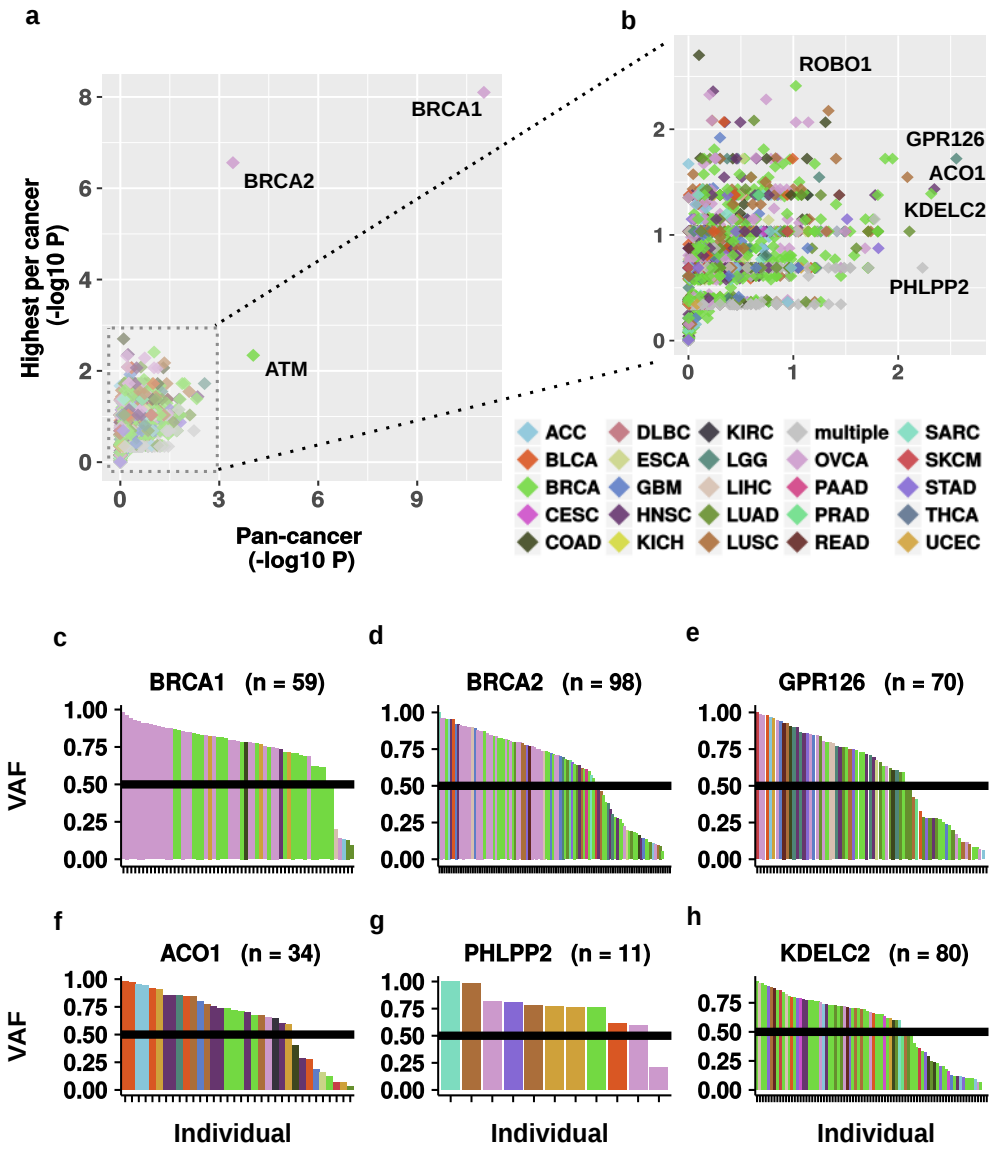

Figure 4.2: **Two-hit enrichment pan- and per-cancer type.** (a) Plotting the highest per-cancer P value (-$\log_{10}$) against the pan-cancer P value (-$\log_{10}$) for each gene. Each gene is colored by cancer type where it is the most enriched. If the a genes greatest enrichment occurred in two or 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) *KDELC2*. 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.

Gene	One-hit count	Two-hit count	Total	One-way textitP value
BRCA1	5	54	59	9.5 x 10^{-12}
ATM	25	60	85	9.3 x 10^{-5}
BRCA2	32	66	98	3.8 x 10^{-4}
GPR126	23	47	70	2.8 x 10^{-3}
ACO1	9	25	34	4.5 x 10^{-3}
KDELC2	28	52	80	4.8 x 10^{-3}
PHLPP2	1	10	11	5.9 x 10^{-3}
DBR1	0	7	7	7.8 x 10^{-3}

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 x 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 x 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 x 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 x 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 and 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.

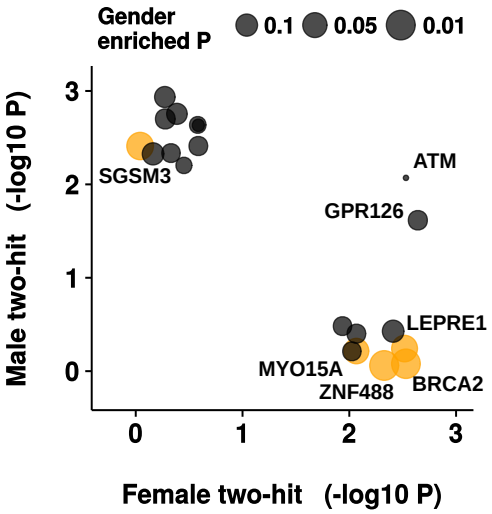

Figure 4.3: **Top male and female candidates display sexually dimorphism.** The union of the top ten two-hit candidates from both males and females. Gender enriched P values were calculated using a Fishers exact test which compared one- and two-hit counts in males versus females. Circle size is proportional 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 (*GPR126*, *ACO1*, *PHLPP2*, and *KDELC2*), and cells were subsequently measured for proliferative capacity. Specifically, in MCF10A cells, siRNA to *KDELC2* ($P = 1.8$ x10^{-3}, Welch two-sample T-test) and *PHLPP2* ($P = 1.3$ x 10^{-10}) caused significant increases in proliferation (Fig. 4.4a). siRNA to *PHLPP2* caused a borderline significant proliferation increase in MRC-5 cells ($P = 0.063$). Cell type independent effects are consistent with *PHLPP2*'s

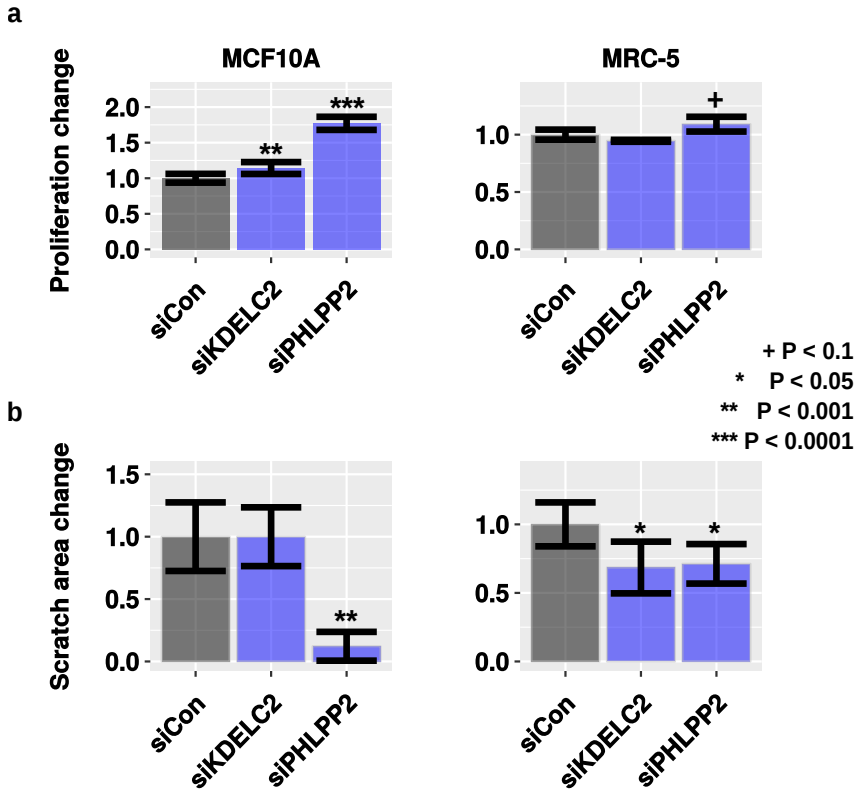

Figure 4.4: **Knockdown of candidate two-hit genes induces cancerous phenotypes.** (a) Proliferation and (b) scratch assays performed in MCF10A (left) and MRC-5 (right) cell lines after *KDELC2* and *PHLPP2* knockdown with siRNA. The mean the standard deviation are provided with each bar. Results that were significantly different (Welch two-sample T-test) from control siRNA (siCon) are indicated within each panel.

known tumor suppressive activity (Fig. 4.4a) [434, 435].

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.

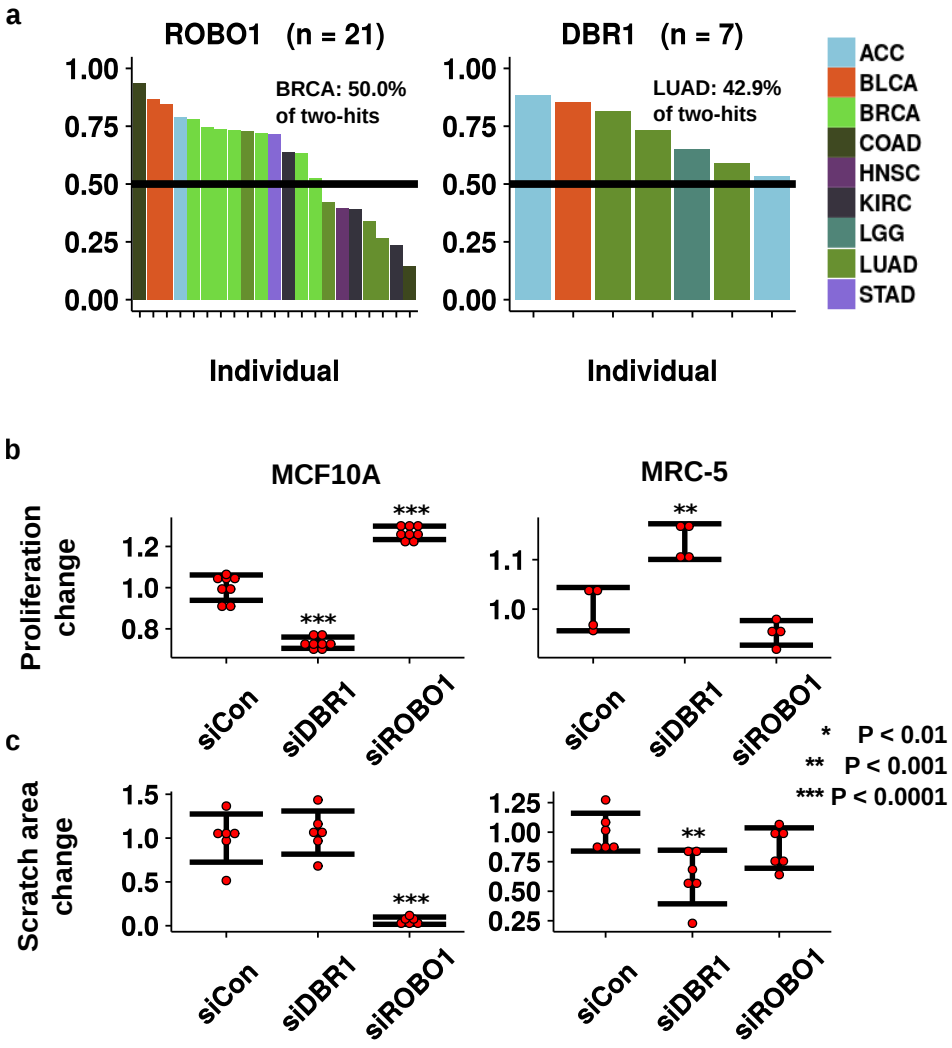

Figure 4.5: ***ROBO1* and *DBR1* show cell type-specific phenotypes.** (a) Waterfall plots depicting the variant allele fraction (VAF) for all one- and two-hit cases in *ROBO1* (left) and *DBR1* (right). Black, horizontal bars indicate a VAF of 0.5. Cases with a VAF below and above this threshold are 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

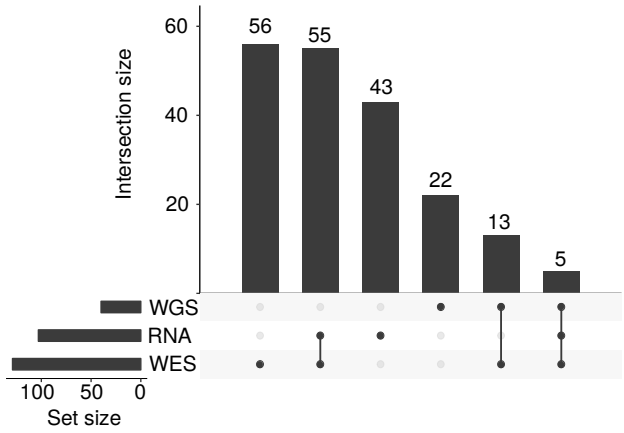

Figure 5.1: **The number of Nigerian samples with each NGS data type.** WGS: Whole-genome sequencing, WES: whole-exome sequencing, RNA: mRNA sequencing.

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 $P < 1.0 \times ^{-4}$) 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

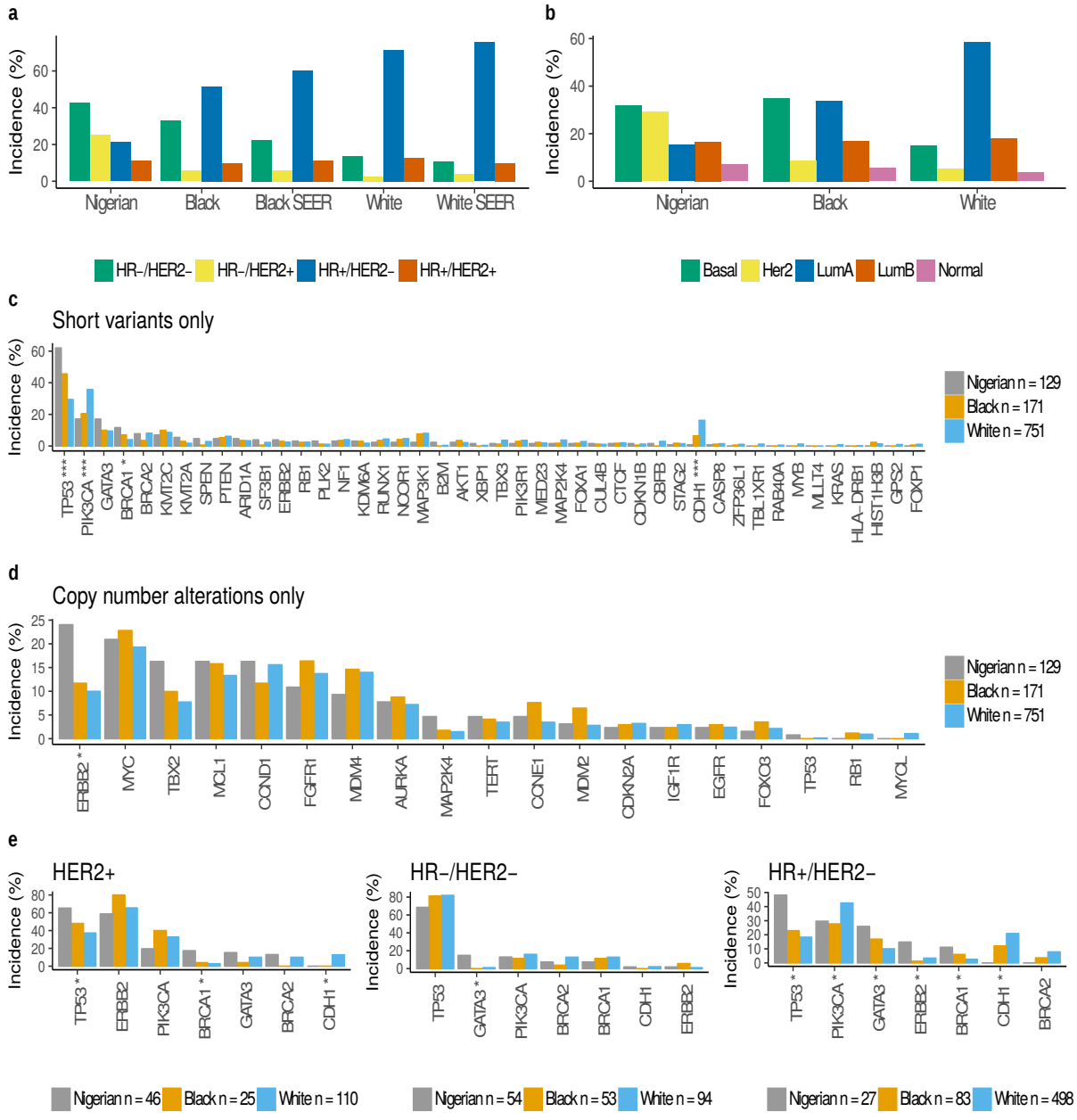

Figure 5.2: **Landscape of breast cancer in Nigerians compared to Black and White Americans.** (a) Incidence of IHC subtypes in the Nigerian, the Black and White cohorts from TCGA, and in the SEER database. (b) Incidence of PAM50 subtypes in Nigerians, Blacks and Whites. (c) Comparison of the frequencies 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).

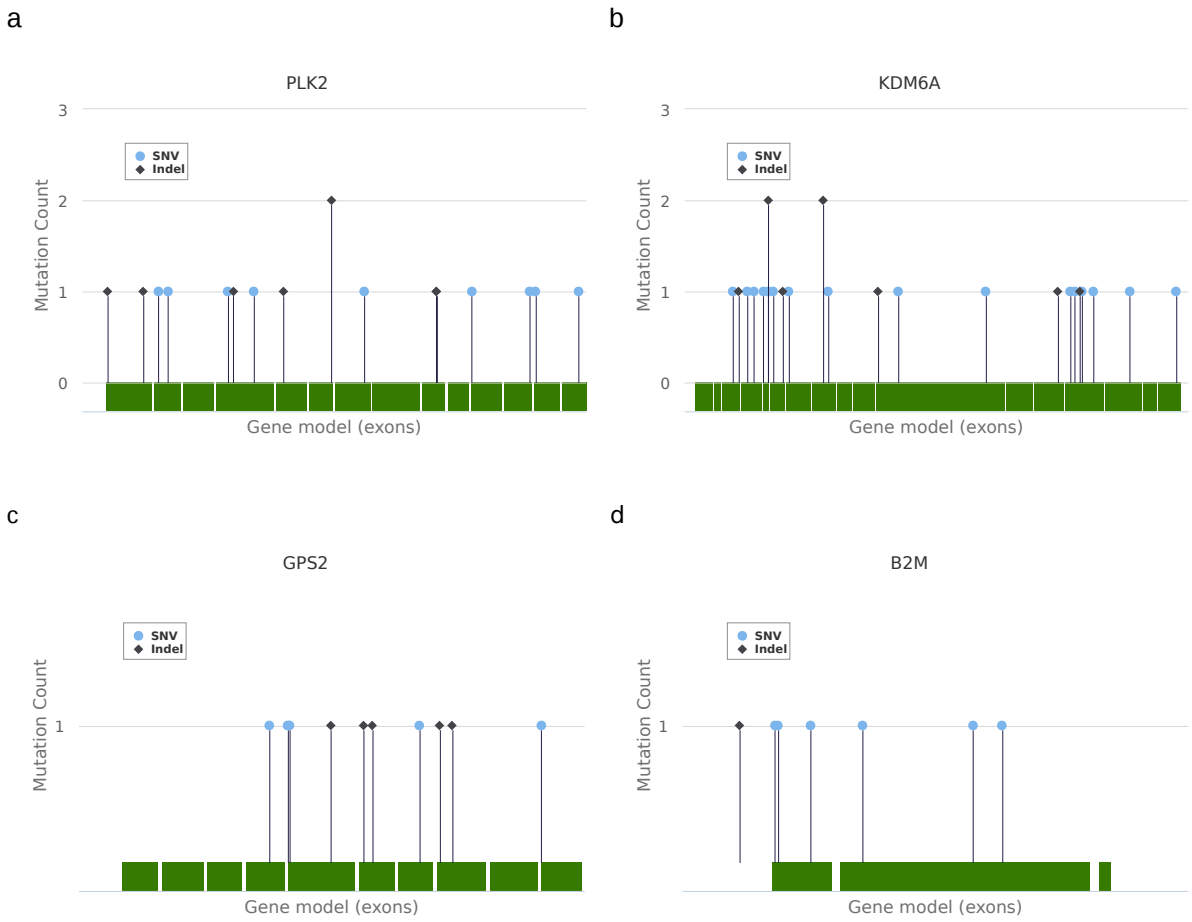

Figure 5.3: **Lollipop plots for novel significantly mutated breast cancer genes.** Protein-altering SNVs and indels for (a) *PLK2*, (b) *KDM6A*, (c) *GPS2*, and (d) *B2M*. The start position of a deletion in *B2M* falls outside of the first exon; however, that deletion is represented in panel d since it spans part of the first exon.

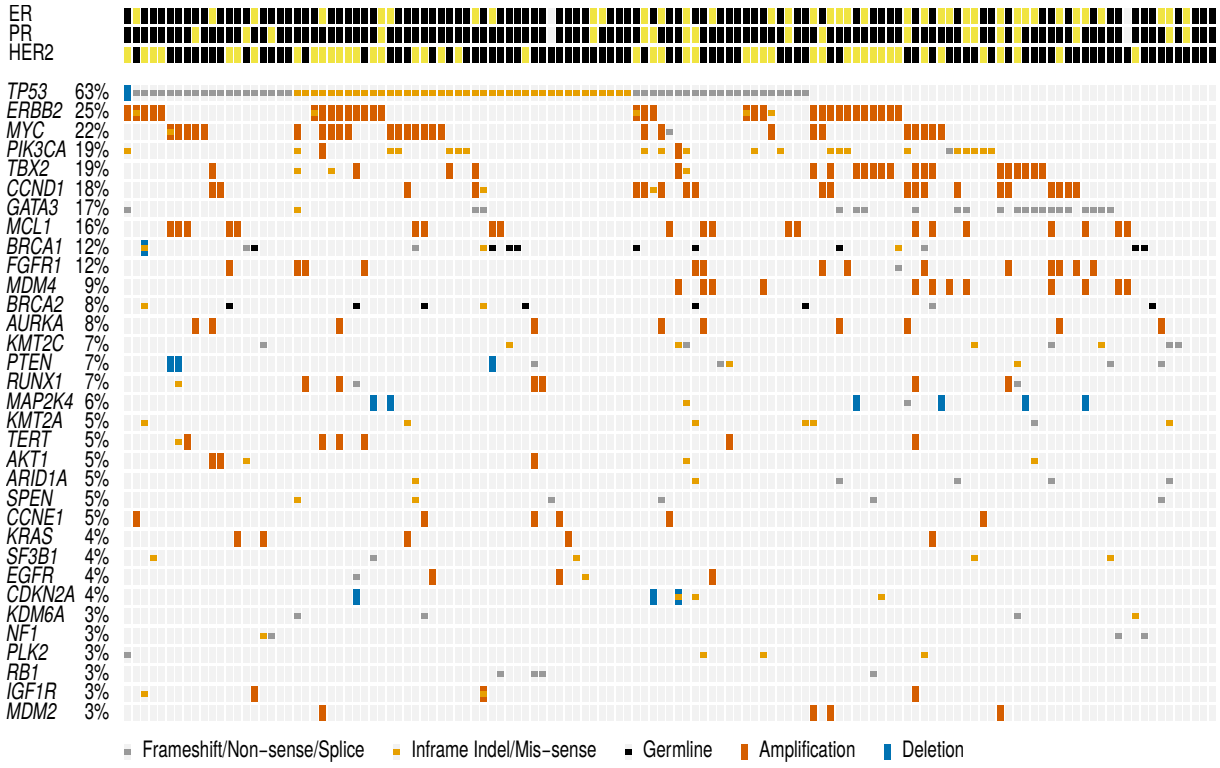

Figure 5.4: **Oncoprint of short mutations and CNAs in Nigerians.** Recurrently mutated genes (Tables 5.5 and 5.6) that were altered least 3% of Nigerians are shown.

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 reflects 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

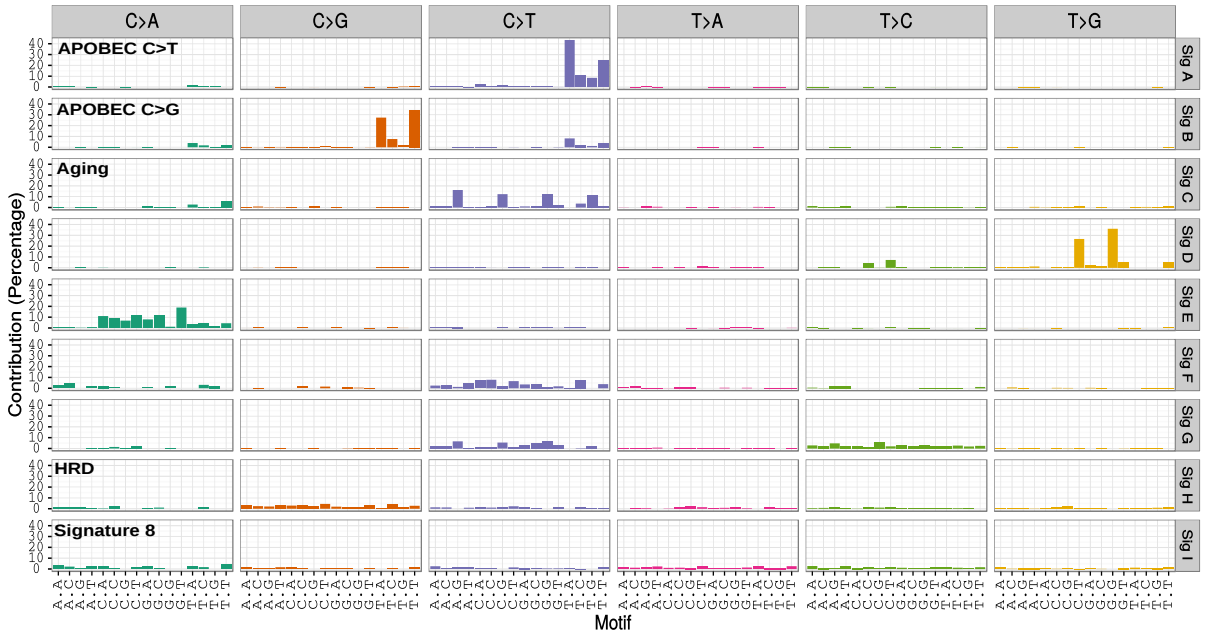

Figure 5.5: **Tri-nucleotide substitution patterns of nine inferred mutation signatures.** Nine mutation signatures were jointly estimated from 500 exomes and 122 whole genomes using non-negative matrix factorization. All possible substitutions are represented within their tri-nucleotide context. Bars depict the percentage to which each tri-nucleotide substitution contributes to a given signature. APOBEC C>T, APOBEC C>G, aging, and HRD mutation signatures as well as COSMiC signature 8 are denoted.

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 ($P = 3.5 \text{ x } 10^{-9}$, 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 ($P = 2.2 \text{ x } 10^{-15}$) (Fig. 5.9a-d). Consistent with previous work [476], HER2+ tumors had the highest contributions from APOBEC C>T and C>G signatures ($P = 1.6 \text{ x } 10^{-8}$ and $P = 9.1 \text{ x } 10^{-4}$, respectively) (Fig. 5.8b and

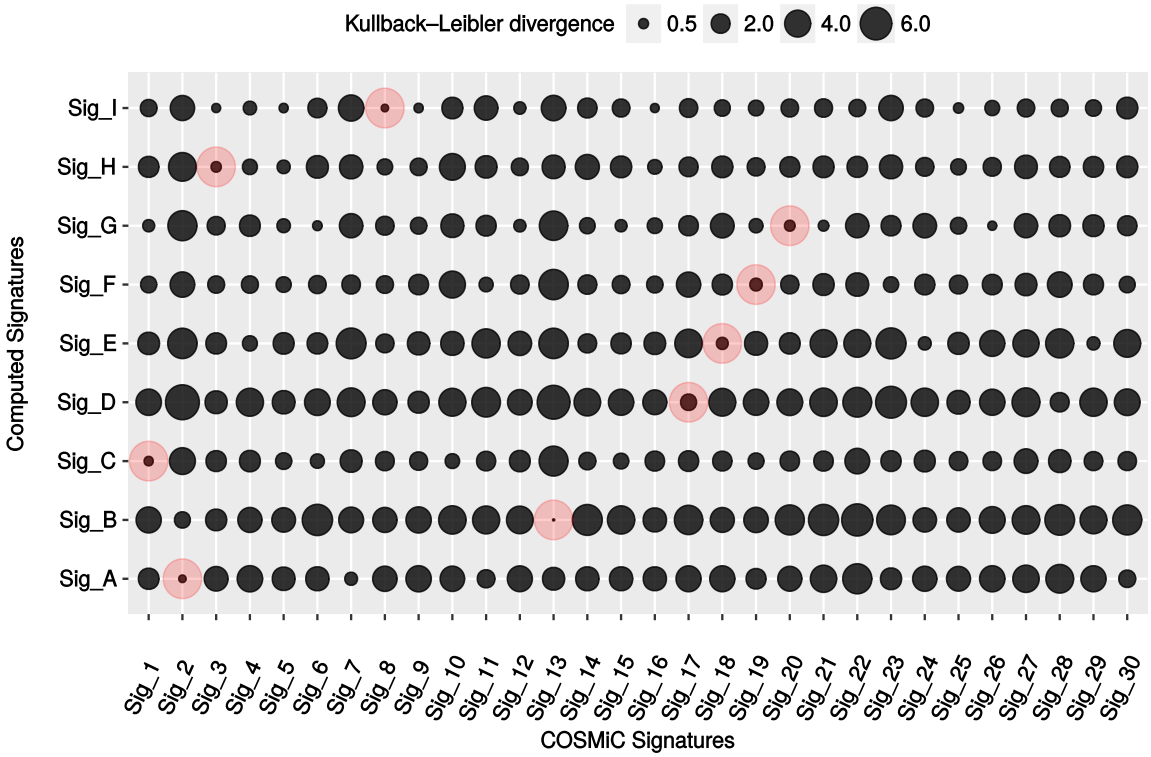

Figure 5.6: **Derived mutation signatures compared to COSMiC mutation signatures and correlation between WES and WGS signature contributions.** Kullback-Leibler divergence was calculated pairwise for derived and COSMiC mutation signatures. For each derived signature, the smallest divergence value — which indicates the most similar COSMiC signature — is denoted by a pink circle.

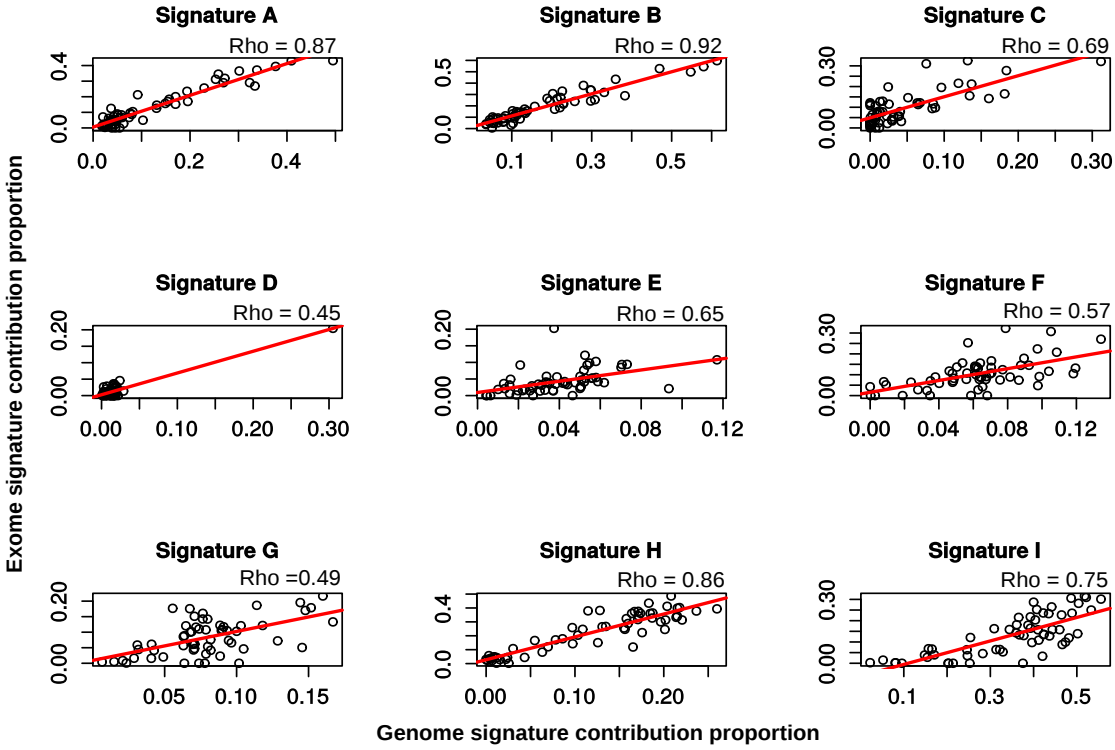

Figure 5.7: **Correlations between WES and WGS mutation signature contributions.** Scatterplots of nine mutation signature contributions between WES and WGS for 59 individuals. Spearman correlation was calculated for each signature with Rho depicted on each plot. The derived signature is denoted above each plot.

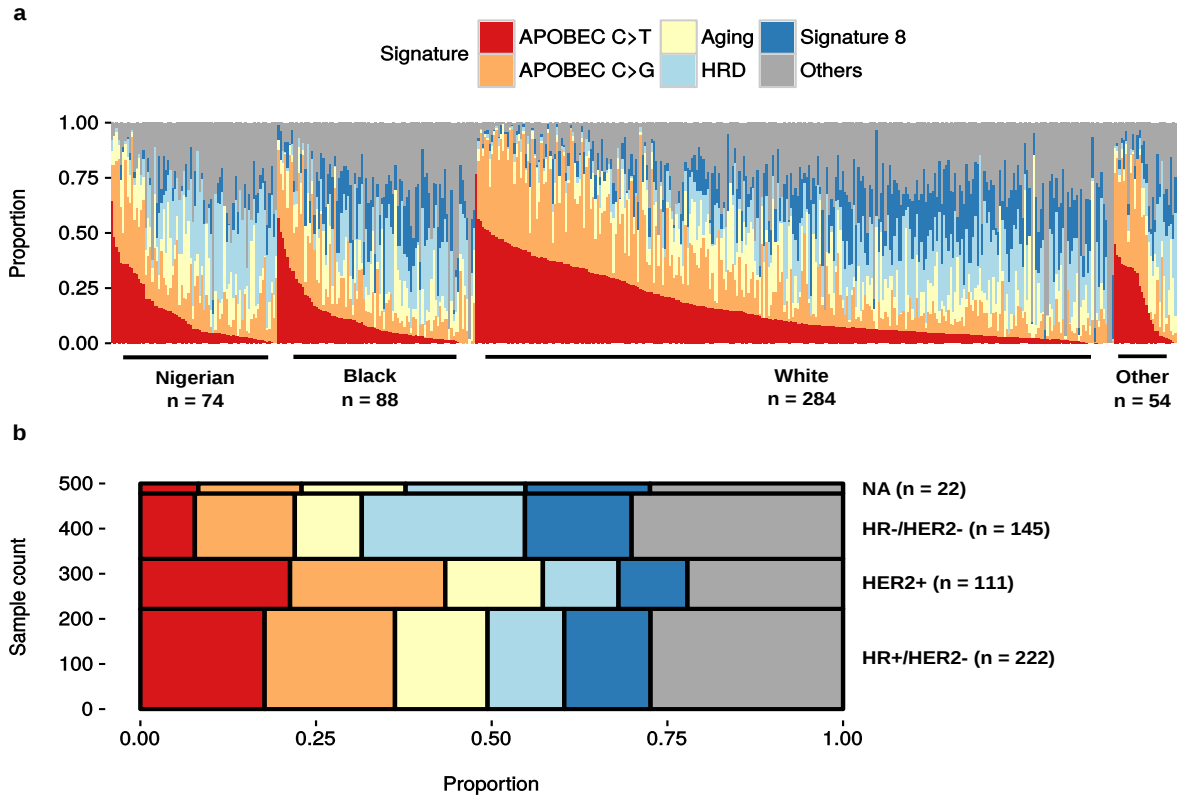

Figure 5.8: **Mutation signature contributions across race/ethnicity and subtype.** (a) The contribution (proportion) of mutation signatures (Signatures D, E, F, and G are combined into "Other") within each individual. Individuals are partitioned by race/ethnicity and ordered by APOBEC C>T signature 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.

Fig. 5.9e-f). 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 x 10^{-7}) [476, 477]. *TP53* mutations were associated with higher HRD contributions (BH P = 2.1 x 10^{-13}, MWU), higher missense mutation burden (BH P = 6.5 x 10^{-45}), and increased copy number segmentation (BH P = 2.0 x 10^{-43}) (Fig. 5.10a). In contrast, *CDH1* or *PIK3CA* mutations — which frequently co-occur (P = 3.8 x 10^{-8}) — were associated with lower HRD contributions (*CDH1* BH P = 5.2 x 10^{-11}; *PIK3CA* BH P = 2.1 x 10^{-17}) in addition to higher contributions from APOBEC C>T (BH P = 3.0 x 10^{-9}; BH P = 3.8 x 10^{-17}) and C>G (BH P = 1.7 x 10^{-4}; BH P = 2.1 x 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 x 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 x 10^{-6}), Nigerians and Whites (P = 4.6 x 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 x 10^{-7}), which was recapitulated using WGS data (P= 6.9 x 10^{-3}) (Fig. 5.13a-b). These subtype differences were not observed for either Nigerians or Blacks.

