8-2022

Haplotype-Informed Allelic Imbalance Detection From Rna In Cancer

Zuhal Ozcan

HAPLOTYPE-INFORMED ALLELIC IMBALANCE DETECTION FROM RNA IN CANCER

by

Zuhal Ozcan, B.Sc.

APPROVED:

_____
Paul Scheet, Ph.D.
Advisory Professor

_____
Yasminka A. Jakubek-Swartzlander, Ph.D.

_____
Eduardo Vilar-Sanchez, M.D., Ph.D.

_____
Swathi Arur, Ph.D.

_____
Chad Huff, Ph.D.

_____
Yin Liu, Ph.D.

APPROVED:

_____
Dean, The University of Texas
MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

HAPLOTYPE-INFORMED ALLELIC IMBALANCE DETECTION FROM RNA IN CANCER

A

Dissertation

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

by

Zuhal Ozcan, B.Sc.
Houston, Texas

August, 2022

To my parents Hayriye and Ali Ozcan, and sister Nihal Ozcan, for their unconditional love, encouragement, and support.

To my best friend and fiancé, Vincent Liguori, who has been by my side through my ups and downs and never stopped believing in me and being excited about my work.

Acknowledgments

I would like to thank my advisor, Paul Scheet, for sharing his wisdom and knowledge with me, and supporting my career development, especially by welcoming me to his lab when I needed a second home. With his guidance, I have learned a great deal about cancer biology and statistics, and improved my bioinformatics skills. The scientific foundation I built under his mentorship will help shape the rest of my career. I am grateful for the opportunities I was given to participate in collaborations; for helping me set my priorities when I was busy with multiple projects and interviews, and for believing in me and encouraging me when I faced challenges. I am also very grateful for the many conversations we had where I got to enjoy his great sense of humor and even had a chance to hear him sing.

I would like to thank Sasha Jakubek for her endless time and support, and for being there for me whenever I needed. She has been a second sister to me. I greatly appreciate her patience and desire for teaching she showed when I bombarded her with questions.

I thank Dr. Vilar for giving me the opportunity to collaborate with him and his group. Our fruitful collaborations have provided me insights into hereditary colorectal cancer syndromes and helped me gain new skills.

I would like to thank Anthony San Lucas for the many insightful and stimulating conversations, and his helpful advice about hapLOHseq.

I would like to thank all my current and previous lab mates, specifically Justin Wong, Kyle Chang, Smruthy Sivakumar, and Jerry Fowler for their helpful advice and discussions.

I would like to thank Dr. Shelley Barton for her inspirational mentorship and support since 2013.

I would like to thank my Advisory Committee members Drs. Chad Huff, Swathi Arur, and Yin Liu for helping me shape my dissertation project and supporting my career development. I am also grateful to them for their valuable time and advice.

Finally, I would like to thank my parents Hayriye and Ali Ozcan, and sister, Nihal Ozcan for their love, support, and encouragement, and my fiancé Vincent Liguori, who has been supporting me throughout my PhD journey.

HAPLOTYPE-INFORMED ALLELIC IMBALANCE DETECTION FROM RNA IN CANCER

Zuhal Ozcan, B.Sc.

Advisory Professor: Paul Scheet, Ph.D.

Abstract

Comprehensive genomic and transcriptomic characterization of tumors has uncovered enrichment for distinct aneuploidy and expression patterns, demonstrating the utility of molecular based classification of cancers and their subtypes. Existing cohorts with transcriptomic profiling from next-generation sequencing contain an untapped potential to also relate genomics with rich clinical phenotypes. Yet, derivation of somatic copy number and expression profiles from analyses of RNA has remained elusive. Further, DNA analysis in these cohorts is not always feasible due to limited tissue availability or financial constraints. Here, we present a statistical approach that overcomes these challenges using haplotype information to aid detection of somatic chromosomal copy number alterations (SCNAs), which result in allelic imbalance, i.e., deviations from the expected 1-to-1 allelic ratios at heterozygous loci. We initially applied a native version of our method to 1,970 tumor samples from 7 sites in The Cancer Genome Atlas (TCGA), inferring genotypes directly from RNA-sequencing (RNA-seq). This resulted in an SCNA detection rate of 68%. Encouraged by this, we next leveraged large public genetic reference data and array derived germline genotypes, from matched blood samples, to impute millions of germline variants for 4,942 patients across 28 TCGA cancer sites, resulting in improved genotype calling and haplotype inference. This latter approach increased our power for tumor SCNA

detection from RNA-seq to 85%, while maintaining a false positive rate of ~5%. SCNA burden inferred from RNA-seq was highly correlated (R = 0.92) with "gold standard" DNA derived estimates. To demonstrate the approach's potential clinical utility, we replicated SCNA features associated with clinical subtypes of breast cancer from RNA-seq successfully. Following this work, we investigated the role of the phenomenon of X-inactivation in female carcinogenesis through a comprehensive profiling of allelic imbalance observed in the X chromosome using tumor samples from the females in the TCGA breast cancer cohort. We observed higher rates of chromosome-level allelic imbalance for the X chromosome, both in comparison to the autosomes (derived from RNA) and the X chromosome derived from DNA, suggesting these are epigenetically driven by X-inactivation. Additionally, our results are in line with the findings from the literature that indicate loss of X chromosome inactivation's role in female carcinogenesis through association with more aggressive and basal-like subtype of breast cancer. Taken together, our results suggest a substantial improvement over existing methods, allowing for comprehensive studies of SCNA from RNA-seq and opening avenues for cost effective large-scale studies of tumors, as well as elucidating epigenetically driven mechanisms' contribution to carcinogenesis.

Table of Contents

List of Figures

List of Tables

Abbreviations

ACC: Adrenocortical carcinoma

aCGH: Array comparative genomic hybridization

BAF: B allele frequencies

BLCA: Bladder urothelial carcinoma

BRCA: Breast invasive carcinoma

CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma

CHOL: Cholangiocarcinoma

CNAs: Copy number alterations

cn-LOH: Copy neutral LOH

COAD: Colon adenocarcinoma

DGV: Database of Genomic Variants

ESCA: Esophageal carcinoma

FN: False negative

FP: False positive

FPR: False positive rate

GATK: Genome Analysis Toolkit

GBM: Glioblastoma multiforme

GDC: Genomic Data Commons

HLA: Human leukocyte antigen

HMM: Hidden Markov model

HNSC: Head and neck squamous cell carcinoma

HR: Hazard ratio

HRC: Haplotype Reference Consortium

ICGC: International Cancer Genome Consortium

KICH: Kidney chromophobe

KIRC: Kidney renal clear cell carcinoma

KIRP: Kidney renal papillary cell carcinoma

LGG: Brain lower grade glioma

LIHC: Liver hepatocellular carcinoma

LOH: Loss of heterozygosity

LUAD: Lung adenocarcinoma

LUSC: Lung squamous cell carcinoma

MCF: Mutant cell fraction

MESO: Mesothelioma

MIS: Michigan Imputation Server

Mb: Megabases

NCI: National Cancer Institute

NGS: Next generation sequencing

OV: Ovarian serous cystadenocarcinoma

PAAD: Pancreatic adenocarcinoma

PCPG: Pheochromocytoma and paraganglioma

PRAD: Prostate adenocarcinoma

READ: Rectum adenocarcinoma

RIN: RNA integrity number

RNA-seq: RNA-sequencing

SCNAs: Somatic chromosomal copy number alterations

scRNA-seq: Single cell RNA-seq

SKCM: Skin cutaneous melanoma

SNP: Single nucleotide polymorphism

STAD: Stomach adenocarcinoma

TARGET: Therapeutically Applicable Research to Generate Effective Treatments

TCGA: The Cancer Genome Atlas

TGCT: Testicular germ cell tumors

THCA: Thyroid carcinoma

TN: True negative

TP: True positive

TPR: True positive rate

TSG: Tumor suppressor genes

UCS: Uterine carcinosarcoma

UVM: Uveal melanoma

WES: Whole exome sequencing

XIC: X chromosome inactivation center

XIST: X-inactive specific transcript

CHAPTER 1

INTRODUCTION

1.1 Background

As the second leading cause of death in the U.S. after heart diseases, cancer constitutes a major public health issue. In 2022 alone, 1.9 million new cases of cancer are expected to be diagnosed and 609,360 people are expected to die of cancer [5]. Furthermore, it is anticipated that 1 in 3 women and 1 in 2 men will develop the disease in their lifetime. The top 5 leading cancer types for women are lung and bronchus, breast, colon and rectum, pancreas, and ovary cancer, whereas for men, the order is lung and bronchus, prostate, colon and rectum, pancreas, and liver and intrahepatic bile duct cancers [5].

Cancer is a genetic disease in which a group of cells, which originally derived from a single abnormal cell, gain a selective advantage, and grow and divide uncontrollably by disobeying the regular rules of cell division. Although DNA replication is strictly regulated to ensure the correctness of the process, mistakes happen at a rate of $\sim 10^{-9}$ per base pair per cell division for humans [6]. These altered DNA sequences are called mutations. In addition to the mistakes during DNA replication, external factors in the environment, such as tobacco consumption and exposure to radiation and certain chemicals, may lead to mutations as well.

Even though not all mutations are harmful, certain mutations, such as mutations in proto-oncogenes, which are the genes encoding proteins that function to stimulate cell cycle division, may cause overexpression and excessive cell proliferation. Similarly, mutations in tumor suppressor genes (TSG), which are the genes encoding proteins that regulate DNA repair, restrain proliferation, and induce apoptosis, may cause inactivation and lead cells to grow abnormally.

Usually, a series of mutations, involving the mutations in proto-oncogenes and tumor suppressor genes, are responsible from tumorigenesis [7]. Genomic instability, as one of the (evolving) hallmarks of cancer, may stem from the accumulation of such mutations which provide the genetic diversity essential for carcinogenesis. Gene mutations, chromosomal alterations, and genomic rearrangements are examples that occur in genomically unstable cells [8]. Characterizing these alterations may be helpful for shedding light on the molecular mechanisms of tumor initiation and progression.

Genomic instability most often results in allelic imbalance, which is defined as a deviation from the expected 1-to-1 ratio of inherited parental alleles, or haplotypes. Normally, a person inherits two alleles, one copy from each of their parents. Here, I will define an allele to be an allelic variant or dosage at a single base pair position in the consensus reference genome. At a germline heterozygous locus, these alleles can arbitrarily be labeled as allele 'A' and 'B'. Duplication of one of these alleles (AAB or ABB), causes allelic imbalance as the inherited parental haplotypes ratio become 2-to-1 (AAB) or 1-to-2 (ABB) and deviates from the normal 1-to-1 A:B ratio in at least some portion of the cells in a heterogeneous mixture. Similarly, deletion of one of the alleles (A leading to 1:0 or B leading to 0:1) causes a disturbance to the 1-to-1 A:B ratio and gives rise to loss of heterozygosity (LOH) and therefore allelic imbalance. A special type of LOH in which the normal copy number remains unchanged is copy neutral LOH (cn-LOH). Loss of one parental copy and duplication of the other causes cn-LOH (AA leading to 2:0 or BB leading to 0:2).

For individuals with hereditary cancer predisposition and who have a germline mutation in certain genes, e.g., TSGs *RB1* and *BRCA1*, LOH is associated with loss of the wild type allele (functional copy) as it leaves no TSG for protection. cn-LOH is thought to serve as the 'second hit' in the Knudson hypothesis. According to the Knudson hypothesis, for the majority of the

TSGs, a phenotypic change requires both alleles to be inactivated or 'hit' through a genetic or epigenetic modification [9].

Beyond the alterations in individual genes, aneuploidy, which results from loss or gain of chromosome arms and/ or whole chromosomes, may contribute to cancer initiation and progression as well. Studies have shown that nearly 90% of solid tumor exhibit aneuploidy, ranging between 26% and 99% depending on the tumor type [10]. For solid tumors, typically, 25% of the genome is affected by copy number changes through chromosome arm (or whole chromosome) alterations with a median of 5 losses and 3 gains of chromosome arms [11,12]. Furthermore, site specific aneuploidy patterns for different cancers have been observed in a plethora of studies [10,12–15], suggesting that carcinogenesis is driven by specific aneuploidies; for instance, common occurrence of gain of chromosome arm 12p in testicular germ cell tumors [16] and gain of chromosome 7 and loss of chromosome 10 across glioblastomas that do not carry IDH mutations [17].

Usually, to detect genomic regions that exhibit allelic imbalance, DNA-based methods, e.g., array comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) array, are used. The strategy behind these traditional allelic imbalance detection methods is to identify deviations of (within sample) allele frequencies from 0.5. When considering the genome wide distribution of B allele frequencies (BAF; a measure of the signal from B allele over the sum of A and B alleles) at germline heterozygous loci, for genomic regions that do not harbor an allelic imbalance event, one would expect to observe BAF distributed around 0.5. In contrast, in the presence of allelic imbalance events, one would expect to observe a band of separation indicating a deviation of the BAFs from 0.5 as one band would result from the elevated frequencies of the alleles of one haplotype and another distinct band from the decreased allele frequencies (Fig. 1.1).

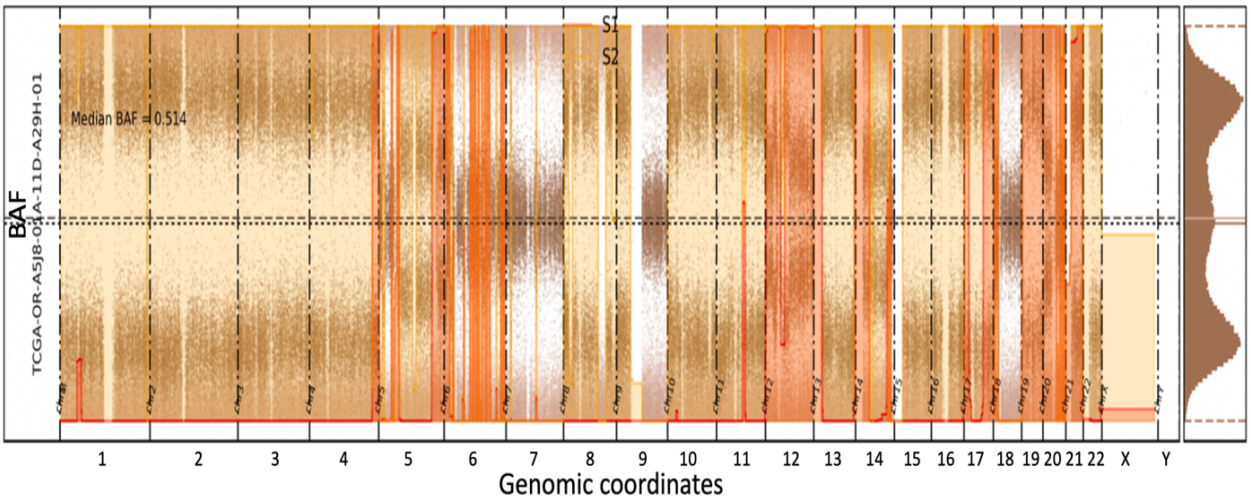


Fig. 1.1. Allelic imbalance detection from B allele frequencies (BAF). Visualization of hapLOH [1]output for tumor sample TCGA-OR-A5J8-01A-11D-A29H-01. B allele frequencies at heterozygous sites ranging from 0 to 1 are shown with points throughout the genome. The orange and yellow lines overlaid with the BAFs represent the posterior probabilities of allelic imbalance (orange and yellow lines for different levels of allelic imbalance). The band separation in the BAFs are visible where allelic imbalance is detected, such as chromosomes 1-4 as opposed to chromosome 7, which does not exhibit allelic imbalance.

The magnitude of deviation from 0.5 depends on the event type, e.g., gain, loss, or cn-LOH, and mutant cell fraction (MCF), i.e., the fraction of cells harboring the chromosomal alteration. If the MCF is high enough, e.g., 100% tumor sample, the band separation will be visually identifiable and there is no need for a special algorithm to determine allelic imbalance events. However, as MCF decreases, e.g., 20% tumor sample, the ability to detect the events visually disappears. For that reason, detection of subtle allelic imbalances poses a challenge to the traditional allelic imbalance identification methods. Most methods [18–21] for detecting tumor associated allelic imbalances have low power at MCFs below 10%-15%.

To address the issues regarding identification of subtle allelic imbalance events from SNP array data, Vattathil et al. developed a method named hapLOH [1]. Using the same principles of detection of the shifts from 0.5 in the BAFs at heterozygous loci, hapLOH expands this by taking haplotype information into consideration and searches for deviations of haplotype allele frequencies from 0.5. Estimating germline haplotypes and then searching for allelic imbalance

among the haplotypes makes the subtle events more distinguishable. hapLOH effectively identifies higher level imbalance, i.e., the two bands of BAF sufficiently diverged, such as in cn-LOH events, at MCFs as low as 4% and at MCFs 7% and higher, it successfully discerns the difference between genomic regions with lower, i.e., the two bands of BAF overlap, such as in such deletion events, and higher imbalances.