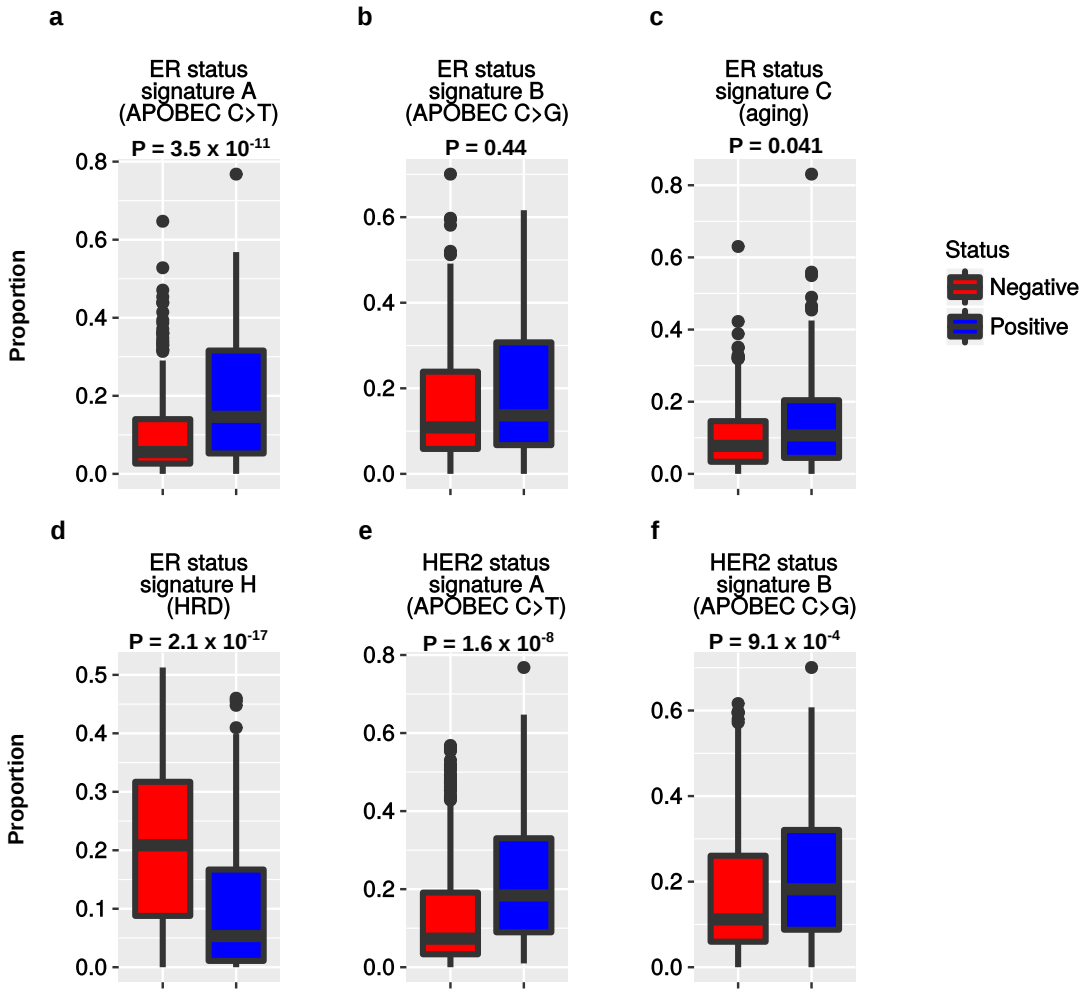

Figure 5.9: **Mutation signature contributions between tumors positive and negative for IHC markers.** Boxplots represent (a) APOBEC C>T, (b) APOBEC C>G, (c) aging, and (d) HRD signatures partitioned by ER status. Similarly, contributions from (e) APOBEC C>T and (f) APOBEC C>G signatures between HER2 positive and negative tumors. P values shown were calculated via Mann-Whitney U.

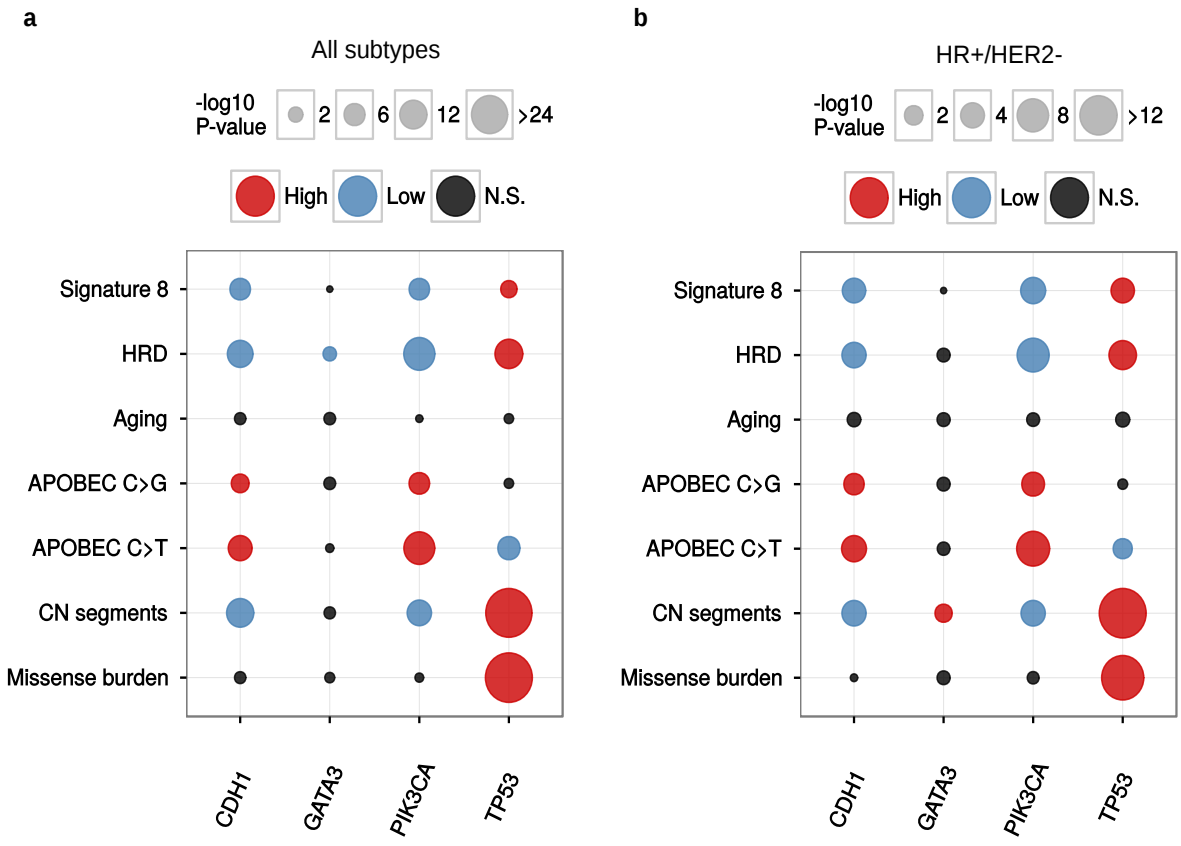

Figure 5.10: **Associations between genome-wide oncogenic features and the mutation status of common driver genes.** Dot plot depicting the relationships between mutation status in *TP53*, *PIK3CA*, *CDH1*, and *GATA3* and mutation signatures (APOBEC C>T, APOBEC C>G, aging, HRD, and signature 8), missense 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.

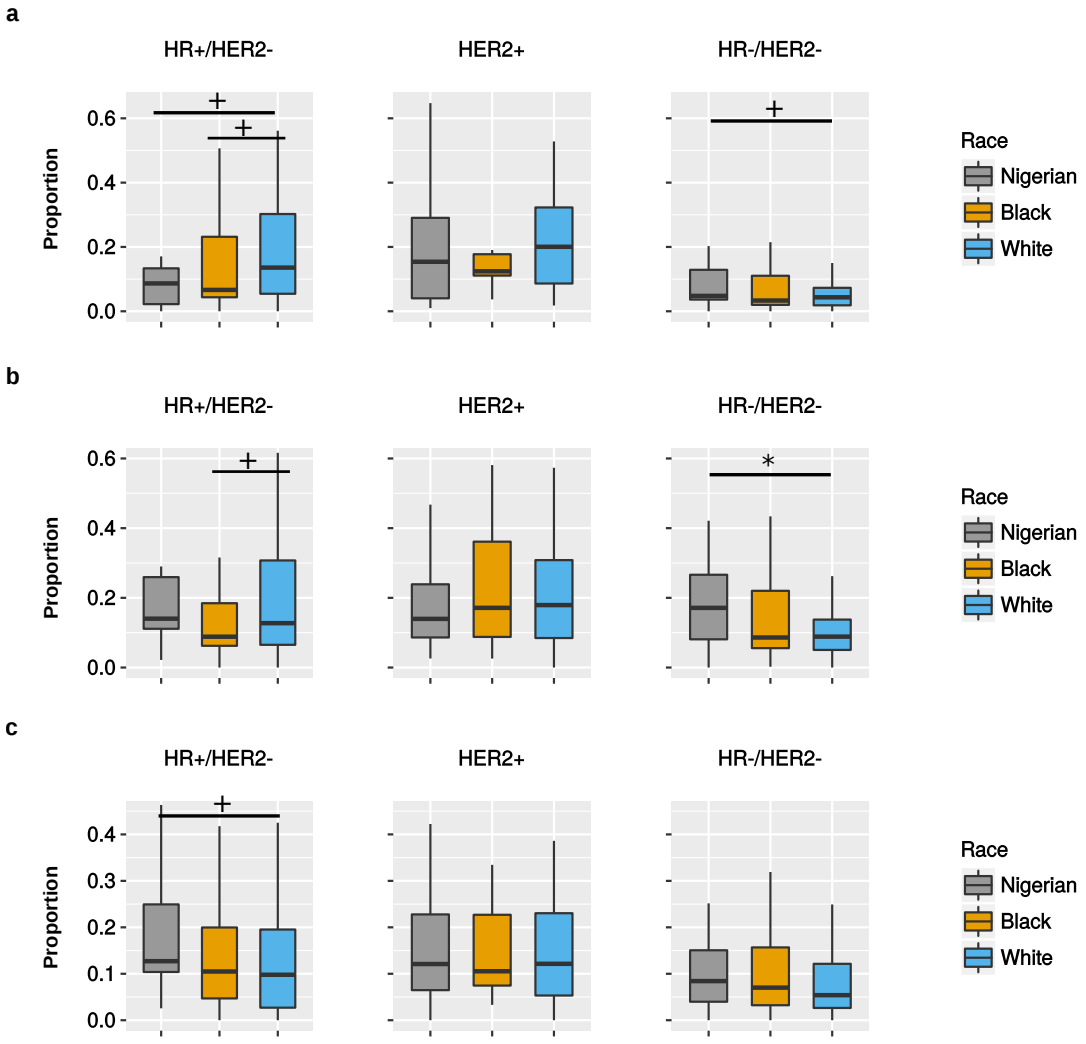

Figure 5.11: **The proportion of APOBEC C>T, APOBEC C>G, and aging signatures by race/ethnicity and IHC subtype using WES.** The proportion of APOBEC C>T, APOBEC C>G, and aging signatures by race/ethnicity and IHC subtype using WES. Differences in (a) APOBEC C>T, (b) APOBEC C>G, and (c) aging signatures 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 ≤ 0.05.

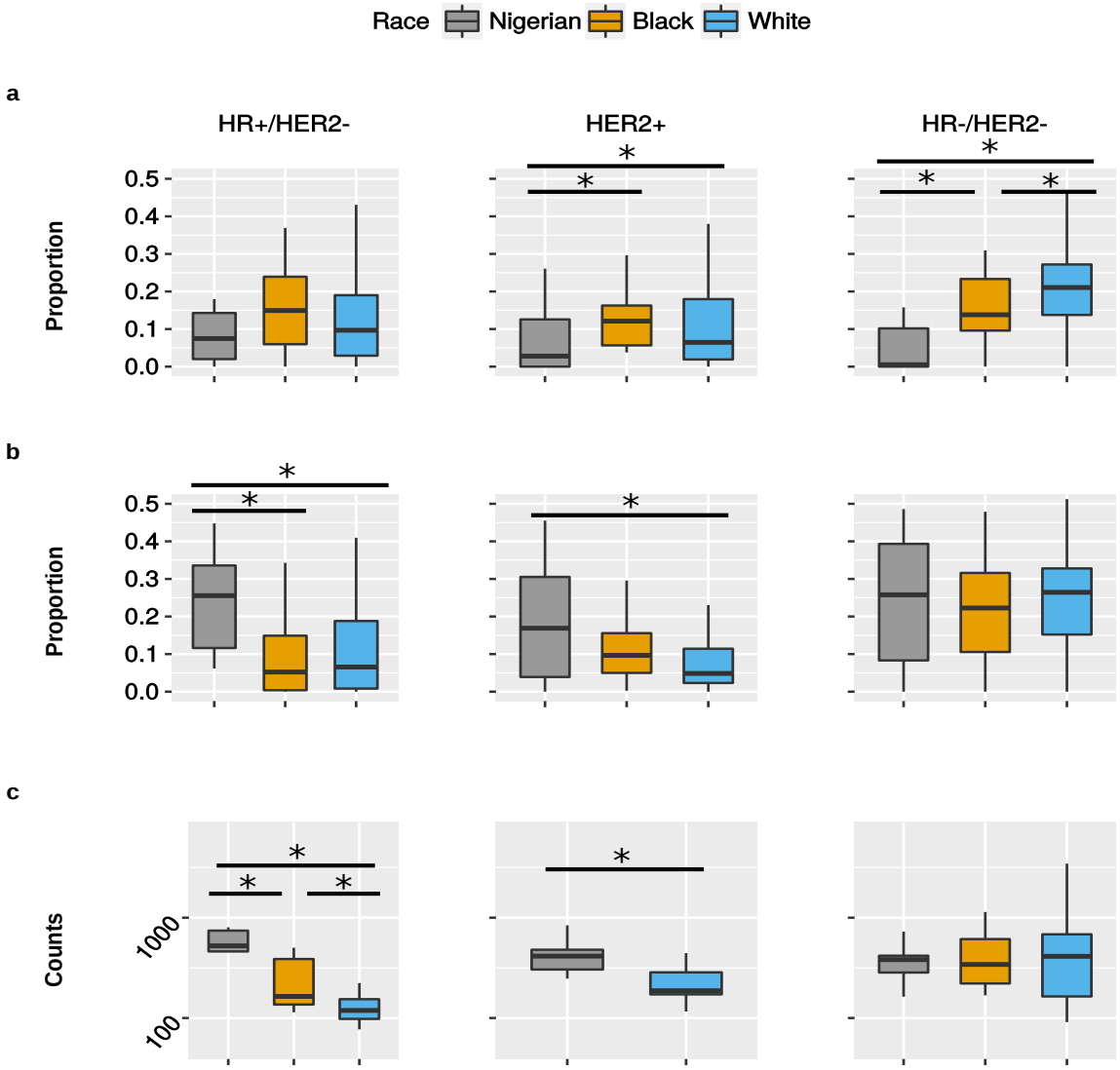

Figure 5.12: **Mutation signature contributions and structural variant counts partitioned by race/ethnicity and IHC subtype.** Mutation signature contributions from (a) HRD and (b) signature 8 subdivided by race/ethnicity and IHC subtype. (c) Boxplots representing the number of SVs identified across 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.

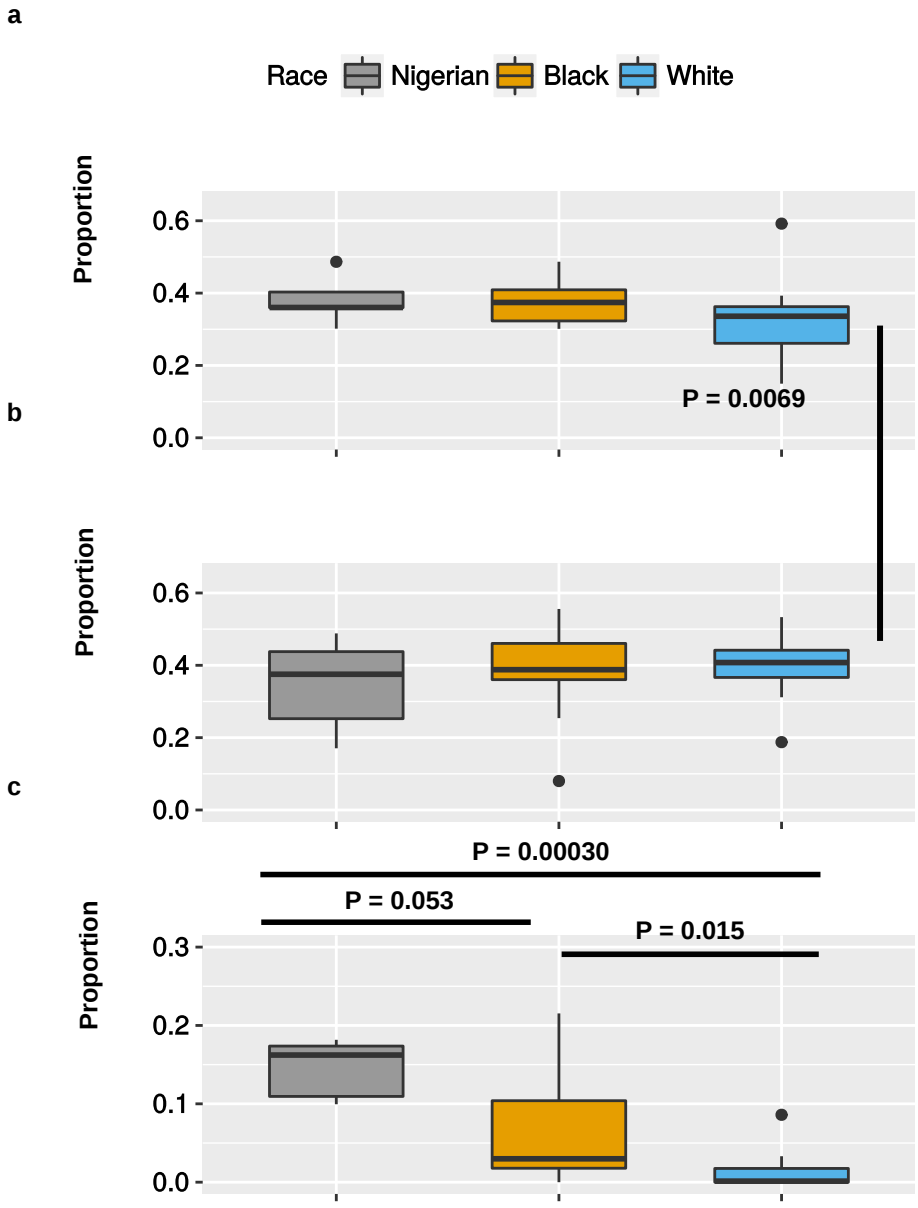

Figure 5.13: **Mutation signature contributions by race/ethnicity using WGS.** Boxplots of WGS signature 8 contributions for (a) HR+/HER2- and (b) HR-/HER2- malignancies. (c) The proportion of HRD signature within HR+/HER2- malignancies. Racial/ethnic differences across subtypes were assessed using Kruskal-Wallis 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 x 10^{-4}) and White (P = 1.6 x 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 [485] and basal-like breast cancers [474]. In this same set of genomes, Nigerians had more SVs than both Black (P = 0.03) and White cohorts (P = 2.8 x 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 x 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 *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 *CDH1* or *PIK3CA* mutations ("*CDH1/PIK3CA*"), *TP53* or *BRCA1/2* mutations ("*TP53/BRCA1/BRCA2*"), mutations from both aforementioned categories ("Both"), or mutations in neither of the aforementioned categories ("Neither"). APOBEC contributions were significantly higher in *CDH1/PIK3CA* compared to the *TP53/BRCA1/BRCA2* ($P = 1.8 \times 10^{-6}$, Dunn's test) and Neither ($P = 7.2 \times 10^{-9}$) groups. Tumors harboring mutations from both groups ("Both") had lower APOBEC contributions than *CDH1/PIK3CA* ($P = 0.11$), yet higher than *TP53/BRCA1/BRCA2* ($P = 5.0 \times 10^{-3}$) (Fig. 5.14b). In contrast, *TP53/BRCA1/BRCA2* had significantly higher HRD contributions than all other groups (P *CDH1/PIK3CA* $= 9.9 \times 10^{-15}$; Both $= 1.6 \times 10^{-4}$; Neither $= 2.1 \times 10^{-4}$), while *CDH1/PIK3CA* had significantly lower contributions than all other groups (P Both $= 3.5 \times 10^{-3}$; 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 *TP53/BRCA1/BRCA2* (Figure 5-14 a-c), suggesting there may be other mechanisms, such as *BRCA1/2* methylation [486], 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 — *TP53/BRCA1/BRCA2* and Neither — encompassed 70.3% of Nigerians and 66.3% of Blacks but only 47.7% of Whites ($P = 1.2 \times 10^{-3}$, Chi-squared test) (Fig. 5.14d). This suggests that individuals with African ancestry

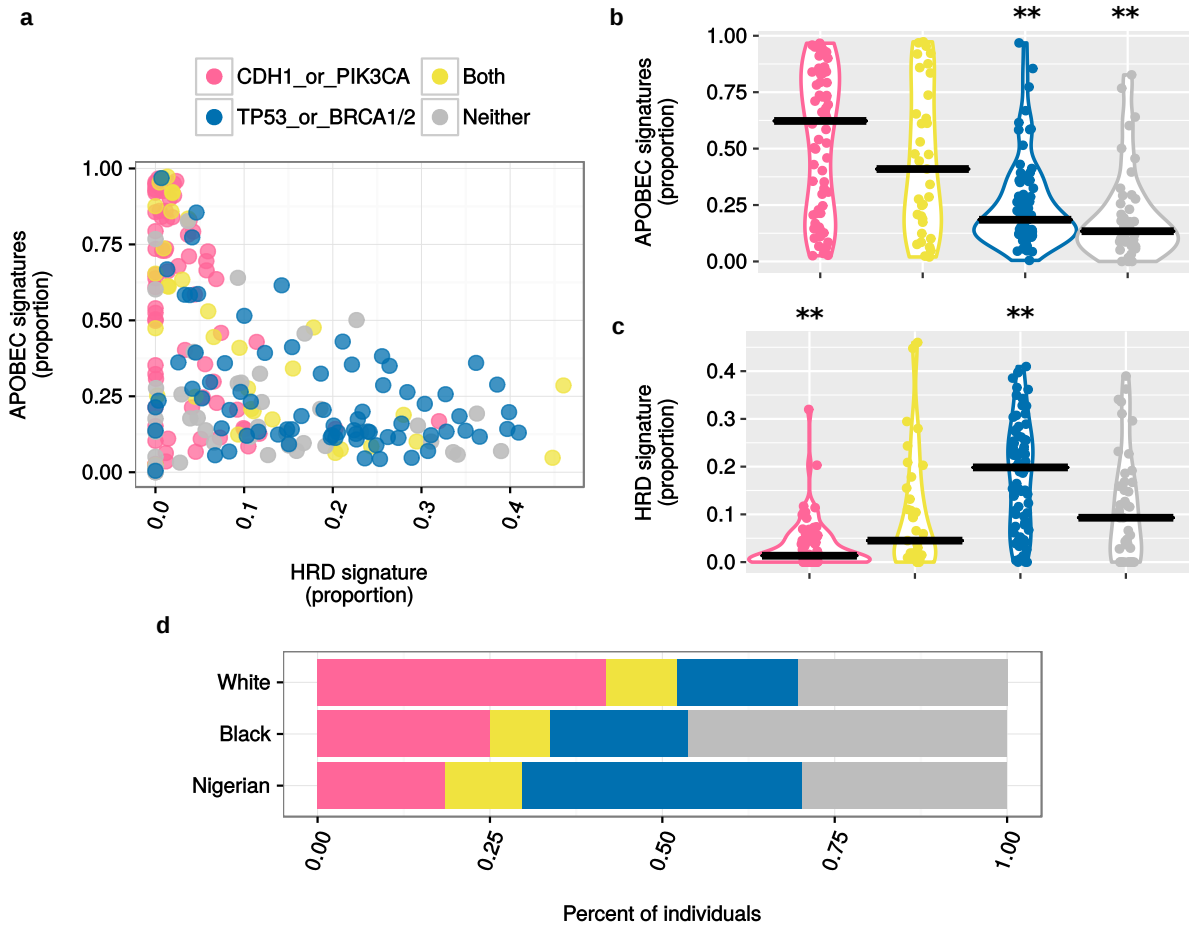

Figure 5.14: **Driver genes associate with APOBEC and HRD signature balance in HR+/HER2- breast cancer.** (a) For each malignancy, the proportion of APOBEC signatures (sum of APOBEC C>T and C>G) by the proportion of HRD is shown. Each patient is colored based on harboring a *CDH1* or *PIK3CA* mutation (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.

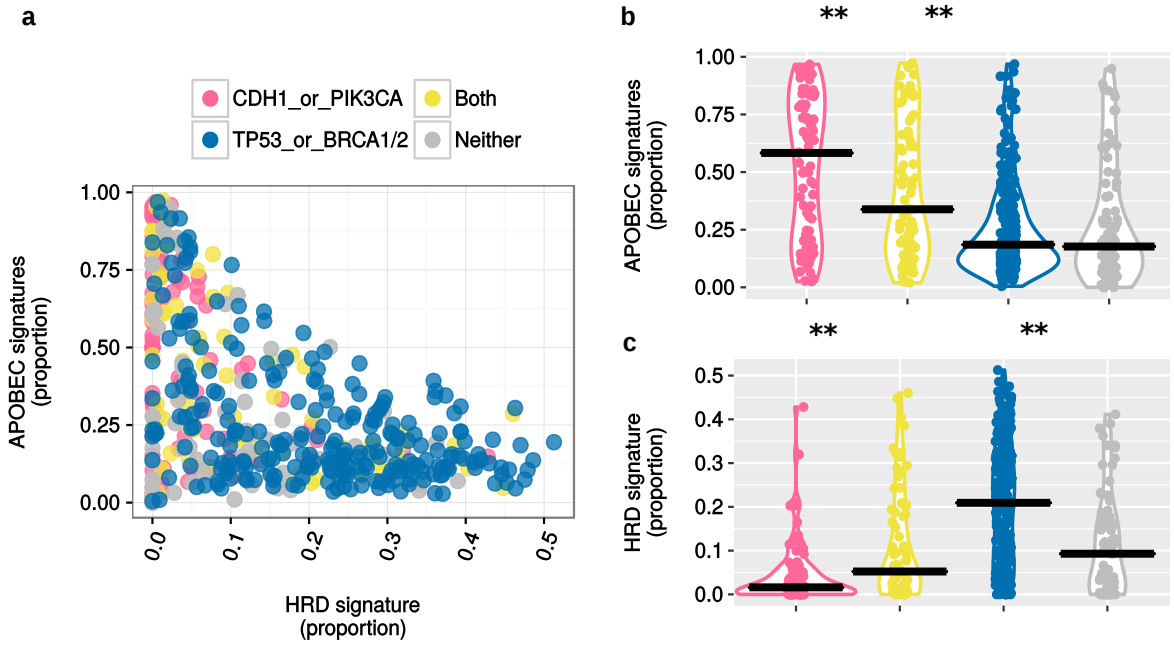

Figure 5.15: **Driver genes associate with APOBEC and HRD signature balance across all breast cancer IHC subtypes.** (a) For each malignancy, the proportion of APOBEC signatures (sum of APOBEC C>T and C>G) by the proportion of HRD is shown. Each patient is colored based on harboring a *CDH1* or *PIK3CA* 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. 5.12b and Fig. 5.13c). 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 $P < 1.0 \text{ x } 10^{-4}$, ANOVA). Racial differences adjusted for PAM50 subtype, however, were modest (Fig. 5.16b). The Cytotoxic cell signature ($P = 4.0 \text{ x } 10^{-3}$) 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-

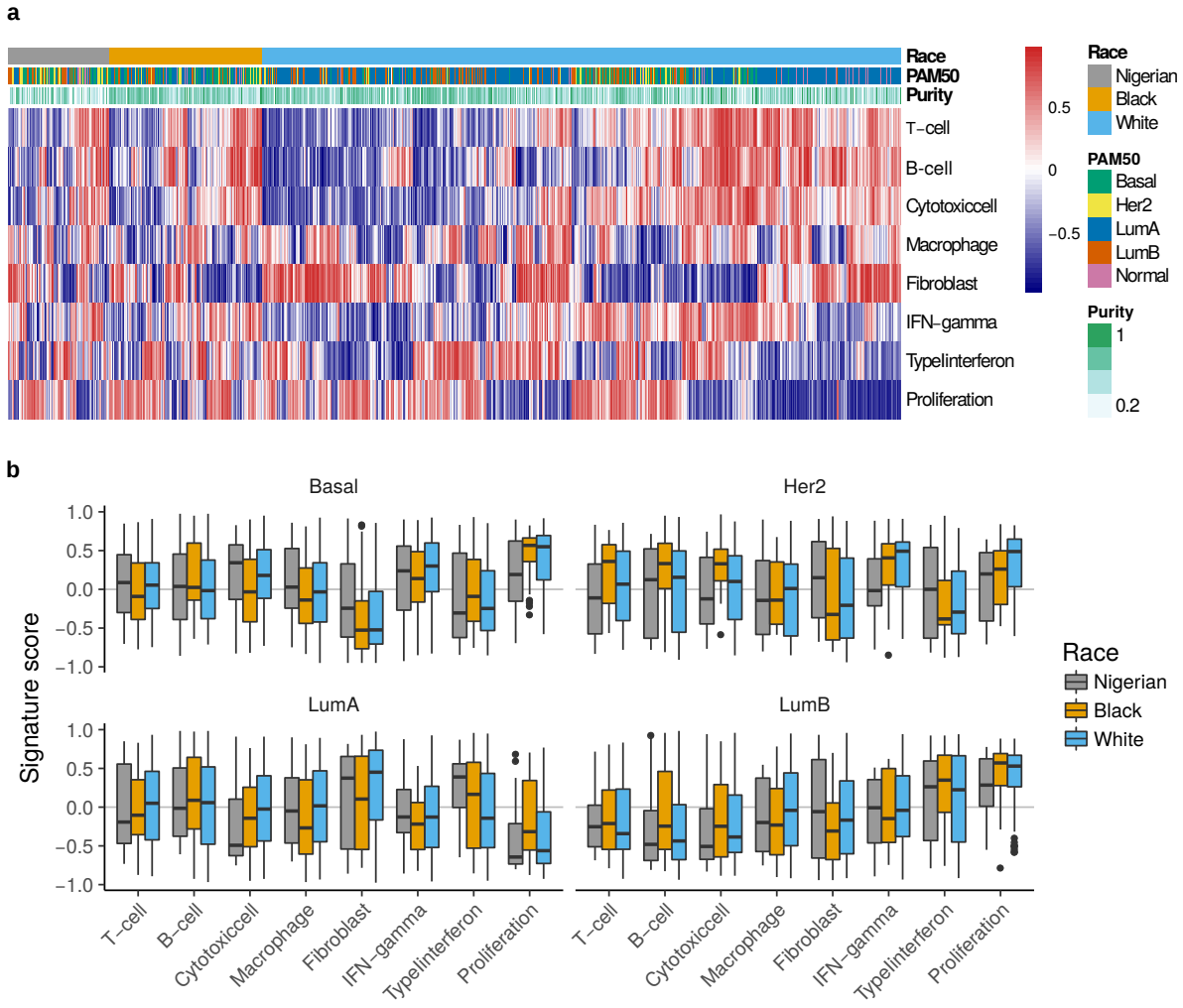

Figure 5.16: **Gene signatures of immune cell infiltration.** (a) Heatmap visualizing gene signature activation in all 1,040 patients with RNA-seq data from the combined Nigerian, Black and White cohort. High signature scores (red) indicate high overall expression of genes from a given signature, whereas 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).

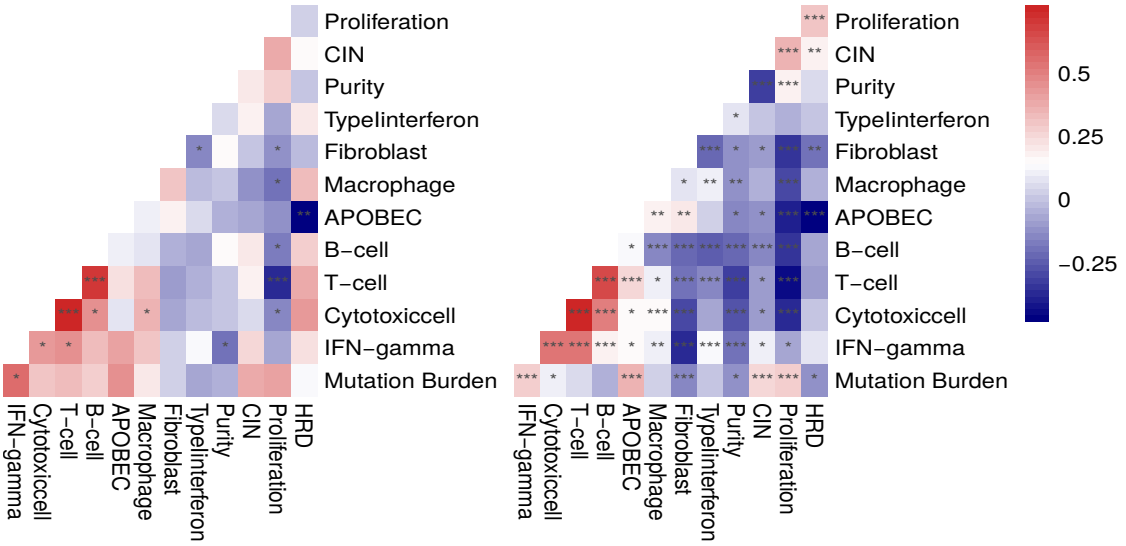

Figure 5.17: **Pairwise Pearson correlation of immune signatures and potential predictors of response to immunotherapy.** Potential predictors include APOBEC (C>T and C>G combined) and HRD signatures, chromosome instability (CIN), and mutation burden. The Nigerian data is shown in the leftmost panel 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 $P < 1.0 \times 10^{-4}$), and, consistent with recent reports, APOBEC contributions were associated with increased T-cell infiltration (Rho = 0.25, BH $P < 1.0 \times 10^{-4}$) [492]. Conversely, CIN positively correlated with mutation burden (Rho = 0.28, BH $P < 1.0 \times 10^{-4}$) 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 *PLK2* and *KDM6A* as novel significantly mutated genes in breast cancer, both with a propensity to be altered in HER2+ tumors. *PLK2* is a cell cycle regulator and presumed tumor suppressor, while *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]. *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 *KDM6A* [507]. In addition, *KDM6A* might be required for a luminal-to-basal phenotypic switch [508]. Depletion of *GPS2* has been shown to promote cell proliferation in MCF-7 breast cancer cell lines [509]. *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 as 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 asses 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 $P = 5.2 \times 10^{-14}$, Dunn's test) and those with other (non-E542K/E545K) *PIK3CA* mutations (BH $P = 9.2 \times 10^{-6}$). 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`

```
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

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
FASTQ1 --secondOut FASTQ2 --unpairedOut /dev/null 2>> LOG_FILE

bwa mem  -M -t 32 -R "READ_GROUP" /path/to/human_g1k_v37.fasta FASTQ1
FASTQ2 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. --noef 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 "hereditary" 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) [554] 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) [555] 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 $\geq$ 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:

$$y_i = \beta_0 + \beta_{burden} X_{burden_i} + \epsilon$$

Race adjusted:

$$y_i = \beta_0 + \beta_{burden} X_{burden_i} + \beta_{race} X_{race_i} + \epsilon$$

Cancer type adjusted:

$$y_i = \beta_0 + \beta_{burden} X_{burden_i} + \beta_{type} X_{type_i} + \epsilon$$

Race and cancer type adjusted:

$$y_i = \beta_0 + \beta_{burden} X_{burden_i} + \beta_{race} X_{race_i} + \beta_{type} X_{type_i} + \epsilon$$

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:

$$y_i = \beta_0 + \beta_{BRCA1} X_{BRCA1_i} + \beta_{BRCA2} X_{BRCA2_i} + \beta_{HighBurden} X_{HighBurden_i} + \epsilon$$

Race adjusted:

$$y_i = \beta_0 + \beta_{BRCA1} X_{BRCA1_i} + \beta_{BRCA2} X_{BRCA2_i} + \beta_{HighBurden} X_{HighBurden_i} + \beta_{race} X_{race_i} + \epsilon$$

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"

and

"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 a 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:

$$Odds_{TCGA} = \sum_i AC_{TCGA_i} / (\sum_i AN_{TCGA_i} \sum_i AC_{TCGA_i})$$

$$Odds_{Non-TCGA} = \sum_i AC_{Non-TCGA_i} / (\sum_i AN_{Non-TCGA_i} \sum_i AC_{Non-TCGA_i})$$

$$OR = Odds_{TCGA} / Odds_{Non-TCGA}$$

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 ($P \leq 5 \text{ x } 10^{-8}$), 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

24hrs. Following cisplatin incubation, the cisplatin was removed and growth media was added for 24hrs. Cells were then fixed with 4% paraformaldehyde in PBS +calcium/magnesium for 20min, quenched in 100mM NH4Cl for 10 minutes, then fixed in 0.5% Triton-X 100 for 15min. Cells were then blocked in 5% powdered milk in TBS-T (blotto) for 1hr. A rabbit primary antibody to yH2AX (Bethyl Laboratories A300-081A) and blotto were incubated on the cells overnight at 4C. Following washes, cells were incubated with Donkey Anti-Rabbit 488 (ABCAM ab150073) for 1hr 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 yH2AX staining within single cells. DNA damage was calculated as the yH2AX 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 yH2AX.

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 $\leq$ 0.005" AND "ExAC_AF $\leq$ 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) [577] and Lumpy Express (v0.2.10) [578]. 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 ($\geq$ 50% African ancestry), genomic White ($\geq$ 90% European ancestry), and genomic Asian ($\geq$ 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 previously 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. Permuted 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

[1] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 4 March 2011.

[2] Mary C White, Dawn M Holman, Jennifer E Boehm, Lucy A Peipins, Melissa Grossman, and S Jane Henley. Age and cancer risk: a potentially modifiable relationship. *Am. J. Prev. Med.*, 46(3 Suppl 1):S7–15, March 2014.

[3] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2017. *CA Cancer J. Clin.*, 67(1):7–30, January 2017.

[4] Freddie Bray, Ahmedin Jemal, Nathan Grey, Jacques Ferlay, and David Forman. Global cancer transitions according to the human development index (2008-2030): a population-based study. *Lancet Oncol.*, 13(8):790–801, August 2012.

[5] Merel Kimman, Rosana Norman, Stephen Jan, David Kingston, and Mark Woodward. The burden of cancer in member countries of the association of southeast asian nations (ASEAN). *Asian Pac. J. Cancer Prev.*, 13(2):411–420, 2012.

[6] Danny R Youlden, Susanna M Cramb, and Peter D Baade. The international epidemiology of lung cancer: geographical distribution and secular trends. *J. Thorac. Oncol.*, 3(8):819–831, August 2008.

[7] Rengaswamy Sankaranarayanan, Kunnambath Ramadas, and You-Lin Qiao. Managing the changing burden of cancer in asia. *BMC Med.*, 12:3, 8 January 2014.

[8] Rebecca L Siegel, Stacey A Fedewa, William F Anderson, Kimberly D Miller, Jiemin Ma, Philip S Rosenberg, and Ahmedin Jemal. Colorectal cancer incidence patterns in the united states, 1974-2013. *J. Natl. Cancer Inst.*, 109(8), 1 August 2017.

[9] C R Smittenaar, K A Petersen, K Stewart, and N Moitt. Cancer incidence and mortality projections in the UK until 2035. *Br. J. Cancer*, 115(9):1147–1155, 25 October 2016.

[10] Joshua D Schiffman, Paul G Fisher, and Peter Gibbs. Early detection of cancer: past, present, and future. *Am Soc Clin Oncol Educ Book*, pages 57–65, 2015.

[11] Richard M Hoffman, David Espey, and Robert L Rhyne. A public-health perspective on screening colonoscopy. *Expert Rev. Anticancer Ther.*, 11(4):561–569, April 2011.

[12] Doris Schopper and Chris de Wolf. How effective are breast cancer screening programmes by mammography? review of the current evidence. *Eur. J. Cancer*, 45(11):1916–1923, July 2009.

[13] Siang Yong Tan and Yvonne Tatsumura. George papanicolaou (1883-1962): Discoverer of the pap smear. *Singapore Med. J.*, 56(10):586–587, October 2015.

[14] S C Hiom. Diagnosing cancer earlier: reviewing the evidence for improving cancer survival. *Br. J. Cancer*, 112 Suppl 1:S1–5, 31 March 2015.

[15] Anthony B Miller, Claus Wall, Cornelia J Baines, Ping Sun, Teresa To, and Steven A Narod. Twenty five year follow-up for breast cancer incidence and mortality of the canadian national breast screening study: randomised screening trial. *BMJ*, 348:g366, 11 February 2014.

[16] Amit G Singal, Anjana Pillai, and Jasmin Tiro. Early detection, curative treatment, and survival rates for hepatocellular carcinoma surveillance in patients with cirrhosis: a meta-analysis. *PLoS Med.*, 11(4):e1001624, April 2014.

[17] Stacy Loeb, Marc A Bjurlin, Joseph Nicholson, Teuvo L Tammela, David F Penson, H Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *Eur. Urol.*, 65(6):1046–1055, June 2014.

[18] Eveline A M Heijnsdijk, Elisabeth M Wever, Anssi Auvinen, Jonas Hugosson, Stefano Ciatto, Vera Nelen, Maciej Kwiatkowski, Arnauld Villers, Alvaro Páez, Sue M Moss, Marco Zappa, Teuvo L J Tammela, Tuukka Mäkinen, Sigrid Carlsson, Ida J Korfage, Marie-Louise Essink-Bot, Suzie J Otto, Gerrit Draisma, Chris H Bangma, Monique J Roobol, Fritz H Schröder, and Harry J de Koning. Quality-of-life effects of prostate-specific antigen screening. *N. Engl. J. Med.*, 367(7):595–605, 16 August 2012.

[19] Robert A Smith, Kimberly S Andrews, Durado Brooks, Stacey A Fedewa, Deana Manassaram-Baptiste, Debbie Saslow, Otis W Brawley, and Richard C Wender. Cancer screening in the united states, 2017: A review of current american cancer society guidelines and current issues in cancer screening. *CA Cancer J. Clin.*, 67(2):100–121, March 2017.

[20] Hannah K Weir, Robert N Anderson, Sallyann M Coleman King, Ashwini Soman, Trevor D Thompson, Yuling Hong, Bjorn Moller, and Steven Leadbetter. Heart disease and cancer deaths - trends and projections in the united states, 1969-2020. *Prev. Chronic Dis.*, 13:E157, 17 November 2016.

[21] Sara Fernandes-Taylor, Taiwo Adesoye, and Joan R Bloom. Managing psychosocial issues faced by young women with breast cancer at the time of diagnosis and during active treatment. *Curr. Opin. Support. Palliat. Care*, 9(3):279–284, September 2015.

[22] Ruth Curtis, Annmarie Groarke, and Frank Sullivan. Stress and self-efficacy predict psychological adjustment at diagnosis of prostate cancer. *Sci. Rep.*, 4:5569, 4 July 2014.

[23] Samantha Hansford and David G Huntsman. Boveri at 100: Theodor boveri and genetic predisposition to cancer. *J. Pathol.*, 234(2):142–145, October 2014.

[24] A Balmain. Cancer genetics: from boveri and mendel to microarrays. *Nat. Rev. Cancer*, 1(1):77–82, October 2001.

[25] Gary A Koretzky. The legacy of the philadelphia chromosome. *J. Clin. Invest.*, 117(8):2030–2032, August 2007.

[26] P C Nowell and D A Hungerford. Chromosome studies on normal and leukemic human leukocytes. *J. Natl. Cancer Inst.*, 25:85–109, July 1960.

[27] J D Rowley. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405):290–293, 1 June 1973.

[28] Ruibao Ren. Mechanisms of BCR-ABL in the pathogenesis of chronic myelogenous leukaemia. *Nat. Rev. Cancer*, 5(3):172–183, March 2005.

[29] D Stehelin, H E Varmus, J M Bishop, and P K Vogt. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547):170–173, 11 March 1976.

[30] E S Lander, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, J Meldrim, J P Mesirov, C Miranda, W Morris, J Naylor, C Raymond, M Rosetti, R Santos, A Sheridan, C Sougnez, Y Stange-Thomann, N Stojanovic, A Subramanian, D Wyman, J Rogers, J Sulston, R Ainscough, S Beck, D Bentley, J Burton, C Clee, N Carter, A Coulson, R Deadman, P Deloukas, A Dunham, I Dunham, R Durbin, L French, D Grafham, S Gregory, T Hubbard, S Humphray, A Hunt, M Jones, C Lloyd, A McMurray, L Matthews, S Mercer, S Milne, J C Mullikin, A Mungall, R Plumb, M Ross, R Shownkeen, S Sims, R H Waterston, R K Wilson, L W Hillier, J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R

Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowki, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 15 February 2001.

[31] S Bamford, E Dawson, S Forbes, J Clements, R Pettett, A Dogan, A Flanagan, J Teague, P A Futreal, M R Stratton, and R Wooster. The COSMIC (catalogue of somatic mutations in cancer) database and website. *Br. J. Cancer*, 91(2):355–358, 19 July 2004.

[32] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gennady Denisov, Yuan Lin, Jeffrey R MacDonald, Andy Wing Chun Pang, Mary Shago, Timothy B Stockwell, Alexia Tsiamouri, Vineet Bafna, Vikas Bansal, Saul A Kravitz, Dana A Busam, Karen Y Beeson, Tina C McIntosh, Karin A Remington, Josep F Abril, John Gill, Jon Borman, Yu-Hui Rogers, Marvin E Frazier, Stephen W Scherer, Robert L Strausberg, and J Craig Venter. The diploid genome sequence of an individual human. *PLoS Biol.*, 5(10):e254, 4 September 2007.

[33] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, Xavier Gomes, Karrie Tartaro, Faheem Niazi, Cynthia L Turcotte, Gerard P Irzyk, James R Lupski, Craig Chinault, Xing-Zhi Song, Yue Liu, Ye Yuan, Lynne Nazareth, Xiang Qin, Donna M Muzny, Marcel Margulies, George M Weinstock, Richard A Gibbs, and Jonathan M Rothberg. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452(7189):872–876, 17 April 2008.

[34] Cristian Tomasetti, Luigi Marchionni, Martin A Nowak, Giovanni Parmigiani, and Bert Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proc. Natl. Acad. Sci. U. S. A.*, 112(1):118–123, 6 January 2015.

[35] Collin J Tokheim, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein, and Rachel Karchin. Evaluating the evaluation of cancer driver genes. *Proc. Natl. Acad. Sci. U. S. A.*, 113(50):14330–14335, 13 December 2016.

[36] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, Jr, and Kenneth W Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, 29 March 2013.

[37] Eva Y H P Lee and William J Muller. Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect. Biol.*, 2(10):a003236, October 2010.

[38] Todd W Miller. Initiating breast cancer by PIK3CA mutation. *Breast Cancer Res.*, 14(1):301, 7 February 2012.

[39] Emma R Cantwell-Dorris, John J O'Leary, and Orla M Sheils. BRAFV600E: implications for carcinogenesis and molecular therapy. *Mol. Cancer Ther.*, 10(3):385–394, March 2011.

[40] Cyriac Kandoth, Michael D McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, Qunyuan Zhang, Joshua F McMichael, Matthew A Wyczalkowski, Mark D M Leiserson, Christopher A Miller, John S Welch, Matthew J Walter, Michael C Wendl, Timothy J Ley, Richard K Wilson, Benjamin J Raphael, and Li Ding. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, 17 October 2013.

[41] A P Weng. Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia. *Science*, 306(5694):269–271, 2004.

[42] Teresa Davoli, Andrew Wei Xu, Kristen E Mengwasser, Laura M Sack, John C Yoon, Peter J Park, and Stephen J Elledge. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962, 7 November 2013.

[43] Collin Tokheim, Rohit Bhattacharya, Noushin Niknafs, Derek M Gygax, Rick Kim, Michael Ryan, David L Masica, and Rachel Karchin. Exome-Scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.*, 76(13):3719–3731, 1 July 2016.

[44] Helit Cohen, Rotem Ben-Hamo, Moriah Gidoni, Ilana Yitzhaki, Renana Kozol, Alona Zilberberg, and Sol Efroni. Shift in GATA3 functions, and GATA3 mutations, control progression and clinical presentation in breast cancer. *Breast Cancer Res.*, 16(6):464, 20 November 2014.

[45] Motoki Takaku, Sara A Grimm, and Paul A Wade. GATA3 in breast cancer: Tumor suppressor or oncogene? *Gene Expr.*, 16(4):163–168, 2015.

[46] Barbara Mair, Tomasz Konopka, Claudia Kerzendorfer, Katia Sleiman, Sejla Salic, Violeta Serra, Markus K Muellner, Vasiliki Theodorou, and Sebastian M B Nijman. Gain- and Loss-of-Function mutations in the breast cancer gene GATA3 result in differential drug sensitivity. *PLoS Genet.*, 12(9):e1006279, September 2016.

[47] Nicholas J Wang, Zachary Sanborn, Kelly L Arnett, Laura J Bayston, Wilson Liao, Charlotte M Proby, Irene M Leigh, Eric A Collisson, Patricia B Gordon, Lakshmi Jakkula, Sally Pennypacker, Yong Zou, Mimansa Sharma, Jeffrey P North, Swapna S Vemula, Theodora M Mauro, Isaac M Neuhaus, Philip E Leboit, Joe S Hur, Kyunghee

Park, Nam Huh, Pui-Yan Kwok, Sarah T Arron, Pierre P Massion, Allen E Bale, David Haussler, James E Cleaver, Joe W Gray, Paul T Spellman, Andrew P South, Jon C Aster, Stephen C Blacklow, and Raymond J Cho. Loss-of-function mutations in notch receptors in cutaneous and lung squamous cell carcinoma. *Proc. Natl. Acad. Sci. U. S. A.*, 108(43):17761–17766, 25 October 2011.

[48] Margaret S Ebert and Phillip A Sharp. Emerging roles for natural microRNA sponges. *Curr. Biol.*, 20(19):R858–61, 12 October 2010.

[49] Akihiro Fujimoto, Mayuko Furuta, Yasushi Totoki, Tatsuhiko Tsunoda, Mamoru Kato, Yuichi Shiraishi, Hiroko Tanaka, Hiroaki Taniguchi, Yoshiiku Kawakami, Masaki Ueno, Kunihito Gotoh, Shun-Ichi Ariizumi, Christopher P Wardell, Shinya Hayami, Toru Nakamura, Hiroshi Aikata, Koji Arihiro, Keith A Boroevich, Tetsuo Abe, Kaoru Nakano, Kazuhiro Maejima, Aya Sasaki-Oku, Ayako Ohsawa, Tetsuo Shibuya, Hiromi Nakamura, Natsuko Hama, Fumie Hosoda, Yasuhito Arai, Shoko Ohashi, Tomoko Urushidate, Genta Nagae, Shogo Yamamoto, Hiroki Ueda, Kenji Tatsuno, Hidenori Ojima, Nobuyoshi Hiraoka, Takuji Okusaka, Michiaki Kubo, Shigeru Marubashi, Terumasa Yamada, Satoshi Hirano, Masakazu Yamamoto, Hideki Ohdan, Kazuaki Shimada, Osamu Ishikawa, Hiroki Yamaue, Kazuki Chayama, Satoru Miyano, Hiroyuki Aburatani, Tatsuhiro Shibata, and Hidewaki Nakagawa. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.*, 48(5):500–509, May 2016.

[50] Vidisha Tripathi, Zhen Shen, Arindam Chakraborty, Sumanprava Giri, Susan M Freier, Xiaolin Wu, Yongqing Zhang, Myriam Gorospe, Supriya G Prasanth, Ashish Lal, and Kannanganattu V Prasanth. Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS Genet.*, 9(3):e1003368, March 2013.

[51] Rajesha Rupaimoole and Frank J Slack. MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. *Nat. Rev. Drug Discov.*, 16(3):203–222, March 2017.

[52] Masayuki Matsui and David R Corey. Non-coding RNAs as drug targets. *Nat. Rev. Drug Discov.*, 16(3):167–179, March 2017.

[53] Gwyn T Williams and Farzin Farzaneh. Are snoRNAs and snoRNA host genes new players in cancer? *Nat. Rev. Cancer*, 12(2):84–88, 19 January 2012.

[54] Kaiissar Mannoor, Jipei Liao, and Feng Jiang. Small nucleolar RNAs in cancer. *Biochim. Biophys. Acta*, 1826(1):121–128, August 2012.

[55] Zurab Siprashvili, Dan E Webster, Danielle Johnston, Rajani M Shenoy, Alexander J Ungewickell, Aparna Bhaduri, Ross Flockhart, Brian J Zarnegar, Yonglu Che, Francesca Meschi, Joseph D Puglisi, and Paul A Khavari. The noncoding RNAs SNORD50A and SNORD50B bind K-Ras and are recurrently deleted in human cancer. *Nat. Genet.*, 48(1):53–58, January 2016.

[56] Michael L Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46, January 2010.

[57] Janine Meienberg, Rémy Bruggmann, Konrad Oexle, and Gabor Matyas. Clinical sequencing: is WGS the better WES? *Hum. Genet.*, 135(3):359–362, March 2016.

[58] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nat. Rev. Genet.*, 12(11):745–755, 27 September 2011.

[59] Leslie G Biesecker and Robert C Green. Diagnostic clinical genome and exome sequencing. *N. Engl. J. Med.*, 370(25):2418–2425, 19 June 2014.

[60] Aziz Belkadi, Alexandre Bolze, Yuval Itan, Aurélie Cobat, Quentin B Vincent, Alexander Antipenko, Lei Shang, Bertrand Boisson, Jean-Laurent Casanova, and Laurent Abel. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci. U. S. A.*, 112(17):5473–5478, 28 April 2015.

[61] Gabrielle Bertier, Martin Hétu, and Yann Joly. Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. *BMC Med. Genomics*, 9(1):52, 11 August 2016.

[62] Ian R Watson, Koichi Takahashi, P Andrew Futreal, and Lynda Chin. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.*, 14(10):703–718, October 2013.

[63] Elaine R Mardis and Richard K Wilson. Cancer genome sequencing: a review. *Hum. Mol. Genet.*, 18(R2):R163–8, 15 October 2009.

[64] Julia R Pon and Marco A Marra. Driver and passenger mutations in cancer. *Annu. Rev. Pathol.*, 10:25–50, 2015.

[65] Kai Ye, Jiayin Wang, Reyka Jayasinghe, Eric-Wubbo Lameijer, Joshua F McMichael, Jie Ning, Michael D McLellan, Mingchao Xie, Song Cao, Venkata Yellapantula, Kuan-Lin Huang, Adam Scott, Steven Foltz, Beifang Niu, Kimberly J Johnson, Matthijs Moed, P Eline Slagboom, Feng Chen, Michael C Wendl, and Li Ding. Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.*, 22(1):97–104, January 2016.

[66] Benjamin J Raphael. Chapter 6: Structural variation and medical genomics. *PLoS Comput. Biol.*, 8(12):e1002821, 27 December 2012.

[67] Lixing Yang, Lovelace J Luquette, Nils Gehlenborg, Ruibin Xi, Psalm S Haseley, Chih-Heng Hsieh, Chengsheng Zhang, Xiaojia Ren, Alexei Protopopov, Lynda Chin, Raju Kucherlapati, Charles Lee, and Peter J Park. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153(4):919–929, 9 May 2013.

[68] Rameen Beroukhim, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T Mc Henry, Reid M Pinchback, Azra H Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S Lawrence, Barbara A Weir, Kumiko E Tanaka, Derek Y Chiang, Adam J Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J Kaye, Hidefumi Sasaki, Joel E Tepper, Jonathan A Fletcher, Josep Tabernero, José Baselga, Ming-Sound Tsao, Francesca Demichelis, Mark A Rubin, Pasi A Janne, Mark J Daly, Carmelo Nucera, Ross L Levine, Benjamin L Ebert, Stacey Gabriel, Anil K Rustgi, Cristina R Antonescu, Marc Ladanyi, Anthony Letai, Levi A Garraway, Massimo Loda, David G Beer, Lawrence D True, Aikou Okamoto, Scott L Pomeroy, Samuel Singer, Todd R Golub, Eric S Lander, Gad Getz, William R Sellers, and Matthew Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899–905, 18 February 2010.

[69] Darren J Burgess. Epigenetics. dissecting driving DNA methylations. *Nat. Rev. Cancer*, 12(7):448–449, 7 June 2012.

[70] Andrew P Feinberg, Michael A Koldobskiy, and Anita Göndör. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.*, 17(5):284–299, May 2016.

[71] Thomas Helleday, Saeed Eshtad, and Serena Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, 15(9):585–598, September 2014.

[72] Alberto Ciccia and Stephen J Elledge. The DNA damage response: making it safe to play with knives. *Mol. Cell*, 40(2):179–204, 22 October 2010.

[73] David J Araten, David W Golde, Rong H Zhang, Howard T Thaler, Lucia Gargiulo, Rosario Notaro, and Lucio Luzzatto. A quantitative measurement of the human somatic mutation rate. *Cancer Res.*, 65(18):8111–8117, 15 September 2005.

[74] Cristian Tomasetti, Bert Vogelstein, and Giovanni Parmigiani. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci. U. S. A.*, 110(6):1999–2004, 5 February 2013.

[75] Josef Jiricny. The multifaceted mismatch-repair system. *Nat. Rev. Mol. Cell Biol.*, 7(5):335–346, May 2006.

[76] T Lindahl and D E Barnes. Repair of endogenous DNA damage. *Cold Spring Harb. Symp. Quant. Biol.*, 65:127–133, 2000.

[77] Deborah E Barnes and Tomas Lindahl. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu. Rev. Genet.*, 38:445–476, 2004.

[78] Jurgen A Marteijn, Hannes Lans, Wim Vermeulen, and Jan H J Hoeijmakers. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell Biol.*, 15(7):465–481, July 2014.

[79] Orlando D Schärer. Nucleotide excision repair in eukaryotes. *Cold Spring Harb. Perspect. Biol.*, 5(10):a012609, 1 October 2013.

[80] Irene Kamileri, Ismene Karakasilioti, and George A Garinis. Nucleotide excision repair: new tricks with old bricks. *Trends Genet.*, 28(11):566–573, November 2012.

[81] Anuja Mehta and James E Haber. Sources of DNA double-strand breaks and models of recombinational DNA repair. *Cold Spring Harb. Perspect. Biol.*, 6(9):a016428, 7 August 2014.

[82] J Ross Chapman, Martin R G Taylor, and Simon J Boulton. Playing the end game: DNA double-strand break repair pathway choice. *Mol. Cell*, 47(4):497–510, 24 August 2012.

[83] Michael R Lieber. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.*, 79:181–211, 2010.

[84] Aaron R Quinlan and Ira M Hall. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet.*, 28(1):43–53, January 2012.

[85] Romualdo Benigni and Cecilia Bossa. Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology. *Chem. Rev.*, 111(4):2507–2536, 13 April 2011.

[86] Kahlin Cheung-Ong, Guri Giaever, and Corey Nislow. DNA-damaging agents in cancer chemotherapy: serendipity and chemical biology. *Chem. Biol.*, 20(5):648–659, 23 May 2013.

[87] Centers for Disease Control and Prevention (US), National Center for Chronic Disease Prevention and Health Promotion (US), and Office on Smoking and Health (US). *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General*. Centers for Disease Control and Prevention (US), Atlanta (GA), 1 April 2011.

[88] Gerd P Pfeifer, Mikhail F Denissenko, Magali Olivier, Natalia Tretyakova, Stephen S Hecht, and Pierre Hainaut. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, 21(48):7435–7451, 21 October 2002.

[89] M F Denissenko, A Pao, M Tang, and G P Pfeifer. Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science*, 274(5286):430–432, 18 October 1996.

[90] E Eisenstadt, A J Warren, J Porter, D Atkins, and J H Miller. Carcinogenic epoxides of benzo[a]pyrene and cyclopenta[cd]pyrene induce base substitutions via specific transversions. *Proc. Natl. Acad. Sci. U. S. A.*, 79(6):1945–1949, March 1982.

[91] Steven A Roberts, Michael S Lawrence, Leszek J Klimczak, Sara A Grimm, David Fargo, Petar Stojanov, Adam Kiezun, Gregory V Kryukov, Scott L Carter, Gordon Saksena, Shawn Harris, Ruchir R Shah, Michael A Resnick, Gad Getz, and Dmitry A Gordenin. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.*, 45(9):970–976, September 2013.

[92] Michael B Burns, Nuri A Temiz, and Reuben S Harris. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.*, 45(9):977–983, September 2013.

[93] Charles Swanton, Nicholas McGranahan, Gabriel J Starrett, and Reuben S Harris. APOBEC enzymes: Mutagenic fuel for cancer evolution and heterogeneity. *Cancer Discov.*, 5(7):704–712, July 2015.

[94] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, 3(1):246–259, 31 January 2013.

[95] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Marcin Imielinsk, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 22 August 2013.

[96] Ludmil B Alexandrov, Young Seok Ju, Kerstin Haase, Peter Van Loo, Iñigo Martincorena, Serena Nik-Zainal, Yasushi Totoki, Akihiro Fujimoto, Hidewaki Nakagawa, Tatsuhiro Shibata, Peter J Campbell, Paolo Vineis, David H Phillips, and Michael R Stratton. Mutational signatures associated with tobacco smoking in human cancer. *Science*, 354(6312):618–622, 4 November 2016.

[97] Simon N Powell and Lisa A Kachnic. Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. *Oncogene*, 22(37):5784–5791, 1 September 2003.

[98] Junran Zhang. The role of BRCA1 in homologous recombination repair in response to replication stress: significance in tumorigenesis and cancer therapy. *Cell Biosci.*, 3(1):11, 6 February 2013.

[99] Christine S Walsh. Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. *Gynecol. Oncol.*, 137(2):343–350, May 2015.

[100] Rohit Prakash, Yu Zhang, Weiran Feng, and Maria Jasin. Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb. Perspect. Biol.*, 7(4):a016600, 1 April 2015.

[101] Tenghui Chen, Zixing Wang, Wanding Zhou, Zechen Chong, Funda Meric-Bernstam, Gordon B Mills, and Ken Chen. Hotspot mutations delineating diverse mutational signatures and biological utilities across cancer types. *BMC Genomics*, 17 Suppl 2:394, 23 June 2016.

[102] Nnennaya Kanu, Maria Antonietta Cerone, Gerald Goh, Lykourgos-Panagiotis Zalmas, Jirina Bartkova, Michelle Dietzen, Nicholas McGranahan, Rebecca Rogers, Emily K Law, Irina Gromova, Maik Kschischo, Michael I Walton, Olivia W Rossanese, Jiri Bartek, Reuben S Harris, Subramanian Venkatesan, and Charles Swanton. DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. *Genome Biol.*, 17(1):185, 15 September 2016.

[103] Nazneen Rahman. Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–308, 2014.

[104] David E Anderson. Genetic study of breast cancer: Identification of a high risk group. *Cancer*, 34(4):1090–1097, 1974.

[105] Brian MacMahon, Philip Cole, and James Brown. Etiology of human breast cancer: A review2. *JNCI: Journal of the National Cancer Institute*, 50(1):21–42, 1 January 1973.

[106] Henry T Lynch and Anne J Krush. Cancer family "g" revisited: 1895-1970. *Cancer*, 27(6):1505–1511, 1971.

[107] H T Lynch. Hereditary factors in cancer: Study of two large midwestern kindreds. *Arch. Intern. Med.*, 117(2):206, 1 February 1966.

[108] Y K Fung, A L Murphree, A T'Ang, J Qian, S H Hinrichs, and W F Benedict. Structural evidence for the authenticity of the human retinoblastoma gene. *Science*, 236(4809):1657–1661, 26 June 1987.

[109] A Antoniou, P D P Pharoah, S Narod, H A Risch, J E Eyfjord, J L Hopper, N Loman, H Olsson, O Johannsson, A Borg, B Pasini, P Radice, S Manoukian, D M Eccles, N Tang, E Olah, H Anton-Culver, E Warner, J Lubinski, J Gronwald, B Gorski, H Tulinius, S Thorlacius, H Eerola, H Nevanlinna, K Syrjäkoski, O-P Kallioniemi, D Thompson, C Evans, J Peto, F Lalloo, D G Evans, and D F Easton. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am. J. Hum. Genet.*, 72(5):1117–1130, May 2003.

[110] Valentina Silvestri, Daniel Barrowdale, Anna Marie Mulligan, Susan L Neuhausen, Stephen Fox, Beth Y Karlan, Gillian Mitchell, Paul James, Darcy L Thull, Kristin K Zorn, Natalie J Carter, Katherine L Nathanson, Susan M Domchek, Timothy R Rebbeck, Susan J Ramus, Robert L Nussbaum, Olufunmilayo I Olopade, Johanna Rantala, Sook-Yee Yoon, Maria A Caligo, Laura Spugnesi, Anders Bojesen, Inge Sokilde Pedersen, Mads Thomassen, Uffe Birk Jensen, Amanda Ewart Toland, Leigha Senter, Irene L Andrulis, Gord Glendon, Peter J Hulick, Evgeny N Imyanitov, Mark H Greene, Phuong L Mai, Christian F Singer, Christine Rappaport-Fuerhauser, Gero Kramer, Joseph Vijai, Kenneth Offit, Mark Robson, Anne Lincoln, Lauren Jacobs, Eva Machackova, Lenka Foretova, Marie Navratilova, Petra Vasickova, Fergus J Couch, Emily Hallberg, Kathryn J Ruddy, Priyanka Sharma, Sung-Won Kim, kConFab Investigators, Manuel R Teixeira, Pedro Pinto, Marco Montagna, Laura Matricardi, Adalgeir Arason, Oskar Th Johannsson, Rosa B Barkardottir, Anna Jakubowska, Jan Lubinski, Angel Izquierdo, Miguel Angel Pujana, Judith Balmaña, Orland Diez, Gabriella Ivady, Janos Papp, Edith Olah, Ava Kwong, Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), Heli Nevanlinna, Kristiina Aittomäki, Pedro Perez Segura, Trinidad Caldes, Tom Van Maerken, Bruce Poppe, Kathleen B M Claes, Claudine Isaacs, Camille Elan, Christine Lasset, Dominique Stoppa-Lyonnet, Laure Barjhoux, Muriel Belotti, Alfons Meindl, Andrea Gehrig, Christian Sutter, Christoph Engel, Dieter Niederacher, Doris Steinemann, Eric Hahnen, Karin Kast, Norbert Arnold, Raymonda Varon-Mateeva, Dorothea Wand, Andrew K Godwin, D Gareth Evans, Debra Frost, Jo Perkins, Julian Adlard, Louise Izatt, Radka Platte, Ros Eeles, Steve Ellis, EMBRACE, Ute Hamann, Judy Garber, Florentia Fostira, George Fountzilas, Barbara Pasini, Giuseppe Giannini, Piera Rizzolo, Antonio Russo, Laura Cortesi, Laura Papi, Liliana Varesco, Domenico Palli, Ines Zanna, Antonella Savarese, Paolo Radice, Siranoush Manoukian, Bernard Peissel, Monica Barile, Bernardo Bonanni, Alessandra Viel, Valeria Pensotti, Stefania Tommasi, Paolo Peterlongo, Jeffrey N Weitzel, Ana Osorio, Javier Benitez, Lesley McGuffog, Sue Healey, Anne-Marie Gerdes, Bent Ejlertsen, Thomas V O Hansen, Linda Steele, Yuan Chun Ding, Nadine Tung, Ramunas Janavicius, David E Goldgar, Saundra S Buys, Mary B Daly, Anita Bane, Mary Beth Terry, Esther M John, Melissa Southey, Douglas F Easton, Georgia Chenevix-Trench, Antonis C Antoniou, and Laura Ottini. Male breast cancer in BRCA1 and BRCA2 mutation carriers: pathology data from the consortium of investigators of modifiers of BRCA1/2. *Breast Cancer Res.*, 18(1):15, 9 February 2016.

[111] Yu Chuan Tai, Susan Domchek, Giovanni Parmigiani, and Sining Chen. Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.*, 99(23):1811–1814, 5 December 2007.

[112] Elaine A Ostrander and Miriam S Udler. The role of the BRCA2 gene in susceptibility to prostate cancer revisited. *Cancer Epidemiol. Biomarkers Prev.*, 17(8):1843–1848, August 2008.

[113] Yosef Shiloh and Yael Ziv. The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. *Nat. Rev. Mol. Cell Biol.*, 14(4):197–210, April 2013.

[114] Michael Choi, Thomas Kipps, and Razelle Kurzrock. ATM mutations in cancer: Therapeutic implications. *Mol. Cancer Ther.*, 15(8):1781–1791, August 2016.

[115] M B Kastan and D S Lim. The many substrates and functions of ATM. *Nat. Rev. Mol. Cell Biol.*, 1(3):179–186, December 2000.

[116] Charlotte Näslund-Koch, Børge G Nordestgaard, and Stig E Bojesen. Increased risk for other cancers in addition to breast cancer for CHEK2*1100delC heterozygotes estimated from the copenhagen general population study. *J. Clin. Oncol.*, 34(11):1208–1216, 10 April 2016.

[117] H Nevanlinna and J Bartek. The CHEK2 gene and inherited breast cancer susceptibility. *Oncogene*, 25(43):5912–5919, 25 September 2006.

[118] Kate A McBride, Mandy L Ballinger, Emma Killick, Judy Kirk, Martin H N Tattersall, Rosalind A Eeles, David M Thomas, and Gillian Mitchell. Li-Fraumeni syndrome: cancer risk assessment and clinical management. *Nat. Rev. Clin. Oncol.*, 11(5):260–271, May 2014.

[119] Valérie Bonadona, Bernard Bonaïti, Sylviane Olschwang, Sophie Grandjouan, Laetitia Huiart, Michel Longy, Rosine Guimbaud, Bruno Buecher, Yves-Jean Bignon, Olivier Caron, Chrystelle Colas, Catherine Noguès, Sophie Lejeune-Dumoulin, Laurence Olivier-Faivre, Florence Polycarpe-Osaer, Tan Dat Nguyen, Françoise Desseigne, Jean-Christophe Saurin, Pascaline Berthet, Dominique Leroux, Jacqueline Duffour, Sylvie Manouvrier, Thierry Frébourg, Hagay Sobol, Christine Lasset, Catherine Bonaïti-Pellié, and French Cancer Genetics Network. Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in lynch syndrome. *JAMA*, 305(22):2304–2310, 8 June 2011.

[120] Henry T Lynch, Carrie L Snyder, Trudy G Shaw, Christopher D Heinen, and Megan P Hitchins. Milestones of lynch syndrome: 1895-2015. *Nat. Rev. Cancer*, 15(3):181–194, March 2015.

[121] H T Lynch, P M Lynch, S J Lanspa, C L Snyder, J F Lynch, and C R Boland. Review of the lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin. Genet.*, 76(1):1–18, July 2009.

[122] Fay Kastrinos and Sapna Syngal. Recently identified colon cancer predispositions: MYH and MSH6 mutations. *Semin. Oncol.*, 34(5):418–424, October 2007.

[123] Mark A Jenkins, Marina E Croitoru, Neerav Monga, Sean P Cleary, Michelle Cotterchio, John L Hopper, and Steven Gallinger. Risk of colorectal cancer in monoallelic and biallelic carriers of MYH mutations: a population-based case-family study. *Cancer Epidemiol. Biomarkers Prev.*, 15(2):312–314, February 2006.