Following hapLOH, San Lucas et al. developed a method named hapLOHseq [22] for detection of genomic regions that exhibit subtle allelic imbalance from exome sequencing data utilizing the same principles underlying hapLOH. Exceeding the capabilities of the other methods, using data from 80x exome sequencing, hapLOHseq successfully identified allelic imbalance events as small as 10 megabases (Mb) at MCFs 16% and higher. Using 30x whole genome sequencing data, hapLOHseq effectively detected events at MCFs 4% and higher.

RNA-seq data from tumor samples, is often analyzed for quantifying expression analysis, detection of alternatively spliced genes, and identification of novel transcripts as a complementary approach to provide further information on the disease. Although joint analysis of RNA and DNA may often be ideal for copy number detection, it may not always be feasible in all settings. For example, in settings where the amount of tumor sample is limited or financial constraints do not permit additional analysis on a separate platform, the data from RNA may be the only data source available. Under such circumstances, the ability to infer allelic imbalance from RNA is invaluable. Despite the increasing efforts to extend RNA-seq analysis pipelines to provide copy number information, it still has not been fully established. This is mainly due to the challenges specific to analysis of RNA as factors beyond the underlying copy number alterations (CNAs) could affect mRNA quantities, such as the various epigenetic mechanisms that regulate gene expression and non-uniform coverage of the genome.

To summarize, the utility of the data from bulk RNA-seq experiments can be greatly increased by extending the data analysis to include detection of somatic CNAs (SCNAs) and yielding to a more comprehensive tumor characterization in certain settings in which additional analysis on DNA is not feasible either due to the restricted availability of tumor sample and/ or funding related constraints. Furthermore, there is a large number of grossly underutilized data repositories, e.g., The National Cancer Institute (NCI) Genomic Data Commons (GDC) [23], with bulk RNA-seq data across different cancer research programs, including The Cancer Genome Atlas Project (TCGA) [24], Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [25], and International Cancer Genome Consortium (ICGC) [26]. Data from the well phenotyped cohorts of these repositories could be reused for further analysis to increase the utility of these datasets and especially benefit rare cancers that suffer from scarcity of data. Lastly, the well characterized samples would also assist investigation of associations between SCNAs and clinical features.

1.2 Dissertation outline

This dissertation project focuses on the implementation, tuning, testing, and application of hapLOHseq to bulk RNA-seq data in cancer settings. hapLOHseq is the next generation sequencing (NGS) implementation of hapLOH, which is a sensitive statistical software developed by the Scheet lab that utilizes germline haplotypes to identify genomic regions that exhibit allelic imbalance from SNP array data. Throughout the dissertation, RNA based allelic imbalance calls detected with hapLOHseq are compared with DNA based calls identified with hapLOH, which serves as a gold standard. In the remainder of the dissertation, the following organization is going to be used:

In chapter 2, I describe hapLOHseq's application to nearly 5,000 tumor samples across 28 tumor sites in the TCGA, beginning with a discussion on the importance of allelic imbalance as a signature of genomic instability, the motivation for implementing hapLOHseq to RNA-seq data, and the two possible options for obtaining the germline haplotypes from: (1) tumor RNA-seq and (2) imputed genotypes derived from DNA SNP array data from a matched-normal sample such as blood. The subsequent sections of the chapter demonstrate the results from benchmarking of the approach against two other methods available for allelic imbalance detection from bulk RNA-seq, the performance of the approach at the gene, chromosome arm, and genome levels, and the method's potential clinical use with several applications on the TCGA breast cancer cohort.

In chapter 3, I provide guidelines to ensure the user achieves the highest quality (best possible given the data) results and describe the reason for each. The chapter contains several quality control measures for both before and after applying hapLOHseq software to RNA-seq data, including pre-processing of the RNA-seq data to check whether the data meets certain requirements to make high quality allelic imbalance calls and post-processing of the calls to clean up and improve the precision.

In chapter 4, I provide the details of an analysis of hapLOHseq's application to the X chromosome from 600 plus females across the TCGA breast cancer cohort. The chapter aims to characterize signatures of X-inactivation and to investigate the role of X-inactivation, or the lack thereof, in female carcinogenesis, and explores the relationship between several clinical features, such as overall survival, tumor subtype, and tumor biological age, and the status of having an X-inactivation induced imbalance exclusive to RNA, i.e., while there is no allelic imbalance in DNA, epigenetic mechanisms utilized for X-inactivation induces allelic imbalance in RNA.

Finally, in chapter 5, I discuss the limitations of the method in several scenarios and mention potential improvements and extensions as future directions.

CHAPTER 2

CHROMOSOMAL IMBALANCES DETECTED VIA RNA-SEQUENCING IN 28 CANCERS

The contents of this chapter are based on the following publication [27], reprinted with permission, from the journal Bioinformatics:

Z. Ozcan, F.A. San Lucas, J.W. Wong, K. Chang, K.H. Stopsack, J. Fowler, Y.A. Jakubek, P. Scheet, Chromosomal imbalances detected via RNA-sequencing in 28 cancers, Bioinformatics. (2022). https://doi.org/10.1093/bioinformatics/btab861.

2.1 Introduction

Cancer arises as a result of a gradual acquisition of molecular alterations [28]. Genomic instability, a hallmark of cancer [8], leads to DNA alterations, such as somatic copy number alterations (SCNAs), which may span large genomic regions or entire chromosome arms. They can play a key role in the path to tumorigenesis by leading to loss of tumor suppressor genes and/or generating additional copies of oncogenes [9]. SCNAs have been associated with clinical features or outcomes and serve as prognostic indicators [10,29–35]. Hence, detection and genome-wide characterization of SCNAs is a key component for genomic studies of tumor initiation and progression, and of SCNA-associated clinical features and outcomes.

Typically, SCNAs are almost exclusively inferred directly from DNA, measured by technologies such as array comparative genomic hybridization, single nucleotide polymorphism (SNP) DNA microarray or NGS [36–40]. Investigations of RNA, either by microarray or NGS [RNA-sequencing (RNA-seq)], often complement DNA analyses through quantification of gene expression, and identification of novel transcripts and gene fusions [41,42] or point mutations [43–47] to further our understanding of disease. Yet, in many settings, particularly where tumor material or funding is limited, data exist from RNA-seq only. However, the extension of RNA-seq data into SCNA calling has not been as well developed. Inferring SCNAs from RNA-seq data is inherently difficult, since both regulation of expression and underlying DNA copy number will alter the observable quantities of mRNA. In addition, due to the non-uniform coverage of the genome from RNA-seq, it is challenging to differentiate between dynamically varying gene expression and SCNAs.

Recently the relative void of methods to detect SCNAs from RNA has been partially addressed. Most of these methods are exclusively tailored to single cell RNA-seq (scRNA-seq), such as HoneyBADGER [48], CopyKAT [49] and inferCNV [50], while some can be applied to bulk RNA, such as CaSpER [4] and SuperFreq [51]. CaSpER integrates genome-wide total gene expression and allelic signals to detect and visualize SCNAs; SuperFreq also uses both read counts and BAF dispersions for SCNA inference, requiring referent samples to be available for normalization. Another approach for detection of SCNAs from bulk RNA profiling integrated coverage data and tumor-specific SCNA frequency patterns from public, external, data to identify chromosome-arm level aneuploidy, which was in turn assessed for association with prostate cancer outcomes [52]. Yet, these methods do not utilize haplotype information (the genetic makeup of a single chromosome that is passed on from a parent), which has been shown to increase power for SCNA detection in studies with SNP microarray data [1,3,53,54].

We sought to facilitate inference of SCNAs from RNA by applying an approach that utilizes haplotypes for SCNA detection from bulk RNA-seq, opening avenues for joint analysis of aneuploidy and expression from population-scale data. Consideration of haplotype structure implicitly models the signal at multiple genomic loci (or SNP markers) *jointly*, which not only offers an opportunity for increased power, but also requires the patterns to sustain beyond individual transcripts, which may be modulated by factors beyond SCNAs. Our approach enables robust detection of megabase-scale SCNAs that represent gain, loss or copy neutral loss of heterozygosity (cn-LOH) events. The strength of our approach derives from modeling the allelic imbalance at genomic regions affected by SCNAs. Allelic imbalance refers to a deviation from the expected 1:1 ratio of 'A' and 'B' alleles at germline heterozygous (genotype 'AB') loci. Alterations such as deletion (genotype: A- or B-, ratio: 1:0 or 0:1), duplication (genotype: AAB or ABB, ratio: 2:1 or 1:2) and cn-LOH (AA or BB, ratio: 2:0 or 0:2) are representative examples of AI.

In this chapter, I describe our method for detecting large scale SCNAs from tumors using bulk RNA-seq data. I start by explaining the details of the method, including the two possible approaches for obtaining germline genotype calls, which is an essential step for the method. I continue by explaining the possible underlying reasons of the performance improvement achieved with genotype imputation and follow by the performance assessment results evaluated at the gene, chromosome arm, and genome levels by comparing the RNA-seq derived SCNAs with SNP array derived gold standard call set. Furthermore, I also show the method's potential clinical efficacy with examples from the TCGA BRCA cohort and report benchmarking results of the method against two other methods available for detecting SCNAs from bulk RNA-seq.

2.2 Study design

In this study, we demonstrate effective somatic chromosomal copy number alteration identification from RNA-seq, comparing results to those derived from a high-density SNP DNA microarray as a benchmark and so-called 'gold standard' for SCNA detection. We consider scenarios where data are available from RNA-seq only, as well as a complementary scenario where germline DNA data is available from another source such as routinely collected blood. We apply several novel techniques including using RNA-seq for inference of acquired allelic imbalance and the incorporation of genotypes via an imputation step using publicly available large-scale genotype reference data, which improves our performance considerably by enhancing the quality of estimated genotypes and haplotypes. Our results demonstrate that comprehensive and robust inference of megabase-scale SCNAs is possible from bulk RNA-seq.

2.3 Materials and methods

2.3.1 Dataset

RNA-seq BAM files aligned against the human genome build hg38 (GRCh38) and the level 1 raw CEL files from Affymetrix Genome-Wide Human SNP Array 6.0 profiling of 4942 (primary solid) tumor samples across 28 cancer sites in The Cancer Genome Atlas (TCGA) were obtained from the Genomic Data Commons data portal along with BRCA clinical information. The level 1 raw CEL files of the matched-normal (blood) samples across these sites were also downloaded to perform genotype imputation. In addition, for a subset of 7 cancer sites (BRCA, COAD, GBM, LUAD, LUSC, PAAD and PRAD), WES BAM files of 888 (primary solid) tumor samples aligned against the GRCh38 were obtained for comparisons.

2.3.2 Processing of the tumor RNA-seq array data

Our method for the detection of SCNAs relies on the allele-specific signals at germline heterozygous sites. For the purpose of deriving germline genotypes, the sample can come from the tumor itself or from a matched-normal. We explored the utility of using two different sources of data for obtaining germline genotype calls: (i) tumor RNA-seq and (ii) imputed genotypes derived from SNP array data from a matched-normal, specifically blood for the samples to which we had access.

2.3.2.1 Genotyping and phasing

2.3.2.1.1 Approach 1: Genotypes from tumor RNA-seq

For this approach, using tumor RNA-seq, the genotypes were called at sites already known to be polymorphic from large-scale surveys of genetic variation. The Haplotype Reference Consortium (HRC; for individuals of European ancestry) was used as a reference and genotypes were called at these reference sites from the RNA data with the UnifiedGenotyper from Genome Analysis Toolkit [55] (GATK; version 3.6). Subsequently, the genotypes were phased using the MaCH software [56] to reconstruct haplotypes using the set of individual-level genotypes as an internal reference. Singleton SNPs — heterozygous markers that were observed only in one sample at a particular SNP locus within a cancer site — were removed.

2.3.2.1.2 Approach 2: Genotypes imputed

The accuracy of haplotype reconstruction increases with larger reference/internal sample size. Therefore, haplotype reconstruction accuracy is limited particularly for smaller cancer sites when using an internal reference as done in approach 1. For approach 2, we leveraged the available blood genotype data from SNP DNA microarrays, representing genotypes from the matched-normal samples of the TCGA resource (blood). After calling genotypes using the SNP array data of the matched-normal samples with the Birdsuite software [57], the genotypes were prepared for

imputation. To assure the quality of the genotypes submitted for imputation, several quality control steps were performed. To filter out low-quality SNPs, we removed the SNPs that failed Hardy–Weinberg equilibrium test (p-value < 1×10^{-6}), those with missing rate > 5% and excluded monomorphic sites. In addition, samples with greater than 5% missing genotype rate were removed from downstream analyses. Individuals of European ancestry were identified using principal component analysis [EIGENSTRAT [58]] using the genotyped SNPs at 1KG sites with. The cleaned, unphased genotypes from individuals of European ancestry were submitted to the Michigan Imputation Server [59] (MIS), using the hg19 (GRCh37) genome build, the HRC panel (Version r1.1 2016) as the reference, 0.1 R^2 cutoff - which is a metric used to quantify imputation quality - and EUR for population. Consequently, the imputed genotypes and estimated haplotypes for 4942 TCGA samples of European ancestry from 28 cancer sites were downloaded from the MIS and markers with an $R^2 < 0.3$ were removed to remove poorly imputed markers.

In both approaches, the centromeric, human leukocyte antigen (HLA), VDJ and Database of Genomic Variants (DGV) regions were masked for exclusion of putative germline copy number changes.

2.3.2.2 Detection of SCNAs

As noted earlier, we investigate two approaches that differ on how the germline haplotypes are obtained. While the first approach uses tumor RNA to statistically estimate haplotypes, in approach 2, the phased germline genotypes are obtained through genotype imputation performed on a matched-normal from SNP array. The hapLOHseq [22] software identifies the haplotype in excess and using the germline haplotypes, quantifies phase concordance via switch accuracy to determine genomic regions that harbor SCNAs. The hapLOHseq software (version 0.1.2) was applied using the default parameters in both approaches, except –end_param_event (=0.9) and –event_prevalence (=0.05). SCNA calls with <10 markers or those smaller than 2 Mb were

excluded. Furthermore, SCNA calls that contain large (>10 Mb) genomic regions without any heterozygous sites were split up into multiple regions.

2.3.3 Processing of the tumor DNA WES data

To identify SCNAs from WES tumor data, first, the genotypes were called at the HRC sites with UnifiedGenotyper. Second, the genotypes were phased with MaCH software using the set of individual-level genotypes as an internal reference. Third, the hapLOHseq software was used with the default parameters to detect the SCNAs after masking the centromeric, HLA, VDJ and DGV regions and removing singleton SNPs. SCNAs with <10 markers or those smaller than 2 Mb were excluded and the calls that contain large (>10 Mb) regions without any heterozygous sites were split up into multiple regions.

2.3.4 Processing of the tumor SNP array data

To detect SCNAs from SNP array tumor data, first, the genotypes were called using the Birdsuite software, second, the genotypes were phased using MaCH software, and third, the hapLOH software was used with the default parameters to identify regions that harbor SCNAs. Prior to the third step, the markers were mapped from genome build hg19 to hg38 and the centromeric, HLA, VDJ and DGV regions were masked. The SCNA calls detected from the SNP array constitute a gold standard for assessing the performance.

To ensure the consistency in the way the samples were processed, SyQADA [60] was used to automate the pipelines across the three platforms.

2.3.5 Performance assessment

We sought to assess the performance of our method for detection of SCNAs in tumors. To do so, we compared our set of SCNA calls to a gold standard set of SCNA calls from matched-tumor DNA samples processed using arrays, a gold standard set of calls. We contrasted the SCNA call sets at gene-, chromosome arm- and genome levels. At the gene level, we report sensitivity and

specificity. Sensitivity represents the method's power to detect true SCNAs that are identified by the gold standard, while specificity represents the method's ability to correctly identify genes that do not fall within an SCNA region in the gold standard. Therefore, sensitivity (TPR) was calculated as TP / (TP + FN) and specificity (1 - FPR) was calculated as TN / (TN + FP) where TP is true positive, FN is false negative, TN is true negative, and FP is false positive (see Appendix Fig. A1 for details). For each sample, sensitivity and specificity were calculated individually, then median sensitivity and specificity for samples in each TCGA cohort were reported as cohort-level summary statistics. When assessing the method's performance at the chromosome arm level, for each sample, we assessed presence or absence of a chromosome arm-level event, defining an arm-level event as present when at least 50% of the chromosome arm is affected by SCNAs. At the genome level, we calculated genomic burden for each sample, which reflects the percentage of a sample's genome that exhibits SCNAs. For each cancer site independently, we calculated each sample's genomic burden based on RNA-seq-derived SCNA calls and compared with the gold standard derived genomic burden.