[124] M D Gray, J C Shen, A S Kamath-Loeb, A Blank, B L Sopher, G M Martin, J Oshima, and L A Loeb. The werner syndrome protein is a DNA helicase. *Nat. Genet.*, 17(1):100–103, September 1997.

[125] Richarda M de Voer, Marc-Manuel Hahn, Arjen R Mensenkamp, Alexander Hoischen, Christian Gilissen, Arjen Henkes, Liesbeth Spruijt, Wendy A van Zelst-Stams, C Marleen Kets, Eugene T Verwiel, Iris D Nagtegaal, Hans K Schackert, Ad Geurts van Kessel, Nicoline Hoogerbrugge, Marjolijn J L Ligtenberg, and Roland P Kuiper. Deleterious germline BLM mutations and the risk for early-onset colorectal cancer. *Sci. Rep.*, 5:14060, 11 September 2015.

[126] Stephen B Gruber, Nathan A Ellis, Karen K Scott, Ronit Almog, Prema Kolachana, Joseph D Bonner, Tomas Kirchhoff, Lynn P Tomsho, Khedoudja Nafa, Heather Pierce, Marcelo Low, Jaya Satagopan, Hedy Rennert, Helen Huang, Joel K Greenson, Joanna Groden, Beth Rapaport, Jinru Shia, Stephen Johnson, Peter K Gregersen, Curtis C Harris, Jeff Boyd, Gad Rennert, and Kenneth Offit. BLM heterozygosity and the risk of colorectal cancer. *Science*, 297(5589):2013, 20 September 2002.

[127] Martijn F Lutke Holzik, Rolf H Sijmons, Josette Ehm Hoekstra-Weebers, Dirk T Sleijfer, and Harald J Hoekstra. Clinical and genetic aspects of testicular germ cell tumours. *Hered. Cancer Clin. Pract.*, 6(1):3–14, 15 February 2008.

[128] Yang Liu, Lizhong Wang, and Pan Zheng. X-linked tumor suppressors: perplexing inheritance, a unique therapeutic opportunity. *Trends Genet.*, 26(6):260–265, June 2010.

[129] A G Knudson. Mutation and cancer: Statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

[130] S H Friend, R Bernards, S Rogelj, R A Weinberg, J M Rapaport, D M Albert, and T P Dryja. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, 323(6089):643–646, 1986.

[131] I P Tomlinson, R Roylance, and R S Houlston. Two hits revisited again. *J. Med. Genet.*, 38(2):81–85, February 2001.

[132] Georgina L Ryland, Maria A Doyle, David Goode, Samantha E Boyle, David Y H Choong, Simone M Rowley, Jason Li, Australian Ovarian Cancer Study Group, David D L Bowtell, Richard W Tothill, Ian G Campbell, and Kylie L Gorringe. Loss of heterozygosity: what is it good for? *BMC Med. Genomics*, 8:45, 1 August 2015.

[133] Nazneen Rahman and Richard H Scott. Cancer genes associated with phenotypes in monoallelic and biallelic mutation carriers: new lessons from old players. *Hum. Mol. Genet.*, 16 Spec No 1:R60–6, 15 April 2007.

[134] Krishna L Kanchi, Kimberly J Johnson, Charles Lu, Michael D McLellan, Mark D M Leiserson, Michael C Wendl, Qunyuan Zhang, Daniel C Koboldt, Mingchao Xie, Cyriac Kandoth, Joshua F McMichael, Matthew A Wyczalkowski, David E Larson, Heather K Schmidt, Christopher A Miller, Robert S Fulton, Paul T Spellman, Elaine R Mardis, Todd E Druley, Timothy A Graubert, Paul J Goodfellow, Benjamin J Raphael, Richard K Wilson, and Li Ding. Integrated analysis of germline and somatic variants in ovarian cancer. *Nat. Commun.*, 5:3156, 2014.

[135] Charles Lu, Mingchao Xie, Michael C Wendl, Jiayin Wang, Michael D McLellan, Mark D M Leiserson, Kuan-Lin Huang, Matthew A Wyczalkowski, Reyka Jayasinghe, Tapahsama Banerjee, Jie Ning, Piyush Tripathi, Qunyuan Zhang, Beifang Niu, Kai Ye, Heather K Schmidt, Robert S Fulton, Joshua F McMichael, Prag Batra, Cyriac Kandoth, Maheetha Bharadwaj, Daniel C Koboldt, Christopher A Miller, Krishna L Kanchi, James M Eldred, David E Larson, John S Welch, Ming You, Bradley A Ozenberger, Ramaswamy Govindan, Matthew J Walter, Matthew J Ellis, Elaine R Mardis, Timothy A Graubert, John F Dipersio, Timothy J Ley, Richard K Wilson, Paul J Goodfellow, Benjamin J Raphael, Feng Chen, Kimberly J Johnson, Jeffrey D Parvin, and Li Ding. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat. Commun.*, 6:10086, 22 December 2015.

[136] A M Hanby, D P Kelsell, H W Potts, C E Gillett, D T Bishop, N K Spurr, and D M Barnes. Association between loss of heterozygosity of BRCA1 and BRCA2 and morphological attributes of sporadic breast cancer. *Int. J. Cancer*, 88(2):204–208, 15 October 2000.

[137] Imogen Locke, Zsofia Kote-Jarai, Elizabeth Bancroft, Sarah Bullock, Sarah Jugurnauth, Peter Osin, Ashutosh Nerurkar, Louise Izatt, Gabriella Pichert, Gerald P H Gui, and Rosalind A Eeles. Loss of heterozygosity at the BRCA1 and BRCA2 loci detected in ductal lavage fluid from BRCA gene mutation carriers and controls. *Cancer Epidemiol. Biomarkers Prev.*, 15(7):1399–1402, July 2006.

[138] Jose M Silva, Rocio Gonzalez, Mariano Provencio, Gema Dominguez, Jose M Garcia, Isabel Gallego, Jose Palacios, Pilar España, and Felix Bonilla. Loss of heterozygosity in BRCA1 and BRCA2 markers and highgrade malignancy in breast cancer. *Breast Cancer Res. Treat.*, 53(1):9–17, January 1999.

[139] P G Rio, D Pernin, J O Bay, E Albuisson, F Kwiatkowski, M De Latour, D J Bernard-Gallon, and Y J Bignon. Loss of heterozygosity of BRCA1, BRCA2 and ATM genes in sporadic invasive ductal breast carcinoma. *Int. J. Oncol.*, 13(4):849–853, October 1998.

[140] Annegien Broeks, Jos H M Urbanus, Arno N Floore, Ellen C Dahler, Jan G M Klijn, Emiel J Th Rutgers, Peter Devilee, Nicola S Russell, Flora E van Leeuwen, and Laura J van't Veer. ATM-Heterozygous germline mutations contribute to breast Cancer–Susceptibility. *Am. J. Hum. Genet.*, 66(2):494–500, February 2000.

[141] 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.

[142] K M Shannon, P O'Connell, G A Martin, D Paderanga, K Olson, P Dinndorf, and F McCormick. Loss of the normal NF1 allele from the bone marrow of children with type 1 neurofibromatosis and malignant myeloid disorders. *N. Engl. J. Med.*, 330(9):597–601, 3 March 1994.

[143] G S Dite, M A Jenkins, M C Southey, J S Hocking, G G Giles, M R E McCredie, D J Venter, and J L Hopper. Familial risks, Early-Onset breast cancer, and BRCA1 and BRCA2 germline mutations. *JNCI Journal of the National Cancer Institute*, 95(6):448–457, 2003.

[144] A Cassidy, J P Myles, S W Duffy, T Liloglou, and J K Field. Family history and risk of lung cancer: age-at-diagnosis in cases and first-degree relatives. *Br. J. Cancer*, 95(9):1288–1290, 6 November 2006.

[145] Rachel Pearlman, Wendy L Frankel, Benjamin Swanson, Weiqiang Zhao, Ahmet Yilmaz, Kristin Miller, Jason Bacher, Christopher Bigley, Lori Nelsen, Paul J Goodfellow, Richard M Goldberg, Electra Paskett, Peter G Shields, Jo L Freudenheim, Peter P Stanich, Ilene Lattimer, Mark Arnold, Sandya Liyanarachchi, Matthew Kalady, Brandie Heald, Carla Greenwood, Ian Paquette, Marla Prues, David J Draper, Carolyn Lindeman, J Philip Kuebler, Kelly Reynolds, Joanna M Brell, Amy A Shaper, Sameer Mahesh, Nicole Buie, Kisa Weeman, Kristin Shine, Mitchell Haut, Joan Edwards, Shyamal Bastola, Karen Wickham, Karamjit S Khanduja, Rosemary Zacks, Colin C Pritchard, Brian H Shirts, Angela Jacobson, Brian Allen, Albert de la Chapelle, Heather Hampel, and Ohio Colorectal Cancer Prevention Initiative Study Group. Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with Early-Onset colorectal cancer. *JAMA Oncol*, 3(4):464–471, 1 April 2017.

[146] Jennifer K Litton, Kaylene Ready, Huiqin Chen, Angelica Gutierrez-Barrera, Carol J Etzel, Funda Meric-Bernstam, Ana M Gonzalez-Angulo, Huong Le-Petross, Karen Lu, Gabriel N Hortobagyi, and Banu K Arun. Earlier age of onset of BRCA mutation-related cancers in subsequent generations. *Cancer*, 118(2):321–325, 15 January 2012.

[147] 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

BRCA2. australian breast cancer family study. *Cancer Epidemiol. Biomarkers Prev.*, 8(9):741–747, September 1999.

[148] D F Easton, D Ford, and D T Bishop. Breast and ovarian cancer incidence in BRCA1-mutation carriers. breast cancer linkage consortium. *Am. J. Hum. Genet.*, 56(1):265–271, January 1995.

[149] Mandy L Ballinger, David L Goode, Isabelle Ray-Coquard, Paul A James, Gillian Mitchell, Eveline Niedermayr, Ajay Puri, Joshua D Schiffman, Gillian S Dite, Arcadi Cipponi, Robert G Maki, Andrew S Brohl, Ola Myklebost, Eva W Stratford, Susanne Lorenz, Sung-Min Ahn, Jin-Hee Ahn, Jeong Eun Kim, Sue Shanley, Victoria Beshay, Robert Lor Randall, Ian Judson, Beatrice Seddon, Ian G Campbell, Mary-Anne Young, Rajiv Sarin, Jean-Yves Blay, Seán I O'Donoghue, and David M Thomas. Monogenic and polygenic determinants of sarcoma risk: an international genetic study. *Lancet Oncol.*, 17(9):1261–1271, September 2016.

[150] Hannes Helgason, Thorunn Rafnar, Halla S Olafsdottir, Jon G Jonasson, Asgeir Sigurdsson, Simon N Stacey, Adalbjorg Jonasdottir, Laufey Tryggvadottir, Kristin Alexiusdottir, Asgeir Haraldsson, Louise le Roux, Julius Gudmundsson, Hrefna Johannsdottir, Asmundur Oddsson, Arnaldur Gylfason, Olafur T Magnusson, Gisli Masson, Thorvaldur Jonsson, Halla Skuladottir, Daniel F Gudbjartsson, Unnur Thorsteinsdottir, Patrick Sulem, and Kari Stefansson. Loss-of-function variants in ATM confer risk of gastric cancer. *Nat. Genet.*, 47(8):906–910, August 2015.

[151] Timothy R Rebbeck, Nandita Mitra, Fei Wan, Olga M Sinilnikova, Sue Healey, Lesley McGuffog, Sylvie Mazoyer, Georgia Chenevix-Trench, Douglas F Easton, Antonis C Antoniou, Katherine L Nathanson, CIMBA Consortium, Yael Laitman, Anya Kushnir, Shani Paluch-Shimon, Raanan Berger, Jamal Zidan, Eitan Friedman, Hans Ehrencrona, Marie Stenmark-Askmalm, Zakaria Einbeigi, Niklas Loman, Katja Harbst, Johanna Rantala, Beatrice Melin, Dezheng Huo, Olufunmilayo I Olopade, Joyce Seldon, Patricia A Ganz, Robert L Nussbaum, Salina B Chan, Kunle Odunsi, Simon A Gayther, Susan M Domchek, Banu K Arun, Karen H Lu, Gillian Mitchell, Beth Y Karlan, Christine Walsh, Jenny Lester, Andrew K Godwin, Harsh Pathak, Eric Ross, Mary B Daly, Alice S Whittemore, Esther M John, Alexander Miron, Mary Beth Terry, Wendy K Chung, David E Goldgar, Saundra S Buys, Ramunas Janavicius, Laima Tihomirova, Nadine Tung, Cecilia M Dorfling, Elizabeth J van Rensburg, Linda Steele, Susan L Neuhausen, Yuan Chun Ding, Bent Ejlertsen, Anne-Marie Gerdes, Thomas v O Hansen, Teresa Ramón y Cajal, Ana Osorio, Javier Benitez, Javier Godino, Maria-Isabel Tejada, Mercedes Duran, Jeffrey N Weitzel, Kristie A Bobolis, Sharon R Sand, Annette Fontaine, Antonella Savarese, Barbara Pasini, Bernard Peissel, Bernardo Bonanni, Daniela Zaffaroni, Francesca Vignolo-Lutati, Giulietta Scuvera, Giuseppe Giannini, Loris Bernard, Maurizio Genuardi, Paolo Radice, Riccardo Dolcetti, Siranoush Manoukian, Valeria Pensotti, Viviana Gismondi, Drakoulis Yannoukakos, Florentia Fostira, Judy Garber, Diana Torres, Muhammad Usman Rashid, Ute Hamann, Susan Peock, Debra Frost, Radka Platte, D Gareth Evans, Rosalind Eeles, Rosemarie

Davidson, Diana Eccles, Trevor Cole, Jackie Cook, Carole Brewer, Shirley Hodgson, Patrick J Morrison, Lisa Walker, Mary E Porteous, M John Kennedy, Louise Izatt, Julian Adlard, Alan Donaldson, Steve Ellis, Priyanka Sharma, Rita Katharina Schmutzler, Barbara Wappenschmidt, Alexandra Becker, Kerstin Rhiem, Eric Hahnen, Christoph Engel, Alfons Meindl, Stefanie Engert, Nina Ditsch, Norbert Arnold, Hans Jörg Plendl, Christoph Mundhenke, Dieter Niederacher, Markus Fleisch, Christian Sutter, C R Bartram, Nicola Dikow, Shan Wang-Gohrke, Dorothea Gadzicki, Doris Steinemann, Karin Kast, Marit Beer, Raymonda Varon-Mateeva, Andrea Gehrig, Bernhard H Weber, Dominique Stoppa-Lyonnet, Olga M Sinilnikova, Sylvie Mazoyer, Claude Houdayer, Muriel Belotti, Marion Gauthier-Villars, Francesca Damiola, Nadia Boutry-Kryza, Christine Lasset, Hagay Sobol, Jean-Philippe Peyrat, Danièle Muller, Jean-Pierre Fricker, Marie-Agnès Collonge-Rame, Isabelle Mortemousque, Catherine Nogues, Etienne Rouleau, Claudine Isaacs, Anne De Paepe, Bruce Poppe, Kathleen Claes, Kim De Leeneer, Marion Piedmonte, Gustavo Rodriguez, Katie Wakely, John Boggess, Stephanie V Blank, Jack Basil, Masoud Azodi, Kelly-Anne Phillips, Trinidad Caldes, Miguel de la Hoya, Atocha Romero, Heli Nevanlinna, Kristiina Aittomäki, Annemarie H van der Hout, Frans B L Hogervorst, Senno Verhoef, J Margriet Collée, Caroline Seynaeve, Jan C Oosterwijk, Johannes J P Gille, Juul T Wijnen, Encarna B Gómez Garcia, Carolien M Kets, Margreet G E M Ausems, Cora M Aalfs, Peter Devilee, Arjen R Mensenkamp, Ava Kwong, Edith Olah, Janos Papp, Orland Diez, Conxi Lazaro, Esther Darder, Ignacio Blanco, Mónica Salinas, Anna Jakubowska, Jan Lubinski, Jacek Gronwald, Katarzyna Jaworska-Bieniek, Katarzyna Durda, Grzegorz Sukiennicki, Tomasz Huzarski, Tomasz Byrski, Cezary Cybulski, Aleksandra Toloczko-Grabarek, Elżbieta Złowocka-Perłowska, Janusz Menkiszak, Adalgeir Arason, Rosa B Barkardottir, Jacques Simard, Rachel Laframboise, Marco Montagna, Simona Agata, Elisa Alducci, Ana Peixoto, Manuel R Teixeira, Amanda B Spurdle, Min Hyuk Lee, Sue K Park, Sung-Won Kim, Tara M Friebel, Fergus J Couch, Noralane M Lindor, Vernon S Pankratz, Lucia Guidugli, Xianshu Wang, Marc Tischkowitz, Lenka Foretova, Joseph Vijai, Kenneth Offit, Mark Robson, Rohini Rau-Murthy, Noah Kauff, Anneliese Fink-Retter, Christian F Singer, Christine Rappaport, Daphne Gschwantler-Kaulich, Georg Pfeiler, Muy-Kheng Tea, Andreas Berger, Mark H Greene, Phuong L Mai, Evgeny N Imyanitov, Amanda Ewart Toland, Leigha Senter, Anders Bojesen, Inge Sokilde Pedersen, Anne-Bine Skytte, Lone Sunde, Mads Thomassen, Sanne Traasdahl Moeller, Torben A Kruse, Uffe Birk Jensen, Maria Adelaide Caligo, Paolo Aretini, Soo-Hwang Teo, Christina G Selkirk, Peter J Hulick, and Irene Andrulis. Association of type and location of BRCA1 and BRCA2 mutations with risk of breast and ovarian cancer. *JAMA*, 313(13):1347–1361, 7 April 2015.

[152] Nancie Petrucelli, Mary B Daly, and Gerald L Feldman. Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. *Genet. Med.*, 12(5):245–259, May 2010.

[153] Judy E Garber and Kenneth Offit. Hereditary cancer predisposition syndromes. *J. Clin. Oncol.*, 23(2):276–292, 10 January 2005.

[154] Lorelei A Mucci, Jacob B Hjelmborg, Jennifer R Harris, Kamila Czene, David J Havelick, Thomas Scheike, Rebecca E Graff, Klaus Holst, Sören Möller, Robert H Unger, Christina McIntosh, Elizabeth Nuttall, Ingunn Brandt, Kathryn L Penney, Mikael Hartman, Peter Kraft, Giovanni Parmigiani, Kaare Christensen, Markku Koskenvuo, Niels V Holm, Kauko Heikkilä, Eero Pukkala, Axel Skytthe, Hans-Olov Adami, Jaakko Kaprio, and Nordic Twin Study of Cancer (NorTwinCan) Collaboration. Familial risk and heritability of cancer among twins in nordic countries. *JAMA*, 315(1):68–76, 5 January 2016.

[155] Paul D P Pharoah, Alison M Dunning, Bruce A J Ponder, and Douglas F Easton. Association studies for finding cancer-susceptibility genetic variants. *Nat. Rev. Cancer*, 4(11):850–860, November 2004.

[156] Lucia A Hindorff, Elizabeth M Gillanders, and Teri A Manolio. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*, 32(7):945–954, July 2011.

[157] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS Comput. Biol.*, 8(12):e1002822, 27 December 2012.

[158] Tun-Hsiang Yang, Mark Kon, and Charles DeLisi. Genome-wide association studies. *Methods Mol. Biol.*, 939:233–251, 2013.

[159] Kai Wang, Samuel P Dickson, Catherine A Stolle, Ian D Krantz, David B Goldstein, and Hakon Hakonarson. Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.*, 86(5):730–742, 14 May 2010.

[160] Montgomery Slatkin. Linkage disequilibrium–understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.*, 9(6):477–485, June 2008.

[161] Kristin G Ardlie, Leonid Kruglyak, and Mark Seielstad. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, 3(4):299–309, April 2002.

[162] Sarah L Spain and Jeffrey C Barrett. Strategies for fine-mapping complex traits. *Hum. Mol. Genet.*, 24(R1):R111–9, 15 October 2015.

[163] Stacey L Edwards, Jonathan Beesley, Juliet D French, and Alison M Dunning. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, 93(5):779–797, 7 November 2013.

[164] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *Am. J. Hum. Genet.*, 90(1):7–24, 13 January 2012.

[165] N Risch and K Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 13 September 1996.

[166] E S Lander. The new genomics: global views of biology. *Science*, 274(5287):536–539, 25 October 1996.

[167] A Chakravarti. Population genetics–making sense out of sequence. *Nat. Genet.*, 21(1 Suppl):56–60, January 1999.

[168] D E Reich and E S Lander. On the allelic spectrum of human disease. *Trends Genet.*, 17(9):502–510, September 2001.

[169] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, 11(7):499–511, July 2010.

[170] Thomas J Hoffmann and John S Witte. Strategies for imputing and analyzing rare variants in association studies. *Trends Genet.*, 31(10):556–563, October 2015.

[171] Eric Reed, Sara Nunez, David Kulp, Jing Qian, Muredach P Reilly, and Andrea S Foulkes. A guide to genome-wide association analysis and post-analytic interrogation. *Stat. Med.*, 34(28):3769–3792, 10 December 2015.

[172] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res.*, 45(D1):D896–D901, 4 January 2017.

[173] Fergus J Couch, Xianshu Wang, Lesley McGuffog, Andrew Lee, Curtis Olswold, Karoline B Kuchenbaecker, Penny Soucy, Zachary Fredericksen, Daniel Barrowdale, Joe Dennis, Mia M Gaudet, Ed Dicks, Matthew Kosel, Sue Healey, Olga M Sinilnikova, Adam Lee, François Bacot, Daniel Vincent, Frans B L Hogervorst, Susan Peock, Dominique Stoppa-Lyonnet, Anna Jakubowska, kConFab Investigators, Paolo Radice, Rita Katharina Schmutzler, SWE-BRCA, Susan M Domchek, Marion Piedmonte, Christian F Singer, Eitan Friedman, Mads Thomassen, Ontario Cancer Genetics Network, Thomas V O Hansen, Susan L Neuhausen, Csilla I Szabo, Ignacio Blanco, Mark H Greene, Beth Y Karlan, Judy Garber, Catherine M Phelan, Jeffrey N Weitzel, Marco Montagna, Edith Olah, Irene L Andrulis, Andrew K Godwin, Drakoulis Yannoukakos, David E Goldgar, Trinidad Caldes, Heli Nevanlinna, Ana Osorio, Mary Beth Terry, Mary B Daly, Elizabeth J van Rensburg, Ute Hamann, Susan J Ramus, Amanda Ewart Toland, Maria A Caligo, Olufunmilayo I Olopade, Nadine Tung, Kathleen Claes, Mary S Beattie, Melissa C Southey, Evgeny N Imyanitov, Marc Tischkowitz, Ramunas Janavicius, Esther M John, Ava Kwong, Orland Diez, Judith Balmaña, Rosa B Barkardottir, Banu K Arun, Gad Rennert, Soo-Hwang Teo, Patricia A Ganz, Ian Campbell, Annemarie H van der Hout, Carolien H M van Deurzen, Caroline Seynaeve, Encarna B Gómez Garcia, Flora E van Leeuwen, Hanne E J Meijers-Heijboer, Johannes J P Gille, Margreet G E M Ausems, Marinus J Blok, Marjolijn J L Ligtenberg, Matti A Rookus, Peter Devilee, Senno Verhoef, Theo A M van Os,

Juul T Wijnen, HEBON, EMBRACE, Debra Frost, Steve Ellis, Elena Fineberg, Radka Platte, D Gareth Evans, Louise Izatt, Rosalind A Eeles, Julian Adlard, Diana M Eccles, Jackie Cook, Carole Brewer, Fiona Douglas, Shirley Hodgson, Patrick J Morrison, Lucy E Side, Alan Donaldson, Catherine Houghton, Mark T Rogers, Huw Dorkins, Jacqueline Eason, Helen Gregory, Emma McCann, Alex Murray, Alain Calender, Agnès Hardouin, Pascaline Berthet, Capucine Delnatte, Catherine Nogues, Christine Lasset, Claude Houdayer, Dominique Leroux, Etienne Rouleau, Fabienne Prieur, Francesca Damiola, Hagay Sobol, Isabelle Coupier, Laurence Venat-Bouvet, Laurent Castera, Marion Gauthier-Villars, Mélanie Léoné, Pascal Pujol, Sylvie Mazoyer, Yves-Jean Bignon, GEMO Study Collaborators, Elżbieta Złowocka-Perłowska, Jacek Gronwald, Jan Lubinski, Katarzyna Durda, Katarzyna Jaworska, Tomasz Huzarski, Amanda B Spurdle, Alessandra Viel, Bernard Peissel, Bernardo Bonanni, Giulia Melloni, Laura Ottini, Laura Papi, Liliana Varesco, Maria Grazia Tibiletti, Paolo Peterlongo, Sara Volorio, Siranoush Manoukian, Valeria Pensotti, Norbert Arnold, Christoph Engel, Helmut Deissler, Dorothea Gadzicki, Andrea Gehrig, Karin Kast, Kerstin Rhiem, Alfons Meindl, Dieter Niederacher, Nina Ditsch, Hansjoerg Plendl, Sabine Preisler-Adams, Stefanie Engert, Christian Sutter, Raymonda Varon-Mateeva, Barbara Wappenschmidt, Bernhard H F Weber, Brita Arver, Marie Stenmark-Askmalm, Niklas Loman, Richard Rosenquist, Zakaria Einbeigi, Katherine L Nathanson, Timothy R Rebbeck, Stephanie V Blank, David E Cohn, Gustavo C Rodriguez, Laurie Small, Michael Friedlander, Victoria L Bae-Jump, Anneliese Fink-Retter, Christine Rappaport, Daphne Gschwantler-Kaulich, Georg Pfeiler, Muy-Kheng Tea, Noralane M Lindor, Bella Kaufman, Shani Shimon Paluch, Yael Laitman, Anne-Bine Skytte, Anne-Marie Gerdes, Inge Sokilde Pedersen, Sanne Traasdahl Moeller, Torben A Kruse, Uffe Birk Jensen, Joseph Vijai, Kara Sarrel, Mark Robson, Noah Kauff, Anna Marie Mulligan, Gord Glendon, Hilmi Ozcelik, Bent Ejlertsen, Finn C Nielsen, Lars Jønson, Mette K Andersen, Yuan Chun Ding, Linda Steele, Lenka Foretova, Alex Teulé, Conxi Lazaro, Joan Brunet, Miquel Angel Pujana, Phuong L Mai, Jennifer T Loud, Christine Walsh, Jenny Lester, Sandra Orsulic, Steven A Narod, Josef Herzog, Sharon R Sand, Silvia Tognazzo, Simona Agata, Tibor Vaszko, Joellen Weaver, Alexandra V Stavropoulou, Saundra S Buys, Atocha Romero, Miguel de la Hoya, Kristiina Aittomäki, Taru A Muranen, Mercedes Duran, Wendy K Chung, Adriana Lasa, Cecilia M Dorfling, Alexander Miron, BCFR, Javier Benitez, Leigha Senter, Dezheng Huo, Salina B Chan, Anna P Sokolenko, Jocelyne Chiquette, Laima Tihomirova, Tara M Friebel, Bjarni A Agnarsson, Karen H Lu, Flavio Lejbkowicz, Paul A James, Per Hall, Alison M Dunning, Daniel Tessier, Julie Cunningham, Susan L Slager, Chen Wang, Steven Hart, Kristen Stevens, Jacques Simard, Tomi Pastinen, Vernon S Pankratz, Kenneth Offit, Douglas F Easton, Georgia Chenevix-Trench, Antonis C Antoniou, and CIMBA. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet.*, 9(3):e1003212, 27 March 2013.

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

Ernest Hawk, and Xifeng Wu. GWAS-identified colorectal cancer susceptibility loci associated with clinical outcomes. *Carcinogenesis*, 33(7):1327–1331, July 2012.

[175] Maya Ghoussaini, Honglin Song, Thibaud Koessler, Ali Amin Al Olama, Zsofia Kote-Jarai, Kristy E Driver, Karen A Pooley, Susan J Ramus, Susanne Krüger Kjaer, Estrid Hogdall, Richard A DiCioccio, Alice S Whittemore, Simon A Gayther, Graham G Giles, Michelle Guy, Stephen M Edwards, Jonathan Morrison, Jenny L Donovan, Freddie C Hamdy, David P Dearnaley, Audrey T Ardern-Jones, Amanda L Hall, Lynne T O'Brien, Beatrice N Gehr-Swain, Rosemary A Wilkinson, Paul M Brown, John L Hopper, David E Neal, Paul D P Pharoah, Bruce A J Ponder, Rosalind A Eeles, Douglas F Easton, Alison M Dunning, UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology, and UK ProtecT Study Collaborators. Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl. Cancer Inst.*, 100(13):962–966, 2 July 2008.

[176] Konrad Huppi, Jason J Pitt, Brady M Wahlberg, and Natasha J Caplen. The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Front. Genet.*, 3:69, 30 April 2012.

[177] Jason B Wright, Seth J Brown, and Michael D Cole. Upregulation of c-MYC in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol. Cell. Biol.*, 30(6):1411–1420, March 2010.

[178] Nasim Ahmadiyeh, Mark M Pomerantz, Chiara Grisanzio, Paula Herman, Li Jia, Vanessa Almendro, Housheng Hansen He, Myles Brown, X Shirley Liu, Matt Davis, Jennifer L Caswell, Christine A Beckwith, Adam Hills, Laura Macconaill, Gerhard A Coetzee, Meredith M Regan, and Matthew L Freedman. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proc. Natl. Acad. Sci. U. S. A.*, 107(21):9742–9746, 25 May 2010.

[179] Devon M Fitzgerald and Susan M Rosenberg. Driving cancer evolution. *Elife*, 6, 10 March 2017.

[180] Caitriona Holohan, Sandra Van Schaeybroeck, Daniel B Longley, and Patrick G Johnston. Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer*, 13(10):714–726, October 2013.

[181] Lloyd Kelland. The resurgence of platinum-based cancer chemotherapy. *Nat. Rev. Cancer*, 7(8):573–584, August 2007.

[182] Shaloam Dasari and Paul Bernard Tchounwou. Cisplatin in cancer therapy: molecular mechanisms of action. *Eur. J. Pharmacol.*, 740:364–378, 5 October 2014.

[183] Timothy C Johnstone, Ga Young Park, and Stephen J Lippard. Understanding and improving platinum anticancer drugs–phenanthriplatin. *Anticancer Res.*, 34(1):471–476, January 2014.

[184] Ann-Marie Patch, Elizabeth L Christie, Dariush Etemadmoghadam, Dale W Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J Bailey, Karin S Kassahn, Felicity Newell, Michael C J Quinn, Stephen Kazakoff, Kelly Quek, Charlotte Wilhelm-Benartzi, Ed Curry, Huei San Leong, Australian Ovarian Cancer Study Group, Anne Hamilton, Linda Mileshkin, George Au-Yeung, Catherine Kennedy, Jillian Hung, Yoke-Eng Chiew, Paul Harnett, Michael Friedlander, Michael Quinn, Jan Pyman, Stephen Cordner, Patricia O'Brien, Jodie Leditschke, Greg Young, Kate Strachan, Paul Waring, Walid Azar, Chris Mitchell, Nadia Traficante, Joy Hendley, Heather Thorne, Mark Shackleton, David K Miller, Gisela Mir Arnau, Richard W Tothill, Timothy P Holloway, Timothy Semple, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Senel Idrisoglu, Timothy J C Bruxner, Angelika N Christ, Barsha Poudel, Oliver Holmes, Matthew Anderson, Conrad Leonard, Andrew Lonie, Nathan Hall, Scott Wood, Darrin F Taylor, Qinying Xu, J Lynn Fink, Nick Waddell, Ronny Drapkin, Euan Stronach, Hani Gabra, Robert Brown, Andrea Jewell, Shivashankar H Nagaraj, Emma Markham, Peter J Wilson, Jason Ellul, Orla McNally, Maria A Doyle, Ravikiran Vedururu, Collin Stewart, Ernst Lengyel, John V Pearson, Nicola Waddell, Anna deFazio, Sean M Grimmond, and David D L Bowtell. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–494, 28 May 2015.

[185] Satoru Hashimoto, Hirofumi Anai, and Katsuhiro Hanada. Mechanisms of interstrand DNA crosslink repair and human disorders. *Genes Environ*, 38:9, 1 May 2016.

[186] Andrew J Deans and Stephen C West. DNA interstrand crosslink repair and cancer. *Nat. Rev. Cancer*, 11(7):467–480, 24 June 2011.

[187] John M Hinz. Role of homologous recombination in DNA interstrand crosslink repair. *Environ. Mol. Mutagen.*, 51(6):582–603, July 2010.

[188] Dan A Landau, Scott L Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S Lawrence, Carrie Sougnez, Chip Stewart, Andrey Sivachenko, Lili Wang, Youzhong Wan, Wandi Zhang, Sachet A Shukla, Alexander Vartanov, Stacey M Fernandes, Gordon Saksena, Kristian Cibulskis, Bethany Tesar, Stacey Gabriel, Nir Hacohen, Matthew Meyerson, Eric S Lander, Donna Neuberg, Jennifer R Brown, Gad Getz, and Catherine J Wu. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–726, 14 February 2013.

[189] Jan A Burger, Dan A Landau, Amaro Taylor-Weiner, Ivana Bozic, Huidan Zhang, Kristopher Sarosiek, Lili Wang, Chip Stewart, Jean Fan, Julia Hoellenriegel, Mariela Sivina, Adrian M Dubuc, Cameron Fraser, Yulong Han, Shuqiang Li, Kenneth J Livak, Lihua Zou, Youzhong Wan, Sergej Konoplev, Carrie Sougnez, Jennifer R Brown, Lynne V Abruzzo, Scott L Carter, Michael J Keating, Matthew S Davids, William G Wierda, Kristian Cibulskis, Thorsten Zenz, Lillian Werner, Paola Dal Cin, Peter Kharchencko, Donna Neuberg, Hagop Kantarjian, Eric Lander, Stacey Gabriel, Susan

O'Brien, Anthony Letai, David A Weitz, Martin A Nowak, Gad Getz, and Catherine J Wu. Clonal evolution in patients with chronic lymphocytic leukaemia developing resistance to BTK inhibition. *Nat. Commun.*, 7:11589, 20 May 2016.

[190] Euan A Ashley. Towards precision medicine. *Nat. Rev. Genet.*, 17(9):507–522, 16 August 2016.

[191] David M Hyman, Barry S Taylor, and José Baselga. Implementing Genome-Driven oncology. *Cell*, 168(4):584–599, 9 February 2017.

[192] Alison M Schram, Michael F Berger, and David M Hyman. Precision oncology: Charting a path forward to broader deployment of genomic profiling. *PLoS Med.*, 14(2):e1002242, February 2017.

[193] Robert A Smith, Kimberly Andrews, Durado Brooks, Carol E DeSantis, Stacey A Fedewa, Joannie Lortet-Tieulent, Deana Manassaram-Baptiste, Otis W Brawley, and Richard C Wender. Cancer screening in the united states, 2016: A review of current american cancer society guidelines and current issues in cancer screening. *CA Cancer J. Clin.*, 66(2):96–114, March 2016.

[194] José G Guillem, William C Wood, Jeffrey F Moley, Andrew Berchuck, Beth Y Karlan, David G Mutch, Robert F Gagel, Jeffrey Weitzel, Monica Morrow, Barbara L Weber, Francis Giardiello, Miguel A Rodriguez-Bigas, James Church, Stephen Gruber, Kenneth Offit, ASCO, and SSO. ASCO/SSO review of current role of risk-reducing surgery in common hereditary cancer syndromes. *J. Clin. Oncol.*, 24(28):4642–4660, 1 October 2006.

[195] L C Hartmann, D J Schaid, J E Woods, T P Crotty, J L Myers, P G Arnold, P M Petty, T A Sellers, J L Johnson, S K McDonnell, M H Frost, and R B Jenkins. Efficacy of bilateral prophylactic mastectomy in women with a family history of breast cancer. *N. Engl. J. Med.*, 340(2):77–84, 14 January 1999.

[196] Mingyang Song and Edward Giovannucci. Preventable incidence and mortality of carcinoma associated with lifestyle factors among white adults in the united states. *JAMA Oncol*, 2(9):1154–1161, 1 September 2016.

[197] Matthew B Yurgelun, Georgia Chenevix-Trench, and Scott M Lippman. Translating germline cancer risk into precision prevention. *Cell*, 168(4):566–570, 9 February 2017.

[198] D G R Evans. Sensitivity of BRCA1/2 mutation testing in 466 breast/ovarian cancer families. *J. Med. Genet.*, 40(9):107e–107, 1 September 2003.

[199] Christine Sevilla, Jean-Paul Moatti, Claire Julian-Reynier, François Eisinger, Dominique Stoppa-Lyonnet, Brigitte Bressac-de Paillerets, and Hagay Sobol. Testing for BRCA1 mutations: a cost-effectiveness analysis. *Eur. J. Hum. Genet.*, 10(10):599–606, 2002.

[200] M Swift, D Morrell, R B Massey, and C L Chase. Incidence of cancer in 161 families affected by ataxia-telangiectasia. *N. Engl. J. Med.*, 325(26):1831–1836, 26 December 1991.

[201] J Chen and A Lindblom. Germline mutation screening of the STK11/LKB1 gene in familial breast cancer with LOH on 19p. *Clin. Genet.*, 57(5):394–397, May 2000.

[202] Donavan T Cheng, Meera Prasad, Yvonne Chekaluk, Ryma Benayed, Justyna Sadowska, Ahmet Zehir, Aijazuddin Syed, Yan Elsa Wang, Joshua Somar, Yirong Li, Zarina Yelskaya, Donna Wong, Mark E Robson, Kenneth Offit, Michael F Berger, Khedoudja Nafa, Marc Ladanyi, and Liying Zhang. Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med. Genomics*, 10(1):33, 19 May 2017.

[203] Douglas F Easton, Paul D P Pharoah, Antonis C Antoniou, Marc Tischkowitz, Sean V Tavtigian, Katherine L Nathanson, Peter Devilee, Alfons Meindl, Fergus J Couch, Melissa Southey, David E Goldgar, D Gareth R Evans, Georgia Chenevix-Trench, Nazneen Rahman, Mark Robson, Susan M Domchek, and William D Foulkes. Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.*, 372(23):2243–2257, 4 June 2015.

[204] Holly LaDuca, A J Stuenkel, Jill S Dolinsky, Steven Keiles, Stephany Tandy, Tina Pesaran, Elaine Chen, Chia-Ling Gau, Erika Palmaer, Kamelia Shoaepour, Divya Shah, Virginia Speare, Stephanie Gandomi, and Elizabeth Chao. Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. *Genet. Med.*, 16(11):830–837, November 2014.

[205] Tom Walsh, Ming K Lee, Silvia Casadei, Anne M Thornton, Sunday M Stray, Christopher Pennil, Alex S Nord, Jessica B Mandell, Elizabeth M Swisher, and Mary-Claire King. Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.*, 107(28):12629–12633, 13 July 2010.

[206] Kasmintan A Schrader, Donavan T Cheng, Vijai Joseph, Meera Prasad, Michael Walsh, Ahmet Zehir, Ai Ni, Tinu Thomas, Ryma Benayed, Asad Ashraf, Annie Lincoln, Maria Arcila, Zsofia Stadler, David Solit, David M Hyman, David Hyman, Liying Zhang, David Klimstra, Marc Ladanyi, Kenneth Offit, Michael Berger, and Mark Robson. Germline variants in targeted tumor sequencing using matched normal DNA. *JAMA Oncol*, 2(1):104–111, January 2016.

[207] Fergus J Couch, Hermela Shimelis, Chunling Hu, Steven N Hart, Eric C Polley, Jie Na, Emily Hallberg, Raymond Moore, Abigail Thomas, Jenna Lilyquist, Bingjian Feng, Rachel McFarland, Tina Pesaran, Robert Huether, Holly LaDuca, Elizabeth C Chao, David E Goldgar, and Jill S Dolinsky. Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol*, 13 April 2017.

[208] Hongyan Li, Bingjian Feng, Alexander Miron, Xiaoqing Chen, Jonathan Beesley, Emmanuella Bimeh, Daniel Barrowdale, Esther M John, Mary B Daly, Irene L Andrulis, Saundra S Buys, Peter Kraft, kConFab investigators, Heather Thorne, Georgia Chenevix-Trench, Melissa C Southey, Antonis C Antoniou, Paul A James, Mary Beth Terry, Kelly-Anne Phillips, John L Hopper, Gillian Mitchell, and David E Goldgar. Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the breast cancer family registry and kConFab. *Genet. Med.*, 19(1):30–35, January 2017.

[209] Karoline B Kuchenbaecker, Lesley McGuffog, Daniel Barrowdale, Andrew Lee, Penny Soucy, Joe Dennis, Susan M Domchek, Mark Robson, Amanda B Spurdle, Susan J Ramus, Nasim Mavaddat, Mary Beth Terry, Susan L Neuhausen, Rita Katharina Schmutzler, Jacques Simard, Paul D P Pharoah, Kenneth Offit, Fergus J Couch, Georgia Chenevix-Trench, Douglas F Easton, Antonis C Antoniou, Michael Lush, Ute Hamann, Melissa Southey, Esther M John, Wendy K Chung, Mary B Daly, Saundra S Buys, David E Goldgar, Cecilia M Dorfling, Elizabeth J van Rensburg, Yuan Chun Ding, Bent Ejlertsen, Anne-Marie Gerdes, Thomas V O Hansen, Susan Slager, Emily Hallberg, Javier Benitez, Ana Osorio, Nancy Cohen, William Lawler, Jeffrey N Weitzel, Paolo Peterlongo, Valeria Pensotti, Riccardo Dolcetti, Monica Barile, Bernando Bonanni, Jacopo Azzollini, Siranoush Manoukian, Bernard Peissel, Paolo Radice, Antonella Savarese, Laura Papi, Giuseppe Giannini, Florentia Fostira, Irene Konstantopoulou, Julian Adlard, Carole Brewer, Jackie Cook, Rosemarie Davidson, Diana Eccles, Ros Eeles, Steve Ellis, Debra Frost, Shirley Hodgson, Louise Izatt, Fiona Lalloo, Kai-Ren Ong, Andrew K Godwin, Norbert Arnold, Bernd Dworniczak, Christoph Engel, Andrea Gehrig, Eric Hahnen, Jan Hauke, Karin Kast, Alfons Meindl, Deiter Niederacher, Rita Katherina Schmutzler, Raymonda Varon-Mateeva, Shan Wang-Gohrke, Barbara Wappenschmidt, Laure Barjhoux, Marie-Agnes Collonge-Rame, Camille Elan, Lisa Golmard, GEMO Study Collaborators, EMBRACE, Emmanuelle Barouk-Simonet, Fabienne Lesueur, Sylvie Mazoyer, Joanna Sokolowska, Dominique Stoppa-Lyonnet, Claudine Isaacs, Kathleen B M Claes, Bruce Poppe, Miguel de la Hoya, Vanesa Garcia-Barberan, Kristiina Aittomaki, Heli Nevanlinna, Margreet G E M Ausems, J L de Lange, Encarna B Gomez Garcia, Frans B L Hogervorst, HEBON, Carolien M Kets, Hanne E Meijers-Heijboer, Jan C Oosterwijk, Matti A Rookus, Christi J van Asperen, Ans M W van den Ouweland, Helena C van Doorn, Theo A M van Os, Ava Kwong, Edith Olah, Orland Diez, Joan Brunet, Conxi Lazaro, Alex Teule, Jacek Gronwald, Anna Jakubowska, Katarzyna Kaczmarek, Jan Lubinski, Grzegorz Sukiennicki, Rosa B Barkardottir, Jocelyne Chiquette, Simona Agata, Marco Montagna, Manuel R Teixeira, Sue Kyung Park, KConFab Investigators, Curtis Olswold, Marc Tischkowitz, Lenka Foretova, Pragna Gaddam, Joseph Vijai, Georg Pfeiler, Christine Rappaport-Fuerhauser, Christian F Singer, Muy-Kheng M Tea, Mark H Greene, Jennifer T Loud, Gad Rennert, Evgeny N Imyanitov, Peter J Hulick, John L Hays, Marion Piedmonte, Gustavo C Rodriguez, Julie Martyn, Gord Glendon, Anna Marie Mulligan, Irene L Andrulis, Amanda Ewart Toland, Uffe Birk Jenson, Torben A Kruse, Inge Sokilde Pedersen, Mads Thomassen, Maria A Caligo, Soo-Hwang Teo, Raanan Berger,

Eitan Friedman, Yael Laitman, Brita Arver, Ake Borg, Hans Ehrancrona, Johanna Rantala, Olufunmilayo I Olopade, Patricia A Ganz, Robert L Nussbaum, Angela R Bradbury, Susan M Domchek, Katherine L Nathanson, Banu K Arun, Paul James, Beth Y Karlan, Jenny Lester, Jacques Simard, Paul D P Pharoah, Kenneth Offit, Fergus J Couch, Georgia Chenevix-Trench, Douglas F Easton, and Antonis C Antoniou. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.*, 109(7), 1 July 2017.

[210] Robert Szulkin, Thomas Whitington, Martin Eklund, Markus Aly, Rosalind A Eeles, Douglas Easton, Z Sofia Kote-Jarai, Ali Amin Al Olama, Sara Benlloch, Kenneth Muir, Graham G Giles, Melissa C Southey, Liesel M Fitzgerald, Brian E Henderson, Fredrick Schumacher, Christopher A Haiman, Johanna Schleutker, Tiina Wahlfors, Teuvo L J Tammela, Børge G Nordestgaard, Tim J Key, Ruth C Travis, David E Neal, Jenny L Donovan, Freddie C Hamdy, Paul Pharoah, Nora Pashayan, Kay-Tee Khaw, Janet L Stanford, Stephen N Thibodeau, Shannon K McDonnell, Daniel J Schaid, Christiane Maier, Walther Vogel, Manuel Luedeke, Kathleen Herkommer, Adam S Kibel, Cezary Cybulski, Jan Lubiński, Wojciech Kluźniak, Lisa Cannon-Albright, Hermann Brenner, Katja Butterbach, Christa Stegmaier, Jong Y Park, Thomas Sellers, Hui-Yi Lin, Hui-Yi Lim, Chavdar Slavov, Radka Kaneva, Vanio Mitev, Jyotsna Batra, Judith A Clements, Australian Prostate Cancer BioResource, Amanda Spurdle, Manuel R Teixeira, Paula Paulo, Sofia Maia, Hardev Pandha, Agnieszka Michael, Andrzej Kierzek, Practical Consortium, Henrik Gronberg, and Fredrik Wiklund. Prediction of individual genetic risk to prostate cancer using a polygenic score. *Prostate*, 75(13):1467–1474, September 2015.

[211] Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.*, 17(7):392–406, July 2016.

[212] M J E Frampton, P Law, K Litchfield, E J Morris, D Kerr, C Turnbull, I P Tomlinson, and R S Houlston. Implications of polygenic risk for personalised colorectal cancer screening. *Ann. Oncol.*, 27(3):429–434, March 2016.

[213] S Leedham and I Tomlinson. The continuum model of selection in human tumors: General paradigm or niche product? *Cancer Res.*, 72(13):3131–3134, 2012.

[214] Britta Weigelt, Felipe C Geyer, and Jorge S Reis-Filho. Histological types of breast cancer: how special are they? *Mol. Oncol.*, 4(3):192–208, June 2010.

[215] D Craig Allred. Issues and updates: evaluating estrogen receptor-alpha, progesterone receptor, and HER2 in breast cancer. *Mod. Pathol.*, 23 Suppl 2:S52–9, May 2010.

[216] Kornelia Polyak and Otto Metzger Filho. SnapShot: breast cancer. *Cancer Cell*, 22(4):562–562.e1, 16 October 2012.

[217] George W Sledge, Eleftherios P Mamounas, Gabriel N Hortobagyi, Harold J Burstein, Pamela J Goodwin, and Antonio C Wolff. Past, present, and future challenges in breast cancer treatment. *J. Clin. Oncol.*, 32(19):1979–1986, 1 July 2014.

[218] M Clemons, S Danson, and A Howell. Tamoxifen (nolvadex): a review. *Cancer Treat. Rev.*, 28(4):165–180, August 2002.

[219] Valentina I Petkov, Dave P Miller, Nadia Howlader, Nathan Gliner, Will Howe, Nicola Schussler, Kathleen Cronin, Frederick L Baehner, Rosemary Cress, Dennis Deapen, Sally L Glaser, Brenda Y Hernandez, Charles F Lynch, Lloyd Mueller, Ann G Schwartz, Stephen M Schwartz, Antoinette Stroup, Carol Sweeney, Thomas C Tucker, Kevin C Ward, Charles Wiggins, Xiao-Cheng Wu, Lynne Penberthy, and Steven Shak. Breast-cancer-specific mortality in patients treated based on the 21-gene assay: a SEER population-based study. *npj Breast Cancer*, 2(1):e0128345, 8 December 2016.

[220] Rebecca Dent, Maureen Trudeau, Kathleen I Pritchard, Wedad M Hanna, Harriet K Kahn, Carol A Sawka, Lavina A Lickley, Ellen Rawlinson, Ping Sun, and Steven A Narod. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clin. Cancer Res.*, 13(15 Pt 1):4429–4434, 1 August 2007.

[221] Hagen Kennecke, Rinat Yerushalmi, Ryan Woods, Maggie Chon U Cheang, David Voduc, Caroline H Speers, Torsten O Nielsen, and Karen Gelmon. Metastatic behavior of breast cancer subtypes. *J. Clin. Oncol.*, 28(20):3271–3277, 10 July 2010.

[222] C Dilara Savci-Heijink, Hans Halfwerk, Gerrit K J Hooijer, Hugo M Horlings, Jelle Wesseling, and Marc J van de Vijver. Retrospective analysis of metastatic behaviour of breast cancer subtypes. *Breast Cancer Res. Treat.*, 150(3):547–557, April 2015.

[223] J Crown, J O'Shaughnessy, and G Gullo. Emerging targeted therapies in triple-negative breast cancer. *Ann. Oncol.*, 23 Suppl 6:vi56–65, August 2012.

[224] Justin M Balko, Luis J Schwarz, Na Luo, Mónica V Estrada, Jennifer M Giltnane, Daniel Dávila-González, Kai Wang, Violeta Sánchez, Phillip T Dean, Susan E Combs, Donna Hicks, Joseph A Pinto, Melissa D Landis, Franco D Doimi, Roman Yelensky, Vincent A Miller, Phillip J Stephens, David L Rimm, Henry Gómez, Jenny C Chang, Melinda E Sanders, Rebecca S Cook, and Carlos L Arteaga. Triple-negative breast cancers with amplification of JAK2 at the 9p24 locus demonstrate JAK2-specific dependence. *Sci. Transl. Med.*, 8(334):334ra53, 13 April 2016.

[225] Giuseppe Palma, Giuseppe Frasci, Andrea Chirico, Emanuela Esposito, Claudio Siani, Carmela Saturnino, Claudio Arra, Gennaro Ciliberto, Antonio Giordano, and Massimiliano D'Aiuto. Triple negative breast cancer: looking for the missing link between biology and treatments. *Oncotarget*, 6(29):26560–26574, 29 September 2015.

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

Sherman, Anna Marie Mulligan, Fergus J Couch, Christoph Engel, Lesley McGuffog, Sue Healey, Olga M Sinilnikova, Melissa C Southey, Mary Beth Terry, David Goldgar, Frances O'Malley, Esther M John, Ramunas Janavicius, Laima Tihomirova, Thomas V O Hansen, Finn C Nielsen, Ana Osorio, Alexandra Stavropoulou, Javier Benítez, Siranoush Manoukian, Bernard Peissel, Monica Barile, Sara Volorio, Barbara Pasini, Riccardo Dolcetti, Anna Laura Putignano, Laura Ottini, Paolo Radice, Ute Hamann, Muhammad U Rashid, Frans B Hogervorst, Mieke Kriege, Rob B van der Luijt, HEBON, Susan Peock, Debra Frost, D Gareth Evans, Carole Brewer, Lisa Walker, Mark T Rogers, Lucy E Side, Catherine Houghton, EMBRACE, Joellen Weaver, Andrew K Godwin, Rita K Schmutzler, Barbara Wappenschmidt, Alfons Meindl, Karin Kast, Norbert Arnold, Dieter Niederacher, Christian Sutter, Helmut Deissler, Doroteha Gadzicki, Sabine Preisler-Adams, Raymonda Varon-Mateeva, Ines Schönbuchner, Heidrun Gevensleben, Dominique Stoppa-Lyonnet, Muriel Belotti, Laure Barjhoux, GEMO Study Collaborators, Claudine Isaacs, Beth N Peshkin, Trinidad Caldes, Miguel de la Hoya, Carmen Cañadas, Tuomas Heikkinen, Päivi Heikkilä, Kristiina Aittomäki, Ignacio Blanco, Conxi Lazaro, Joan Brunet, Bjarni A Agnarsson, Adalgeir Arason, Rosa B Barkardottir, Martine Dumont, Jacques Simard, Marco Montagna, Simona Agata, Emma D'Andrea, Max Yan, Stephen Fox, kConFab Investigators, Timothy R Rebbeck, Wendy Rubinstein, Nadine Tung, Judy E Garber, Xianshu Wang, Zachary Fredericksen, Vernon S Pankratz, Noralane M Lindor, Csilla Szabo, Kenneth Offit, Rita Sakr, Mia M Gaudet, Christian F Singer, Muy-Kheng Tea, Christine Rappaport, Phuong L Mai, Mark H Greene, Anna Sokolenko, Evgeny Imyanitov, Amanda Ewart Toland, Leigha Senter, Kevin Sweet, Mads Thomassen, Anne-Marie Gerdes, Torben Kruse, Maria Caligo, Paolo Aretini, Johanna Rantala, Anna von Wachenfeld, Karin Henriksson, SWE-BRCA Collaborators, Linda Steele, Susan L Neuhausen, Robert Nussbaum, Mary Beattie, Kunle Odunsi, Lara Sucheston, Simon A Gayther, Kate Nathanson, Jenny Gross, Christine Walsh, Beth Karlan, Georgia Chenevix-Trench, Douglas F Easton, Antonis C Antoniou, and Consortium of Investigators of Modifiers of BRCA1/2. Pathology of breast and ovarian cancers among BRCA1 and BRCA2 mutation carriers: results from the consortium of investigators of modifiers of BRCA1/2 (CIMBA). *Cancer Epidemiol. Biomarkers Prev.*, 21(1):134–147, January 2012.

[227] Alexandra J van den Broek, Marjanka K Schmidt, Laura J van 't Veer, Rob A E M Tollenaar, and Flora E van Leeuwen. Worse breast cancer prognosis of BRCA1/BRCA2 mutation carriers: what's the evidence? a systematic review with meta-analysis. *PLoS One*, 10(3):e0120189, 27 March 2015.

[228] Balázs Győrffy, Christos Hatzis, Tara Sanft, Erin Hofstatter, Bilge Aktas, and Lajos Pusztai. Multigene prognostic tests in breast cancer: past, present, future. *Breast Cancer Res.*, 17:11, 27 January 2015.

[229] Brett Wallden, James Storhoff, Torsten Nielsen, Naeem 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,

Charles M Perou, Philip S Bernard, and Joel S Parker. Development and verification of the PAM50-based prosigna breast cancer gene signature assay. *BMC Med. Genomics*, 8:54, 22 August 2015.

[230] Muaiad Kittaneh, Alberto J Montero, and Stefan Glück. Molecular profiling for breast cancer: a comprehensive review. *Biomark. Cancer*, 5:61–70, 29 October 2013.

[231] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 18 April 2012.

[232] Philip J Stephens, David J McBride, Meng-Lay Lin, Ignacio Varela, Erin D Pleasance, Jared T Simpson, Lucy A Stebbings, Catherine Leroy, Sarah Edkins, Laura J Mudie, Chris D Greenman, Mingming Jia, Calli Latimer, Jon W Teague, King Wai Lau, John Burton, Michael A Quail, Harold Swerdlow, Carol Churcher, Rachael Natrajan, Anieta M Sieuwerts, John W M Martens, Daniel P Silver, Anita Langerød, Hege E G Russnes, John A Foekens, Jorge S Reis-Filho, Laura van 't Veer, Andrea L Richardson, Anne-Lise Børresen-Dale, Peter J Campbell, P Andrew Futreal, and Michael R Stratton. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, 462(7276):1005–1010, 24 December 2009.

[233] Göran Jönsson, Johan Staaf, Johan Vallon-Christersson, Markus Ringnér, Karolina Holm, Cecilia Hegardt, Haukur Gunnarsson, Rainer Fagerholm, Carina Strand, Bjarni A Agnarsson, Outi Kilpivaara, Lena Luts, Päivi Heikkilä, Kristiina Aittomäki, Carl Blomqvist, Niklas Loman, Per Malmström, Håkan Olsson, Oskar Th Johannsson, Adalgeir Arason, Heli Nevanlinna, Rosa B Barkardottir, and Ake Borg. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res.*, 12(3):R42, 24 June 2010.

[234] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 4 October 2012.

[235] Tomas Reinert, Everardo D Saad, Carlos H Barrios, and José Bines. Clinical implications of ESR1 mutations in hormone Receptor-Positive advanced breast cancer. *Front. Oncol.*, 7:26, 15 March 2017.

[236] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, Reanne Bowlby, Hui Shen, Sikander Hayat, Robert Fieldhouse, Susan C Lester,

Gary M K Tse, Rachel E Factor, Laura C Collins, Kimberly H Allison, Yunn-Yi Chen, Kristin Jensen, Nicole B Johnson, Steffi Oesterreich, Gordon B Mills, Andrew D Cherniack, Gordon Robertson, Christopher Benz, Chris Sander, Peter W Laird, Katherine A Hoadley, Tari A King, TCGA Research Network, and Charles M Perou. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, 8 October 2015.

[237] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B Brinkman, Sandro Morganella, Miriam R Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A Foekens, Moritz Gerstung, Gerrit K J Hooijer, Se Jin Jang, David R Jones, Hyung-Yong Kim, Tari A King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A Purdie, Keiran Raine, Kamna Ramakrishnan, F Germán Rodríguez-González, Gilles Romieu, Anieta M Sieuwerts, Peter T Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T Ueno, Christos Sotiriou, Alain Viari, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, John W M Martens, Anne-Lise Børresen-Dale, Andrea L Richardson, Gu Kong, Gilles Thomas, and Michael R Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2 June 2016.

[238] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 23 January 2014.

[239] Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John Marshall, Lucy A Stebbings, Andrew Menzies, Sancha Martin, Kenric Leung, Lina Chen, Catherine Leroy, Manasa Ramakrishna, Richard Rance, King Wai Lau, Laura J Mudie, Ignacio Varela, David J McBride, Graham R Bignell, Susanna L Cooke, Adam Shlien, John Gamble, Ian Whitmore, Mark Maddison, Patrick S Tarpey, Helen R Davies, Elli Papaemmanuil, Philip J Stephens, Stuart McLaren, Adam P Butler, Jon W Teague, Göran Jönsson, Judy E Garber, Daniel Silver, Penelope Miron, Aquila Fatima, Sandrine Boyault, Anita Langerød, Andrew Tutt, John W M Martens, Samuel A J R Aparicio, Åke Borg, Anne Vincent Salomon, Gilles Thomas, Anne-Lise Børresen-Dale, Andrea L Richardson, Michael S Neuberger, P Andrew Futreal, Peter J Campbell,

Michael R Stratton, and Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, 25 May 2012.

[240] Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Sieuwerts, Peter T Simpson, Tari A King, Keiran Raine, Jorunn E Eyfjord, Gu Kong, Åke Borg, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, Anne-Lise Børresen-Dale, John W M Martens, Paul N Span, Sunil R Lakhani, Anne Vincent-Salomon, Christos Sotiriou, Andrew Tutt, Alastair M Thompson, Steven Van Laere, Andrea L Richardson, Alain Viari, Peter J Campbell, Michael R Stratton, and Serena Nik-Zainal. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.*, 23(4):517–525, April 2017.

[241] Stephen Henderson, Ankur Chakravarthy, Xiaoping Su, Chris Boshoff, and Tim Robert Fenton. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep.*, 7(6):1833–1841, 26 June 2014.

[242] Nicholas McGranahan, Francesco Favero, Elza C de Bruin, Nicolai Juul Birkbak, Zoltan Szallasi, and Charles Swanton. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci. Transl. Med.*, 7(283):283ra54, 15 April 2015.

[243] Claudia R Baquet, Shiraz I Mishra, Patricia Commiskey, Gary L Ellison, and Mary DeShields. Breast cancer epidemiology in blacks and whites: disparities in incidence, mortality, survival rates and histology. *J. Natl. Med. Assoc.*, 100(5):480–488, May 2008.

[244] Lu Chen and Christopher I Li. Racial disparities in breast cancer diagnosis and treatment by hormone receptor and HER2 status. *Cancer Epidemiol. Biomarkers Prev.*, 24(11):1666–1672, November 2015.

[245] Lisa C Richardson, S Jane Henley, Jacqueline W Miller, Greta Massetti, and Cheryll C Thomas. Patterns and trends in Age-Specific Black-White differences in breast cancer incidence and mortality - united states, 1999-2014. *MMWR Morb. Mortal. Wkly. Rep.*, 65(40):1093–1098, 14 October 2016.

[246] Eric C Dietze, Christopher Sistrunk, Gustavo Miranda-Carboni, Ruth O'Regan, and Victoria L Seewaldt. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat. Rev. Cancer*, 15(4):248–254, April 2015.

[247] Katie M O'Brien, Stephen R Cole, Chiu-Kit Tse, Charles M Perou, Lisa A Carey, William D Foulkes, Lynn G Dressler, Joseph Geradts, and Robert C Millikan. Intrinsic breast tumor subtypes, race, and long-term survival in the carolina breast cancer study. *Clin. Cancer Res.*, 16(24):6100–6110, 15 December 2010.

[248] David R Williams, Selina A Mohammed, and Alexandra E Shields. Understanding and effectively addressing breast cancer in african american women: Unpacking the social context. *Cancer*, 122(14):2138–2149, 15 July 2016.

[249] Z Huang, S E Hankinson, G A Colditz, M J Stampfer, D J Hunter, J E Manson, C H Hennekens, B Rosner, F E Speizer, and W C Willett. Dual effects of weight and weight gain on breast cancer risk. *JAMA*, 278(17):1407–1411, 5 November 1997.

[250] C M Friedenreich. Physical activity and cancer prevention: from observational to intervention research. *Cancer Epidemiol. Biomarkers Prev.*, 10(4):287–301, April 2001.

[251] Mette Kalager, Marvin Zelen, Frøydis Langmark, and Hans-Olov Adami. Effect of screening mammography on breast-cancer mortality in norway. *N. Engl. J. Med.*, 363(13):1203–1210, 23 September 2010.

[252] Harald Weedon-Fekjær, Pål R Romundstad, and Lars J Vatten. Modern mammography screening and breast cancer mortality: population study. *BMJ*, 348:g3701, 17 June 2014.

[253] Evan R Myers, Patricia Moorman, Jennifer M Gierisch, Laura J Havrilesky, Lars J Grimm, Sujata Ghate, Brittany Davidson, Ranee Chatterjee Mongtomery, Matthew J Crowley, Douglas C McCrory, Amy Kendrick, and Gillian D Sanders. Benefits and harms of breast cancer screening: A systematic review. *JAMA*, 314(15):1615–1634, 20 October 2015.

[254] M Kalager, M Løberg, M Bretthauer, and H-O Adami. Comparative analysis of breast cancer mortality following mammography screening in denmark and norway. *Ann. Oncol.*, 25(6):1137–1143, June 2014.

[255] Rebecca Smith-Bindman, Diana L Miglioretti, Nicole Lurie, Linn Abraham, Rachel Ballard Barbash, Jodi Strzelczyk, Mark Dignan, William E Barlow, Cherry M Beasley, and Karla Kerlikowske. Does utilization of screening mammography explain racial and ethnic differences in breast cancer? *Ann. Intern. Med.*, 144(8):541–553, 18 April 2006.

[256] V W Chen, P Correa, R J Kurman, X C Wu, J W Eley, D Austin, H Muss, C P Hunter, C Redmond, and M Sobhan. Histological characteristics of breast carcinoma in blacks and whites. *Cancer Epidemiol. Biomarkers Prev.*, 3(2):127–135, March 1994.

[257] David N Danforth, Jr. Disparities in breast cancer outcomes between caucasian and african american women: a model for describing the relationship of biological and nonbiological factors. *Breast Cancer Res.*, 15(3):208, 27 June 2013.

[258] Mary A Gerend and Manacy Pai. Social determinants of Black-White disparities in breast cancer mortality: a review. *Cancer Epidemiol. Biomarkers Prev.*, 17(11):2913–2923, November 2008.

[259] Stephanie B Wheeler, Katherine E Reeder-Hayes, and Lisa A Carey. Disparities in breast cancer treatment and outcomes: biological, social, and health system determinants and opportunities for research. *Oncologist*, 18(9):986–993, 12 August 2013.

[260] Zachary D Stephens, Skylar Y Lee, Faraz Faghri, Roy H Campbell, Chengxiang Zhai, Miles J Efron, Ravishankar Iyer, Michael C Schatz, Saurabh Sinha, and Gene E Robinson. Big data: Astronomical or genomical? *PLoS Biol.*, 13(7):e1002195, July 2015.

[261] Gordon Bell, Tony Hey, and Alex Szalay. Computer science. beyond the data deluge. *Science*, 323(5919):1297–1298, 6 March 2009.

[262] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 42–47. IEEE, May 2013.

[263] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of ?big data? on cloud computing: Review and open research issues. *Inf. Syst.*, 47:98–115, January 2015.

[264] Chris A Mattmann. Computing: A vision for data science. *Nature*, 493(7433):473–475, 2013.

[265] Benjamin J Kelly, James R Fitch, Yangqiu Hu, Donald J Corsmeier, Huachun Zhong, Amy N Wetzel, Russell D Nordquist, David L Newsom, and Peter White. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biol.*, 16:6, 20 January 2015.

[266] Colby Chiang, Ryan M Layer, Gregory G Faust, Michael R Lindberg, David B Rose, Erik P Garrison, Gabor T Marth, Aaron R Quinlan, and Ira M Hall. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat. Methods*, 12(10):966–968, October 2015.

[267] Megan J Puckelwartz, Lorenzo L Pesce, Viswateja Nelakuditi, Lisa Dellefave-Castillo, Jessica R Golbus, Sharlene M Day, Thomas P Cappola, Gerald W Dorn, 2nd, Ian T Foster, and Elizabeth M McNally. Supercomputing for the parallelization of whole genome analysis. *Bioinformatics*, 30(11):1508–1513, 1 June 2014.

[268] Riyue Bao, Kyle Hernandez, Lei Huang, Wenjun Kang, Elizabeth Bartom, Kenan Onel, Samuel Volchenboum, and Jorge Andrade. ExScalibur: A High-Performance Cloud-Enabled suite for whole exome germline and somatic mutation identification. *PLoS One*, 10(8):e0135800, 13 August 2015.

[269] Hugo Y K Lam, Cuiping Pan, Michael J Clark, Phil Lacroute, Rui Chen, Rajini Haraksingh, Maeve O'Huallachain, Mark B Gerstein, Jeffrey M Kidd, Carlos D Bustamante, and Michael Snyder. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat. Biotechnol.*, 30(3):226–229, 7 March 2012.

[270] Mohamed Abouelhoda, Shadi Alaa Issa, and Moustafa Ghanem. Tavaxy: integrating taverna and galaxy workflows with cloud computing support. *BMC Bioinformatics*, 13:77, 4 May 2012.

[271] Tyler W H Backman and Thomas Girke. systemPipeR: NGS workflow and report generation environment. *BMC Bioinformatics*, 17:388, 20 September 2016.

[272] Enis Afgan, Dannon Baker, Marius van den Beek, Daniel Blankenberg, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Greg Von Kuster, Eric Rasche, Nicola Soranzo, Nitesh Turaga, James Taylor, Anton Nekrutenko, and Jeremy Goecks. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, 44(W1):W3–W10, 8 July 2016.

[273] Jeffrey G Reid, Andrew Carroll, Narayanan Veeraraghavan, Mahmoud Dahdouli, Andreas Sundquist, Adam English, Matthew Bainbridge, Simon White, William Salerno, Christian Buhay, Fuli Yu, Donna Muzny, Richard Daly, Geoff Duyk, Richard A Gibbs, and Eric Boerwinkle. Launching genomics into the cloud: deployment of mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics*, 15:30, 29 January 2014.

[274] Michael Wilde, Mihael Hategan, Justin M Wozniak, Ben Clifford, Daniel S Katz, and Ian Foster. Swift: A language for distributed parallel scripting. *Parallel Comput.*, 37(9):633–652, 2011.

[275] Christopher Wilks, Melissa S Cline, Erich Weiler, Mark Diehkans, Brian Craft, Christy Martin, Daniel Murphy, Howdy Pierce, John Black, Donavan Nelson, Brian Litzinger, Thomas Hatton, Lori Maltbie, Michael Ainsworth, Patrick Allen, Linda Rosewood, Elizabeth Mitchell, Bradley Smith, Jim Warner, John Groboske, Haifang Telc, Daniel Wilson, Brian Sanford, Hannes Schmidt, David Haussler, and Daniel Maltbie. The cancer genomics hub (CGHub): overcoming cancer through the power of torrential data. *Database*, 2014, 29 September 2014.

[276] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 1 October 2015.

[277] 1000 Genomes Project Consortium, Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard A Gibbs, Matt E Hurles, and Gil A McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 28 October 2010.

[278] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, 375(12):1109–1112, 22 September 2016.

[279] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[280] statgen. statgen/bamutil. `https://github.com/statgen/bamUtil`. Accessed: 2017-6-25.

[281] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9):1297–1303, September 2010.

[282] Diimitrios Georgakopoulos, Mark Hornick, and Amit Sheth. An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, 3(2):119–153, 1995.

[283] W M P van der Aalst, A H M ter Hofstede, B Kiepuszewski, and A P Barros. Workflow patterns. *Distributed and Parallel Databases*, 14(1):5–51, 1 July 2003.

[284] Nick Russell, Arthur H M ter Hofstede, David Edmond, and Wil M P van der Aalst. Workflow data patterns: Identification, representation and tool support. In *Conceptual Modeling – ER 2005*, pages 353–368. Springer, Berlin, Heidelberg, 24 October 2005.

[285] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 15 July 2009.

[286] Jean-François Pineau, Yves Robert, and Frédéric Vivien. The impact of heterogeneity on master-slave scheduling. *Parallel Comput.*, 34(3):158–176, 2008.

[287] M Sullivan and D Anderson. Marionette: a system for parallel distributed programming using a master/slave model. In *[1989] Proceedings. The 9th International Conference on Distributed Computing Systems*.

[288] Sanaz Mostaghim, Jurgen Branke, Andrew Lewis, and Hartmut Schmeck. Parallel multi-objective optimization using Master-Slave model on heterogeneous resources. In *2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence)*, 2008.

[289] Matthias Hovestadt, Odej Kao, Axel Keller, and Achim Streit. Scheduling in HPC resource management systems: Queuing vs. planning. In Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn, editors, *Job Scheduling Strategies for Parallel Processing*, volume 2862 of *Lecture Notes in Computer Science*, pages 1–20. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[290] Klaus Krauter, Rajkumar Buyya, and Muthucumaru Maheswaran. A taxonomy and survey of grid resource management systems for distributed computing. *Softw. Pract. Exp.*, 32(2):135–164, February 2002.

[291] M Hategan, J Wozniak, and K Maheshwari. Coasters: Uniform resource provisioning and access for clouds and grids. In *2011 Fourth IEEE International Conference on Utility and Cloud Computing*, 2011.

[292] Cesare Pautasso and Gustavo Alonso. Parallel computing patterns for grid workflows. In *2006 Workshop on Workflows in Support of Large-Scale Science*, pages 1–10. IEEE, June 2006.

[293] Bertram Ludäscher, Ilkay Altintas, Chad Berkley, Dan Higgins, Efrat Jaeger, Matthew Jones, Edward A Lee, Jing Tao, and Yang Zhao. Scientific workflow management and the kepler system. *Concurr. Comput.*, 18(10):1039–1065, 2006.

[294] Nick Russell, Wil M P van der Aalst, Arthur H M ter Hofstede, and David Edmond. Workflow resource patterns: Identification, representation and tool support. In *Lecture Notes in Computer Science*, pages 216–232. 2005.

[295] Heinz Stockinger, Marco Pagni, Lorenzo Cerutti, and Laurent Falquet. Grid approach to embarrassingly parallel CPU-Intensive bioinformatics problems. In *2006 Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)*, 2006.

[296] A Litvinova, C Engelmann, and S L Scott. A proactive fault tolerance framework for High-Performance computing. In *Parallel and Distributed Computing and Networks*, Calgary,AB,Canada, 2010. ACTAPRESS.

[297] H Nakamura, T Hayashida, M Kondo, Y Tajima, M Imai, and T Nanya. Skewed checkpointing for tolerating multi-node failures. In *Proceedings of the 23rd IEEE International Symposium on Reliable Distributed Systems, 2004.*, pages 116–125. IEEE, 2004.

[298] Lars Feuk, Andrew R Carson, and Stephen W Scherer. Structural variation in the human genome. *Nat. Rev. Genet.*, 7(2):85–97, February 2006.

[299] Ryan L Collins, Harrison Brand, Claire E Redin, Carrie Hanscom, Caroline Antolik, Matthew R Stone, Joseph T Glessner, Tamara Mason, Giulia Pregno, Naghmeh Dorrani, Giorgia Mandrile, Daniela Giachino, Danielle Perrin, Cole Walsh, Michelle Cipicchio, Maura Costello, Alexei Stortchevoi, Joon-Yong An, Benjamin B Currall, Catarina M Seabra, Ashok Ragavendran, Lauren Margolin, Julian A Martinez-Agosto, Diane Lucente, Brynn Levy, Stephan J Sanders, Ronald J Wapner, Fabiola Quintero-Rivera, Wigard Kloosterman, and Michael E Talkowski. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. *Genome Biol.*, 18(1):36, 6 March 2017.

[300] Scaling behavior of short read sequence aligners. `http://en.community.dell.com/techcenter/high-performance-computing/b/genomics/archive/2016/07/26/scaling-behavior-of-short-read-sequence-aligners`. Accessed: 2017-6-26.

[301] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, April 2012.

[302] Derek W Barnett, Erik K Garrison, Aaron R Quinlan, Michael P Strömberg, and Gabor T Marth. BamTools: a c++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):1691–1692, 15 June 2011.

[303] John Ousterhout and Fred Douglis. Beating the I/O bottleneck: A case for log-structured file systems. *Oper. Syst. Rev.*, 23(1):11–28, January 1989.

[304] Feng Wang, Qin Xin, Bo Hong, Scott A Brandt, Ethan Miller, Darrell Long, and T McLarty. File system workload analysis for large scale scientific computing applications. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 2004.

[305] Kosar and Tevfik. *Data Intensive Distributed Computing: Challenges and Solutions for Large-scale Information Management: Challenges and Solutions for Large-scale Information Management*. IGI Global, 31 January 2012.

[306] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, SOSP '03, pages 29–43, New York, NY, USA, 2003. ACM.

[307] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, October 2013.

[308] UK10K Consortium, Klaudia Walter, Josine L Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, John R B Perry, Changjiang Xu, Marta Futema, Daniel Lawson, Valentina Iotchkova, Stephan Schiffels, Audrey E Hendricks, Petr Danecek, Rui Li, James Floyd, Louise V Wain, Inês Barroso, Steve E Humphries, Matthew E Hurles, Eleftheria Zeggini, Jeffrey C Barrett, Vincent Plagnol, J Brent Richards, Celia M T Greenwood, Nicholas J Timpson, Richard Durbin, and Nicole Soranzo. The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571):82–90, 1 October 2015.

[309] Lincoln D Stein, Bartha M Knoppers, Peter Campbell, Gad Getz, and Jan O Korbel. Data analysis: Create a cloud commons. *Nature*, 523(7559):149–151, 9 July 2015.

[310] Jill C Mwenifumbo and Marco A Marra. Cancer genome-sequencing study design. *Nat. Rev. Genet.*, 14(5):321–332, May 2013.

[311] Pierre Martinez, Nicolai Juul Birkbak, Marco Gerlinger, Nicholas McGranahan, Rebecca A Burrell, Andrew J Rowan, Tejal Joshi, Rosalie Fisher, James Larkin, Zoltan Szallasi, and Charles Swanton. Parallel evolution of tumour subclones mimics diversity between tumours. *J. Pathol.*, 230(4):356–364, August 2013.

[312] Noushin Niknafs, Violeta Beleva-Guthrie, Daniel Q Naiman, and Rachel Karchin. SubClonal hierarchy inference from somatic mutations: Automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput. Biol.*, 11(10):e1004416, October 2015.

[313] Christopher J Ricketts and W Marston Linehan. Intratumoral heterogeneity in kidney cancer. *Nat. Genet.*, 46(3):214–215, March 2014.

[314] Marco Gerlinger, Andrew J Rowan, Stuart Horswell, James Larkin, David Endesfelder, Eva Gronroos, Pierre Martinez, Nicholas Matthews, Aengus Stewart, Patrick Tarpey, Ignacio Varela, Benjamin Phillimore, Sharmin Begum, Neil Q McDonald, Adam Butler, David Jones, Keiran Raine, Calli Latimer, Claudio R Santos, Mahrokh Nohadani, Aron C Eklund, Bradley Spencer-Dene, Graham Clark, Lisa Pickering, Gordon Stamp, Martin Gore, Zoltan Szallasi, Julian Downward, P Andrew Futreal, and Charles Swanton. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.*, 366(10):883–892, 8 March 2012.

[315] Li Ding, Matthew J Ellis, Shunqiang Li, David E Larson, Ken Chen, John W Wallis, Christopher C Harris, Michael D McLellan, Robert S Fulton, Lucinda L Fulton, Rachel M Abbott, Jeremy Hoog, David J Dooling, Daniel C Koboldt, Heather Schmidt, Joelle Kalicki, Qunyuan Zhang, Lei Chen, Ling Lin, Michael C Wendl, Joshua F McMichael, Vincent J Magrini, Lisa Cook, Sean D McGrath, Tammi L Vickery, Elizabeth Appelbaum, Katherine Deschryver, Sherri Davies, Therese Guintoli, Li Lin, Robert Crowder, Yu Tao, Jacqueline E Snider, Scott M Smith, Adam F Dukes, Gabriel E Sanderson, Craig S Pohl, Kim D Delehaunty, Catrina C Fronick, Kimberley A Pape, Jerry S Reed, Jody S Robinson, Jennifer S Hodges, William Schierding, Nathan D Dees, Dong Shen, Devin P Locke, Madeline E Wiechert, James M Eldred, Josh B Peck, Benjamin J Oberkfell, Justin T Lolofie, Feiyu Du, Amy E Hawkins, Michelle D O'Laughlin, Kelly E Bernard, Mark Cunningham, Glendoria Elliott, Mark D Mason, Dominic M Thompson, Jr, Jennifer L Ivanovich, Paul J Goodfellow, Charles M Perou, George M Weinstock, Rebecca Aft, Mark Watson, Timothy J Ley, Richard K Wilson, and Elaine R Mardis. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature*, 464(7291):999–1005, 15 April 2010.

[316] Gunes Gundem, Peter Van Loo, Barbara Kremeyer, Ludmil B Alexandrov, Jose M C Tubio, Elli Papaemmanuil, Daniel S Brewer, Heini M L Kallio, Gunilla Högnäs, Matti Annala, Kati Kivinummi, Victoria Goody, Calli Latimer, Sarah O'Meara, Kevin J Dawson, William Isaacs, Michael R Emmert-Buck, Matti Nykter, Christopher Foster, Zsofia Kote-Jarai, Douglas Easton, Hayley C Whitaker, ICGC Prostate UK Group, David E Neal, Colin S Cooper, Rosalind A Eeles, Tapio Visakorpi, Peter J Campbell,

Ultan McDermott, David C Wedge, and G Steven Bova. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 16 April 2015.

[317] Jens G Lohr, Viktor A Adalsteinsson, Kristian Cibulskis, Atish D Choudhury, Mara Rosenberg, Peter Cruz-Gordillo, Joshua M Francis, Cheng-Zhong Zhang, Alex K Shalek, Rahul Satija, John J Trombetta, Diana Lu, Naren Tallapragada, Narmin Tahirova, Sora Kim, Brendan Blumenstiel, Carrie Sougnez, Alarice Lowe, Bang Wong, Daniel Auclair, Eliezer M Van Allen, Mari Nakabayashi, Rosina T Lis, Gwo-Shu M Lee, Tiantian Li, Matthew S Chabot, Amy Ly, Mary-Ellen Taplin, Thomas E Clancy, Massimo Loda, Aviv Regev, Matthew Meyerson, William C Hahn, Philip W Kantoff, Todd R Golub, Gad Getz, Jesse S Boehm, and J Christopher Love. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.*, 32(5):479–484, May 2014.

[318] Priscilla K Brastianos, Scott L Carter, Sandro Santagata, Daniel P Cahill, Amaro Taylor-Weiner, Robert T Jones, Eliezer M Van Allen, Michael S Lawrence, Peleg M Horowitz, Kristian Cibulskis, Keith L Ligon, Josep Tabernero, Joan Seoane, Elena Martinez-Saez, William T Curry, Ian F Dunn, Sun Ha Paek, Sung-Hye Park, Aaron McKenna, Aaron Chevalier, Mara Rosenberg, Frederick G Barker, 2nd, Corey M Gill, Paul Van Hummelen, Aaron R Thorner, Bruce E Johnson, Mai P Hoang, Toni K Choueiri, Sabina Signoretti, Carrie Sougnez, Michael S Rabin, Nancy U Lin, Eric P Winer, Anat Stemmer-Rachamimov, Matthew Meyerson, Levi Garraway, Stacey Gabriel, Eric S Lander, Rameen Beroukhim, Tracy T Batchelor, José Baselga, David N Louis, Gad Getz, and William C Hahn. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.*, 5(11):1164–1177, November 2015.