2.3.6 Comparison to other methods

We followed the recommended workflow for SuperFreq of first applying VarScan2 [61] for variant identification, followed by SuperFreq itself. RNA-seq from two adjacent-to-tumor breast samples in the BRCA resource were supplied to SuperFreq for normalization. Results from the *TP53* analysis of COAD samples were taken directly from their curated data in their GitLab page. Results from CaSpER were obtained directly from what they had curated previously, using their stored and available R data frames. We then summarized results by gene through direct tabulation or applying BEDTools intersect.

2.4 Results

To detect somatic (acquired) copy number alterations (SCNAs) using RNA, we applied a haplotype-based approach. In brief, hapLOHseq detects regions of the genome where the signal at heterozygous sites reflects one of the estimated haplotypes for that individual. A deviation from the expected 1:1 ratio of maternal to paternal DNA indicates a relative over-representation of one of the parental chromosomes, signaling the presence of an SCNA. This approach has been applied successfully to bulk DNA analyses of various tissues [54,62,63].

Here, we explore the potential of this method for detection of large-scale SCNAs from NGS data of bulk RNA (RNA-seq). To do so, we obtained RNA-seq from seven large cancer sets in TCGA: Breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD) and prostate adenocarcinoma (PRAD). To identify genomic regions that harbor SCNAs, we applied hapLOHseq. To assess the accuracy and potential of this approach, we compare SCNAs inferred from RNA-seq to high-confidence SCNA calls detected from DNA SNP microarray data, which have been documented previously [3] with an estimated false-positive rate < 3% [63]. For these purposes, since the RNA and DNA were derived from the same tissue (or tumor), we treat the SNP microarray results as a gold standard. For a subset of samples, we also applied hapLOHseq to whole exome sequencing of DNA (WES) to help interpret the results and assess where deficiencies may be attributed to technology, bioinformatic approaches or inherent limitations of inference from specific nucleic acids.

2.4.1 SCNA detection from RNA-seq

We compared and contrasted the RNA-seq derived SCNAs with the gold standard SCNAs with units of analysis as gene, chromosome arm and entire genome (i.e., burden), focusing here on

gene-level summaries. Table 2.1 [first row: 'hapLOHseq (RNA-seq)'] shows the sensitivity and specificity for SCNA detection solely using tumor RNA for seven cancer sites. Sensitivity measures the method's power to detect true SCNAs (defined as those identified by the gold standard) and specificity measures the method's ability to correctly identify genomic regions where there is no SCNA detected by the gold standard. We obtained a generally high concordance between the events called from RNA-seq with those called from SNP arrays at a gene level, and this held for the other units of analysis as well (Figs. 2.1 and 2.2). Our SCNA detection rate (sensitivity) from RNA-seq was highest for the GBM data at 79%, whereas the sensitivity for PRAD was markedly lower at 45%. We observed high specificities for all seven cancer sites, ranging between 89% and 99%.

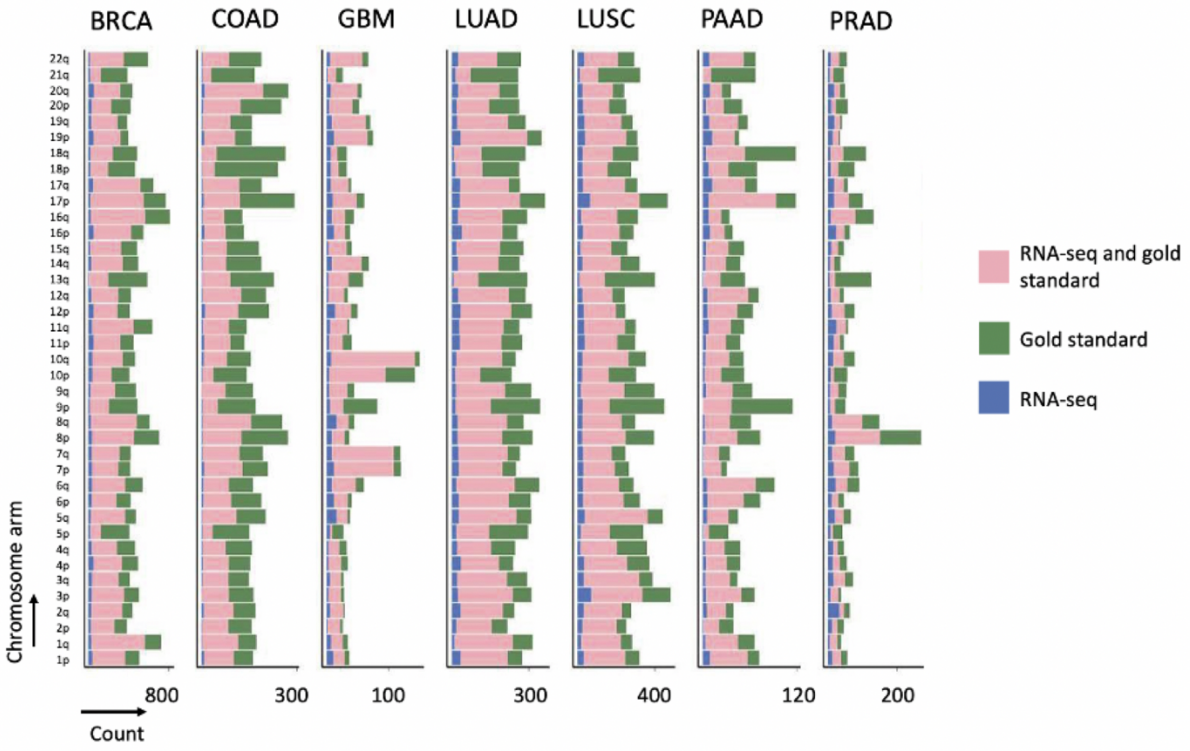


Fig. 2.1. Chromosome arm level concordance assessment summaries across 7 TCGA sites. We identified chromosome arms that were spanned at least 50% by SCNAs. The distribution of the non-acrocentric autosomal chromosome arms (n=39) across the cancer sites are shown. For each site, a stacked bar plot of the number of samples with concordance specific chromosome arm level SCNAs are shown for all 39 chromosome arms.

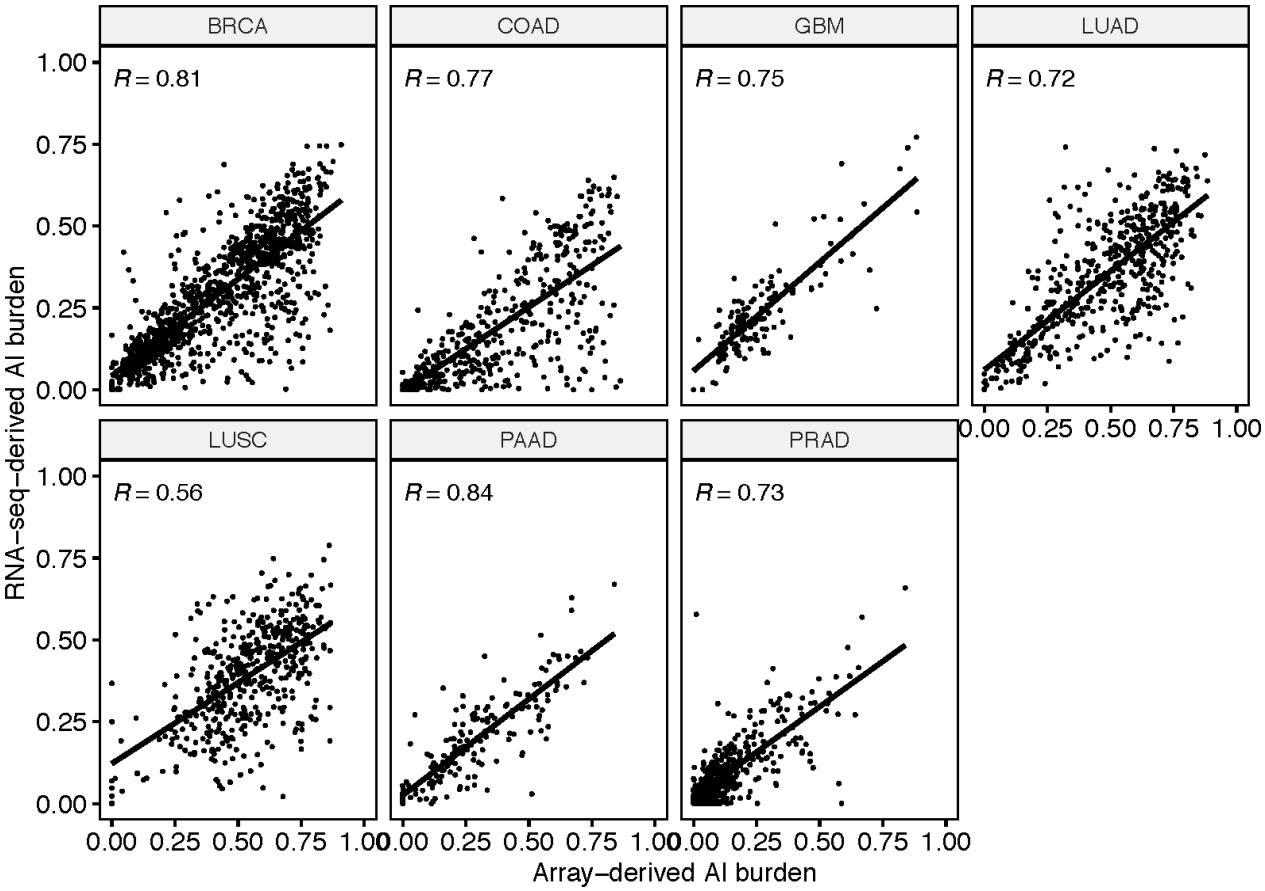


Fig. 2.2. Concordance between RNA-seq and SNP array (gold standard) derived genomic burden. We identified regions that exhibit SCNA for each sample. Scatterplots representing the concordance are shown for each site grouped by cancer site.

	BRCA		COAD		GBM		LUAD		LUSC		PAAD		PRAD	
	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec	Sens	Spec
hapLOHseq (RNA-seq)	71%	94%	47%	99%	79%	93%	74%	91%	73%	89%	69%	97%	45%	97%
hapLOHseq (RNA-seq + imputed genotypes)	84%	92%	79%	97%	89%	92%	84%	90%	88%	88%	80%	95%	66%	94%
hapLOHseq (WES)	93%	89%	94%	92%	94%	96%	89%	92%	91%	90%	94%	89%	76%	97%

Table 2.1. Gene level performance assessment. We evaluated the method at the gene level by comparing SCNA status of genes between the RNA-seq-derived SCNAs and the gold standard (array-based analysis) for seven cohorts in the TCGA. Sens = Sensitivity; the proportion of genes covered by an SCNA in the gold standard that were also identified by the listed approach, Spec = Specificity; the proportion of genes that are not covered by an SCNA event in the gold standard that were also not inferred to be covered by an SCNA by the listed approach.

To assess which factors drive the lower than average performance in the PRAD cohort, we further investigated the samples with the highest sensitivities. We observed that the samples with the highest quartile of sensitivity had the highest number of heterozygous sites and the highest haplotype accuracies (Fig. 2.3). After grouping the samples into quartiles based on their sensitivity, we statistically compared the groups with the Kruskal–Wallis test and noted that the groups were significantly different when compared both by the number of heterozygous sites (p-value < 1e-8) and phase accuracy (p-value = 0.008).

To help understand potential limitations of our approach, we compared our results to those obtained from the application of hapLOHseq to TCGA DNA WES [Table 2.1; third row: 'hapLOHseq (WES)']. Not surprisingly, since it assays the DNA directly, WES consistently achieved higher sensitivities at similar specificities across sites, including PRAD with 76% sensitivity in comparison to 45% from RNA-seq. Although we observed a lower performance for RNA-seq than for WES, our results demonstrate that there exists sufficient information in the RNA-seq for generally accurate inference of SCNAs. Encouraged by this, we explored further the power of RNA-derived SCNA profiling approaches.

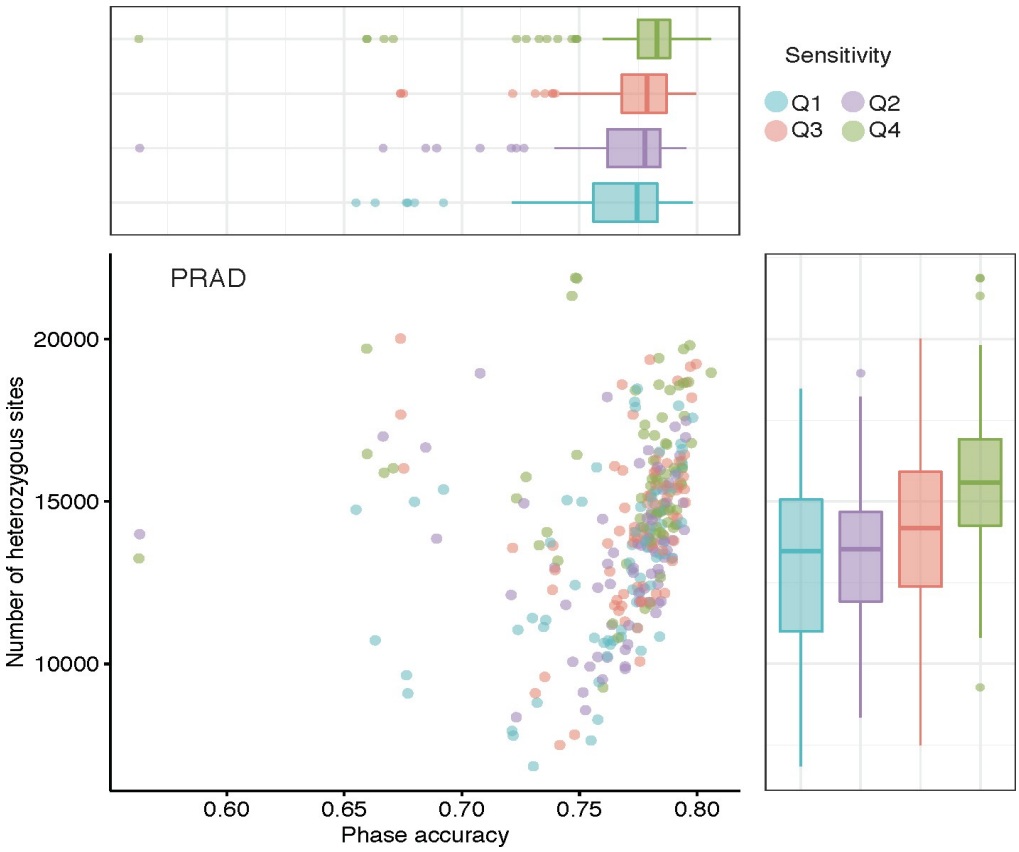


Fig. 2.3. Factors driving low performance in the TCGA PRAD cohort. A scatterplot showing the distribution of sample sensitivities varying based on phase accuracy and number of heterozygous sites. Samples were grouped into quartiles according to their sensitivities.

2.4.2 SCNA detection from RNA-seq and imputation-based haplotype inference

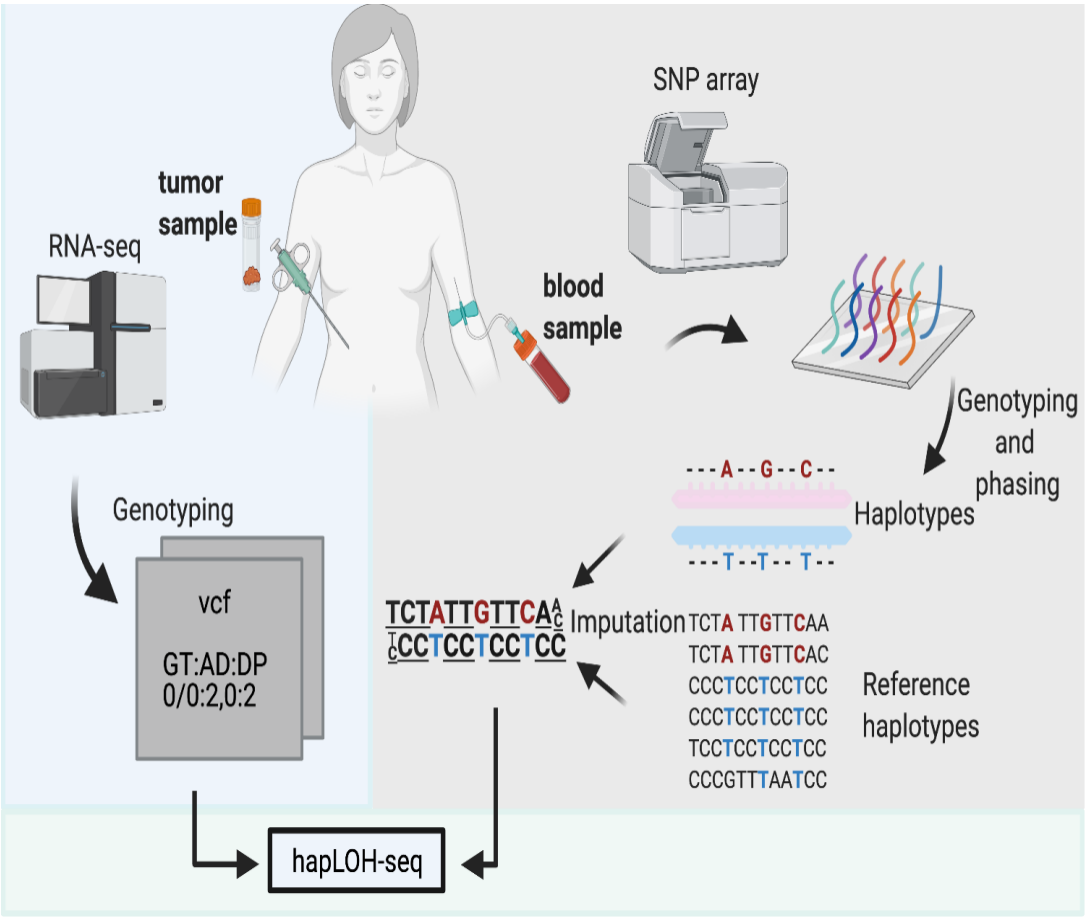


Fig. 2.4. Graphical illustration of the imputation approach in our pipeline.

We hypothesized that the standard genotyping (variant calling) pipelines for NGS, along with modest reference sizes for haplotype phasing, were holding back the potential of our approach. To address this, we imputed germline genotypes and haplotypes from large-scale reference data using the optimized workflows in the Michigan Imputation Server (MIS). We leveraged the genotype data from a matched-blood sample, available for most participants in TCGA (Fig. 2.4). While this sample does not provide any direct information about SCNAs of the tumor, it does provide more accurate identification of heterozygous sites and estimated haplotypes, central to our approach, but without the need to extract DNA from the tumor.

We applied hapLOHseq for detection of tumor SCNAs using RNA-seq signal combined with the more accurate haplotype information, as detailed in the methods. With genotype imputation, in addition to obtaining more accurate haplotype estimation, an increase in the number of heterozygous sites per sample was observed (Fig. 2.5). Furthermore, we also noticed that the new heterozygous sites filled in some segments of the genome that did not harbor any heterozygous sites before (Fig. 2.6). Excitingly for several samples, we even noted that the imputation approach enabled identification of whole chromosome arm level SCNA events that could not be detected before (Fig. 2.7).

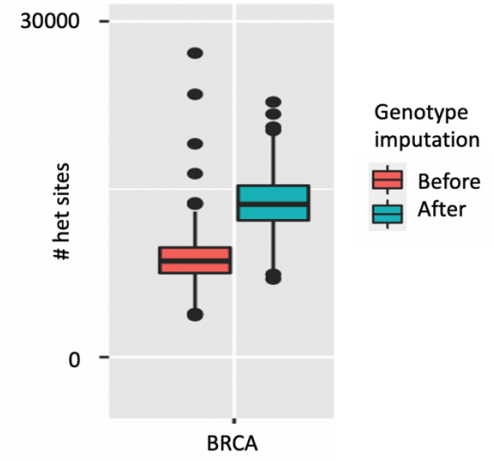


Fig. 2.5. Increase in the number of heterozygous sites. A boxplot representing the improvement of heterozygous sites per sample in the TCGA BRCA cohort with genotype imputation.

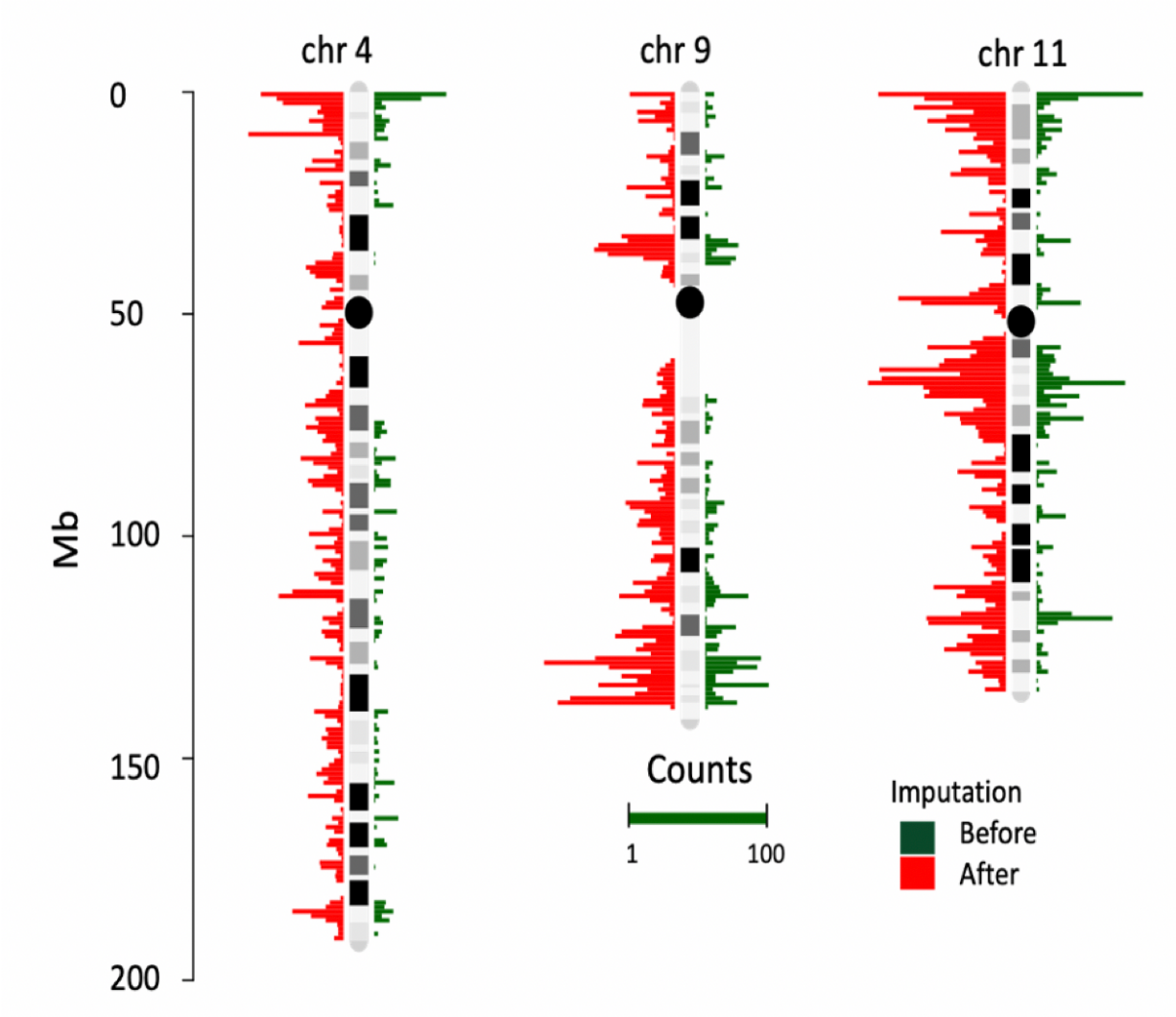


Fig. 2.6. Histogram of the number of heterozygous sites before and after genotype imputation. The distribution of the number of heterozygous sites for BRCA sample TCGA-D8-A1JA focusing on chromosomes 4, 7, and 11, where the most significant improvement has been observed.

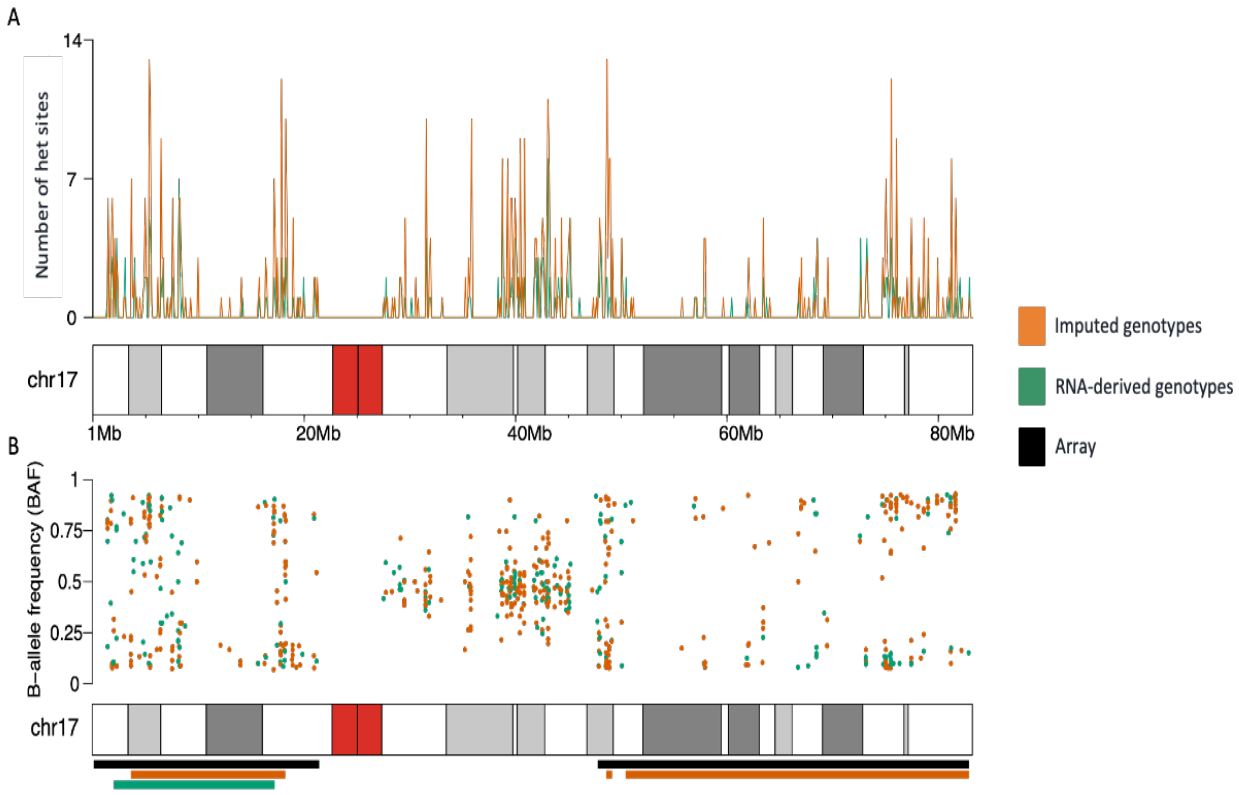


Fig. 2.7. Improvement in SCNA event calling achieved with genotype imputation. Genotype imputation enables detection of a previously missed chromosome arm level event. Distribution of the number of heterozygous sites (top panel) and B allele frequencies with bars underneath representing the SCNAs detected with and without genotype imputation along with the SNP array (gold standard) for chromosome 17 of sample TCGA-A6-6648 (bottom panel).

A systematic evaluation demonstrated that the inclusion of imputed genotype calls and high-quality haplotypes provided a substantial improvement in overall SCNA detection [Table 2.1; second row: 'hapLOHseq' (RNA-seq + imputed genotypes)]. After the imputation approach, sensitivities improved across all cancer sites, ranging from 10% to 32% in absolute increase, i.e., BRCA (13%), COAD (32%), GBM (10%), LUAD (10%), LUSC (15%), PAAD (11%) and PRAD (21%), while high specificities remained similar.

In comparison to WES, the imputation approach has lower sensitivity with similar specificity. For instance, for the BRCA cohort, the imputation approach's sensitivity is 9% lower while specificity is 3% higher. These results indicated the potential for inference of SCNAs purely from RNA-seq, so long as there exist sufficiently informative germline genotypes. To

comprehensively characterize the potential for SCNA inference from RNA-seq across a range of tissue types, we applied this approach to the remaining cancer sites from the TCGA for which data existed. Study abbreviations for all 28 TCGA cohorts investigated in this study are shown in Table 2.2.

Tumor site (abbreviation) (sample size)	Sensitivity	Specificity
Adrenocortical carcinoma (ACC) (n=58)	89%	93%
Bladder urothelial carcinoma (BLCA) (n=261)	83%	90%
Breast invasive carcinoma (BRCA) (n=641)	84%	92%
Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) (n=150)	88%	94%
Cholangiocarcinoma (CHOL) (n=26)	85%	95%
Colon adenocarcinoma (COAD) (n=256)	79%	97%
Esophageal carcinoma (ESCA) (n=54)	91%	79%
Glioblastoma multiforme (GBM) (n=99)	89%	92%
Head and neck squamous cell carcinoma (HNSC) (n=346)	86%	94%
Kidney chromophobe (KICH) (n=6)	94%	95%
Kidney renal clear cell carcinoma (KIRC) (n=56)	90%	94%
Kidney renal papillary cell carcinoma (KIRP) (n=143)	92%	95%
Brain lower grade glioma (LGG) (n=388)	79%	96%
Liver hepatocellular carcinoma (LIHC) (n=115)	82%	95%
Lung adenocarcinoma (LUAD) (n=312)	84%	90%
Lung squamous cell carcinoma (LUSC) (n=221)	88%	88%
Mesothelioma (MESO) (n=76)	87%	95%
Ovarian serous cystadenocarcinoma (OV) (n=249)	87%	87%
Pancreatic adenocarcinoma (PAAD) (n=124)	80%	95%
Pheochromocytoma and paraganglioma (PCPG) (n=133)	89%	96%
Prostate adenocarcinoma (PRAD) (n=317)	66%	94%
Rectum adenocarcinoma (READ) (n=133)	80%	96%
Skin cutaneous melanoma (SKCM) (n=87)	86%	94%
Stomach adenocarcinoma (STAD) (n=190)	86%	87%
Testicular germ cell tumors (TGCT) (n=116)	88%	93%
Thyroid carcinoma (THCA) (n=276)	85%	94%
Uterine carcinosarcoma (UCS) (n=34)	85%	90%
Uveal melanoma (UVM) (n=75)	87%	98%

Table 2.2. Gene level performance summaries across 28 cancer sites. For each cancer site, the study abbreviation, number of samples analyzed in the cohort, and median gene level sensitivity and specificity are shown.

Across all cancer sites, the median SCNA size was 31.28 Mb, approximately equal to the median SCNA size (31.31 Mb) detected by the gold standard. The median number of SCNA events detected per sample was 20 (gold standard: 19). The highest frequency of SCNA calls per sample were observed in esophageal carcinoma (ESCA), ovarian serous cystadenocarcinoma (OV), LUSC, uterine carcinosarcoma (UCS) and bladder urothelial carcinoma (BLCA) with each with a median of 30 SCNAs or more per sample, consistent with the gold standard (Table 2.3). In contrast, thyroid carcinoma (THCA) and uveal melanoma (UVM) had the fewest number of SCNAs per sample with each site having <10 SCNAs per sample. These two sites were also ranked as having the lowest median number of SCNAs by the gold standard.

Tumor site	RNA-seq	Array
ACC	26	20
BLCA	30	30
BRCA	25	23
CESC	20	16
CHOL	23	19
COAD	18	19
ESCA	38	32
GBM	16	13
HNSC	24	22
KICH	17	6
KIRC	11	6
KIRP	12	6
LGG	10	8
LIHC	16	15
LUAD	29	29
LUSC	33	32
MESO	19	15
OV	36	41
PAAD	16	15
PRAD	10	11
SKCM	20	16
STAD	29	28
THCA	6	0
UCS	30	35
PCPG	10	6
READ	23	24
TGCT	28	24
UVM	9	5

Table 2.3. Number of SCNAs. Number of SCNAs per sample across 28 TCGA cohorts (median for each cohort).

Table 2.2 contains per-cancer site gene-level summaries of sensitivity and specificity for 28 cancer sites, comprising 4942 samples in the TCGA, after applying the imputation workflow. At the gene level, our imputation-based approach achieved an 85% median sensitivity and 94% median specificity (all genes, all samples). Across the sites, median sensitivities ranged from 66% to 94%, with median specificities between 79% and 98%. With the exception of ESCA, specificity was always greater than or equal to the sensitivity for a given site. At 79%, 79% and 66%, COAD, brain lower grade glioma (LGG) and PRAD (respectively) were the only cancer sites with sensitivity below 80%. Interestingly, kidney cancers (kidney chromophobe: KICH; kidney renal clear cell carcinoma: KIRC; and kidney renal papillary cell carcinoma: KIRP) were the three cohorts that we observed the best performances for with 94%, 90% and 92% sensitivity and 95%, 94% and 95% specificity.

We evaluated the method at the chromosome arm level as well (Fig. 2.8). For each chromosome arm, we assessed the concordance with the gold standard calls and the results indicate that the majority of the true arm-level SCNAs were inferred correctly across all cancer sites. However, several, such as 5p, 9p, 13q and 21q were missed with RNA consistently across the cohorts. We also note that the majority of the chromosome arm-level SCNAs inferred from RNA-seq in the THCA cohort were not present in the gold standard set.