[319] Muhammed Murtaza, Sarah-Jane Dawson, Katherine Pogrebniak, Oscar M Rueda, Elena Provenzano, John Grant, Suet-Feung Chin, Dana W Y Tsui, Francesco Marass, Davina Gale, H Raza Ali, Pankti Shah, Tania Contente-Cuomo, Hossein Farahani, Karey Shumansky, Zoya Kingsbury, Sean Humphray, David Bentley, Sohrab P Shah, Matthew Wallis, Nitzan Rosenfeld, and Carlos Caldas. Multifocal clonal evolution characterized using circulating tumour DNA in a case of metastatic breast cancer. *Nat. Commun.*, 6:8760, 4 November 2015.

[320] Melissa A Troester, Katherine A Hoadley, Monica D?Arcy, Andrew D Cherniack, Chip Stewart, Daniel C Koboldt, A Gordon Robertson, Swapna Mahurkar, Hui Shen, Matthew D Wilkerson, Rupninder Sandhu, Nicole B Johnson, Kimberly H Allison, Andrew H Beck, Christina Yau, Jay Bowen, Margi Sheth, E Shelley Hwang, Charles M Perou, Peter W Laird, Li Ding, and Christopher C Benz. DNA defects, epigenetics, and gene expression in cancer-adjacent breast: a study from the cancer genome atlas. *npj Breast Cancer*, 2(1):721, 4 December 2016.

[321] Lei Cai, Wei Yuan, Zhou Zhang, Lin He, and Kuo-Chen Chou. In-depth comparison of somatic point mutation callers based on different tumor next-generation sequencing depth data. *Sci. Rep.*, 6:36540, 22 November 2016.

[322] Maurizio Callari, Stephen-John Sammut, Leticia De Mattos-Arruda, Alejandra Bruna, Oscar M Rueda, Suet-Feung Chin, and Carlos Caldas. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Med.*, 9(1):35, 18 April 2017.

[323] David L Goode, Sally M Hunter, Maria A Doyle, Tao Ma, Simone M Rowley, David Choong, Georgina L Ryland, and Ian G Campbell. A simple consensus approach improves somatic mutation prediction accuracy. *Genome Med.*, 5(9):90, 30 September 2013.

[324] Tyler S Alioto, Ivo Buchhalter, Sophia Derdak, Barbara Hutter, Matthew D Eldridge, Eivind Hovig, Lawrence E Heisler, Timothy A Beck, Jared T Simpson, Laurie Tonon, Anne-Sophie Sertier, Ann-Marie Patch, Natalie Jäger, Philip Ginsbach, Ruben Drews, Nagarajan Paramasivam, Rolf Kabbe, Sasithorn Chotewutmontri, Nicolle Diessl, Christopher Previti, Sabine Schmidt, Benedikt Brors, Lars Feuerbach, Michael Heinold, Susanne Gröbner, Andrey Korshunov, Patrick S Tarpey, Adam P Butler, Jonathan Hinton, David Jones, Andrew Menzies, Keiran Raine, Rebecca Shepherd, Lucy Stebbings, Jon W Teague, Paolo Ribeca, Francesc Castro Giner, Sergi Beltran, Emanuele Raineri, Marc Dabad, Simon C Heath, Marta Gut, Robert E Denroche, Nicholas J Harding, Takafumi N Yamaguchi, Akihiro Fujimoto, Hidewaki Nakagawa, Víctor Quesada, Rafael Valdés-Mas, Sigve Nakken, Daniel Vodák, Lawrence Bower, Andrew G Lynch, Charlotte L Anderson, Nicola Waddell, John V Pearson, Sean M Grimmond, Myron Peto, Paul Spellman, Minghui He, Cyriac Kandoth, Semin Lee, John Zhang, Louis Létourneau, Singer Ma, Sahil Seth, David Torrents, Liu Xi, David A Wheeler, Carlos López-Otín, Elías Campo, Peter J Campbell, Paul C Boutros, Xose S Puente, Daniela S Gerhard, Stefan M Pfister, John D McPherson, Thomas J Hudson, Matthias Schlesner, Peter Lichter, Roland Eils, David T W Jones, and Ivo G Gut. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.*, 6:10001, 9 December 2015.

[325] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219, March 2013.

[326] Christopher T Saunders, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 15 July 2012.

[327] Daniel C Koboldt, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher A Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576, March 2012.

[328] Garrick Staples. TORQUE resource manager. In *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, SC '06, New York, NY, USA, 2006. ACM.

[329] R Henderson and D Tweten. Portable batch system: External reference specification. Technical report, 1996.

[330] Andy B Yoo, Morris A Jette, and Mark Grondona. SLURM: Simple linux utility for resource management. In *Lecture Notes in Computer Science*, pages 44–60. 2003.

[331] Gentzsch and W. Sun grid engine : Towards creating a compute power grid. *Proc. first IEEEACM International Symposium on Cluster Computing and the Grid, IEEE Computer Society, 2001*, pages 35–36, 2001.

[332] Marc Cohen, Kathryn Hurley, and Paul Newson. *Google Compute Engine: Managing Secure and Scalable Cloud Computing*. "O'Reilly Media, Inc.", 15 December 2014.

[333] Simon Ostermann, Alexandria Iosup, Nezih Yigitbasi, Radu Prodan, Thomas Fahringer, and Dick Epema. A performance analysis of EC2 cloud computing services for scientific computing. *Cloud computing*, pages 115–131, 2010.

[334] M Mesnier, G R Ganger, and E Riedel. Storage area networking - object-based storage. *IEEE Commun. Mag.*, 41(8):84–90, August 2003.

[335] Van Tran, Jacky Keung, Anna Liu, and Alan Fekete. Application migration to cloud: a taxonomy of critical factors. In *Proceeding of the 2nd international workshop on Software engineering for cloud computing - SECLOUD '11*, page 22, New York, New York, USA, 2011. ACM Press.

[336] Keith R Jackson, Lavanya Ramakrishnan, Krishna Muriki, Shane Canon, Shreyas Cholia, John Shalf, Harvey J Wasserman, and Nicholas J Wright. Performance analysis of high performance computing applications on the amazon web services cloud. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, pages 159–168. IEEE, November 2010.

[337] G R Joubert, H Leather, and M Parsons. *Parallel Computing: On the Road to Exascale*. IOS Press, 28 April 2016.

[338] Rob F Van der Wijngaart, Abdullah Kayi, Jeff R Hammond, Gabriele Jost, Tom St John, Srinivas Sridharan, Timothy G Mattson, John Abercrombie, and Jacob Nelson. Comparing runtime systems with exascale ambitions using the parallel research kernels. In *ISC*, pages 321–339. researchgate.net, 2016.

[339] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, WGS500 Consortium, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, 46(8):912–918, August 2014.

[340] Benjamin J Evans and Martijn Verburg. *The Well-grounded Java Developer: Vital Techniques of Java 7 and Polyglot Programming*. Manning Publications, 2013.

[341] Judy E Garber and Kenneth Offit. Hereditary cancer predisposition syndromes. *J. Clin. Oncol.*, 23(2):276–292, 10 January 2005.

[342] A Brandt, J Lorenzo Bermejo, J Sundquist, and K Hemminki. Age of onset in familial cancer. *Ann. Oncol.*, 19(12):2084–2088, December 2008.

[343] Olivia Fletcher and Richard S Houlston. Architecture of inherited susceptibility to common cancer. *Nat. Rev. Cancer*, 10(5):353–361, May 2010.

[344] Lorelei A Mucci, Jacob B Hjelmborg, Jennifer R Harris, Kamila Czene, David J Havelick, Thomas Scheike, Rebecca E Graff, Klaus Holst, Sören Möller, Robert H Unger, Christina McIntosh, Elizabeth Nuttall, Ingunn Brandt, Kathryn L Penney, Mikael Hartman, Peter Kraft, Giovanni Parmigiani, Kaare Christensen, Markku Koskenvuo, Niels V Holm, Kauko Heikkilä, Eero Pukkala, Axel Skytthe, Hans-Olov Adami, Jaakko Kaprio, and Nordic Twin Study of Cancer (NorTwinCan) Collaboration. Familial risk and heritability of cancer among twins in nordic countries. *JAMA*, 315(1):68–76, 5 January 2016.

[345] Kyriaki Michailidou, Per Hall, Anna Gonzalez-Neira, Maya Ghoussaini, Joe Dennis, Roger L Milne, Marjanka K Schmidt, Jenny Chang-Claude, Stig E Bojesen, Manjeet K Bolla, Qin Wang, Ed Dicks, Andrew Lee, Clare Turnbull, Nazneen Rahman, Breast and Ovarian Cancer Susceptibility Collaboration, Olivia Fletcher, Julian Peto, Lorna Gibson, Isabel Dos Santos Silva, Heli Nevanlinna, Taru A Muranen, Kristiina Aittomäki, Carl Blomqvist, Kamila Czene, Astrid Irwanto, Jianjun Liu, Quinten Waisfisz, Hanne Meijers-Heijboer, Muriel Adank, Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), Rob B van der Luijt, Rebecca Hein, Norbert Dahmen, Lars Beckman, Alfons Meindl, Rita K Schmutzler, Bertram Müller-Myhsok, Peter Lichtner, John L Hopper, Melissa C Southey, Enes Makalic, Daniel F Schmidt, Andre G Uitterlinden, Albert Hofman, David J Hunter, Stephen J Chanock, Daniel Vincent, François Bacot, Daniel C Tessier, Sander Canisius, Lodewyk F A Wessels, Christopher A Haiman, Mitul Shah, Robert Luben, Judith Brown, Craig Luccarini, Nils Schoof, Keith Humphreys, Jingmei Li, Børge G Nordestgaard, Sune F Nielsen, Henrik Flyger, Fergus J Couch, Xianshu Wang, Celine Vachon, Kristen N Stevens, Diether Lambrechts, Matthieu Moisse, Robert Paridaens, Marie-Rose Christiaens, Anja Rudolph, Stefan Nickels, Dieter Flesch-Janys, Nichola Johnson, Zoe Aitken, Kirsimari Aaltonen, Tuomas Heikkinen, Annegien Broeks, Laura J Van't Veer, C Ellen van der Schoot, Pascal Guénel, Thérèse Truong, Pierre Laurent-Puig, Florence Menegaux, Frederik Marme, Andreas Schneeweiss, Christof Sohn, Barbara Burwinkel, M Pilar Zamora, Jose Ignacio Arias Perez, Guillermo Pita, M Rosario Alonso, Angela Cox, Ian W Brock, Simon S Cross, Malcolm W R Reed, Elinor J Sawyer, Ian Tomlinson, Michael J Kerin, Nicola Miller, Brian E Henderson, Fredrick Schumacher, Loic Le Marchand, Irene L Andrulis, Julia A Knight, Gord Glendon, Anna Marie Mulligan, kConFab Investigators, Australian Ovarian Cancer Study Group, Annika Lindblom, Sara Margolin, Maartje J Hooning, Antoinette Hollestelle, Ans M W van den

Ouweland, Agnes Jager, Quang M Bui, Jennifer Stone, Gillian S Dite, Carmel Apicella, Helen Tsimiklis, Graham G Giles, Gianluca Severi, Laura Baglietto, Peter A Fasching, Lothar Haeberle, Arif B Ekici, Matthias W Beckmann, Hermann Brenner, Heiko Müller, Volker Arndt, Christa Stegmaier, Anthony Swerdlow, Alan Ashworth, Nick Orr, Michael Jones, Jonine Figueroa, Jolanta Lissowska, Louise Brinton, Mark S Goldberg, France Labrèche, Martine Dumont, Robert Winqvist, Katri Pylkäs, Arja Jukkola-Vuorinen, Mervi Grip, Hiltrud Brauch, Ute Hamann, Thomas Brüning, GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network, Paolo Radice, Paolo Peterlongo, Siranoush Manoukian, Bernardo Bonanni, Peter Devilee, Rob A E M Tollenaar, Caroline Seynaeve, Christi J van Asperen, Anna Jakubowska, Jan Lubinski, Katarzyna Jaworska, Katarzyna Durda, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M Hartikainen, Natalia V Bogdanova, Natalia N Antonenkova, Thilo Dörk, Vessela N Kristensen, Hoda Anton-Culver, Susan Slager, Amanda E Toland, Stephen Edge, Florentia Fostira, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Keitaro Matsuo, Hidemi Ito, Hiroji Iwata, Aiko Sueta, Anna H Wu, Chiu-Chen Tseng, David Van Den Berg, Daniel O Stram, Xiao-Ou Shu, Wei Lu, Yu-Tang Gao, Hui Cai, Soo Hwang Teo, Cheng Har Yip, Sze Yee Phuah, Belinda K Cornes, Mikael Hartman, Hui Miao, Wei Yen Lim, Jen-Hwei Sng, Kenneth Muir, Artitaya Lophatananon, Sarah Stewart-Brown, Pornthep Siriwanarangsan, Chen-Yang Shen, Chia-Ni Hsiung, Pei-Ei Wu, Shian-Ling Ding, Suleeporn Sangrajrang, Valerie Gaborieau, Paul Brennan, James McKay, William J Blot, Lisa B Signorello, Qiuyin Cai, Wei Zheng, Sandra Deming-Halverson, Martha Shrubsole, Jirong Long, Jacques Simard, Montse Garcia-Closas, Paul D P Pharoah, Georgia Chenevix-Trench, Alison M Dunning, Javier Benitez, and Douglas F Easton. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.*, 45(4):353–61, 361e1–2, April 2013.

[346] Rosalind A Eeles, Ali Amin Al Olama, Sara Benlloch, Edward J Saunders, Daniel A Leongamornlert, Malgorzata Tymrakiewicz, Maya Ghoussaini, Craig Luccarini, Joe Dennis, Sarah Jugurnauth-Little, Tokhir Dadaev, David E Neal, Freddie C Hamdy, Jenny L Donovan, Ken Muir, Graham G Giles, Gianluca Severi, Fredrik Wiklund, Henrik Gronberg, Christopher A Haiman, Fredrick Schumacher, Brian E Henderson, Loic Le Marchand, Sara Lindstrom, Peter Kraft, David J Hunter, Susan Gapstur, Stephen J Chanock, Sonja I Berndt, Demetrius Albanes, Gerald Andriole, Johanna Schleutker, Maren Weischer, Federico Canzian, Elio Riboli, Tim J Key, Ruth C Travis, Daniele Campa, Sue A Ingles, Esther M John, Richard B Hayes, Paul D P Pharoah, Nora Pashayan, Kay-Tee Khaw, Janet L Stanford, Elaine A Ostrander, Lisa B Signorello, Stephen N Thibodeau, Dan Schaid, Christiane Maier, Walther Vogel, Adam S Kibel, Cezary Cybulski, Jan Lubinski, Lisa Cannon-Albright, Hermann Brenner, Jong Y Park, Radka Kaneva, Jyotsna Batra, Amanda B Spurdle, Judith A Clements, Manuel R Teixeira, Ed Dicks, Andrew Lee, Alison M Dunning, Caroline Baynes, Don Conroy, Melanie J Maranian, Shahana Ahmed, Koveela Govindasami, Michelle Guy, Rosemary A Wilkinson, Emma J Sawyer, Angela Morgan, David P Dearnaley, Alan Horwich, Robert A Huddart, Vincent S Khoo, Christopher C Parker, Nicholas J Van As,

Christopher J Woodhouse, Alan Thompson, Tim Dudderidge, Chris Ogden, Colin S Cooper, Artitaya Lophatananon, Angela Cox, Melissa C Southey, John L Hopper, Dallas R English, Markus Aly, Jan Adolfsson, Jiangfeng Xu, Siqun L Zheng, Meredith Yeager, Rudolf Kaaks, W Ryan Diver, Mia M Gaudet, Mariana C Stern, Roman Corral, Amit D Joshi, Ahva Shahabi, Tiina Wahlfors, Teuvo L J Tammela, Anssi Auvinen, Jarmo Virtamo, Peter Klarskov, Børge G Nordestgaard, M Andreas Røder, Sune F Nielsen, Stig E Bojesen, Afshan Siddiq, Liesel M Fitzgerald, Suzanne Kolb, Erika M Kwon, Danielle M Karyadi, William J Blot, Wei Zheng, Qiuyin Cai, Shannon K McDonnell, Antje E Rinckleb, Bettina Drake, Graham Colditz, Dominika Wokolorczyk, Robert A Stephenson, Craig Teerlink, Heiko Muller, Dietrich Rothenbacher, Thomas A Sellers, Hui-Yi Lin, Chavdar Slavov, Vanio Mitev, Felicity Lose, Srilakshmi Srinivasan, Sofia Maia, Paula Paulo, Ethan Lange, Kathleen A Cooney, Antonis C Antoniou, Daniel Vincent, François Bacot, Daniel C Tessier, COGS–Cancer Research UK GWAS–ELLIPSE (part of GAME-ON) Initiative, Australian Prostate Cancer Bioresource, UK Genetic Prostate Cancer Study Collaborators/British Association of Urological Surgeons' Section of Oncology, UK ProtecT (Prostate testing for cancer and Treatment) Study Collaborators, PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium, Zsofia Kote-Jarai, and Douglas F Easton. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.*, 45(4):385–91, 391e1–2, April 2013.

[347] M J E Frampton, P Law, K Litchfield, E J Morris, D Kerr, C Turnbull, I P Tomlinson, and R S Houlston. Implications of polygenic risk for personalised colorectal cancer screening. *Ann. Oncol.*, 27(3):429–434, March 2016.

[348] Nasim Mavaddat, Paul D P Pharoah, Kyriaki Michailidou, Jonathan Tyrer, Mark N Brook, Manjeet K Bolla, Qin Wang, Joe Dennis, Alison M Dunning, Mitul Shah, Robert Luben, Judith Brown, Stig E Bojesen, Børge G Nordestgaard, Sune F Nielsen, Henrik Flyger, Kamila Czene, Hatef Darabi, Mikael Eriksson, Julian Peto, Isabel Dos-Santos-Silva, Frank Dudbridge, Nichola Johnson, Marjanka K Schmidt, Annegien Broeks, Senno Verhoef, Emiel J Rutgers, Anthony Swerdlow, Alan Ashworth, Nick Orr, Minouk J Schoemaker, Jonine Figueroa, Stephen J Chanock, Louise Brinton, Jolanta Lissowska, Fergus J Couch, Janet E Olson, Celine Vachon, Vernon S Pankratz, Diether Lambrechts, Hans Wildiers, Chantal Van Ongeval, Erik van Limbergen, Vessela Kristensen, Grethe Grenaker Alnæs, Silje Nord, Anne-Lise Borresen-Dale, Heli Nevanlinna, Taru A Muranen, Kristiina Aittomäki, Carl Blomqvist, Jenny Chang-Claude, Anja Rudolph, Petra Seibold, Dieter Flesch-Janys, Peter A Fasching, Lothar Haeberle, Arif B Ekici, Matthias W Beckmann, Barbara Burwinkel, Frederik Marme, Andreas Schneeweiss, Christof Sohn, Amy Trentham-Dietz, Polly Newcomb, Linda Titus, Kathleen M Egan, David J Hunter, Sara Lindstrom, Rulla M Tamimi, Peter Kraft, Nazneen Rahman, Clare Turnbull, Anthony Renwick, Sheila Seal, Jingmei Li, Jianjun Liu, Keith Humphreys, Javier Benitez, M Pilar Zamora, Jose Ignacio Arias Perez, Primitiva Menéndez, Anna Jakubowska, Jan Lubinski, Katarzyna Jaworska-Bieniek,

Katarzyna Durda, Natalia V Bogdanova, Natalia N Antonenkova, Thilo Dörk, Hoda Anton-Culver, Susan L Neuhausen, Argyrios Ziogas, Leslie Bernstein, Peter Devilee, Robert A E M Tollenaar, Caroline Seynaeve, Christi J van Asperen, Angela Cox, Simon S Cross, Malcolm W R Reed, Elza Khusnutdinova, Marina Bermisheva, Darya Prokofyeva, Zalina Takhirova, Alfons Meindl, Rita K Schmutzler, Christian Sutter, Rongxi Yang, Peter Schürmann, Michael Bremer, Hans Christiansen, Tjoung-Won Park-Simon, Peter Hillemanns, Pascal Guénel, Thérèse Truong, Florence Menegaux, Marie Sanchez, Paolo Radice, Paolo Peterlongo, Siranoush Manoukian, Valeria Pensotti, John L Hopper, Helen Tsimiklis, Carmel Apicella, Melissa C Southey, Hiltrud Brauch, Thomas Brüning, Yon-Dschun Ko, Alice J Sigurdson, Michele M Doody, Ute Hamann, Diana Torres, Hans-Ulrich Ulmer, Asta Försti, Elinor J Sawyer, Ian Tomlinson, Michael J Kerin, Nicola Miller, Irene L Andrulis, Julia A Knight, Gord Glendon, Anna Marie Mulligan, Georgia Chenevix-Trench, Rosemary Balleine, Graham G Giles, Roger L Milne, Catriona McLean, Annika Lindblom, Sara Margolin, Christopher A Haiman, Brian E Henderson, Fredrick Schumacher, Loic Le Marchand, Ursula Eilber, Shan Wang-Gohrke, Maartje J Hooning, Antoinette Hollestelle, Ans M W van den Ouweland, Linetta B Koppert, Jane Carpenter, Christine Clarke, Rodney Scott, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M Hartikainen, Hermann Brenner, Volker Arndt, Christa Stegmaier, Aida Karina Dieffenbach, Robert Winqvist, Katri Pylkäs, Arja Jukkola-Vuorinen, Mervi Grip, Kenneth Offit, Joseph Vijai, Mark Robson, Rohini Rau-Murthy, Miriam Dwek, Ruth Swann, Katherine Annie Perkins, Mark S Goldberg, France Labrèche, Martine Dumont, Diana M Eccles, William J Tapper, Sajjad Rafiq, Esther M John, Alice S Whittemore, Susan Slager, Drakoulis Yannoukakos, Amanda E Toland, Song Yao, Wei Zheng, Sandra L Halverson, Anna González-Neira, Guillermo Pita, M Rosario Alonso, Nuria Álvarez, Daniel Herrero, Daniel C Tessier, Daniel Vincent, Francois Bacot, Craig Luccarini, Caroline Baynes, Shahana Ahmed, Mel Maranian, Catherine S Healey, Jacques Simard, Per Hall, Douglas F Easton, and Montserrat Garcia-Closas. Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.*, 107(5), May 2015.

[349] Simon Leedham and Ian Tomlinson. The continuum model of selection in human tumors: general paradigm or niche product? *Cancer Res.*, 72(13):3131–3134, 1 July 2012.

[350] Daniel G MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, Stephen B Montgomery, Cornelis A Albers, Zhengdong D Zhang, Donald F Conrad, Gerton Lunter, Hancheng Zheng, Qasim Ayub, Mark A DePristo, Eric Banks, Min Hu, Robert E Handsaker, Jeffrey A Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H A Barnes, Clara Amid, Denise R Carvalho-Silva, Alexandra H Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David N Cooper, Yali Xue, Irene Gallego Romero, 1000 Genomes Project Consortium, Jun Wang, Yingrui Li, Richard A Gibbs, Steven A McCarroll, Emmanouil T Dermitzakis, Jonathan K

Pritchard, Jeffrey C Barrett, Jennifer Harrow, Matthew E Hurles, Mark B Gerstein, and Chris Tyler-Smith. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 17 February 2012.

[351] Loukas Moutsianas, Vineeta Agarwala, Christian Fuchsberger, Jason Flannick, Manuel A Rivas, Kyle J Gaulton, Patrick K Albers, GoT2D Consortium, Gil McVean, Michael Boehnke, David Altshuler, and Mark I McCarthy. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet.*, 11(4):e1005165, April 2015.

[352] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, 42(Database issue):D980–5, January 2014.

[353] Jinghui Zhang, Michael F Walsh, Gang Wu, Michael N Edmonson, Tanja A Gruber, John Easton, Dale Hedges, Xiaotu Ma, Xin Zhou, Donald A Yergeau, Mark R Wilkinson, Bhavin Vadodaria, Xiang Chen, Rose B McGee, Stacy Hines-Dowell, Regina Nuccio, Emily Quinn, Sheila A Shurtleff, Michael Rusch, Aman Patel, Jared B Becksfort, Shuoguo Wang, Meaghann S Weaver, Li Ding, Elaine R Mardis, Richard K Wilson, Amar Gajjar, David W Ellison, Alberto S Pappo, Ching-Hon Pui, Kim E Nichols, and James R Downing. Germline mutations in predisposition genes in pediatric cancer. *N. Engl. J. Med.*, 373(24):2336–2346, 10 December 2015.

[354] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 23 January 2014.

[355] Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, Taru Tukiainen, Daniel P Birnbaum, Jack A Kosmicki, Laramie E Duncan, Karol Estrada, Fengmei Zhao, James Zou, Emma Pierce-Hoffman, Joanne Berghout, David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sullivan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J

Daly, Daniel G MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 18 August 2016.

[356] Kasmintan A Schrader, Donavan T Cheng, Vijai Joseph, Meera Prasad, Michael Walsh, Ahmet Zehir, Ai Ni, Tinu Thomas, Ryma Benayed, Asad Ashraf, Annie Lincoln, Maria Arcila, Zsofia Stadler, David Solit, David M Hyman, David Hyman, Liying Zhang, David Klimstra, Marc Ladanyi, Kenneth Offit, Michael Berger, and Mark Robson. Germline variants in targeted tumor sequencing using matched normal DNA. *JAMA Oncol*, 2(1):104–111, January 2016.

[357] Thomas Paul Slavin, Mariana Niell-Swiller, Ilana Solomon, Bita Nehoray, Christina Rybak, Kathleen R Blazer, and Jeffrey N Weitzel. Clinical application of multigene panels: Challenges of Next-Generation counseling and cancer risk management. *Front. Oncol.*, 5:208, 29 September 2015.

[358] Andrew G Renehan, Marcel Zwahlen, and Matthias Egger. Adiposity and cancer risk: new mechanistic insights from epidemiology. *Nat. Rev. Cancer*, 15(8):484–498, August 2015.

[359] Andrew G Renehan, Margaret Tyson, Matthias Egger, Richard F Heller, and Marcel Zwahlen. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet*, 371(9612):569–578, 16 February 2008.

[360] Mandy L Ballinger, David L Goode, Isabelle Ray-Coquard, Paul A James, Gillian Mitchell, Eveline Niedermayr, Ajay Puri, Joshua D Schiffman, Gillian S Dite, Arcadi Cipponi, Robert G Maki, Andrew S Brohl, Ola Myklebost, Eva W Stratford, Susanne Lorenz, Sung-Min Ahn, Jin-Hee Ahn, Jeong Eun Kim, Sue Shanley, Victoria Beshay, Robert Lor Randall, Ian Judson, Beatrice Seddon, Ian G Campbell, Mary-Anne Young, Rajiv Sarin, Jean-Yves Blay, Seán I O'Donoghue, David M Thomas, and International Sarcoma Kindred Study. Monogenic and polygenic determinants of sarcoma risk: an international genetic study. *Lancet Oncol.*, 17(9):1261–1271, September 2016.

[361] Sébastien Jacquemont, Bradley P Coe, Micha Hersch, Michael H Duyzend, Niklas Krumm, Sven Bergmann, Jacques S Beckmann, Jill A Rosenfeld, and Evan E Eichler. A higher mutational burden in females supports a "female protective model" in neurodevelopmental disorders. *Am. J. Hum. Genet.*, 94(3):415–425, 6 March 2014.

[362] Matteo Fumagalli, Ida Moltke, Niels Grarup, Fernando Racimo, Peter Bjerregaard, Marit E Jørgensen, Thorfinn S Korneliussen, Pascale Gerbault, Line Skotte, Allan Linneberg, Cramer Christensen, Ivan Brandslund, Torben Jørgensen, Emilia Huerta-Sánchez, Erik B Schmidt, Oluf Pedersen, Torben Hansen, Anders Albrechtsen, and Rasmus Nielsen. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254):1343–1347, 18 September 2015.

[363] Ryan K C Yuen, Bhooma Thiruvahindrapuram, Daniele Merico, Susan Walker, Kristiina Tammimies, Ny Hoang, Christina Chrysler, Thomas Nalpathamkalam, Giovanna Pellecchia, Yi Liu, Matthew J Gazzellone, Lia D'Abate, Eric Deneault, Jennifer L Howe, Richard S C Liu, Ann Thompson, Mehdi Zarrei, Mohammed Uddin, Christian R Marshall, Robert H Ring, Lonnie Zwaigenbaum, Peter N Ray, Rosanna Weksberg, Melissa T Carter, Bridget A Fernandez, Wendy Roberts, Peter Szatmari, and Stephen W Scherer. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nat. Med.*, 21(2):185–191, February 2015.

[364] Mikko Muona, Samuel F Berkovic, Leanne M Dibbens, Karen L Oliver, Snezana Maljevic, Marta A Bayly, Tarja Joensuu, Laura Canafoglia, Silvana Franceschetti, Roberto Michelucci, Salla Markkinen, Sarah E Heron, Michael S Hildebrand, Eva Andermann, Frederick Andermann, Antonio Gambardella, Paolo Tinuper, Laura Licchetta, Ingrid E Scheffer, Chiara Criscuolo, Alessandro Filla, Edoardo Ferlazzo, Jamil Ahmad, Adeel Ahmad, Betul Baykan, Edith Said, Meral Topcu, Patrizia Riguzzi, Mary D King, Cigdem Ozkara, Danielle M Andrade, Bernt A Engelsen, Arielle Crespel, Matthias Lindenau, Ebba Lohmann, Veronica Saletti, João Massano, Michael Privitera, Alberto J Espay, Birgit Kauffmann, Michael Duchowny, Rikke S Møller, Rachel Straussberg, Zaid Afawi, Bruria Ben-Zeev, Kaitlin E Samocha, Mark J Daly, Steven Petrou, Holger Lerche, Aarno Palotie, and Anna-Elina Lehesjoki. A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nat. Genet.*, 47(1):39–46, January 2015.

[365] Alexander Hoischen, Niklas Krumm, and Evan E Eichler. Prioritization of neurodevelopmental disease genes by discovery of new mutations. *Nat. Neurosci.*, 17(6):764–772, June 2014.

[366] Julien Thevenon, Mathieu Milh, François Feillet, Judith St-Onge, Yannis Duffourd, Clara Jugé, Agathe Roubertie, Delphine Héron, Cyril Mignot, Emmanuel Raffo, Bertrand Isidor, Sandra Wahlen, Damien Sanlaville, Nathalie Villeneuve, Véronique Darmency-Stamboul, Annick Toutain, Mathilde Lefebvre, Mondher Chouchane, Frédéric Huet, Arnaud Lafon, Anne de Saint Martin, Gaetan Lesca, Salima El Chehadeh, Christel Thauvin-Robinet, Alice Masurel-Paulet, Sylvie Odent, Laurent Villard, Christophe Philippe, Laurence Faivre, and Jean-Baptiste Rivière. Mutations in SLC13A5 cause autosomal-recessive epileptic encephalopathy with seizure onset in the first days of life. *Am. J. Hum. Genet.*, 95(1):113–120, 3 July 2014.

[367] Seth A Ament, Szabolcs Szelinger, Gustavo Glusman, Justin Ashworth, Liping Hou, Nirmala Akula, Tatyana Shekhtman, Judith A Badner, Mary E Brunkow, Denise E Mauldin, Anna-Barbara Stittrich, Katherine Rouleau, Sevilla D Detera-Wadleigh, John I Nurnberger, Jr, Howard J Edenberg, Elliot S Gershon, Nicholas Schork, Bipolar Genome Study, Nathan D Price, Richard Gelinas, Leroy Hood, David Craig, Francis J McMahon, John R Kelsoe, and Jared C Roach. Rare variants in neuronal excitability genes influence risk for bipolar disorder. *Proc. Natl. Acad. Sci. U. S. A.*, 112(11):3576–3581, 17 March 2015.

[368] G McMichael, M N Bainbridge, E Haan, M Corbett, A Gardner, S Thompson, B W M van Bon, C L van Eyk, J Broadbent, C Reynolds, M E O'Callaghan, L S Nguyen, D L Adelson, R Russo, S Jhangiani, H Doddapaneni, D M Muzny, R A Gibbs, J Gecz, and A H MacLennan. Whole-exome sequencing points to considerable genetic heterogeneity of cerebral palsy. *Mol. Psychiatry*, 20(2):176–182, February 2015.

[369] K Joeri van der Velde, Joël Kuiper, Bryony A Thompson, John-Paul Plazzer, Gert van Valkenhoef, Mark de Haan, Jan D H Jongbloed, Cisca Wijmenga, Tom J de Koning, Kristin M Abbott, Richard Sinke, Amanda B Spurdle, Finlay Macrae, Maurizio Genuardi, Rolf H Sijmons, Morris A Swertz, and InSiGHT Group. Evaluation of CADD scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization. *Hum. Mutat.*, 36(7):712–719, July 2015.

[370] Asta Försti, Abhishek Kumar, Nagarajan Paramasivam, Matthias Schlesner, Calogerina Catalano, Dagmara Dymerska, Jan Lubinski, Roland Eils, and Kari Hemminki. Pedigree based DNA sequencing pipeline for germline genomes of cancer families. *Hered. Cancer Clin. Pract.*, 14:16, 9 August 2016.

[371] H Hu, S A Haas, J Chelly, H Van Esch, M Raynaud, A P M de Brouwer, S Weinert, G Froyen, S G M Frints, F Laumonnier, T Zemojtel, M I Love, H Richard, A-K Emde, M Bienek, C Jensen, M Hambrock, U Fischer, C Langnick, M Feldkamp, W Wissink-Lindhout, N Lebrun, L Castelnau, J Rucci, R Montjean, O Dorseuil, P Billuart, T Stuhlmann, M Shaw, M A Corbett, A Gardner, S Willis-Owen, C Tan, K L Friend, S Belet, K E P van Roozendaal, M Jimenez-Pocquet, M-P Moizard, N Ronce, R Sun, S O'Keeffe, R Chenna, A van Bömmel, J Göke, A Hackett, M Field, L Christie, J Boyle, E Haan, J Nelson, G Turner, G Baynam, G Gillessen-Kaesbach, U Müller, D Steinberger, B Budny, M Badura-Stronka, A Latos-Bieleńska, L B Ousager, P Wieacker, G Rodríguez Criado, M-L Bondeson, G Annerén, A Dufke, M Cohen, L Van Maldergem, C Vincent-Delorme, B Echenne, B Simon-Bouy, T Kleefstra, M Willemsen, J-P Fryns, K Devriendt, R Ullmann, M Vingron, K Wrogemann, T F Wienker, A Tzschach, H van Bokhoven, J Gecz, T J Jentsch, W Chen, H-H Ropers, and V M Kalscheuer. X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Mol. Psychiatry*, 21(1):133–148, January 2016.

[372] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, Heidi L Rehm, and ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genet. Med.*, 17(5):405–424, May 2015.

[373] 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

Witmer, Hua Ling, Kurt Hetrick, Lee Watkins, Karynne E Patterson, Frederic Reinier, Elizabeth Blue, Donna Muzny, Martin Kircher, Kaya Bilguvar, Francesc López-Giráldez, V Reid Sutton, Holly K Tabor, Suzanne M Leal, Murat Gunel, Shrikant Mane, Richard A Gibbs, Eric Boerwinkle, Ada Hamosh, Jay Shendure, James R Lupski, Richard P Lifton, David Valle, Deborah A Nickerson, Centers for Mendelian Genomics, and Michael J Bamshad. The genetic basis of mendelian phenotypes: Discoveries, challenges, and opportunities. *Am. J. Hum. Genet.*, 97(2):199–215, 6 August 2015.

[374] Tarjinder Singh, Mitja I Kurki, David Curtis, Shaun M Purcell, Lucy Crooks, Jeremy McRae, Jaana Suvisaari, Himanshu Chheda, Douglas Blackwood, Gerome Breen, Olli Pietiläinen, Sebastian S Gerety, Muhammad Ayub, Moira Blyth, Trevor Cole, David Collier, Eve L Coomber, Nick Craddock, Mark J Daly, John Danesh, Marta Di-Forti, Alison Foster, Nelson B Freimer, Daniel Geschwind, Mandy Johnstone, Shelagh Joss, Georg Kirov, Jarmo Körkkö, Outi Kuismin, Peter Holmans, Christina M Hultman, Conrad Iyegbe, Jouko Lönnqvist, Minna Männikkö, Steve A McCarroll, Peter McGuffin, Andrew M McIntosh, Andrew McQuillin, Jukka S Moilanen, Carmel Moore, Robin M Murray, Ruth Newbury-Ecob, Willem Ouwehand, Tiina Paunio, Elena Prigmore, Elliott Rees, David Roberts, Jennifer Sambrook, Pamela Sklar, David St Clair, Juha Veijola, James T R Walters, Hywel Williams, Swedish Schizophrenia Study, INTERVAL Study, DDD Study, UK10 K Consortium, Patrick F Sullivan, Matthew E Hurles, Michael C O'Donovan, Aarno Palotie, Michael J Owen, and Jeffrey C Barrett. Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders. *Nat. Neurosci.*, 19(4):571–577, April 2016.

[375] Matthieu Deschamps, Guillaume Laval, Maud Fagny, Yuval Itan, Laurent Abel, Jean-Laurent Casanova, Etienne Patin, and Lluis Quintana-Murci. Genomic signatures of selective pressures and introgression from archaic hominins at human innate immunity genes. *Am. J. Hum. Genet.*, 98(1):5–21, 7 January 2016.

[376] Marcello Niceta, Emilia Stellacci, Karen W Gripp, Giuseppe Zampino, Maria Kousi, Massimiliano Anselmi, Alice Traversa, Andrea Ciolfi, Deborah Stabley, Alessandro Bruselles, Viviana Caputo, Serena Cecchetti, Sabrina Prudente, Maria T Fiorenza, Carla Boitani, Nicole Philip, Dmitriy Niyazov, Chiara Leoni, Takaya Nakane, Kim Keppler-Noreuil, Stephen R Braddock, Gabriele Gillessen-Kaesbach, Antonio Palleschi, Philippe M Campeau, Brendan H L Lee, Celio Pouponnot, Lorenzo Stella, Gianfranco Bocchinfuso, Nicholas Katsanis, Katia Sol-Church, and Marco Tartaglia. Mutations impairing GSK3-Mediated MAF phosphorylation cause cataract, deafness, intellectual disability, seizures, and a down syndrome-like facies. *Am. J. Hum. Genet.*, 96(5):816–825, 7 May 2015.

[377] Valeria D'Argenio, Maria Valeria Esposito, Jean Ann Gilder, Giulia Frisso, and Francesco Salvatore. Should a BRCA2 stop codon human variant, usually considered a polymorphism, be classified as a predisposing mutation? *Cancer*, 120(10):1594–1595, 15 May 2014.

[378] Sean T Martin, Hiroyuki Matsubayashi, Carmelle D Rogers, Juliet Philips, Fergus J Couch, Kieran Brune, Charles J Yeo, Scott E Kern, Ralph H Hruban, and Michael Goggins. Increased prevalence of the BRCA2 polymorphic stop codon K3326X among individuals with familial pancreatic cancer. *Oncogene*, 24(22):3652–3656, 19 May 2005.

[379] M R Akbari, R Malekzadeh, D Nasrollahzadeh, D Amanian, F Islami, S Li, I Zandvakili, R Shakeri, M Sotoudeh, K Aghcheli, R Salahi, A Pourshams, S Semnani, P Boffetta, S M Dawsey, P Ghadirian, and S A Narod. Germline BRCA2 mutations and the risk of esophageal squamous cell carcinoma. *Oncogene*, 27(9):1290–1296, 21 February 2008.

[380] Matthew F Rudd, Emily L Webb, Athena Matakidou, Gabrielle S Sellick, Richard D Williams, Helen Bridle, Tim Eisen, Richard S Houlston, and GELCAPS Consortium. Variants in the GH-IGF axis confer susceptibility to lung cancer. *Genome Res.*, 16(6):693–701, June 2006.

[381] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 23 January 2014.

[382] A G Knudson, Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.*, 68(4):820–823, April 1971.

[383] Amy M Dworkin, Andrew D Spearman, Stephanie Y Tseng, Kevin Sweet, and Amanda Ewart Toland. Methylation not a frequent "second hit" in tumors with germline BRCA mutations. *Fam. Cancer*, 8(4):339–346, 2 April 2009.

[384] Da Pang, Yashuang Zhao, Weinan Xue, Ming Shan, Yanbo Chen, Youxue Zhang, Guoqiang Zhang, Feng Liu, Dalin Li, and Yanmei Yang. Methylation profiles of the BRCA1 promoter in hereditary and sporadic breast cancer among han chinese. *Med. Oncol.*, 29(3):1561–1568, September 2012.

[385] Paula Silva Felicio, Matias Eliseo Melendez, Lidia Maria Rebolho Batista Arantes, Ligia Maria Kerr, Dirce Maria Carraro, Rebeca Silveira Grasel, Natalia Campacci, Cristovam Scapulatempo-Neto, Gabriela Carvalho Fernandes, Ana Carolina de Carvalho, and Edenir Inêz Palmero. Genetic and epigenetic characterization of the BRCA1 gene in brazilian women at-risk for hereditary breast cancer. *Oncotarget*, 8(2):2850–2862, 10 January 2017.

[386] David Mossman and Rodney J Scott. Epimutations, inheritance and causes of aberrant DNA methylation in cancer. *Hered. Cancer Clin. Pract.*, 4(2):75–80, 15 May 2006.

[387] M Ollikainen, U Hannelius, C M Lindgren, W M Abdel-Rahman, J Kere, and P Peltomäki. Mechanisms of inactivation of MLH1 in hereditary nonpolyposis colorectal carcinoma: a novel approach. *Oncogene*, 26(31):4541–4549, 5 July 2007.

[388] I P Tomlinson, R Roylance, and R S Houlston. Two hits revisited again. *J. Med. Genet.*, 38(2):81–85, February 2001.

[389] A J W Paige. Redefining tumour suppressor genes: exceptions to the two-hit hypothesis. *Cell. Mol. Life Sci.*, 60(10):2147–2163, October 2003.

[390] A G Knudson. Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer*, 1(2):157–162, November 2001.

[391] Kate A McBride, Mandy L Ballinger, Emma Killick, Judy Kirk, Martin H N Tattersall, Rosalind A Eeles, David M Thomas, and Gillian Mitchell. Li-Fraumeni syndrome: cancer risk assessment and clinical management. *Nat. Rev. Clin. Oncol.*, 11(5):260–271, May 2014.

[392] J Hall, M Lee, B Newman, J Morrow, L Anderson, B Huey, and M King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684–1689, 1990.

[393] Y Miki, J Swensen, D Shattuck-Eidens, P A Futreal, K Harshman, S Tavtigian, Q Liu, C Cochran, L M Bennett, and W Ding. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182):66–71, 7 October 1994.

[394] L S Friedman, E A Ostermeyer, C I Szabo, P Dowd, E D Lynch, S E Rowell, and M C King. Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nat. Genet.*, 8(4):399–404, December 1994.

[395] R Wooster, S L Neuhausen, J Mangion, Y Quirk, D Ford, N Collins, K Nguyen, S Seal, T Tran, and D Averill. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*, 265(5181):2088–2090, 30 September 1994.

[396] Nazneen Rahman. Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–308, 16 January 2014.

[397] Mary-Claire King, Joan H Marks, Jessica B Mandell, and New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science*, 302(5645):643–646, 24 October 2003.

[398] J P Struewing, P Hartge, S Wacholder, S M Baker, M Berlin, M McAdams, M M Timmerman, L C Brody, and M A Tucker. The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among ashkenazi jews. *N. Engl. J. Med.*, 336(20):1401–1408, 15 May 1997.

[399] Nicholas J Roberts, Yuchen Jiao, Jun Yu, Levy Kopelovich, Gloria M Petersen, Melissa L Bondy, Steven Gallinger, Ann G Schwartz, Sapna Syngal, Michele L Cote, Jennifer Axilbund, Richard Schulick, Syed Z Ali, James R Eshleman, Victor E Velculescu, Michael Goggins, Bert Vogelstein, Nickolas Papadopoulos, Ralph H Hruban, Kenneth W Kinzler, and Alison P Klein. ATM mutations in patients with hereditary pancreatic cancer. *Cancer Discov.*, 2(1):41–46, January 2012.

[400] Michael Choi, Thomas Kipps, and Razelle Kurzrock. ATM mutations in cancer: Therapeutic implications. *Mol. Cancer Ther.*, 15(8):1781–1791, August 2016.

[401] Nazneen Rahman. Mainstreaming genetic testing of cancer predisposition genes. *Clin. Med.*, 14(4):436–439, August 2014.

[402] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, Judy H Cho, Alan E Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N Rotimi, Montgomery Slatkin, David Valle, Alice S Whittemore, Michael Boehnke, Andrew G Clark, Evan E Eichler, Greg Gibson, Jonathan L Haines, Trudy F C Mackay, Steven A McCarroll, and Peter M Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 8 October 2009.

[403] Evan E Eichler, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, 11(6):446–450, June 2010.

[404] Lucia A Hindorff, Elizabeth M Gillanders, and Teri A Manolio. Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis*, 32(7):945–954, July 2011.

[405] Greg Gibson. Rare and common variants: twenty arguments. *Nat. Rev. Genet.*, 13(2):135–145, 18 January 2012.

[406] Lluis Quintana-Murci. Understanding rare and common diseases in the context of human evolution. *Genome Biol.*, 17(1):225, 7 November 2016.

[407] Daniel G MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, Stephen B Montgomery, Cornelis A Albers, Zhengdong D Zhang, Donald F Conrad, Gerton Lunter, Hancheng Zheng, Qasim Ayub, Mark A DePristo, Eric Banks, Min Hu, Robert E Handsaker, Jeffrey A Rosenfeld, Menachem Fromer, Mike Jin, Xinmeng Jasmine Mu, Ekta Khurana, Kai Ye, Mike Kay, Gary Ian Saunders, Marie-Marthe Suner, Toby Hunt, If H A Barnes, Clara Amid, Denise R Carvalho-Silva, Alexandra H Bignell, Catherine Snow, Bryndis Yngvadottir, Suzannah Bumpstead, David N Cooper, Yali Xue, Irene Gallego Romero, 1000 Genomes Project Consortium, Jun Wang, Yingrui Li, Richard A Gibbs, Steven A McCarroll, Emmanouil T Dermitzakis, Jonathan K Pritchard, Jeffrey C Barrett, Jennifer Harrow, Matthew E Hurles, Mark B Gerstein, and Chris Tyler-Smith. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 17 February 2012.

[408] Paul L Auer and Guillaume Lettre. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.*, 7(1):16, 23 February 2015.

[409] Seunggeung Lee, Gonçalo R Abecasis, Michael Boehnke, and Xihong Lin. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, 95(1):5–23, 3 July 2014.

[410] Catherine A Shu, Malcolm C Pike, Anjali R Jotwani, Tara M Friebel, Robert A Soslow, Douglas A Levine, Katherine L Nathanson, Jason A Konner, Angela G Arnold, Faina Bogomolniy, Fanny Dao, Narciso Olvera, Elizabeth K Bancroft, Deborah J Goldfrank, Zsofia K Stadler, Mark E Robson, Carol L Brown, Mario M Leitao, Jr, Nadeem R Abu-Rustum, Carol A Aghajanian, Joanne L Blum, Susan L Neuhausen, Judy E Garber, Mary B Daly, Claudine Isaacs, Rosalind A Eeles, Patricia A Ganz, Richard R Barakat, Kenneth Offit, Susan M Domchek, Timothy R Rebbeck, and Noah D Kauff. Uterine cancer after Risk-Reducing salpingo-oophorectomy without hysterectomy in women with BRCA mutations. *JAMA Oncol*, 2(11):1434–1440, 1 November 2016.

[411] Tim K Shen, Theodoros N Teknos, Amanda E Toland, Leigha Senter, and Rebecca Nagy. Salivary gland cancer in BRCA-positive families: a retrospective review. *JAMA Otolaryngol. Head Neck Surg.*, 140(12):1213–1217, December 2014.

[412] Jin Hee Kim, Heon Kim, Kye Young Lee, Kang-Hyeon Choe, Jeong-Seon Ryu, Ho Il Yoon, Sook Whan Sung, Keun-Young Yoo, and Yun-Chul Hong. Genetic polymorphisms of ataxia telangiectasia mutated affect lung cancer risk. *Hum. Mol. Genet.*, 15(7):1181–1186, 1 April 2006.

[413] M Ahmed and N Rahman. ATM and breast cancer susceptibility. *Oncogene*, 25(43):5906–5911, 25 September 2006.

[414] Hannes Helgason, Thorunn Rafnar, Halla S Olafsdottir, Jon G Jonasson, Asgeir Sigurdsson, Simon N Stacey, Adalbjorg Jonasdottir, Laufey Tryggvadottir, Kristin Alexiusdottir, Asgeir Haraldsson, Louise le Roux, Julius Gudmundsson, Hrefna Johannsdottir, Asmundur Oddsson, Arnaldur Gylfason, Olafur T Magnusson, Gisli Masson, Thorvaldur Jonsson, Halla Skuladottir, Daniel F Gudbjartsson, Unnur Thorsteinsdottir, Patrick Sulem, and Kari Stefansson. Loss-of-function variants in ATM confer risk of gastric cancer. *Nat. Genet.*, 47(8):906–910, August 2015.

[415] Joseph S Pitula, Kathryn M Deck, Stephen L Clarke, Sheila A Anderson, Aparna Vasanthakumar, and Richard S Eisenstein. Selective inhibition of the citrate-to-isocitrate reaction of cytosolic aconitase by phosphomimetic mutation of serine-711. *Proc. Natl. Acad. Sci. U. S. A.*, 101(30):10907–10912, 27 July 2004.

[416] Isao Hirano, Satoki Nakamura, Daisuke Yokota, Takaaki Ono, Kazuyuki Shigeno, Shinya Fujisawa, Kaori Shinjo, and Kazunori Ohnishi. Depletion of pleckstrin homology domain leucine-rich repeat protein phosphatases 1 and 2 by Bcr-Abl promotes chronic myelogenous leukemia cell proliferation through continuous phosphorylation of akt isoforms. *J. Biol. Chem.*, 284(33):22155–22165, 14 August 2009.

[417] J Liu, H L Weiss, P Rychahou, L N Jackson, B M Evers, and T Gao. Loss of PHLPP expression in colon cancer: role in proliferation and tumorigenesis. *Oncogene*, 28(7):994–1004, 19 February 2009.

[418] Haishan Huang, Xiaofu Pan, Honglei Jin, Yang Li, Lin Zhang, Caili Yang, Pei Liu, Ya Liu, Lili Chen, Jingxia Li, Junlan Zhu, Xingruo Zeng, Kai Fu, Guorong Chen, Jimin Gao, and Chuanshu Huang. PHLPP2 downregulation contributes to lung carcinogenesis following B[a]P/B[a]PDE exposure. *Clin. Cancer Res.*, 21(16):3783–3793, 15 August 2015.

[419] Dawid G Nowak, Hyejin Cho, Tali Herzka, Kaitlin Watrud, Daniel V DeMarco, Victoria M Y Wang, Serif Senturk, Christof Fellmann, David Ding, Tumas Beinortas, David Kleinman, Muhan Chen, Raffaella Sordella, John E Wilkinson, Mireia Castillo-Martin, Carlos Cordon-Cardo, Brian D Robinson, and Lloyd C Trotman. MYC drives Pten/Trp53-Deficient proliferation and metastasis due to IL6 secretion and AKT suppression via PHLPP2. *Cancer Discov.*, 5(6):636–651, June 2015.

[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.

[421] M Tevfik Dorak and Ebru Karpuzoglu. Gender differences in cancer susceptibility: an inadequately addressed issue. *Front. Genet.*, 3:268, 28 November 2012.

[422] Carole Ober, Dagan A Loisel, and Yoav Gilad. Sex-specific genetic architecture of human disease. *Nat. Rev. Genet.*, 9(12):911–922, December 2008.

[423] Andrea Clocchiatti, Elisa Cora, Yosra Zhang, and G Paolo Dotto. Sexual dimorphism in cancer. *Nat. Rev. Cancer*, 16(5):330–339, May 2016.

[424] Yuan Yuan, Lingxiang Liu, Hu Chen, Yumeng Wang, Yanxun Xu, Huzhang Mao, Jun Li, Gordon B Mills, Yongqian Shu, Liang Li, and Han Liang. Comprehensive characterization of molecular differences in cancer between male and female patients. *Cancer Cell*, 29(5):711–722, 9 May 2016.

[425] Hae-Mi Woo, Hong-Joon Park, Jeong-In Baek, Mi-Hyun Park, Un-Kyung Kim, Borum Sagong, and Soo Kyung Koo. Whole-exome sequencing identifies MYO15A mutations as a cause of autosomal recessive nonsyndromic hearing loss in korean families. *BMC Med. Genet.*, 14:72, 17 July 2013.

[426] N Liburd, M Ghosh, S Riazuddin, S Naz, S Khan, Z Ahmed, S Riazuddin, Y Liang, P S Menon, T Smith, A C Smith, K S Chen, J R Lupski, E R Wilcox, L Potocki, and T B Friedman. Novel mutations of MYO15A associated with profound deafness in consanguineous families and moderately severe hearing loss in a patient with Smith-Magenis syndrome. *Hum. Genet.*, 109(5):535–541, November 2001.

[427] Flavia Palombo, Nadia Al-Wardy, Guido Alberto Gnecchi Ruscone, Manuela Oppo, Mohammed Nasser Al Kindi, Andrea Angius, Khalsa Al Lamki, Giorgia Girotto, Tania Giangregorio, Matteo Benelli, Alberto Magi, Marco Seri, Paolo Gasparini, Francesco Cucca, Marco Sazzini, Mazin Al Khabori, Tommaso Pippucci, and Giovanni Romeo. A novel founder MYO15A frameshift duplication is the major cause of genetic hearing loss in oman. *J. Hum. Genet.*, 62(2):259–264, February 2017.

[428] Sara Lindström, Deborah J Thompson, Andrew D Paterson, Jingmei Li, Gretchen L Gierach, Christopher Scott, Jennifer Stone, Julie A Douglas, Isabel dos Santos-Silva, Pablo Fernandez-Navarro, Jajini Verghase, Paula Smith, Judith Brown, Robert Luben, Nicholas J Wareham, Ruth J F Loos, John A Heit, V Shane Pankratz, Aaron Norman, Ellen L Goode, Julie M Cunningham, Mariza deAndrade, Robert A Vierkant, Kamila Czene, Peter A Fasching, Laura Baglietto, Melissa C Southey, Graham G Giles, Kaanan P Shah, Heang-Ping Chan, Mark A Helvie, Andrew H Beck, Nicholas W Knoblauch, Aditi Hazra, David J Hunter, Peter Kraft, Marina Pollan, Jonine D Figueroa, Fergus J Couch, John L Hopper, Per Hall, Douglas F Easton, Norman F Boyd, Celine M Vachon, and Rulla M Tamimi. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. *Nat. Commun.*, 5:5303, 24 October 2014.

[429] Tan Tan, Kai Zhang, and Wenjun Chen Sun. Genetic variants of ESR1 and SGSM3 are associated with the susceptibility of breast cancer in the chinese population. *Breast Cancer*, 24(3):369–374, May 2017.

[430] Chaoqun Wang, Hua Zhao, Xiankun Zhao, Jiao Wan, Dayong Wang, Wanli Bi, Xinghong Jiang, and Yuzhen Gao. Association between an insertion/deletion polymorphism within 3'UTR of SGSM3 and risk of hepatocellular carcinoma. *Tumour Biol.*, 35(1):295–301, January 2014.

[431] Anja Rudolph, Peter A Fasching, Sabine Behrens, Ursula Eilber, Manjeet K Bolla, Qin Wang, Deborah Thompson, Kamila Czene, Judith S Brand, Jingmei Li, Christopher Scott, V Shane Pankratz, Kathleen Brandt, Emily Hallberg, Janet E Olson, Adam Lee, Matthias W Beckmann, Arif B Ekici, Lothar Haeberle, Gertraud Maskarinec, Loic Le Marchand, Fredrick Schumacher, Roger L Milne, Julia A Knight, Carmel Apicella, Melissa C Southey, Miroslav K Kapuscinski, John L Hopper, Irene L Andrulis, Graham G Giles, Christopher A Haiman, Kay-Tee Khaw, Robert Luben, Per Hall, Paul D P Pharoah, Fergus J Couch, Douglas F Easton, Isabel Dos-Santos-Silva, Celine Vachon, and Jenny Chang-Claude. A comprehensive evaluation of interaction between genetic variants and use of menopausal hormone therapy on mammographic density. *Breast Cancer Res.*, 17:110, 16 August 2015.

[432] Cristina R Ferrone, Douglas A Levine, Laura H Tang, Peter J Allen, William Jarnagin, Murray F Brennan, Kenneth Offit, and Mark E Robson. BRCA germline mutations in jewish patients with pancreatic adenocarcinoma. *J. Clin. Oncol.*, 27(3):433–438, 20 January 2009.

[433] Niall G Howlett, Toshiyasu Taniguchi, Susan Olson, Barbara Cox, Quinten Waisfisz, Christine De Die-Smulders, Nicole Persky, Markus Grompe, Hans Joenje, Gerard Pals, Hideyuki Ikeda, Edward A Fox, and Alan D D'Andrea. Biallelic inactivation of BRCA2 in fanconi anemia. *Science*, 297(5581):606–609, 26 July 2002.

[434] Haoming Xia, Jianting Long, Ruifen Zhang, Xiaosong Yang, and Zhefu Ma. MiR-32 contributed to cell proliferation of human breast cancer cells by suppressing of PHLPP2 expression. *Biomed. Pharmacother.*, 75:105–110, October 2015.

[435] John Brognard, Emma Sierecki, Tianyan Gao, and Alexandra C Newton. PHLPP and a second isoform, PHLPP2, differentially attenuate the amplitude of akt signaling by regulating distinct akt isoforms. *Mol. Cell*, 25(6):917–931, 23 March 2007.

[436] K B Chapman and J D Boeke. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell*, 65(3):483–492, 3 May 1991.

[437] M Fahad Khalid, Masad J Damha, Stewart Shuman, and Beate Schwer. Structure-function analysis of yeast RNA debranching enzyme (dbr1), a manganese-dependent phosphodiesterase. *Nucleic Acids Res.*, 33(19):6349–6360, 7 November 2005.

[438] B Han, H K Park, T Ching, J Panneerselvam, H Wang, Y Shen, J Zhang, L Li, R Che, L Garmire, and P Fei. Human DBR1 modulates the recycling of snRNPs to affect alternative RNA splicing and contributes to the suppression of cancer development. *Oncogene*, 15 May 2017.

[439] Jeffry D Sander and J Keith Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, 32(4):347–355, April 2014.

[440] Katja Brose, Kimberly S Bland, Kuan Hong Wang, David Arnott, William Henzel, Corey S Goodman, Marc Tessier-Lavigne, and Thomas Kidd. Slit proteins bind robo receptors and have an evolutionarily conserved role in repulsive axon guidance. *Cell*, 96(6):795–806, March 1999.

[441] W Yuan, L Zhou, J H Chen, J Y Wu, Y Rao, and D M Ornitz. The mouse SLIT family: secreted ligands for ROBO expressed in patterns that suggest a role in morphogenesis and axon guidance. *Dev. Biol.*, 212(2):290–306, 15 August 1999.

[442] Rishi K Gara, Sonam Kumari, Aditya Ganju, Murali M Yallapu, Meena Jaggi, and Subhash C Chauhan. Slit/Robo pathway: a promising therapeutic target for cancer. *Drug Discov. Today*, 20(1):156–164, January 2015.

[443] Po-Hao Chang, Wendy W Hwang-Verslues, Yi-Cheng Chang, Chun-Chin Chen, Michael Hsiao, Yung-Ming Jeng, King-Jen Chang, Eva Y-H P Lee, Jin-Yuh Shew, and Wen-Hwa Lee. Activation of robo1 signaling of breast cancer cells by slit2 from stromal fibroblast restrains tumorigenesis via blocking PI3K/Akt/β-catenin pathway. *Cancer Res.*, 72(18):4652–4661, 15 September 2012.

[444] Fengxia Qin, Huikun Zhang, Li Ma, Xiaoli Liu, Kun Dai, Wenliang Li, Feng Gu, Li Fu, and Yongjie Ma. Low expression of slit2 and robo1 is associated with poor prognosis and brain-specific metastasis of breast cancer patients. *Sci. Rep.*, 5:14430, 24 September 2015.

[445] Ashraf Dallol, Eva Forgacs, Alonso Martinez, Yoshitaka Sekido, Rosemary Walker, Takeshi Kishida, Pamela Rabbitts, Eamonn R Maher, John D Minna, and Farida Latif. Tumour specific promoter region methylation of the human homologue of the drosophila roundabout gene DUTT1 (ROBO1) in human cancers. *Oncogene*, 21(19):3020–3028, 2 May 2002.

[446] Rolando A R Villacis, Francine B Abreu, Priscila M Miranda, Maria A C Domingues, Dirce M Carraro, Erika M M Santos, Victor P Andrade, Benedito M Rossi, Maria I Achatz, and Silvia R Rogatto. ROBO1 deletion as a novel germline alteration in breast and colorectal cancer patients. *Tumour Biol.*, 37(3):3145–3153, March 2016.

[447] Wei Jin, Ying Chen, Gen-Hong Di, Penelope Miron, Yi-Feng Hou, Hui Gao, and Zhi-Ming Shao. Estrogen receptor (ER) beta or p53 attenuates ERalpha-mediated transcriptional activation on the BRCA2 promoter. *J. Biol. Chem.*, 283(44):29671–29680, 31 October 2008.

[448] William D Foulkes, Kelly Metcalfe, Ping Sun, Wedad M Hanna, Henry T Lynch, Parviz Ghadirian, Nadine Tung, Olufunmilayo I Olopade, Barbara L Weber, Jane McLennan, Ivo A Olivotto, Louis R Bégin, and Steven A Narod. Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: the influence of age, grade, and histological type. *Clin. Cancer Res.*, 10(6):2029–2034, 15 March 2004.

[449] Mehadad Noruzinia, Isabelle Coupier, and Pascal Pujol. Is BRCA1/BRCA2-related breast carcinogenesis estrogen dependent? *Cancer*, 104(8):1567–1574, 15 October 2005.

[450] David C Johnson, Niels Weinhold, Jonathan S Mitchell, Bowang Chen, Martin Kaiser, Dil B Begum, Jens Hillengass, Uta Bertsch, Walter A Gregory, David Cairns, Graham H Jackson, Asta Försti, Jolanta Nickel, Per Hoffmann, Markus M Nöethen, Owen W Stephens, Bart Barlogie, Faith E Davis, Kari Hemminki, Hartmut Goldschmidt, Richard S Houlston, and Gareth J Morgan. Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma. *Nat. Commun.*, 7:10290, 8 January 2016.

[451] Chen Wu, Dong Li, Weihua Jia, Zhibin Hu, Yifeng Zhou, Dianke Yu, Tong Tong, Mingrong Wang, Dongmei Lin, Yan Qiao, Yuling Zhou, Jiang Chang, Kan Zhai, Menghan Wang, Lixuan Wei, Wen Tan, Hongbing Shen, Yixin Zeng, and Dongxin Lin. Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. *Nat. Genet.*, 45(6):632–638, June 2013.

[452] Capucine Van Rechem, Joshua C Black, Patricia Greninger, Yang Zhao, Carlos Donado, Paul D Burrowes, Brendon Ladd, David C Christiani, Cyril H Benes, and Johnathan R Whetstine. A coding single-nucleotide polymorphism in lysine demethylase KDM4A associates with increased sensitivity to mTOR inhibitors. *Cancer Discov.*, 5(3):245–254, March 2015.

[453] Peter A Fasching, Paul D P Pharoah, Angela Cox, Heli Nevanlinna, Stig E Bojesen, Thomas Karn, Annegien Broeks, Flora E van Leeuwen, Laura J van't Veer, Renate Udo, Alison M Dunning, Dario Greco, Kristiina Aittomäki, Carl Blomqvist, Mitul Shah, Børge G Nordestgaard, Henrik Flyger, John L Hopper, Melissa C Southey, Carmel Apicella, Montserrat Garcia-Closas, Mark Sherman, Jolanta Lissowska, Caroline Seynaeve, Petra E A Huijts, Rob A E M Tollenaar, Argyrios Ziogas, Arif B Ekici, Claudia Rauh, Arto Mannermaa, Vesa Kataja, Veli-Matti Kosma, Jaana M Hartikainen, Irene L Andrulis, Hilmi Ozcelik, Anna-Marie Mulligan, Gord Glendon, Per Hall, Kamila Czene, Jianjun Liu, Jenny Chang-Claude, Shan Wang-Gohrke, Ursula Eilber, Stefan Nickels, Thilo Dörk, Maria Schiekel, Michael Bremer, Tjoung-Won Park-Simon, Graham G Giles, Gianluca Severi, Laura Baglietto, Maartje J Hooning, John W M Martens, Agnes Jager, Mieke Kriege, Annika Lindblom, Sara Margolin, Fergus J Couch, Kristen N Stevens, Janet E Olson, Matthew Kosel, Simon S Cross, Sabapathy P Balasubramanian, Malcolm W R Reed, Alexander Miron, Esther M John, Robert Winqvist, Katri Pylkäs, Arja Jukkola-Vuorinen, Saila Kauppila, Barbara Burwinkel, Frederik Marme, Andreas Schneeweiss, Christof Sohn, Georgia Chenevix-Trench, kConFab Investigators, Diether Lambrechts, Anne-Sophie Dieudonne, Sigrid Hatse, Erik van Limbergen, Javier Benitez, Roger L Milne, M Pilar Zamora, José Ignacio Arias Pérez, Bernardo Bonanni, Bernard Peissel, Bernard Loris, Paolo Peterlongo, Preetha Rajaraman, Sara J Schonfeld, Hoda Anton-Culver, Peter Devilee, Matthias W Beckmann, Dennis J Slamon, Kelly-Anne Phillips, Jonine D Figueroa, Manjeet K Humphreys, Douglas F Easton, and Marjanka K Schmidt. The role of genetic breast cancer susceptibility variants as prognostic factors. *Hum. Mol. Genet.*, 21(17):3926–3939, 1 September 2012.

[454] W D Cook and B J McCaw. Accommodating haploinsufficient tumor suppressor genes in knudson's model. *Oncogene*, 19(30):3434–3438, 13 July 2000.

[455] Sudheer Kumar Gara, Li Jia, Maria J Merino, Sunita K Agarwal, Lisa Zhang, Maggie Cam, Dhaval Patel, and Electron Kebebew. Germline HABP2 mutation causing familial nonmedullary thyroid cancer. *N. Engl. J. Med.*, 373(5):448–455, 30 July 2015.

[456] F P Li and J F Fraumeni, Jr. Soft-tissue sarcomas, breast cancer, and other neoplasms. a familial syndrome? *Ann. Intern. Med.*, 71(4):747–752, October 1969.

[457] F P Li and J F Fraumeni, Jr. Rhabdomyosarcoma in children: epidemiologic study and identification of a familial cancer syndrome. *J. Natl. Cancer Inst.*, 43(6):1365–1373, December 1969.

[458] D Malkin, F P Li, L C Strong, J F Fraumeni, Jr, C E Nelson, D H Kim, J Kassel, M A Gryka, F Z Bischoff, and M A Tainsky. Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, 250(4985):1233–1238, 30 November 1990.

[459] D Liaw, D J Marsh, J Li, P L Dahia, S I Wang, Z Zheng, S Bose, K M Call, H C Tsou, M Peacocke, C Eng, and R Parsons. Germline mutations of the PTEN gene in cowden disease, an inherited breast and thyroid cancer syndrome. *Nat. Genet.*, 16(1):64–67, May 1997.

[460] Robert Pilarski, Randall Burt, Wendy Kohlman, Lana Pho, Kristen M Shannon, and Elizabeth Swisher. Cowden syndrome and the PTEN hamartoma tumor syndrome: systematic review and revised diagnostic criteria. *J. Natl. Cancer Inst.*, 105(21):1607–1616, 6 November 2013.

[461] K E Nichols, D Malkin, J E Garber, J F Fraumeni, Jr, and F P Li. Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers. *Cancer Epidemiol. Biomarkers Prev.*, 10(2):83–87, February 2001.

[462] Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. Global cancer statistics, 2012. *CA Cancer J. Clin.*, 65(2):87–108, March 2015.

[463] Dezheng Huo, Francis Ikpatt, Andrey Khramtsov, Jean-Marie Dangou, Rita Nanda, James Dignam, Bifeng Zhang, Tatyana Grushko, Chunling Zhang, Olayiwola Oluwasola, David Malaka, Sani Malami, Abayomi Odetunde, Adewumi O Adeoye, Festus Iyare, Adeyinka Falusi, Charles M Perou, and Olufunmilayo I Olopade. Population differences in breast cancer: survey in indigenous african women reveals over-representation of triple-negative breast cancer. *J. Clin. Oncol.*, 27(27):4515–4521, 20 September 2009.

[464] Kelly Servick. Breast cancer. breast cancer: a world of differences. *Science*, 343(6178):1452–1453, 28 March 2014.

[465] Lisa A Newman, James Mason, David Cote, Yael Vin, Kathryn Carolin, David Bouwman, and Graham A Colditz. African-American ethnicity, socioeconomic status, and breast cancer survival: a meta-analysis of 14 studies involving over 10,000 African-American and 40,000 white american patients with carcinoma of the breast. *Cancer*, 94(11):2844–2854, 1 June 2002.

[466] Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, 121(7):2750–2767, July 2011.

[467] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 4 October 2012.

[468] Philip J Stephens, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R Bignell, Lucy R Yates, Elli Papaemmanuil, David Beare, Adam Butler, Angela Cheverton, John Gamble, Jonathan Hinton, Mingming Jia, Alagu Jayakumar, David Jones, Calli Latimer, King Wai Lau, Stuart McLaren, David J McBride, Andrew Menzies, Laura Mudie, Keiran Raine, Roland Rad, Michael Spencer Chapman, Jon Teague, Douglas Easton, Anita Langerød, Oslo Breast Cancer Consortium (OSBREAC), Ming Ta Michael Lee, Chen-Yang Shen, Benita Tan Kiat Tee, Bernice Wong Huimin, Annegien Broeks, Ana Cristina Vargas, Gulisa Turashvili, John Martens, Aquila Fatima, Penelope Miron, Suet-Feung Chin, Gilles Thomas, Sandrine Boyault, Odette Mariani, Sunil R Lakhani, Marc van de Vijver, Laura van 't Veer, John Foekens, Christine Desmedt, Christos Sotiriou, Andrew Tutt, Carlos Caldas, Jorge S Reis-Filho, Samuel A J R Aparicio, Anne Vincent Salomon, Anne-Lise Børresen-Dale, Andrea L Richardson, Peter J Campbell, P Andrew Futreal, and Michael R Stratton. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–404, 16 May 2012.

[469] Shantanu Banerji, Kristian Cibulskis, Claudia Rangel-Escareno, Kristin K Brown, Scott L Carter, Abbie M Frederick, Michael S Lawrence, Andrey Y Sivachenko, Carrie Sougnez, Lihua Zou, Maria L Cortes, Juan C Fernandez-Lopez, Shouyong Peng, Kristin G Ardlie, Daniel Auclair, Veronica Bautista-Piña, Fujiko Duke, Joshua Francis, Joonil Jung, Antonio Maffuz-Aziz, Robert C Onofrio, Melissa Parkin, Nam H Pho, Valeria Quintanar-Jurado, Alex H Ramos, Rosa Rebollar-Vega, Sergio Rodriguez-Cuevas, Sandra L Romero-Cordoba, Steven E Schumacher, Nicolas Stransky, Kristin M Thompson, Laura Uribe-Figueroa, Jose Baselga, Rameen Beroukhim, Kornelia Polyak, Dennis C Sgroi, Andrea L Richardson, Gerardo Jimenez-Sanchez, Eric S Lander, Stacey B Gabriel, Levi A Garraway, Todd R Golub, Jorge Melendez-Zajgla, Alex Toker, Gad Getz, Alfredo Hidalgo-Miranda, and Matthew Meyerson. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409, 20 June 2012.

[470] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 18 April 2012.

[471] Sohrab P Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui 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, Alireza

Heravi-Moussavi, Jamie Rosner, Daniel Lai, Inanc Birol, Richard Varhol, Angela Tam, Noreen Dhalla, Thomas Zeng, Kevin Ma, Simon K Chan, Malachi Griffith, Annie Moradian, S-W Grace Cheng, Gregg B Morin, Peter Watson, Karen Gelmon, Stephen Chia, Suet-Feung Chin, Christina Curtis, Oscar M Rueda, Paul D Pharoah, Sambasivarao Damaraju, John Mackey, Kelly Hoon, Timothy Harkins, Vasisht Tadigotla, Mahvash Sigaroudinia, Philippe Gascard, Thea Tlsty, Joseph F Costello, Irmtraud M Meyer, Connie J Eaves, Wyeth W Wasserman, Steven Jones, David Huntsman, Martin Hirst, Carlos Caldas, Marco A Marra, and Samuel Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–399, 4 April 2012.

[472] Yong Wang, Jill Waters, Marco L Leung, Anna Unruh, Whijae Roh, Xiuqing Shi, Ken Chen, Paul Scheet, Selina Vattathil, Han Liang, Asha Multani, Hong Zhang, Rui Zhao, Franziska Michor, Funda Meric-Bernstam, and Nicholas E Navin. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, 512(7513):155–160, 14 August 2014.

[473] Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, Dana W Y Tsui, Bin Liu, Sarah-Jane Dawson, Jean Abraham, Helen Northen, John F Peden, Abhik Mukherjee, Gulisa Turashvili, Andrew R Green, Steve McKinney, Arusha Oloumi, Sohrab Shah, Nitzan Rosenfeld, Leigh Murphy, David R Bentley, Ian O Ellis, Arnie Purushotham, Sarah E Pinder, Anne-Lise Børresen-Dale, Helena M Earl, Paul D Pharoah, Mark T Ross, Samuel Aparicio, and Carlos Caldas. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.*, 7:11479, 10 May 2016.

[474] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B Brinkman, Sandro Morganella, Miriam R Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A Foekens, Moritz Gerstung, Gerrit K J Hooijer, Se Jin Jang, David R Jones, Hyung-Yong Kim, Tari A King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A Purdie, Keiran Raine, Kamna Ramakrishnan, F Germán Rodríguez-González, Gilles Romieu, Anieta M Sieuwerts, Peter T Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, Åke Borg, Naoto T Ueno, Christos Sotiriou, Alain Viari, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, John W M Martens, Anne-Lise Børresen-Dale, Andrea L

Richardson, Gu Kong, Gilles Thomas, and Michael R Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2 June 2016.

[475] Sandro Morganella, Ludmil B Alexandrov, Dominik Glodzik, Xueqing Zou, Helen Davies, Johan Staaf, Anieta M Sieuwerts, Arie B Brinkman, Sancha Martin, Manasa Ramakrishna, Adam Butler, Hyung-Yong Kim, Åke Borg, Christos Sotiriou, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Hendrik G Stunnenberg, Marc J van de Vijver, John W M Martens, Anne-Lise Børresen-Dale, Andrea L Richardson, Gu Kong, Gilles Thomas, Julian Sale, Cristina Rada, Michael R Stratton, Ewan Birney, and Serena Nik-Zainal. The topography of mutational processes in breast cancer genomes. *Nat. Commun.*, 7:11383, 2 May 2016.

[476] Steven A Roberts, Michael S Lawrence, Leszek J Klimczak, Sara A Grimm, David Fargo, Petar Stojanov, Adam Kiezun, Gregory V Kryukov, Scott L Carter, Gordon Saksena, Shawn Harris, Ruchir R Shah, Michael A Resnick, Gad Getz, and Dmitry A Gordenin. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.*, 45(9):970–976, September 2013.

[477] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Marcin Imielinsk, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van 't Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elías Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell, and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 22 August 2013.

[478] Daniel E Spratt, Tiffany Chan, Levi Waldron, Corey Speers, Felix Y Feng, Olorunseun O Ogunwobi, and Joseph R Osborne. Racial/Ethnic disparities in genomic sequencing. *JAMA Oncol*, 2(8):1070–1074, 1 August 2016.

[479] Carlos D Bustamante, Esteban González Burchard, and Francisco M De la Vega. Genomics for the world. *Nature*, 475(7355):163–165, 13 July 2011.

[480] Alice B Popejoy and Stephanie M Fullerton. Genomics is failing on diversity. *Nature*, 538(7624):161–164, 13 October 2016.

[481] Amanda Eng, Valerie McCormack, and Isabel dos Santos-Silva. Receptor-defined subtypes of breast cancer in indigenous populations in africa: a systematic review and meta-analysis. *PLoS Med.*, 11(9):e1001720, September 2014.

[482] Serena Liao, Ryan J Hartmaier, Kandace P McGuire, Shannon L Puhalla, Soumya Luthra, Uma R Chandran, Tianzhou Ma, Rohit Bhargava, Francesmary Modugno, Nancy E Davidson, Steve Benz, Adrian V Lee, George C Tseng, and Steffi Oesterreich. The molecular landscape of premenopausal breast cancer. *Breast Cancer Res.*, 17:104, 7 August 2015.

[483] Manikandan Periyasamy, Hetal Patel, Chun-Fui Lai, Van T M Nguyen, Ekaterina Nevedomskaya, Alison Harrod, Roslin Russell, Judit Remenyi, Anna Maria Ochocka, Ross S Thomas, Frances Fuller-Pace, Balázs Győrffy, Carlos Caldas, Naveenan Navaratnam, Jason S Carroll, Wilbert Zwart, R Charles Coombes, Luca Magnani, Laki Buluwela, and Simak Ali. APOBEC3B-Mediated cytidine deamination is required for estrogen receptor action in breast cancer. *Cell Rep.*, 13(1):108–121, 6 October 2015.

[484] Yanfeng Zhang, Ryan Delahanty, Xingyi Guo, Wei Zheng, and Jirong Long. Integrative genomic analysis reveals functional diversification of APOBEC gene family in breast cancer. *Hum. Genomics*, 9:34, 18 December 2015.

[485] Ann-Marie Patch, Elizabeth L Christie, Dariush Etemadmoghadam, Dale W Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J Bailey, Karin S Kassahn, Felicity Newell, Michael C J Quinn, Stephen Kazakoff, Kelly Quek, Charlotte Wilhelm-Benartzi, Ed Curry, Huei San Leong, Australian Ovarian Cancer Study Group, Anne Hamilton, Linda Mileshkin, George Au-Yeung, Catherine Kennedy, Jillian Hung, Yoke-Eng Chiew, Paul Harnett, Michael Friedlander, Michael Quinn, Jan Pyman, Stephen Cordner, Patricia O'Brien, Jodie Leditschke, Greg Young, Kate Strachan, Paul Waring, Walid Azar, Chris Mitchell, Nadia Traficante, Joy Hendley, Heather Thorne, Mark Shackleton, David K Miller, Gisela Mir Arnau, Richard W Tothill, Timothy P Holloway, Timothy Semple, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Senel Idrisoglu, Timothy J C Bruxner, Angelika N Christ, Barsha Poudel, Oliver Holmes, Matthew Anderson, Conrad Leonard, Andrew Lonie, Nathan Hall, Scott Wood, Darrin F Taylor, Qinying Xu, J Lynn Fink, Nick Waddell, Ronny Drapkin, Euan Stronach, Hani Gabra, Robert Brown, Andrea Jewell, Shivashankar H Nagaraj, Emma Markham, Peter J Wilson, Jason Ellul, Orla McNally, Maria A Doyle, Ravikiran Vedururu, Collin Stewart, Ernst Lengyel, John V Pearson, Nicola Waddell, Anna deFazio, Sean M Grimmond, and David D L Bowtell. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–494, 28 May 2015.

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

erts, Peter T Simpson, Tari A King, Keiran Raine, Jorunn E Eyfjord, Gu Kong, Åke Borg, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, Anne-Lise Børresen-Dale, John W M Martens, Paul N Span, Sunil R Lakhani, Anne Vincent-Salomon, Christos Sotiriou, Andrew Tutt, Alastair M Thompson, Steven Van Laere, Andrea L Richardson, Alain Viari, Peter J Campbell, Michael R Stratton, and Serena Nik-Zainal. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.*, 23(4):517–525, April 2017.

[487] Katie M O'Brien, Stephen R Cole, Chiu-Kit Tse, Charles M Perou, Lisa A Carey, William D Foulkes, Lynn G Dressler, Joseph Geradts, and Robert C Millikan. Intrinsic breast tumor subtypes, race, and long-term survival in the carolina breast cancer study. *Clin. Cancer Res.*, 16(24):6100–6110, 15 December 2010.

[488] Sasha E Stanton and Mary L Disis. Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *J Immunother Cancer*, 4:59, 18 October 2016.

[489] Antonella Sistigu, Takahiro Yamazaki, Erika Vacchelli, Kariman Chaba, David P Enot, Julien Adam, Ilio Vitale, Aicha Goubar, Elisa E Baracco, Catarina Remédios, Laetitia Fend, Dalil Hannani, Laetitia Aymeric, Yuting Ma, Mireia Niso-Santano, Oliver Kepp, Joachim L Schultze, Thomas Tüting, Filippo Belardelli, Laura Bracci, Valentina La Sorsa, Giovanna Ziccheddu, Paola Sestili, Francesca Urbani, Mauro Delorenzi, Magali Lacroix-Triki, Virginie Quidville, Rosa Conforti, Jean-Philippe Spano, Lajos Pusztai, Vichnou Poirier-Colame, Suzette Delaloge, Frederique Penault-Llorca, Sylvain Ladoire, Laurent Arnould, Joanna Cyrta, Marie-Charlotte Dessoliers, Alexander Eggermont, Marco E Bianchi, Mikael Pittet, Camilla Engblom, Christina Pfirschke, Xavier Préville, Gilles Uzè, Robert D Schreiber, Melvyn T Chow, Mark J Smyth, Enrico Proietti, Fabrice André, Guido Kroemer, and Laurence Zitvogel. Cancer cell-autonomous contribution of type I interferon signaling to the efficacy of chemotherapy. *Nat. Med.*, 20(11):1301–1309, November 2014.