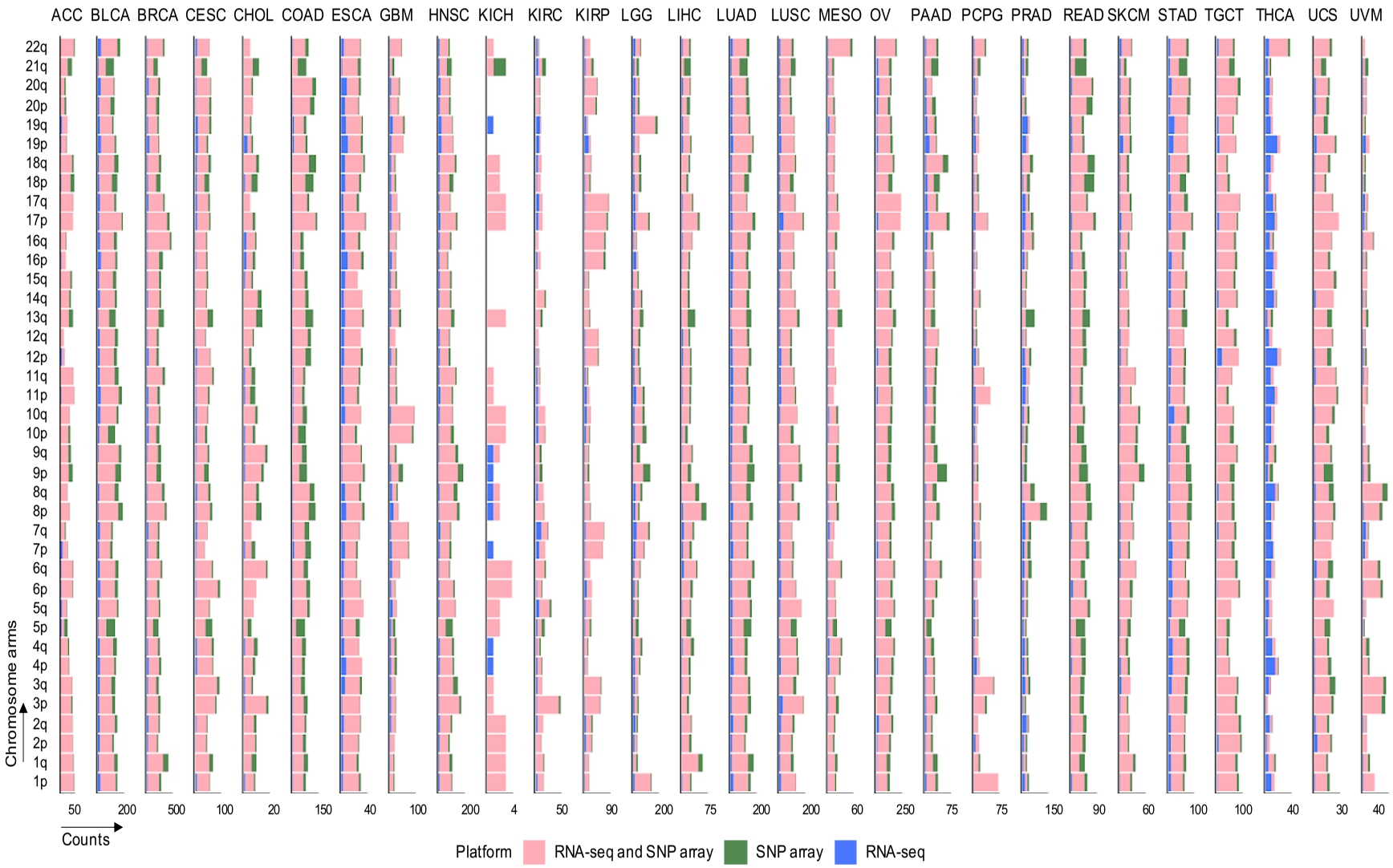


Fig. 2.8. Chromosome arm level concordance assessment summaries across 28 cancer sites. We identified chromosome arms that were spanned by SCNAs (≥50%) and for each arm we evaluated the concordance between RNA-seq and gold standard. The distribution of the non-acrocentric autosomal chromosome arms (n=39) across the cancer sites are shown. For each site, a stacked bar plot of the number of samples with concordance specific chromosome arm level SCNAs are shown for all 39 chromosome arms.

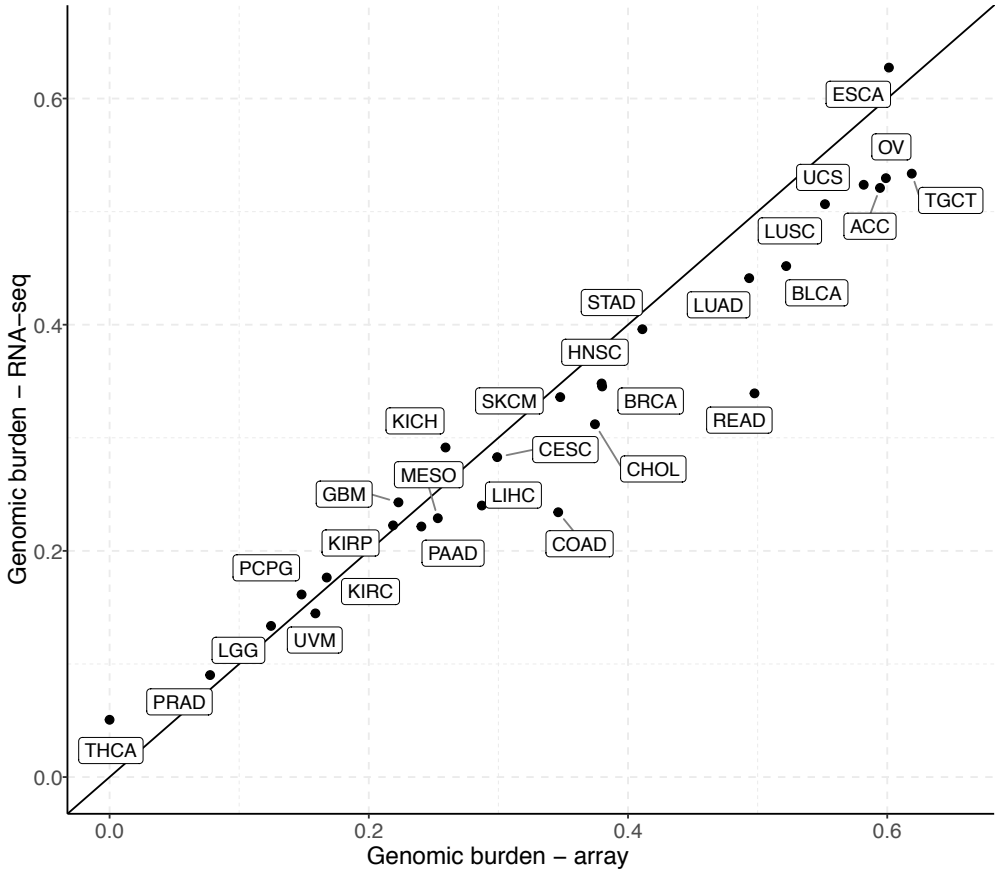


Fig. 2.9. Concordance assessment at genome level "genomic burden" across 28 cancer sites in the TCGA. Genomic burden is defined as the fraction of the genome that is affected by SCNAs. A scatter plot demonstrating the concordance between RNA-seq- and gold standard-derived genomic burden (median) for each cancer site is shown.

To evaluate the SCNA patterns at a whole genome level, we calculated 'genomic burden' – the proportion of a sample's genome exhibiting SCNAs. Marginally, we observed a median 0.28 genomic burden among all samples across all cancer sites, compared with 0.31 from DNA microarrays. Further, we investigated the patterns of genomic burden per cancer site. The highest genomic burden was observed in ESCA (0.63), followed by OV (0.53) and TGCT (0.53). THCA was the lowest (0.05), followed by PRAD (0.09). Median genomic burden for all cancer sites along with the corresponding array-derived genomic burden are shown in Figure 2.9. Next, we assessed the correlation of the RNA-derived genomic burden with the gold standard-derived genomic burden at a sample level, resulting in a very strong positive correlation (R=0.92; Fig.

2.10). When grouped by cancer site, most of the sites exhibited correlations larger than 0.9 (Fig. 2.11).

SCNA calls discovered by RNA-seq that were not detected in the DNA (gold standard) are putative false positives in our analysis. However, some appear to be false negatives in the gold standard set. To attempt to discriminate between these, we examined in more detail the COAD data, where these putative calls made up 6% of all SCNAs discovered in RNA-seq. We explicitly tested the genomic regions of the putative false positive SCNAs in the corresponding SNP array data with a specific binomial test leveraging phase concordance. Among the 6%, we found that approximately one-fourth (23%) of the calls were validated in the SNP array data (at a p-value < 0.05), indicative that the reported false positive rates are modestly over-estimated.

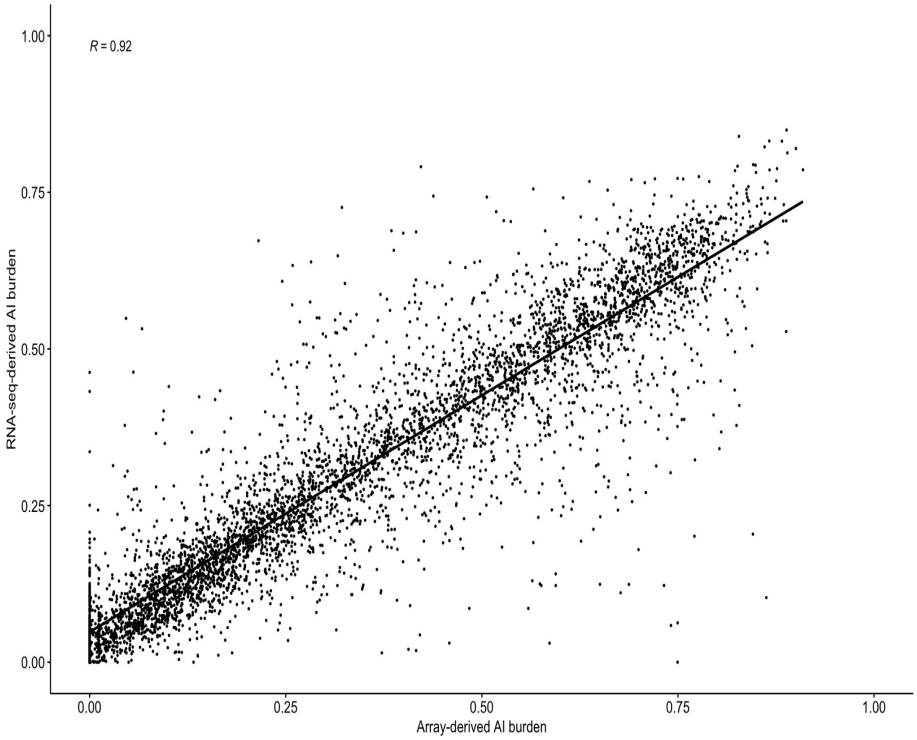


Fig. 2.10. Concordance assessment at genome level across the samples in 28 cancer sites in the TCGA. A scatter plot demonstrating the concordance between RNA-seq- and gold standard-derived genomic burden (median) for each sample is shown.

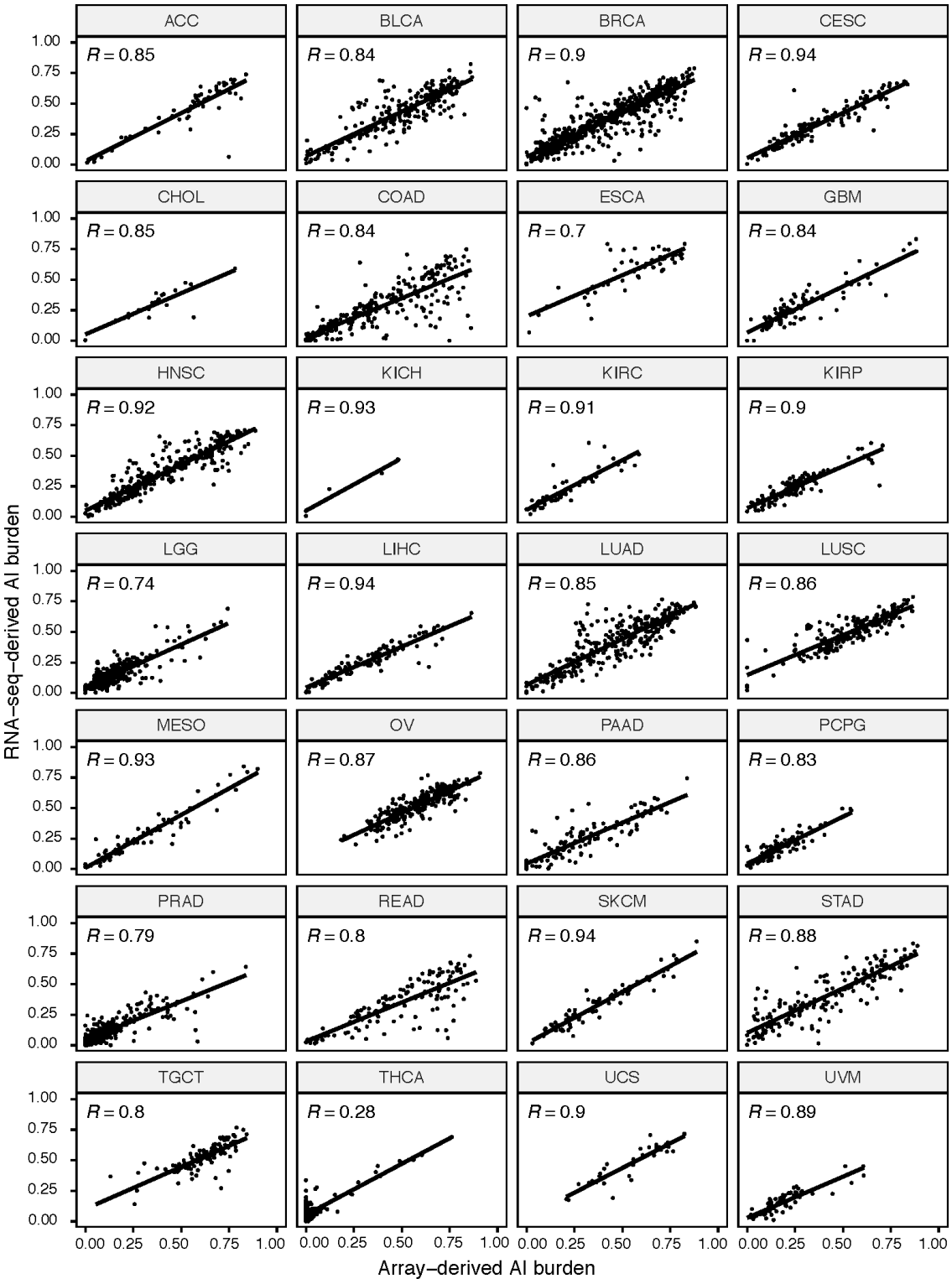


Fig. 2.11. Concordance assessment at genome level for 28 cancers the TCGA. For each site, a scatterplot representing genome level concordance is shown.

2.4.3 Comparison to other methods for bulk RNA-seq

While most methods for SCNA inference from RNA-seq have been designed for single cell data, we were able to conduct a detailed comparison with one state-of-the-art method for bulk RNA. Table 2.5 contains a summary of results from our method, hapLOHseq and CaSpER, for BRCA and GBM, sites analyzed in the original paper for CaSpER. We compared both methods to the gold standard. Compared with the gold standard calls at the gene level, for both cancers, hapLOHseq offered a superior performance, with a substantial increase in sensitivities with absolute gains of 30% and 42%. Against CaSpER's own benchmark, the methods appeared more similar, with an edge to hapLOHseq in detection but at some cost in specificity (Table 2.6). Fig. 2.12 demonstrates performance comparison between hapLOHseq and CaSpER at sample level.

Method	BRCA (n=77)		GBM (n=98)	
	Sens	Spec	Sens	Spec
hapLOHseq (RNA-seq)	74%	94%	81%	93%
hapLOHseq (RNA-seq + imputed genotypes)	85%	91%	89%	92%
CaSpER	43%	82%	59%	95%

Table 2.4. hapLOHseq and CaSpER comparison. hapLOHseq and CaSpER performance evaluation. Rows 1-3 show performance results at the gene level obtained by comparing each method to the gold standard [3].

We were also able to successfully run SuperFreq on a subset of the BRCA samples. From analysis of 12 samples where we had results from both SuperFreq and hapLOHseq, the sensitivity for hapLOHseq was 85% versus 77% for SuperFreq (specificities were 95% and 98%, respectively). Fig. 2.13 demonstrates performance comparison between hapLOHseq and SuperFreq at sample level. The authors of SuperFreq demonstrated high sensitivities analyzing *TP53* alterations in high mutant cell fraction settings for COAD. We were able to detect SCNAs in *TP53* in this set with hapLOHseq at an equivalent rate but without the need for additional RNA samples for normalization.

Method	BRCA (n=77)		GBM (n=98)	
	Sensitivity	Specificity	Sensitivity	Specificity
hapLOHseq (RNA-seq)	67%	82%	78%	88%
hapLOHseq (RNA-seq + imputed genotypes)	75%	77%	84%	86%
CaSpER	60%	85%	73%	95%

Table 2.5. hapLOHseq and CaSpER performance evaluation against CaSpER's gold standard. Rows 1-3 show performance results at the gene level obtained by comparing each method to the gold standardused in Harmanci *et al* [4].

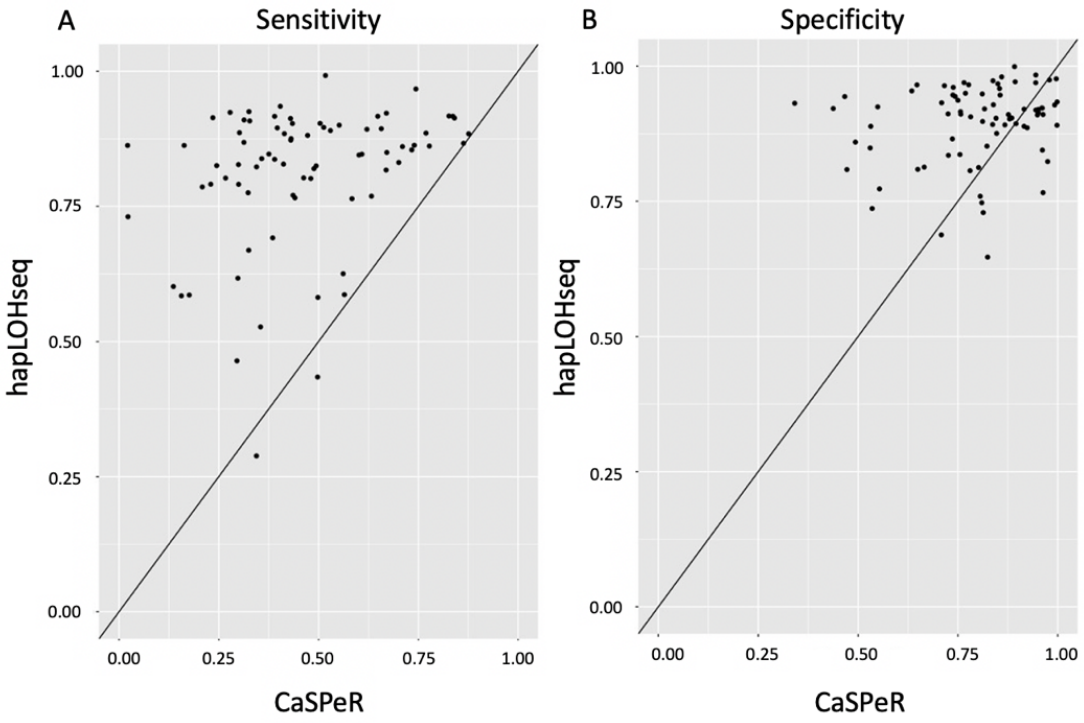


Fig. 2.12. hapLOHseq and CaSpER performance comparison. (A) Sensitivity, (B) Specificity.

Each sample is represented with a black dot.

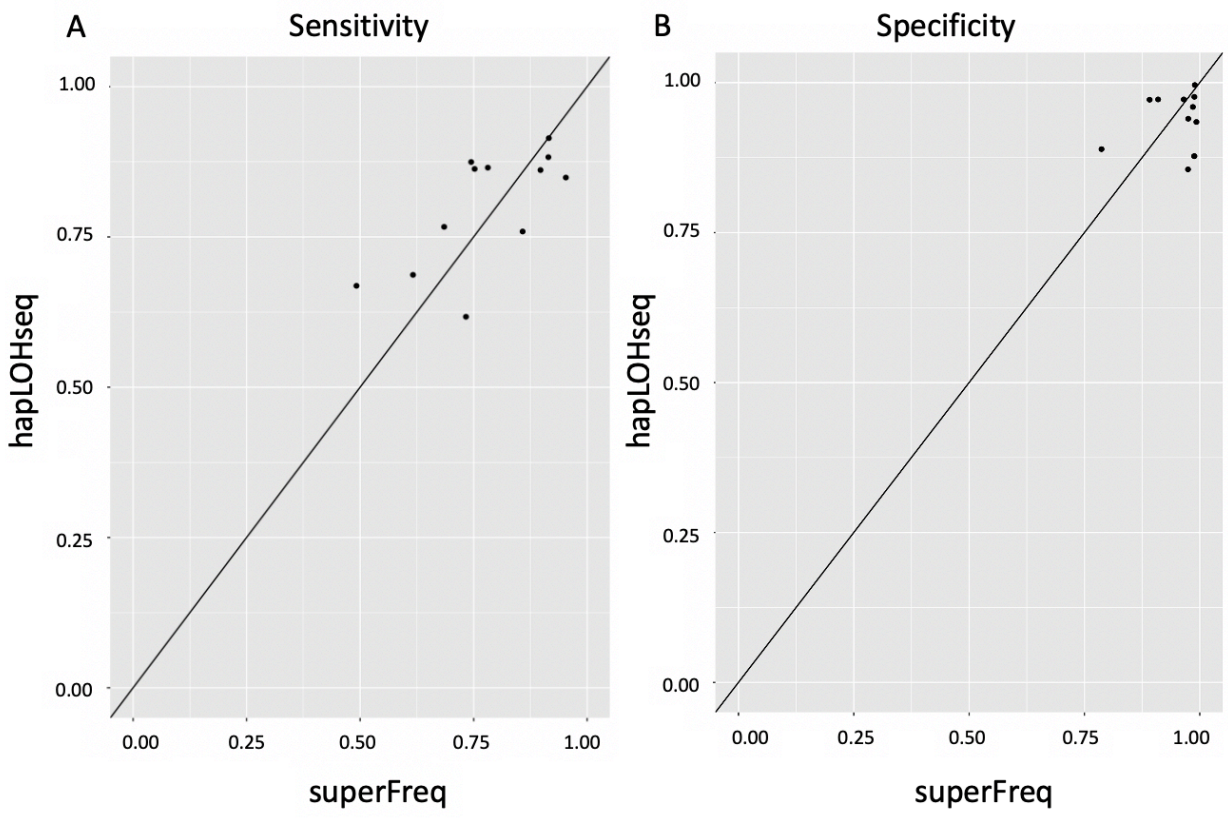


Fig. 2.13. hapLOHseq and superFreq performance comparison. (A) Sensitivity, (B) Specificity. Each sample is represented with a black dot.

2.4.4 Translational/ prognostic use

Finally, to demonstrate a translational potential, we examined the portability of conclusions from others' analyses of DNA to ours from RNA-seq. In an example in breast cancer, we recapitulated the distinct genomic burden distributions across different subtypes previously observed in TCGA BRCA data [2], demonstrating that those of a basal subtype are characterized by high genomic burden in comparison to the others (Fig. 2.14A). We observed that the samples in the basal subtype that have more than 40% of their genome altered comprise 84% of all basal samples, consistent with the previous report. Furthermore, analyzing the RNA-seq, we were also powered to observe that chromosome arm 5q is more frequently altered in the basal subtype, whereas chromosome arms 1q and 16q are more frequently altered in the luminal subtypes, consistent with

the previous findings from analyzing DNA directly [2] (Fig. 2.14B). Finally, we specifically investigated the concordance between hapLOHseq and gold standard results for five genes that are frequently affected by CNA events for the BRCA cohort, i.e., *BRCA1/2*, *PTEN*, *RB1* and *TP53*. We showed that hapLOHseq obtained promising results that have a potential clinical use with 93% sensitivity and 85% specificity for *BRCA1*, 78% sensitivity and 95% specificity for *BRCA2*, 85% sensitivity and 88% specificity for *PTEN*, 87% sensitivity and 94% specificity for *RB1* and 90% sensitivity and 86% specificity for *TP53*. Sample-level concordance assessment for each of the genes is shown in Fig. 2.14C.

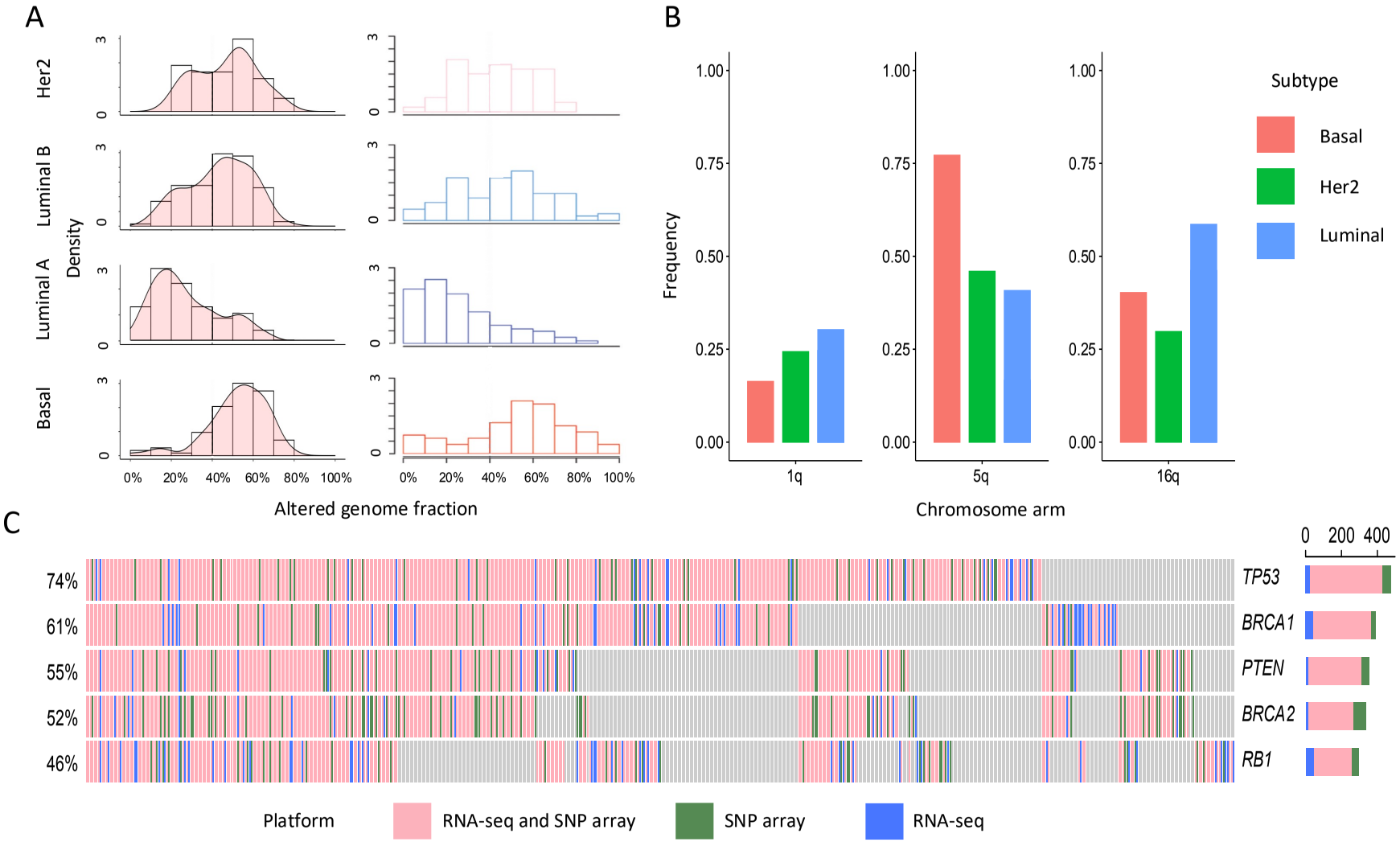


Fig. 2.14. Clinical efficacy of hapLOHseq results demonstrated using TCGA BRCA cohort. A) Recapitulating the genomic burden distribution across different subtypes: left: from the supplementary material of the TCGA BRCA paper (Cancer Genome Atlas Network, 2012), right: hapLOHseq results; histogram of sample genomic burden across the cohort grouped by subtypes. B) Frequency of chromosome arm level alterations in 1q, 5q, and 16q as a fraction of number of samples across different subtypes. C) Concordance assessment for the five genes that are frequently affected by SCNA events. Rows represent the genes and columns represent the samples in the cohort.

2.5 Discussion

In this study, we detect and characterize the genomic landscapes of SCNAs from tumor bulk RNA-seq using a haplotype-aware statistical method. We proposed two approaches that differ in the way germline genotypes are obtained for subsequent analysis of 'B allele' frequencies (BAFs). While the first approach solely uses RNA-seq from tumor to estimate the haplotypes, the second approach leverages available or potentially collectible SNP array (or equivalent) data from a matched-normal sample to achieve higher accuracies in genotyping and haplotype reconstruction through a popular imputation pipeline.

In an analysis of 28 cancer sites from TCGA, our method achieved high sensitivity for SCNA detection with the imputation approach (85% versus 68%), retaining high specificity as well (~95%). Summaries of SCNA genomic burden were sufficiently high as to potentially obviate the need for analyzing DNA. In sites with lower sensitivities, such as PRAD, analyses of DNA exome sequencing reflected difficulties as well, indicating challenges for such sites more specific to targeted sequencing data.

Our imputation approach addresses difficulties associated with genotype calling from RNA-seq, e.g., due to non-uniform coverage, which in turn provides highly accurate and phased genotypes. Indeed, we explored these factors as direct contributors to improved performance in PRAD. Our germline heterozygote identification could improve other methods for bulk RNA analysis, as well, such as CaSpER and SuperFreq. In different statistical implementations, each of these combine information from not only total read counts but also BAF dispersion at heterozygotes. Whereas SuperFreq relies on external data for normalization (e.g., paired normal RNA samples), CaSpER's approach works on a sample-by-sample basis, as does hapLOHseq. We conducted a detailed comparison to CaSpER, observing higher sensitivities with our

haplotype-based approach. Ultimately, getting the absolute performance characteristics will depend on improved gold standard datasets. We note that these methods use information orthogonal to that leveraged by hapLOHseq and thus may offer improvements when applied in combination, or integrated for joint analyses -- an area of future study.

Blood, buccal or adjacent normal samples can serve as a surrogate of the germline. The first two are non-surgical and more easily collected, whereas the third may be available for some specimens. Array-based genotyping of these samples presents an economical approach for improved tumor SCNA characterization. This is feasible for existing clinical cohorts with banked patient blood samples or biobanks with existing genotype data. In our exhibition, we focused on individuals of European ancestry since the MIS had the specific reference panel. This results in a 'best case' scenario given current resources. However, efforts such as TOPMed [64] will generate high-density genotype panels for individuals of non-European ancestry. Our approach further highlights the need for genetic panels of high diversity in biomedical research.

Our approach, as we have demonstrated here, does not attempt to detect balanced duplications, i.e., those where maternal and paternal chromosome segments are present in equal ratios. This may be overcome through integration of coverage data with our existing approach, which would be feasible for large balanced duplications as shown previously [4,52]. While any method for RNA-seq will have natural limitations in detecting alterations that do not span expressed genes, over larger regions limiting factors will be mitigated or averaged toward genome levels. Further, molecular alterations caused by focal SCNAs of key cancer drivers may be detectable through traditional RNA analyses of altered gene expression, including at the pathway level, and/ or identification of specific transcripts.

In summary, we show that the proposed haplotype-based approach for RNA-derived SCNA calls is robust for detection of megabase-scale somatic mutations. Overall, detection rates were

generally higher than 85% at specificities high enough for de novo discoveries and assessments of genomic associations with clinical phenotypes, across malignancies. Indeed, we successfully recapitulated SCNA features associated with clinical subtypes of breast cancer. Our findings show that our method can be used to increase the utility of bulk RNA-seq by allowing for a more comprehensive molecular profiling of tumors in settings where DNA analysis is impractical due to limited tissue sample availability or financial constraints and enables secondary analyses of existing data from high-value clinical cohorts.

CHAPTER 3

QUALITY CONTROL AND BEST PRACTICES

3.1. Introduction

In this chapter, I discuss quality control measures that can be taken before running hapLOHseq and 'best practices' to follow while running hapLOHseq and after obtaining event calls to ensure obtaining optimal results, although each setting is unique and would require reconsideration of parameters.