[490] Michael J Campbell, Nathan Y Tonlaar, Elisabeth R Garwood, Dezheng Huo, Dan H Moore, Andrey I Khramtsov, Afred Au, Frederick Baehner, Yinghua Chen, David O Malaka, Amy Lin, Oyinlolu O Adeyanju, Shihong Li, Can Gong, Michael McGrath, Olufunmilayo I Olopade, and Laura J Esserman. Proliferating macrophages associated with high grade, hormone receptor negative breast cancer and poor clinical outcome. *Breast Cancer Res. Treat.*, 128(3):703–711, August 2011.

[491] C A Adisa, N Eleweke, Au A Alfred, M J Campbell, R Sharma, O Nseyo, V Tandon, R Mukhtar, A Greninger, J Di Risi, and L J Esserman. Biology of breast cancer in nigerian women: a pilot study. *Ann. Afr. Med.*, 11(3):169–175, July 2012.

[492] Marcel Smid, F Germán Rodríguez-González, Anieta M Sieuwerts, Roberto Salgado, Wendy J C Prager-Van der Smissen, Michelle van der Vlugt-Daane, Anne van Galen, Serena Nik-Zainal, Johan Staaf, Arie B Brinkman, Marc J van de Vijver, Andrea L Richardson, Aquila Fatima, Kim Berentsen, Adam Butler, Sancha Martin, Helen R

Davies, Reno Debets, Marion E Meijer-Van Gelder, Carolien H M van Deurzen, Gaëtan MacGrogan, Gert G G M Van den Eynden, Colin Purdie, Alastair M Thompson, Carlos Caldas, Paul N Span, Peter T Simpson, Sunil R Lakhani, Steven Van Laere, Christine Desmedt, Markus Ringnér, Stefania Tommasi, Jorunn Eyford, Annegien Broeks, Anne Vincent-Salomon, P Andrew Futreal, Stian Knappskog, Tari King, Gilles Thomas, Alain Viari, Anita Langerød, Anne-Lise Børresen-Dale, Ewan Birney, Hendrik G Stunnenberg, Mike Stratton, John A Foekens, and John W M Martens. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nat. Commun.*, 7:12910, 26 September 2016.

[493] Teresa Davoli, Hajime Uno, Eric C Wooten, and Stephen J Elledge. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, 355(6322), 20 January 2017.

[494] Naiyer A Rizvi, Matthew D Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J Havel, William Lee, Jianda Yuan, Phillip Wong, Teresa S Ho, Martin L Miller, Natasha Rekhtman, Andre L Moreira, Fawzia Ibrahim, Cameron Bruggeman, Billel Gasmi, Roberta Zappasodi, Yuka Maeda, Chris Sander, Edward B Garon, Taha Merghoub, Jedd D Wolchok, Ton N Schumacher, and Timothy A Chan. Cancer immunology. mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*, 348(6230):124–128, 3 April 2015.

[495] Dezheng Huo, Hai Hu, Suhn K Rhie, Eric R Gamazon, Andrew D Cherniack, Jianfang Liu, Toshio F Yoshimatsu, Jason J Pitt, Katherine A Hoadley, Melissa Troester, Yuanbin Ru, Tara Lichtenberg, Lori A Sturtz, Carl S Shelley, Christopher C Benz, Gordon B Mills, Peter W Laird, Craig D Shriver, Charles M Perou, and Olufunmilayo I Olopade. Comparison of breast cancer molecular features and survival by african and european ancestry in the cancer genome atlas. *JAMA Oncol*, 4 May 2017.

[496] Peter Bailey, David K Chang, Katia Nones, Amber L Johns, Ann-Marie Patch, Marie-Claude Gingras, David K Miller, Angelika N Christ, Tim J C Bruxner, Michael C Quinn, Craig Nourse, L Charles Murtaugh, Ivon Harliwong, Senel Idrisoglu, Suzanne Manning, Ehsan Nourbakhsh, Shivangi Wani, Lynn Fink, Oliver Holmes, Venessa Chin, Matthew J Anderson, Stephen Kazakoff, Conrad Leonard, Felicity Newell, Nick Waddell, Scott Wood, Qinying Xu, Peter J Wilson, Nicole Cloonan, Karin S Kassahn, Darrin Taylor, Kelly Quek, Alan Robertson, Lorena Pantano, Laura Mincarelli, Luis N Sanchez, Lisa Evers, Jianmin Wu, Mark Pinese, Mark J Cowley, Marc D Jones, Emily K Colvin, Adnan M Nagrial, Emily S Humphrey, Lorraine A Chantrill, Amanda Mawson, Jeremy Humphris, Angela Chou, Marina Pajic, Christopher J Scarlett, Andreia V Pinho, Marc Giry-Laterriere, Ilse Rooman, Jaswinder S Samra, James G Kench, Jessica A Lovell, Neil D Merrett, Christopher W Toon, Krishna Epari, Nam Q Nguyen, Andrew Barbour, Nikolajs Zeps, Kim Moran-Jones, Nigel B Jamieson, Janet S Graham, Fraser Duthie, Karin Oien, Jane Hair, Robert Grützmann, Anirban Maitra, Christine A Iacobuzio-Donahue, Christopher L Wolfgang, Richard A Morgan, Rita T Lawlor, Vincenzo Corbo, Claudio Bassi, Borislav Rusev, Paola Capelli, Roberto Salvia,

Giampaolo Tortora, Debabrata Mukhopadhyay, Gloria M Petersen, Australian Pancreatic Cancer Genome Initiative, Donna M Munzy, William E Fisher, Saadia A Karim, James R Eshleman, Ralph H Hruban, Christian Pilarsky, Jennifer P Morton, Owen J Sansom, Aldo Scarpa, Elizabeth A Musgrove, Ulla-Maja Hagbo Bailey, Oliver Hofmann, Robert L Sutherland, David A Wheeler, Anthony J Gill, Richard A Gibbs, John V Pearson, Nicola Waddell, Andrew V Biankin, and Sean M Grimmond. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, 531(7592):47–52, 3 March 2016.

[497] Yi-Bo Gao, Zhao-Li Chen, Jia-Gen Li, Xue-Da Hu, Xue-Jiao Shi, Zeng-Miao Sun, Fan Zhang, Zi-Ran Zhao, Zi-Tong Li, Zi-Yuan Liu, Yu-Da Zhao, Jian Sun, Cheng-Cheng Zhou, Ran Yao, Su-Ya Wang, Pan Wang, Nan Sun, Bai-Hua Zhang, Jing-Si Dong, Yue Yu, Mei Luo, Xiao-Li Feng, Su-Sheng Shi, Fang Zhou, Feng-Wei Tan, Bin Qiu, Ning Li, Kang Shao, Li-Jian Zhang, Lan-Jun Zhang, Qi Xue, Shu-Geng Gao, and Jie He. Genetic landscape of esophageal squamous cell carcinoma. *Nat. Genet.*, 46(10):1097–1102, October 2014.

[498] Yaoting Gui, Guangwu Guo, Yi Huang, Xueda Hu, Aifa Tang, Shengjie Gao, Renhua Wu, Chao Chen, Xianxin Li, Liang Zhou, Minghui He, Zesong Li, Xiaojuan Sun, Wenlong Jia, Jinnong Chen, Shangming Yang, Fangjian Zhou, Xiaokun Zhao, Shengqing Wan, Rui Ye, Chaozhao Liang, Zhisheng Liu, Peide Huang, Chunxiao Liu, Hui Jiang, Yong Wang, Hancheng Zheng, Liang Sun, Xingwang Liu, Zhimao Jiang, Dafei Feng, Jing Chen, Song Wu, Jing Zou, Zhongfu Zhang, Ruilin Yang, Jun Zhao, Congjie Xu, Weihua Yin, Zhichen Guan, Jiongxian Ye, Hong Zhang, Jingxiang Li, Karsten Kristiansen, Michael L Nickerson, Dan Theodorescu, Yingrui Li, Xiuqing Zhang, Songgang Li, Jian Wang, Huanming Yang, Jun Wang, and Zhiming Cai. Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat. Genet.*, 43(9):875–878, 7 August 2011.

[499] Timothy F Burns, Peiwen Fei, Kimberly A Scata, David T Dicker, and Wafik S El-Deiry. Silencing of the novel p53 target gene Snk/Plk2 leads to mitotic catastrophe in paclitaxel (taxol)-exposed cells. *Mol. Cell. Biol.*, 23(16):5556–5571, August 2003.

[500] Helen M Coley, Eleftheria Hatzimichael, Sarah Blagden, Iain McNeish, Alastair Thompson, Tim Crook, and Nelofer Syed. Polo like kinase 2 tumour suppressor and cancer biomarker: new perspectives on drug sensitivity/resistance in ovarian cancer. *Oncotarget*, 3(1):78–83, January 2012.

[501] Nelofer Syed, Helen M Coley, Jalid Sehouli, Dominique Koensgen, Alexander Mustea, Peter Szlosarek, Iain McNeish, Sarah P Blagden, Peter Schmid, David P Lovell, Eleftheria Hatzimichael, and Tim Crook. Polo-like kinase plk2 is an epigenetic determinant of chemosensitivity and clinical outcomes in ovarian cancer. *Cancer Res.*, 71(9):3317–3327, 1 May 2011.

[502] Elizabeth M Matthew, Lori S Hart, Aristotelis Astrinidis, Arunasalam Navaraj, Nathan G Dolloff, David T Dicker, Elizabeth P Henske, and Wafik S El-Deiry. The p53

target plk2 interacts with TSC proteins impacting mTOR signaling, tumor growth and chemosensitivity under hypoxic conditions. *Cell Cycle*, 8(24):4168–4175, 15 December 2009.

[503] Vishal Kothari, Iris Wei, Sunita Shankar, Shanker Kalyana-Sundaram, Lidong Wang, Linda W Ma, Pankaj Vats, Catherine S Grasso, Dan R Robinson, Yi-Mi Wu, Xuhong Cao, Diane M Simeone, Arul M Chinnaiyan, and Chandan Kumar-Sinha. Outlier kinase expression by RNA sequencing as targets for precision therapy. *Cancer Discov.*, 3(3):280–293, March 2013.

[504] Abel Gonzalez-Perez, Alba Jene-Sanz, and Nuria Lopez-Bigas. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. *Genome Biol.*, 14(9):r106, 2013.

[505] Maria Vittoria Dieci, Veronika Smutná, Véronique Scott, Guangliang Yin, Ran Xu, Philippe Vielh, Marie-Christine Mathieu, Cécile Vicier, Melanie Laporte, Francoise Drusch, Valentina Guarneri, Pierfranco Conte, Suzette Delaloge, Ludovic Lacroix, Olivia Fromigué, Fabrice André, and Celine Lefebvre. Whole exome sequencing of rare aggressive breast cancer histologies. *Breast Cancer Res. Treat.*, 156(1):21–32, February 2016.

[506] Jiali Zhuang and Zhiping Weng. Local sequence assembly reveals a high-resolution profile of somatic structural variations in 97 cancer genomes. *Nucleic Acids Res.*, 43(17):8146–8156, 30 September 2015.

[507] Gijs van Haaften, Gillian L Dalgliesh, Helen Davies, Lina Chen, Graham Bignell, Chris Greenman, Sarah Edkins, Claire Hardy, Sarah O'Meara, Jon Teague, Adam Butler, Jonathan Hinton, Calli Latimer, Jenny Andrews, Syd Barthorpe, Dave Beare, Gemma Buck, Peter J Campbell, Jennifer Cole, Simon Forbes, Mingming Jia, David Jones, Chai Yin Kok, Catherine Leroy, Meng-Lay Lin, David J McBride, Mark Maddison, Simon Maquire, Kirsten McLay, Andrew Menzies, Tatiana Mironenko, Lee Mulderrig, Laura Mudie, Erin Pleasance, Rebecca Shepherd, Raffaella Smith, Lucy Stebbings, Philip Stephens, Gurpreet Tang, Patrick S Tarpey, Rachel Turner, Kelly Turrell, Jennifer Varian, Sofie West, Sara Widaa, Paul Wray, V Peter Collins, Koichi Ichimura, Simon Law, John Wong, Siu Tsan Yuen, Suet Yi Leung, Giovanni Tonon, Ronald A DePinho, Yu-Tzu Tai, Kenneth C Anderson, Richard J Kahnoski, Aaron Massie, Sok Kean Khoo, Bin Tean Teh, Michael R Stratton, and P Andrew Futreal. Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. *Nat. Genet.*, 41(5):521–523, May 2009.

[508] Ying Su, Ashim Subedee, Noga Bloushtain-Qimron, Virginia Savova, Marcin Krzystanek, Lewyn Li, Andriy Marusyk, Doris P Tabassum, Alexander Zak, Mary Jo Flacker, Mei Li, Jessica J Lin, Saraswati Sukumar, Hiromu Suzuki, Henry Long, Zoltan Szallasi, Alexander Gimelbrant, Reo Maruyama, and Kornelia Polyak. Somatic cell fusions reveal extensive heterogeneity in basal-like breast cancer. *Cell Rep.*, 11(10):1549–1563, 16 June 2015.

[509] Xiwen Cheng and Hung-Ying Kao. G protein pathway suppressor 2 (GPS2) is a transcriptional corepressor important for estrogen receptor alpha-mediated transcriptional regulation. *J. Biol. Chem.*, 284(52):36395–36404, 25 December 2009.

[510] Carolina Pereira, Pol Gimenez-Xavier, Eva Pros, Maria J Pajares, Massimo Moro, Antonio Gomez, Alejandro Navarro, Enric Condom, Sebastian Moran, Gonzalo Gómez-López, Osvaldo Graña, Miriam Rubio-Camarillo, Alex Martinez-Martí, Jun Yokota, Julian Carretero, Jose M Galbis, Ernest Nadal, David G Pisano, Gabriella Sozzi, Enriqueta Felip, Luis M Montuenga, Luca Roz, Alberto Villanueva, and Montse Sanchez-Cespedes. GENOMIC PROFILING OF PATIENT-DERIVED XENOGRAFTS FOR LUNG CANCER IDENTIFIES B2M INACTIVATION IMPAIRING IMMUNORECOGNITION. *Clin. Cancer Res.*, 21 December 2016.

[511] V Vanderpuye, S Grover, N Hammad, PoojaPrabhakar, H Simonds, F Olopade, and D C Stefan. An update on the management of breast cancer in africa. *Infect. Agent. Cancer*, 12:13, 14 February 2017.

[512] Lesley J Scott. Niraparib: First global approval. *Drugs*, 77(9):1029–1034, June 2017.

[513] Asher Mullard. PARP inhibitors plough on. *Nat. Rev. Drug Discov.*, 16(4):229, 30 March 2017.

[514] Jessica S Brown, Brent O'Carrigan, Stephen P Jackson, and Timothy A Yap. Targeting DNA repair in cancer: Beyond PARP inhibitors. *Cancer Discov.*, 7(1):20–37, January 2017.

[515] Amir Sonnenblick, Evandro de Azambuja, Hatem A Azim, Jr, and Martine Piccart. An update on PARP inhibitors–moving to the adjuvant setting. *Nat. Rev. Clin. Oncol.*, 12(1):27–41, January 2015.

[516] Ken Y Lin and W Lee Kraus. PARP inhibitors for cancer therapy. *Cell*, 169(2):183, 6 April 2017.

[517] Kristine M Frizzell and W Lee Kraus. PARP inhibitors and the treatment of breast cancer: beyond BRCA1/2? *Breast Cancer Res.*, 11(6):111, 26 November 2009.

[518] Rodrigo Dienstmann, In Sock Jang, Brian Bot, Stephen Friend, and Justin Guinney. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discov.*, 5(2):118–123, February 2015.

[519] Kishore Guda, Martina L Veigl, Vinay Varadan, Arman Nosrati, Lakshmeswari Ravi, James Lutterbaugh, Lydia Beard, James K V Willson, W David Sedwick, Zhenghe John Wang, Neil Molyneaux, Alexander Miron, Mark D Adams, Robert C Elston, Sanford D Markowitz, and Joseph E Willis. Novel recurrently mutated genes in african american colon cancers. *Proc. Natl. Acad. Sci. U. S. A.*, 112(4):1149–1154, 27 January 2015.

[520] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manage.*, 35(2):137–144, April 2015.

[521] Paul Muir, Shantao Li, Shaoke Lou, Daifeng Wang, Daniel J Spakowicz, Leonidas Salichos, Jing Zhang, George M Weinstock, Farren Isaacs, Joel Rozowsky, and Mark Gerstein. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.*, 17:53, 23 March 2016.

[522] A G Knudson, Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.*, 68(4):820–823, April 1971.

[523] Brenda K Edwards, Anne-Michelle Noone, Angela B Mariotto, Edgar P Simard, Francis P Boscoe, S Jane Henley, Ahmedin Jemal, Hyunsoon Cho, Robert N Anderson, Betsy A Kohler, Christie R Eheman, and Elizabeth M Ward. Annual report to the nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer*, 120(9):1290–1314, 1 May 2014.

[524] Christine Marosi and Marcus Köller. Challenge of cancer in the elderly. *ESMO Open*, 1(3):e000020, 12 April 2016.

[525] Nazneen Rahman. Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–308, 16 January 2014.

[526] Fergus J Couch, Hermela Shimelis, Chunling Hu, Steven N Hart, Eric C Polley, Jie Na, Emily Hallberg, Raymond Moore, Abigail Thomas, Jenna Lilyquist, Bingjian Feng, Rachel McFarland, Tina Pesaran, Robert Huether, Holly LaDuca, Elizabeth C Chao, David E Goldgar, and Jill S Dolinsky. Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol*, 13 April 2017.

[527] John S Witte and Thomas J Hoffmann. Polygenic modeling of genome-wide association studies: an application to prostate and breast cancer. *OMICS*, 15(6):393–398, June 2011.

[528] Karoline B Kuchenbaecker, Lesley McGuffog, Daniel Barrowdale, Andrew Lee, Penny Soucy, Joe Dennis, Susan M Domchek, Mark Robson, Amanda B Spurdle, Susan J Ramus, Nasim Mavaddat, Mary Beth Terry, Susan L Neuhausen, Rita Katharina Schmutzler, Jacques Simard, Paul D P Pharoah, Kenneth Offit, Fergus J Couch, Georgia Chenevix-Trench, Douglas F Easton, Antonis C Antoniou, Michael Lush, Ute Hamann, Melissa Southey, Esther M John, Wendy K Chung, Mary B Daly, Saundra S Buys, David E Goldgar, Cecilia M Dorfling, Elizabeth J van Rensburg, Yuan Chun Ding, Bent Ejlertsen, Anne-Marie Gerdes, Thomas V O Hansen, Susan Slager, Emily Hallberg, Javier Benitez, Ana Osorio, Nancy Cohen, William Lawler, Jeffrey N Weitzel, Paolo Peterlongo, Valeria Pensotti, Riccardo Dolcetti, Monica Barile, Bernando Bonanni, Jacopo Azzollini, Siranoush Manoukian, Bernard Peissel, Paolo Radice, Antonella Savarese, Laura Papi, Giuseppe Giannini, Florentia Fostira, Irene Konstantopoulou,

Julian Adlard, Carole Brewer, Jackie Cook, Rosemarie Davidson, Diana Eccles, Ros Eeles, Steve Ellis, Debra Frost, Shirley Hodgson, Louise Izatt, Fiona Lalloo, Kai-Ren Ong, Andrew K Godwin, Norbert Arnold, Bernd Dworniczak, Christoph Engel, Andrea Gehrig, Eric Hahnen, Jan Hauke, Karin Kast, Alfons Meindl, Deiter Niederacher, Rita Katherina Schmutzler, Raymonda Varon-Mateeva, Shan Wang-Gohrke, Barbara Wappenschmidt, Laure Barjhoux, Marie-Agnes Collonge-Rame, Camille Elan, Lisa Golmard, GEMO Study Collaborators, EMBRACE, Emmanuelle Barouk-Simonet, Fabienne Lesueur, Sylvie Mazoyer, Joanna Sokolowska, Dominique Stoppa-Lyonnet, Claudine Isaacs, Kathleen B M Claes, Bruce Poppe, Miguel de la Hoya, Vanesa Garcia-Barberan, Kristiina Aittomaki, Heli Nevanlinna, Margreet G E M Ausems, J L de Lange, Encarna B Gomez Garcia, Frans B L Hogervorst, HEBON, Carolien M Kets, Hanne E Meijers-Heijboer, Jan C Oosterwijk, Matti A Rookus, Christi J van Asperen, Ans M W van den Ouweland, Helena C van Doorn, Theo A M van Os, Ava Kwong, Edith Olah, Orland Diez, Joan Brunet, Conxi Lazaro, Alex Teule, Jacek Gronwald, Anna Jakubowska, Katarzyna Kaczmarek, Jan Lubinski, Grzegorz Sukiennicki, Rosa B Barkardottir, Jocelyne Chiquette, Simona Agata, Marco Montagna, Manuel R Teixeira, Sue Kyung Park, KConFab Investigators, Curtis Olswold, Marc Tischkowitz, Lenka Foretova, Pragna Gaddam, Joseph Vijai, Georg Pfeiler, Christine Rappaport-Fuerhauser, Christian F Singer, Muy-Kheng M Tea, Mark H Greene, Jennifer T Loud, Gad Rennert, Evgeny N Imyanitov, Peter J Hulick, John L Hays, Marion Piedmonte, Gustavo C Rodriguez, Julie Martyn, Gord Glendon, Anna Marie Mulligan, Irene L Andrulis, Amanda Ewart Toland, Uffe Birk Jenson, Torben A Kruse, Inge Sokilde Pedersen, Mads Thomassen, Maria A Caligo, Soo-Hwang Teo, Raanan Berger, Eitan Friedman, Yael Laitman, Brita Arver, Ake Borg, Hans Ehrancrona, Johanna Rantala, Olufunmilayo I Olopade, Patricia A Ganz, Robert L Nussbaum, Angela R Bradbury, Susan M Domchek, Katherine L Nathanson, Banu K Arun, Paul James, Beth Y Karlan, Jenny Lester, Jacques Simard, Paul D P Pharoah, Kenneth Offit, Fergus J Couch, Georgia Chenevix-Trench, Douglas F Easton, and Antonis C Antoniou. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.*, 109(7), 1 July 2017.

[529] Andrea Ganna, Kyle Satterstrom, Seyedeh Zekavat, Indraniel Das, Mitja Kurki, Claire Churchhouse, Jessica Alfoldi, Alicia Martin, Aki Havulinna, Andrea Byrnes, Wesley Thompson, Philip Nielsen, Konrad Karczewski, Elmo Saarentaus, Manuel Rivas, Namrata Gupta, Olli Pietilainen, Connor Emdin, Francesco Lescai, Jonas Bybjerg-Grauholm, Jason Flannick, Josep Mercader, Miriam Udler, Markku Laakso, Veikko Salomaa, Christina Hultman, Samuli Ripatti, Eija Hamalainen, Jukka Moilanen, Jarmo Korkko, Outi Kuismin, Merete Nordentoft, David Hougaard, Ole Mors, Thomas Werge, Preben Mortensen, Daniel MacArthur, Mark Daly, Patrick Sullivan, Adam Locke, Aarno Palotie, Anders Borglum, Sekar Kathiresan, and Benjamin Neale. Quantifying the impact of rare and ultra-rare coding variation across the phenotypic spectrum, 2017.

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

huri R Hegde, Elaine Lyon, Brian E Ward, and Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee. ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007. *Genet. Med.*, 10(4):294–300, April 2008.

[531] Jae Yeon Cheon, Jessica Mozersky, and Robert Cook-Deegan. Variants of uncertain significance in BRCA: a harbinger of ethical and policy issues to come? *Genome Med.*, 6(12):121, 19 December 2014.

[532] Charles Lu, Mingchao Xie, Michael C Wendl, Jiayin Wang, Michael D McLellan, Mark D M Leiserson, Kuan-Lin Huang, Matthew A Wyczalkowski, Reyka Jayasinghe, Tapahsama Banerjee, Jie Ning, Piyush Tripathi, Qunyuan Zhang, Beifang Niu, Kai Ye, Heather K Schmidt, Robert S Fulton, Joshua F McMichael, Prag Batra, Cyriac Kandoth, Maheetha Bharadwaj, Daniel C Koboldt, Christopher A Miller, Krishna L Kanchi, James M Eldred, David E Larson, John S Welch, Ming You, Bradley A Ozenberger, Ramaswamy Govindan, Matthew J Walter, Matthew J Ellis, Elaine R Mardis, Timothy A Graubert, John F Dipersio, Timothy J Ley, Richard K Wilson, Paul J Goodfellow, Benjamin J Raphael, Feng Chen, Kimberly J Johnson, Jeffrey D Parvin, and Li Ding. Patterns and functional implications of rare germline variants across 12 cancer types. *Nat. Commun.*, 6:10086, 22 December 2015.

[533] Ingegerd Elvers, Jason Turner-Maier, Ross Swofford, Michele Koltookian, Jeremy Johnson, Chip Stewart, Cheng-Zhong Zhang, Steven E Schumacher, Rameen Beroukhim, Mara Rosenberg, Rachael Thomas, Evan Mauceli, Gad Getz, Federica Di Palma, Jaime F Modiano, Matthew Breen, Kerstin Lindblad-Toh, and Jessica Alföldi. Exome sequencing of lymphomas from three dog breeds reveals somatic mutation patterns reflecting genetic background. *Genome Res.*, 25(11):1634–1645, November 2015.

[534] Eric C Dietze, Christopher Sistrunk, Gustavo Miranda-Carboni, Ruth O'Regan, and Victoria L Seewaldt. Triple-negative breast cancer in African-American women: disparities versus biology. *Nat. Rev. Cancer*, 15(4):248–254, April 2015.

[535] Diana E Ramirez-Ardila, Jean C Helmijr, Maxime P Look, Irene Lurkin, Kirsten Ruigrok-Ritstier, Steven van Laere, Luc Dirix, Fred C Sweep, Paul N Span, Sabine C Linn, John A Foekens, Stefan Sleijfer, Els M J J Berns, and Maurice P H M Jansen. Hotspot mutations in PIK3CA associate with first-line treatment outcome for aromatase inhibitors but not for tamoxifen. *Breast Cancer Res. Treat.*, 139(1):39–49, May 2013.

[536] Toru Mukohara. PI3K mutations in breast cancer: prognostic and therapeutic implications. *Breast Cancer*, 7:111–123, 15 May 2015.

[537] E M Berns, J A Foekens, R Vossen, M P Look, P Devilee, S C Henzen-Logmans, I L van Staveren, W L van Putten, M Inganäs, M E Meijer-van Gelder, C Cornelisse, C J Claassen, H Portengen, B Bakker, and J G Klijn. Complete sequencing of TP53

predicts poor response to systemic therapy of advanced breast cancer. *Cancer Res.*, 60(8):2155–2162, 15 April 2000.

[538] Helen Davies, Dominik Glodzik, Sandro Morganella, Lucy R Yates, Johan Staaf, Xueqing Zou, Manasa Ramakrishna, Sancha Martin, Sandrine Boyault, Anieta M Sieuwerts, Peter T Simpson, Tari A King, Keiran Raine, Jorunn E Eyfjord, Gu Kong, Åke Borg, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, Anne-Lise Børresen-Dale, John W M Martens, Paul N Span, Sunil R Lakhani, Anne Vincent-Salomon, Christos Sotiriou, Andrew Tutt, Alastair M Thompson, Steven Van Laere, Andrea L Richardson, Alain Viari, Peter J Campbell, Michael R Stratton, and Serena Nik-Zainal. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.*, 23(4):517–525, April 2017.

[539] Beth N Peshkin, Michelle L Alabek, and Claudine Isaacs. BRCA1/2 mutations and triple negative breast cancers. *Breast Dis.*, 32(1-2):25–33, 2010.

[540] B G Haffty, A Silber, E Matloff, J Chung, and D Lannin. Racial differences in the incidence of BRCA1 and BRCA2 mutations in a cohort of early onset breast cancer patients: African american compared to white women. *J. Med. Genet.*, 43(2):133–137, February 2006.

[541] Michael J Hall, Julia E Reid, Lynn A Burbidge, Dmitry Pruss, Amie M Deffenbaugh, Cynthia Frye, Richard J Wenstrup, Brian E Ward, Thomas A Scholl, and Walter W Noll. BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer. *Cancer*, 115(10):2222–2233, 15 May 2009.

[542] Jirong Long, Ben Zhang, Lisa B Signorello, Qiuyin Cai, Sandra Deming-Halverson, Martha J Shrubsole, Maureen Sanderson, Joe Dennis, Kyriaki Michailidou, Kyriaki Michailiou, Douglas F Easton, Xiao-Ou Shu, William J Blot, and Wei Zheng. Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. *PLoS One*, 8(4):e58350, 8 April 2013.

[543] Stephen Henderson, Ankur Chakravarthy, Xiaoping Su, Chris Boshoff, and Tim Robert Fenton. APOBEC-mediated cytosine deamination links PIK3CA helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep.*, 7(6):1833–1841, 26 June 2014.

[544] Serena Nik-Zainal, David C Wedge, Ludmil B Alexandrov, Mia Petljak, Adam P Butler, Niccolo Bolli, Helen R Davies, Stian Knappskog, Sancha Martin, Elli Papaemmanuil, Manasa Ramakrishna, Adam Shlien, Ingrid Simonic, Yali Xue, Chris Tyler-Smith, Peter J Campbell, and Michael R Stratton. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat. Genet.*, 46(5):487–491, May 2014.

[545] Michael B Burns, Lela Lackey, Michael A Carpenter, Anurag Rathore, Allison M Land, Brandon Leonard, Eric W Refsland, Delshanee Kotandeniya, Natalia Tretyakova, Jason B Nikas, Douglas Yee, Nuri A Temiz, Duncan E Donohue, Rebecca M McDougle, William L Brown, Emily K Law, and Reuben S Harris. APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*, 494(7437):366–370, 21 February 2013.

[546] Michael Wilde, Mihael Hategan, Justin M Wozniak, Ben Clifford, Daniel S Katz, and Ian Foster. Swift: A language for distributed parallel scripting. *Parallel Comput.*, 37(9):633–652, 2011.

[547] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and 1000 Genome Project Data. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[548] Andy Rimmer, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, WGS500 Consortium, Andrew O M Wilkie, Gil McVean, and Gerton Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, 46(8):912–918, August 2014.

[549] Mohammad Shabbir Hasan, Xiaowei Wu, and Liqing Zhang. Performance evaluation of indel calling tools using real short-read data. *Hum. Genomics*, 9(1), 2015.

[550] William McLaren, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. Deriving the consequences of genomic variants with the ensembl API and SNP effect predictor. *Bioinformatics*, 26(16):2069–2070, 2010.

[551] Peter Van Loo, Silje H Nordgard, Ole Christian Lingjærde, Hege G Russnes, Inga H Rye, Wei Sun, Victor J Weigman, Peter Marynen, Anders Zetterberg, Bjørn Naume, Charles M Perou, Anne-Lise Børresen-Dale, and Vessela N Kristensen. Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U. S. A.*, 107(39):16910–16915, 28 September 2010.

[552] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan F A Grant, Hakon Hakonarson, and Maja Bucan. PennCNV: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, 17(11):1665–1674, November 2007.

[553] Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, 42(Database issue):D980–5, January 2014.

[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,

David N Cooper, Nicole Deflaux, Mark DePristo, Ron Do, Jason Flannick, Menachem Fromer, Laura Gauthier, Jackie Goldstein, Namrata Gupta, Daniel Howrigan, Adam Kiezun, Mitja I Kurki, Ami Levy Moonshine, Pradeep Natarajan, Lorena Orozco, Gina M Peloso, Ryan Poplin, Manuel A Rivas, Valentin Ruano-Rubio, Samuel A Rose, Douglas M Ruderfer, Khalid Shakir, Peter D Stenson, Christine Stevens, Brett P Thomas, Grace Tiao, Maria T Tusie-Luna, Ben Weisburd, Hong-Hee Won, Dongmei Yu, David M Altshuler, Diego Ardissino, Michael Boehnke, John Danesh, Stacey Donnelly, Roberto Elosua, Jose C Florez, Stacey B Gabriel, Gad Getz, Stephen J Glatt, Christina M Hultman, Sekar Kathiresan, Markku Laakso, Steven McCarroll, Mark I McCarthy, Dermot McGovern, Ruth McPherson, Benjamin M Neale, Aarno Palotie, Shaun M Purcell, Danish Saleheen, Jeremiah M Scharf, Pamela Sklar, Patrick F Sullivan, Jaakko Tuomilehto, Ming T Tsuang, Hugh C Watkins, James G Wilson, Mark J Daly, Daniel G MacArthur, and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 18 August 2016.

[555] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, 46(3):310–315, March 2014.

[556] Thomas Paul Slavin, Mariana Niell-Swiller, Ilana Solomon, Bita Nehoray, Christina Rybak, Kathleen R Blazer, and Jeffrey N Weitzel. Clinical application of multigene panels: Challenges of Next-Generation counseling and cancer risk management. *Front. Oncol.*, 5:208, 29 September 2015.

[557] Lisa R Susswein, Megan L Marshall, Rachel Nusbaum, Kristen J Vogel Postula, Scott M Weissman, Lauren Yackowski, Erica M Vaccari, Jeffrey Bissonnette, Jessica K Booker, M Laura Cremona, Federica Gibellini, Patricia D Murphy, Daniel E Pineda-Alvarez, Guido D Pollevick, Zhixiong Xu, Gabi Richard, Sherri Bale, Rachel T Klein, Kathleen S Hruska, and Wendy K Chung. Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. *Genet. Med.*, 18(8):823–832, August 2016.

[558] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, Michael Kuhn, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43(Database issue):D447–52, January 2015.

[559] Clement A Adebamowo, Temidayo O Ogundiran, Adeniyi A Adenipekun, Rasheed A Oyesegun, Oladapo B Campbell, Effiong U Akang, Charles N Rotimi, and Olufunmilayo I Olopade. Obesity and height in urban nigerian women with breast cancer. *Ann. Epidemiol.*, 13(6):455–461, July 2003.

[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

breastfeeding are protective against breast cancer in nigerian women. *Br. J. Cancer*, 98(5):992–996, 11 March 2008.

[561] Dezheng Huo, Francis Ikpatt, Andrey Khramtsov, Jean-Marie Dangou, Rita Nanda, James Dignam, Bifeng Zhang, Tatyana Grushko, Chunling Zhang, Olayiwola Oluwasola, David Malaka, Sani Malami, Abayomi Odetunde, Adewumi O Adeoye, Festus Iyare, Adeyinka Falusi, Charles M Perou, and Olufunmilayo I Olopade. Population differences in breast cancer: survey in indigenous african women reveals overrepresentation of triple-negative breast cancer. *J. Clin. Oncol.*, 27(27):4515–4521, 20 September 2009.

[562] Temidayo O Ogundiran, Dezheng Huo, Adeniyi Adenipekun, Oladapo Campbell, Rasaaq Oyesegun, Effiong Akang, Clement Adebamowo, and Olufunmilayo I Olopade. Case-control study of body size and breast cancer risk in nigerian women. *Am. J. Epidemiol.*, 172(6):682–690, 15 September 2010.

[563] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 4 October 2012.

[564] Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219, March 2013.

[565] Christopher T Saunders, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 15 July 2012.

[566] Alex H Ramos, Lee Lichtenstein, Manaswi Gupta, Michael S Lawrence, Trevor J Pugh, Gordon Saksena, Matthew Meyerson, and Gad Getz. Oncotator: cancer variant annotation tool. *Hum. Mutat.*, 36(4):E2423–9, April 2015.

[567] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham R S Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biol.*, 17(1):122, 6 June 2016.

[568] Giuseppe Narzisi, Jason A O'Rawe, Ivan Iossifov, Han Fang, Yoon-Ha Lee, Zihua Wang, Yiyang Wu, Gholson J Lyon, Michael Wigler, and Michael C Schatz. Accurate de novo and transmitted indel detection in exome-capture data using microassembly. *Nat. Methods*, 11(10):1033–1036, October 2014.

[569] Han Fang, Yiyang Wu, Giuseppe Narzisi, Jason A O'Rawe, Laura T Jimenez Barrón, Julie Rosenbaum, Michael Ronemus, Ivan Iossifov, Michael C Schatz, and Gholson J Lyon. Reducing INDEL calling errors in whole genome and exome sequencing data. *Genome Med.*, 6(10):89, 28 October 2014.

[570] Aleksandr Morgulis, E Michael Gertz, Alejandro A Schäffer, and Richa Agarwala. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, 13(5):1028–1040, June 2006.

[571] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 1 October 2015.

[572] Markus Riester, Angad P Singh, A Rose Brannon, Kun Yu, Catarina D Campbell, Derek Y Chiang, and Michael P Morrissey. PureCN: copy number calling and SNV classification using targeted short read sequencing. *Source Code Biol. Med.*, 11:13, 15 December 2016.

[573] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nat. Rev. Cancer*, 4(3):177–183, March 2004.

[574] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, 12(4):R41, 28 April 2011.

[575] Ronglai Shen and Venkatraman E Seshan. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.*, 44(16):e131, 19 September 2016.

[576] F Favero, T Joshi, A M Marquard, N J Birkbak, M Krzystanek, Q Li, Z Szallasi, and A C Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.*, 26(1):64–70, January 2015.

[577] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 15 September 2012.

[578] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, 15(6):R84, 26 June 2014.

[579] Dezheng Huo, Hai Hu, Suhn K Rhie, Eric R Gamazon, Andrew D Cherniack, Jianfang Liu, Toshio F Yoshimatsu, Jason J Pitt, Katherine A Hoadley, Melissa Troester, Yuanbin Ru, Tara Lichtenberg, Lori A Sturtz, Carl S Shelley, Christopher C Benz, Gordon B Mills, Peter W Laird, Craig D Shriver, Charles M Perou, and Olufunmilayo I Olopade. Comparison of breast cancer molecular features and survival by african and european ancestry in the cancer genome atlas. *JAMA Oncol*, 4 May 2017.

[580] Dezheng Huo, Yonglan Zheng, Temidayo O Ogundiran, Clement Adebamowo, Katherine L Nathanson, Susan M Domchek, Timothy R Rebbeck, Michael S Simon, Esther M John, Anselm Hennis, Barbara Nemesure, Suh-Yuh Wu, M Cristina Leske, Stefan Ambs, Qun Niu, Jing Zhang, Nancy J Cox, and Olufunmilayo I Olopade. Evaluation of 19 susceptibility loci of breast cancer in women of african ancestry. *Carcinogenesis*, 33(4):835–840, April 2012.

[581] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, Adam Kiezun, Peter S Hammerman, Aaron McKenna, Yotam Drier, Lihua Zou, Alex H Ramos, Trevor J Pugh, Nicolas Stransky, Elena Helman, Jaegil Kim, Carrie Sougnez, Lauren Ambrogio, Elizabeth Nickerson, Erica Shefler, Maria L Cortés, Daniel Auclair, Gordon Saksena, Douglas Voet, Michael Noble, Daniel DiCara, Pei Lin, Lee Lichtenstein, David I Heiman, Timothy Fennell, Marcin Imielinski, Bryan Hernandez, Eran Hodis, Sylvan Baca, Austin M Dulak, Jens Lohr, Dan-Avi Landau, Catherine J Wu, Jorge Melendez-Zajgla, Alfredo Hidalgo-Miranda, Amnon Koren, Steven A McCarroll, Jaume Mora, Ryan S Lee, Brian Crompton, Robert Onofrio, Melissa Parkin, Wendy Winckler, Kristin Ardlie, Stacey B Gabriel, Charles W M Roberts, Jaclyn A Biegel, Kimberly Stegmaier, Adam J Bass, Levi A Garraway, Matthew Meyerson, Todd R Golub, Dmitry A Gordenin, Shamil Sunyaev, Eric S Lander, and Gad Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 11 July 2013.

[582] Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 23 January 2014.

[583] Julian S Gehring, Bernd Fischer, Michael Lawrence, and Wolfgang Huber. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, 31(22):3673–3675, 15 November 2015.

[584] Jaegil Kim, Kent W Mouw, Paz Polak, Lior Z Braunstein, Atanas Kamburov, Grace Tiao, David J Kwiatkowski, Jonathan E Rosenberg, Eliezer M Van Allen, Alan D D'Andrea, and Gad Getz. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.*, 48(6):600–606, June 2016.

[585] Jun Hu, Huanying Ge, Matt Newman, and Kejun Liu. OSA: a fast and accurate alignment tool for RNA-Seq. *Bioinformatics*, 28(14):1933–1934, 15 July 2012.

[586] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 1 January 2010.

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

Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 15 March 2012.

[588] Cristóbal Fresno, Germán Alexis González, Gabriela Alejandra Merino, Ana Georgina Flesia, Osvaldo Luis Podhajcer, Andrea Sabina Llera, and Elmer Andrés Fernández. A novel non-parametric method for uncertainty evaluation of correlation-based molecular signatures: its application on PAM50 algorithm. *Bioinformatics*, 33(5):693–700, 1 March 2017.

[589] Sonja Hänzelmann, Robert Castelo, and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*, 14:7, 16 January 2013.