First, I present two sample level metrics that can be used to identify samples with RNA-seq quality issues which can lead to spurious or missed allelic imbalance calls. Second, I evaluate the effect of different imputation quality (R^2) cutoffs on the hapLOHseq results, and discuss the effects of several hapLOHseq parameters on performance. Additionally, as part of the best workflow, I discuss masking of certain genomic regions to exclude potential inherited duplications, and removal of singletons, which are heterozygous markers observed in only one sample at a particular locus in our analysis. Third, I discuss the use of a custom script instead of hapLOHseq's default event caller to determine allelic imbalance event boundaries from posterior probabilities. Fourth, I describe two post-processing steps, i.e., after hapLOHseq detects events, for the removal of inherited duplications and putative false positive calls in genomic regions with a paucity of heterozygous sites. Lastly, I describe several characteristics of the allelic imbalance calls. The goal of presenting these best practices is to aid future studies tune hapLOHseq parameters and pre- and post-processing settings to help achieve study specific sensitivity and specificity goals.

3.2. Quality control

3.2.1. Sample quality

We investigated the relation between RNA integrity number (RIN) score, a sample level RNA quality metric, and hapLOHseq's performance, particularly sensitivity and ability to pick up true allelic imbalance events from RNA. The RIN score ranges between 1 to 10 and quality of the sample increases as the number gets closer to 10. Generally, high-quality RNA samples have a score of 8 or greater, whereas partially fragmented RNA samples have a RIN score between 6 and 8. We assessed the correlation between sensitivity and RIN score for a subset of RNA samples from the LUAD cohort, running hapLOHseq without the use of a paired normal for imputation. Overall, we did not detect low quality RNA samples, as all samples had a RIN score of 7 or greater, and did not observe a correlation between sensitivity and RIN score (Fig. 3.1).

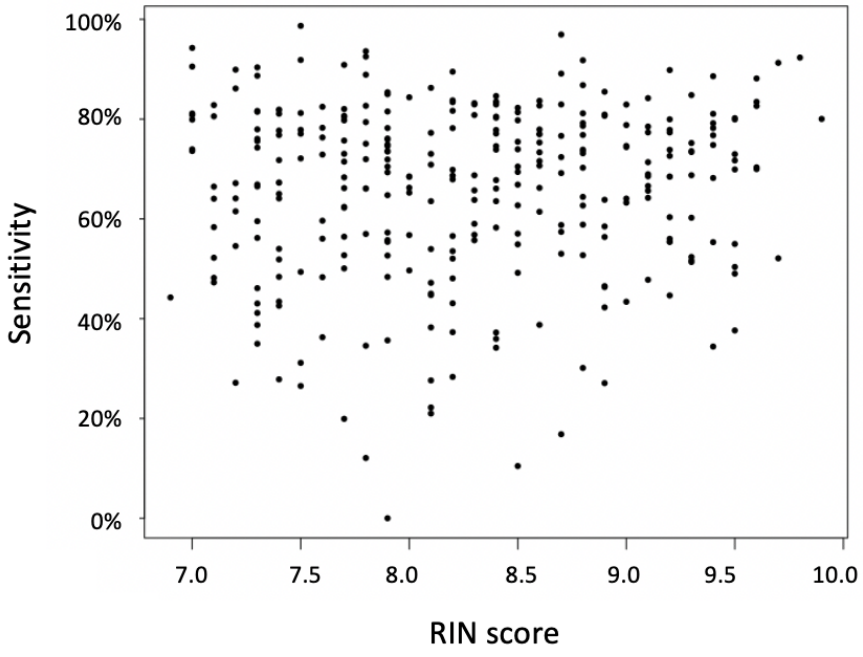


Fig. 3.1. RIN scores of the samples from the TCGA LUAD cohort. The scatterplot illustrates the TCGA LUAD samples' RIN scores and sensitivity values. Each sample is represented with a dot.

Usually, tumor samples are contaminated with normal cells. In these settings, low tumor cellularity poses a challenge in the detection of allelic imbalance in tumor cells. We assessed the relation between sample cellularity and hapLOHseq's performance using the samples in the TCGA PRAD cohort as prostate adenocarcinoma is one of the tumor sites that has low sensitivity when hapLOHseq was run with the use of a paired normal for imputation. Using the clinical data, we categorized 123 samples into 3 cellularity levels, i.e., low, medium, and high cellularity, and performed chromosome arm level concordance analysis for each group separately. The results did not indicate a clear distinction between medium and high cellularity categories and the low cellularity category did not have enough samples to draw a conclusion (Fig. 3.2.).

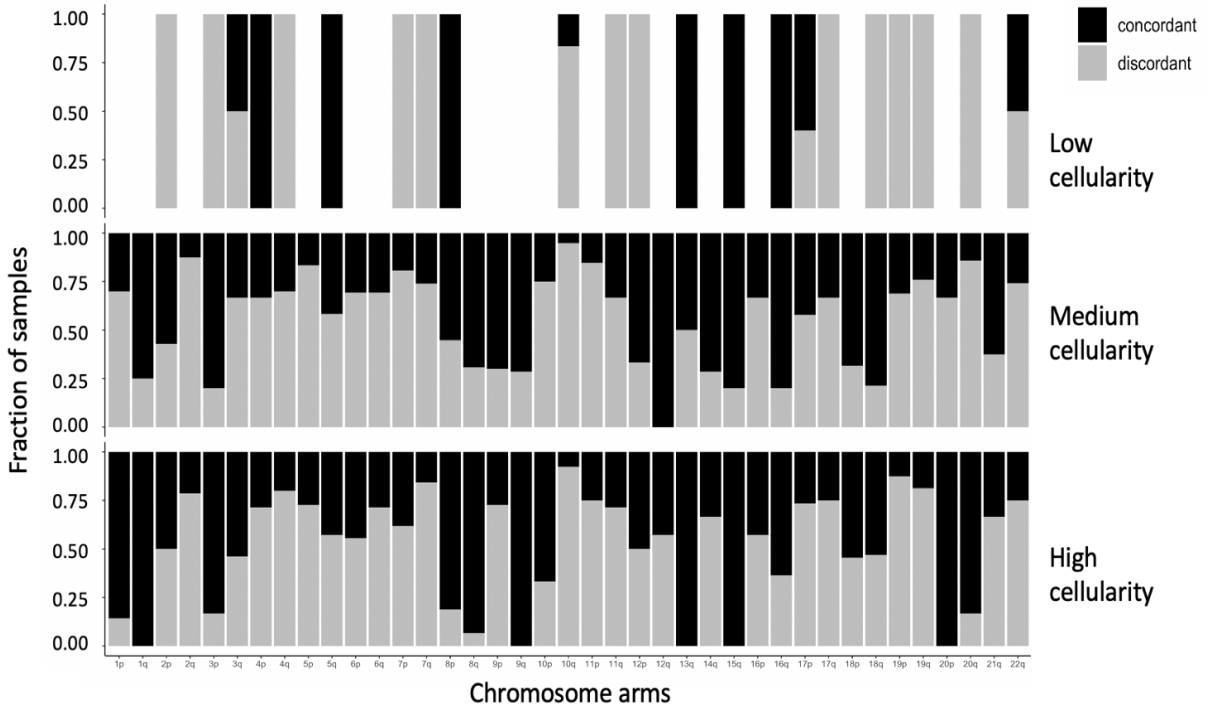


Fig. 3.2. Cellularity of the samples in the TCGA PRAD cohort and concordance between hapLOHseq results and the gold standard. We evaluated cellularity's effect on hapLOHseq's concordance with the gold standard by categorizing the samples into low, medium, and high cellularity groups.

Although the RIN score and sample cellularity did not have a noticeable impact on hapLOHseq performance for these TCGA datasets, their impact should be further assessed in additional datasets samples that have overall lower RIN score and tumor cellularity.

3.3. Best practices

3.3.1. Imputation quality

To ensure genotype imputation quality, we utilized the R^2 metric. R^2 is a quality metric that quantifies quality of imputation for each marker by estimating the correlation between imputed and true genotypes. Throughout our studies, we preferred the commonly used $R^2 < 0.3$ cutoff to remove poorly imputed markers. For a subset consisting of 20 samples from the TCGA LUAD cohort, we tested additional R^2 cutoffs in {0.4, 0.5, 0.7}. Across the 20 samples, 111 allelic imbalance events were detected from the gold standard. We evaluated RNA based calls from hapLOHseq runs set up using different R^2 cutoffs by comparing with the DNA based gold standard calls. The results suggest that increasing the cutoff does not generate more favorable results. Table 3.1. summarizes the results of performance evaluation obtained by intersecting RNA based allelic imbalance events with the gold standard.

	Overlap		
R^2 cutoff	**≥25% reciprocal**	**≥50% reciprocal**	**≥70% reciprocal**
0.3	73%	56%	34%
0.4	72%	55%	31%
0.5	75%	57%	33%
0.7	73%	58%	33%

Table 3.1. The effect of R^2 cutoff on performance. We intersected RNA based allelic imbalance calls obtained from hapLOHseq runs set up with different R^2 cutoffs for genotype imputation quality with the gold standard calls. For each R^2 cutoff, the fraction of RNA based events overlapping with the gold standard is shown for 25%, 50%, and 70% reciprocal overlap, i.e., the fraction of intersection is reciprocal for RNA and DNA.

3.3.2. hapLOHseq run setup

3.3.2.1. hapLOHseq parameters

3.3.2.1.1. Minimum depth

In the studies discussed throughout this dissertation, a minimum depth limit, i.e., through hapLOHseq parameter --min_depth, defined as the minimum depth for markers to be included in HMM, of 10 was applied, which is the default parameter for hapLOHseq. It is possible that other depth limits will produce more favorable results. To assess this, we performed an investigation at other per-marker minimum depths {4,14} for the LUAD cohort. Table 3.2 below summarizes the results. Changing this parameter from 10 to 8 appears to result in similar performance for the hapLOHseq run that uses of a paired normal for imputation. Application to other datasets should be done with caution, as the average coverage may be higher or lower in other settings. For example, when more than 99% of sites have coverage greater than 10, then cut-offs of 4 to 10 would show highly similar performance.

min4		**min6**		**min8**		**min10**		**min12**		**min14**	
Sens	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**
84%	85%	83%	88%	84%	90%	84%	90%	83%	90%	83%	90%

Table 3.2. Assessment of the method using different minimum depth thresholds in the TCGA LUAD cohort. Results are shown for minimum depths 4, 6, 8, 10, 12, and 14.

3.3.2.1.2. Other parameters

In addition to exploring the effects of using non-default values for minimum depth, we also evaluated the consequences of altering several other hapLOHseq parameters, such as --marker_density, --event_prevalence, and --event_mb (see Table 3.3). We tested these parameters on a subset of samples by only altering the parameter and keeping everything else constant.

hapLOHseq parameter	Default	Used
--marker_density	uniform	genomic_pos
--event_prevalence	0.1	0.05
--event_mb	20Mb	10Mb, 30Mb

Table 3.3. hapLOHseq parameters. Values of the non-default parameters used in our hapLOHseq setups.

First, we investigated --marker_density parameter, which defines the strategy for modeling distances between markers. For this parameter, the options are *uniform* (by default) and genomic position (*genomic_pos*). These two options did not make any difference and yielded to identical results. Second, we examined the --event_prevalence parameter, which is an estimate of the proportion of the genome that is expected to be aberrant. For our studies, we used 0.05, instead of the default 0.1. Third, we assessed the impacts of using 10 Mb and 30 Mb, which are non-default values for the --event_mb parameter, a parameter used to provide hapLOHseq the expected allelic imbalance event size (default is 20 Mb). hapLOHseq is typically robust to incorrect specifications of the --event_prevalence and –event_mb parameters, which are the two user specified parameters that are used to set the transition probabilities in the HMM. The results indicated that the differences between the allelic imbalance calls using different --event_mb and --event_prevalence were nearly non-existent.

3.3.2.2. Masking of the HLA, VDJ, and DGV regions

We masked the genomic regions corresponding to the centromere, human leukocyte antigen (HLA), VDJ, and Database of Genomic Variants (DGV) in order to exclude genomic regions with putative germline copy number changes.

The events called by our approach are at megabase scale, while the masked genomic regions are small (DGV region median size was 46Kb, and Table 3.4 below summarizes the sizes of HLA/

VDJ regions). Given that the average SCNA burden is 808 Mb in tumors, masking these regions has negligible impact on the call set. Similarly, when using the hapLOH generated gold standard for comparison to other tools, there would be a negligible impact on specificity and sensitivity estimates. As allelic imbalance signals at these regions could come from naturally occurring DNA rearrangements in immune cells, we feel it is best practice to mask these regions for our haplotype based approach. Here we give a summary of the size of masked regions.

Region	Size (Mb)
IGK	1.4
HLA/MHC	4.9
TRG	0.1
TRB	0.5
TRA & TRD	0.9
IGH	1.3
IGL	0.9

Table 3.4. Sizes of the HLA and VDJ regions. The sizes of the HLA and VDJ regions that were masked in our approach.

To mask these genomic regions, we removed the markers that fall in these regions prior to making allelic imbalance calls, instead of post-processing the allelic imbalance call set to remove the events in these regions. Thus, it is still possible for allelic imbalance calls to span these regions. As part of our standard QC pipeline, we mask centromeres. However, for both SNP arrays and RNA-seq, there are no markers present in the centromeres.

To examine the performance when these variable regions are included, we assessed the impacts of not performing (any) masking in RNA and in the gold standard data. As a result, we obtained marginally higher sensitivities and marginally lower specificities (Table 3.5). For instance, in lung adenocarcinoma, sensitivity was 86% (*vs.* 84% masked), specificity was 87% (*vs.* 90% masked), in lung squamous cell carcinoma, sensitivity was 90% (*vs.* 88% masked), specificity was 85% (*vs.* 88% masked), and in colon adenocarcinoma, sensitivity was 85% (*vs.*

79% masked), specificity was 96% (*vs.* 97% unmasked). However, qualitative interpretation is essentially the same. Since we do not explicitly model the repeat rich nature of these genomic regions, we decided to exclude them in advance as a better practice.

	COAD		LUAD		LUSC	
	Sens	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**
Masked	79%	97%	84%	90%	88%	88%
Unmasked	85%	96%	86%	87%	90%	85%

Table 3.5. Comparison of the masked and unmasked runs. The effects of masking the HLA/ VDJ, and DGV regions in sensitivity and specificity for the TCGA COAD, LUAD, and LUSC cohorts.

3.3.2.3. Removal of singletons

Heterozygous markers that are observed in only one sample at a particular locus are defined as singletons. Since they are observed only once, phasing singletons is not possible unless a large reference panel is used. In addition to the challenges in haplotype estimation, another issue with singletons is that they might have been called heterozygous by mistake. Both the mistakes in phasing and making false heterozygous calls may interfere and be weakening for hapLOHseq as it relies on heterozygous sites and haplotype inference. Therefore, we recommend removal of singletons as part of the best practices workflow.

3.3.3. Using a custom event caller

Instead of using hapLOHseq's default event caller, one can also use their own script (see Appendix B for an example script) to tailor event calling parameters and determine allelic imbalance event boundaries using posterior probabilities from hidden Markov model (HMM). For calling an event, hapLOHseq by default requires posterior probabilities to reach a peak ≥ 0.9 and not drop below 0.5 within the boundaries of an event. Fig. 3.3 below shows the distribution of posterior probabilities of markers for LUAD sample TCGA-05-4244-01A-01R-1107-07, which

is representative of what we observe across TCGA samples. It is not possible to render a smooth ROC from the probabilities to quantify the method's performance through AUROC, as the values cluster at 0 or 1 when using dense, informative genotype data (Fig. 3.4). For illustration, we provide the ROC curve for the LUAD cohort, which shows that tweaking the classifier for our approach does not greatly alter performance (Fig. 3.4). Table 3.6 summarizes the performance results for the LUAD cohort for varying event peak probability cutoffs. According to these results, there is not a significant difference in the performance between using hapLOHseq's default event caller and custom event calling.

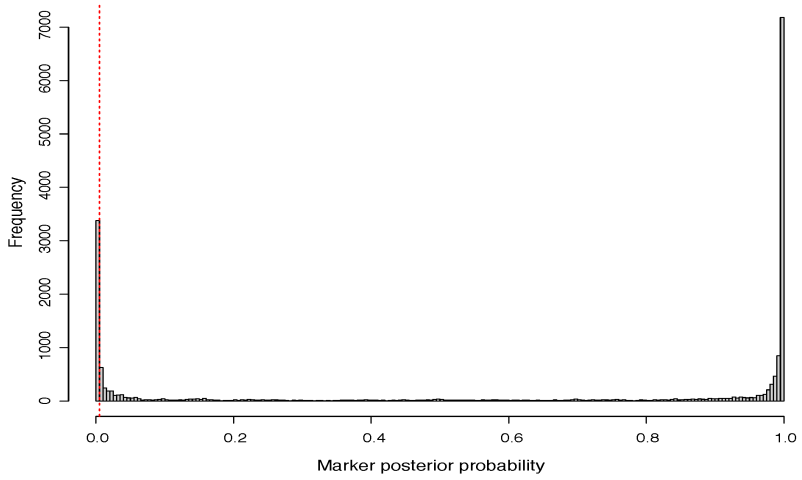


Fig. 3.3. Histogram of posterior probabilities for LUAD sample TCGA-05-4244-01A-01R-1107-07. The red dashed line indicates the 0.005 threshold.

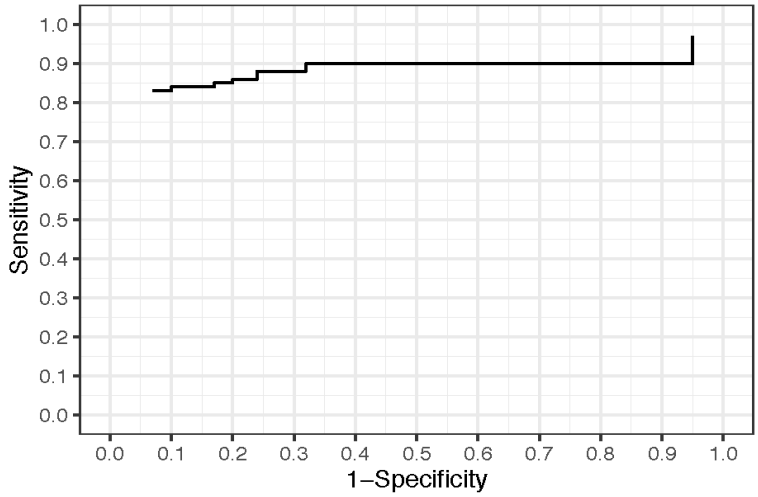


Fig. 3.4. ROC curve for AI classifier. The ROC curve for the TCGA LUAD cohort is shown.

event_boundary	event_peak	Sensitivity	Specificity
0.50	**0.50**	0.84	0.86
0.50	**0.60**	0.84	0.86
0.50	**0.70**	0.84	0.87
0.50	**0.80**	0.84	0.88
0.50	**0.85**	0.84	0.89
0.50	**0.90**	0.84	0.90
0.50	**0.95**	0.83	0.91
0.50	**0.97**	0.83	0.92
0.50	**0.98**	0.83	0.93

Table 3.6. The impact of using a custom event caller to use varying event peak probability cutoffs. The results are shown for the TCGA LUAD cohort.

3.3.4. Post-processing of allelic imbalance events

Post-processing of allelic imbalance calls generated by hapLOHseq is an important step to further improve the precision of events. As part of our standard post-processing pipeline, we removed allelic imbalance events smaller than 2 Mb to exclude potential germline gains and events containing less than 10 heterozygous markers. In addition, we split up events that contain large, i.e., greater than 10 Mb, genomic regions without heterozygous markers into multiple regions. Fig. 3.5 demonstrates an example (TCGA-05-4244-01) from the TCGA LUAD cohort where an allelic imbalance event on chromosome 1 was split up into two regions to omit a large genomic region that does not harbor any heterozygous markers.

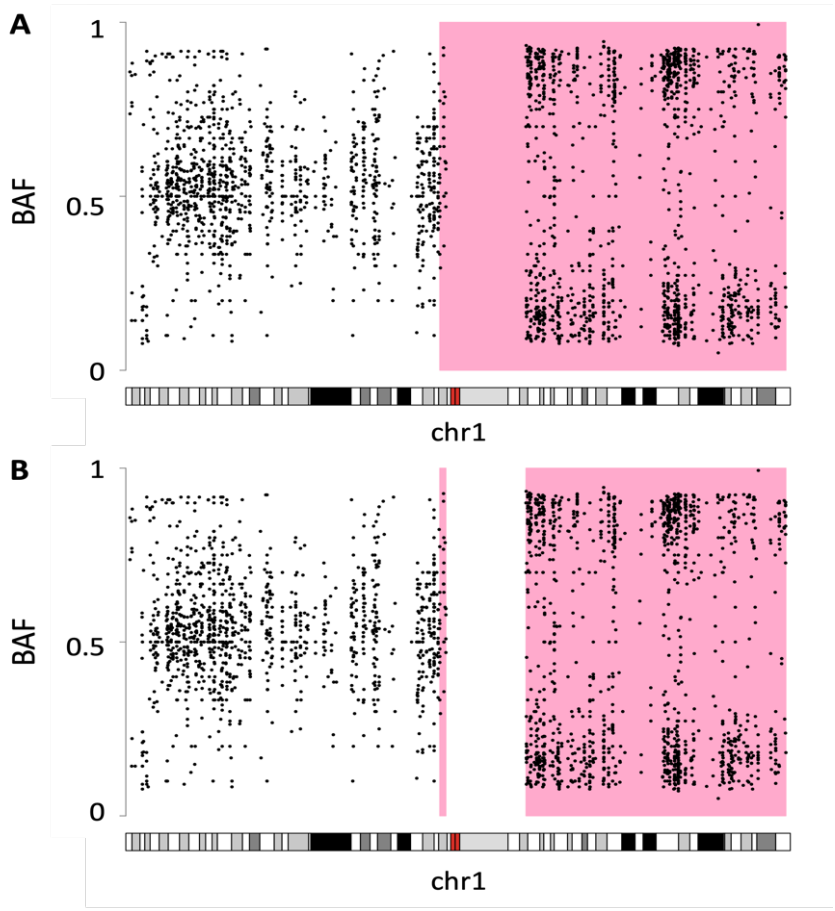


Fig. 3.5. Splitting allelic imbalance events into multiple events to prevent spanning large genomic regions without any heterozygous markers. An allelic imbalance event identified from TCGA LUAD sample TCGA-05-4244-01 on chromosome 1 (A) was split up into two events (B).

3.4. Allelic imbalance events characteristics

3.4.1. Characterization of RNA-exclusive events

We sought to examine the RNA-exclusive events, i.e., the allelic imbalance events that were detected only from RNA and not from the gold standard (DNA), obtained from samples in the TCGA LUAD cohort in order to comprehensively characterize them. Across all the samples in the cohort, we detected 977 RNA-exclusive events in total with a median of 3 events per sample and median size of the RNA-exclusive events was 17.3 Mb. Fig. 3.6 shows the distribution of the RNA-exclusive event sizes. This figure suggests that a large proportion of the RNA-exclusive

calls can be avoided by applying a larger threshold, e.g., in cases where the user is interested in very large - chromosome arm level - events.

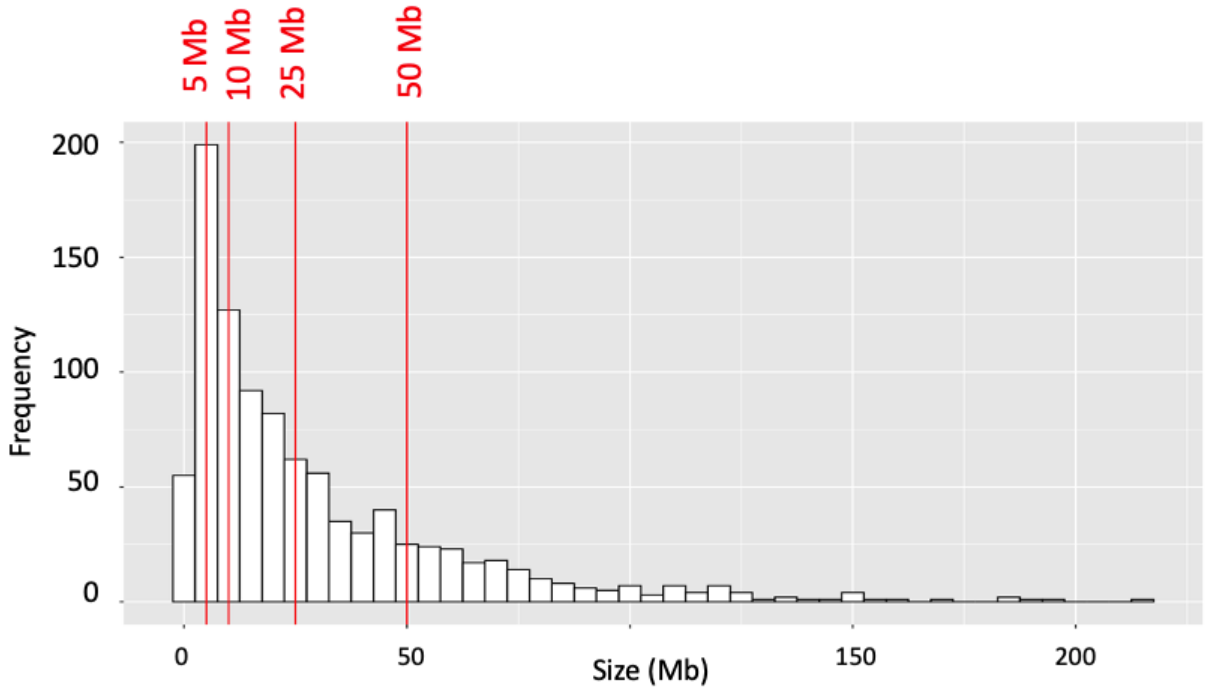


Fig. 3.6. Histogram of the size of the RNA-exclusive allelic imbalance events. Vertical red lines correspond to varying size thresholds.

Although RNA-exclusive calls are considered putative FPs throughout our analyses, some of the calls seemed to be FNs in the gold standard set. Hence, to distinguish between them, we further investigated the RNA-exclusive allelic imbalance events in the TCGA LUAD cohort. We comprehensively tested the genomic regions that harbor the RNA-exclusive events in the matching DNA (gold standard and WES) with a binomial test that leverages switch accuracy to determine whether the genomic region with allelic imbalance detected from RNA has high phase concordance in the corresponding region from DNA, which indicates allelic imbalance (which would confirm the event observed in RNA). The binomial test suggested out of 977 RNA-exclusive events, 294 events were validated in both SNP array and WES, 135 events were validated only in SNP array, and 154 events were validated only in WES. The remaining 394 calls were confirmed as RNA-exclusive as we detected no allelic imbalance signal from either SNP

array or WES in the matching sample. Hence, we conclude that the false positive rate for the LUAD cohort is over-estimated. Fig. 3.7 illustrates the size distribution of the confirmed RNA-exclusive events. Overall, the confirmed RNA-exclusive events appear to be smaller in comparison to all RNA-exclusive events. Finally, we calculated sensitivity and specificity for the LUAD cohort after excluding the confirmed RNA-exclusive events. While sensitivity stayed the same (84%), specificity increased by 2% and reached 92%.

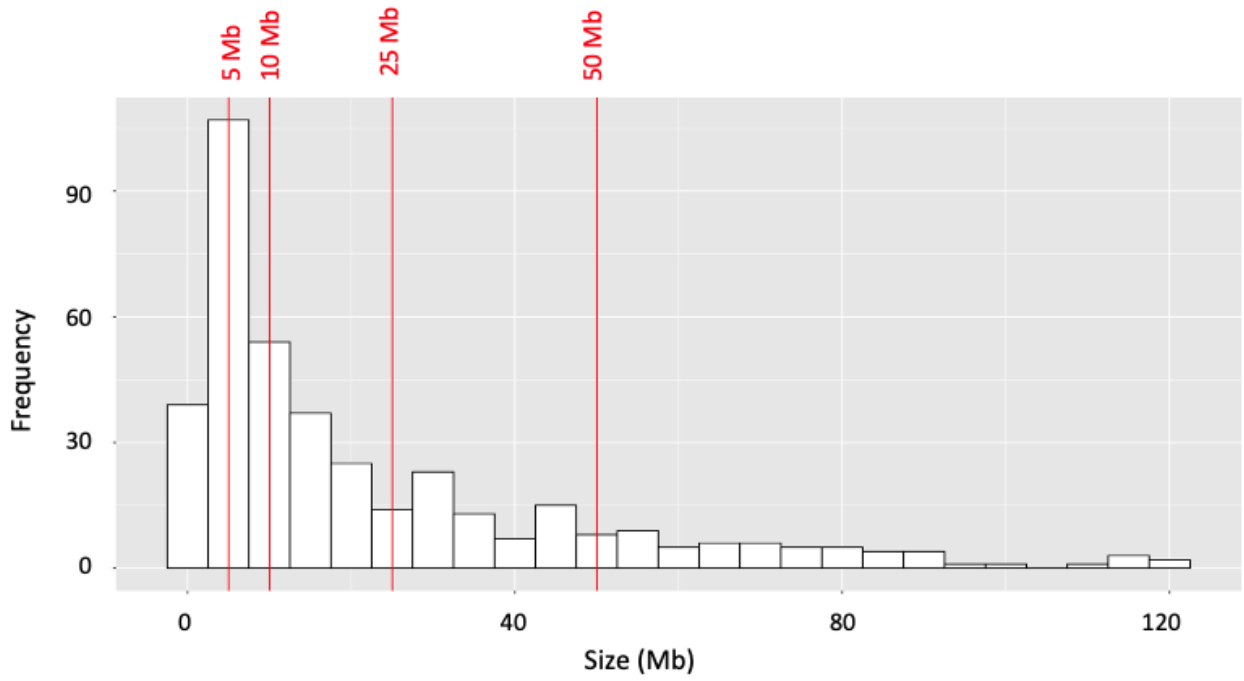


Fig. 3.7. Histogram of the size of the confirmed RNA-exclusive allelic imbalance events. Vertical red lines correspond to varying size thresholds.

3.4.2. Allelic imbalance event types and performance

We assessed our approach's performance for different categories of allelic imbalance (see Table 3.7). We noticed that 'undetermined' events category, i.e., subtle events detected from bulk samples with low MCF that could not be categorized as gain, loss, or cn-LOH, in the gold standard call set was the category of events that was most often failed to be inferred by RNA-based identification, which was most strikingly observed in the pancreatic and prostate adenocarcinomas. Excluding these undetermined (low MCF) events naturally led to higher

sensitivity rates. On the other hand, the best performance was consistently achieved by the cn-LOH event category, likely due to greater BAF perturbations observed from the cn-LOH events.

	COAD		LUAD		LUSC		PAAD		PRAD	
	R	**R+I**	**R**	**R+I**	**R**	**R+I**	**R**	**R+I**	**R**	**R+I**
cn-LOH	51%	85%	79%	90%	79%	91%	81%	89%	61%	81%
Gain	61%	80%	72%	81%	73%	83%	67%	74%	62%	71%
Loss	37%	80%	71%	86%	70%	89%	70%	85%	40%	69%
Undetermined	30%	32%	63%	73%	68%	78%	46%	61%	0%	22%
Overall	49%	79%	72%	84%	71%	88%	67%	80%	45%	66%
Excluding undetermined	49%	82%	74%	85%	71%	88%	71%	81%	48%	69%

Table 3.7. Summary of sensitivity results grouped by allelic imbalance event category. R: RNA-seq, R+I: RNA-seq + imputed genotypes.

3.5. Discussion

In this chapter, I discussed several quality control steps and best practices for the most favorable allelic imbalance detection results using hapLOHseq. I mentioned two sample quality metrics, e.g., RIN score and sample cellularity, and their minor effects on our results. This was most likely because we utilized a high quality repository (TCGA) to obtain our bulk RNA-seq samples, however, I should note that these two metrics should be assessed for other cohorts. Furthermore, I discussed the impact of using high imputation quality (R^2) cutoffs, in addition to the traditional 0.3 cutoff, which did not lead to more favorable results for our data. Next, I discussed several hapLOHseq parameters, including minimum depth parameter. The results of our investigation at other per-marker minimum depths in {4,14} for the TCGA LUAD cohort indicated that changing minimum depth parameter from 10 to 8 resulted in similar performance. However, we advise caution when applying to other datasets, since coverage might differ. Next, we discussed the effects of masking HLA, VDJ, and DGV regions. We found that the qualitative interpretation of

the results remained essentially the same when comparing masked and unmasked hapLOHseq call sets. Then, we showed that removal of singletons leads to performance improvement and provides efficiency in terms of the time it takes to run hapLOHseq. Moreover, we evaluated using a custom event caller to identify allelic imbalance events from posterior probabilities, instead of using hapLOHseq's default event caller and discussed our post processing steps to improve the default allelic imbalance call set. Finally, we discussed the characteristics of RNA-exclusive calls and event categories. These characteristics might be helpful to users to decide when they have certain biological scenarios that they are interested in, e.g., chromosome arm level gain events.

CHAPTER 4

INVESTIGATION OF X-INACTIVATION DRIVEN ALLELIC IMBALANCE PATTERNS IN THE X CHROMOSOME ACROSS THE CANCER GENOME ATLAS BREAST INVASIVE CARCINOMA COHORT

4.1 Introduction

Along with 22 autosomal pairs of chromosomes, human cells almost always contain an additional pair of sex chromosomes. These two sex chromosomes are the X and Y chromosomes. While females typically have two X chromosomes, males typically have one of each sex chromosomes. The X chromosome consists of approximately 155 million base pairs, which corresponds to nearly 5% of the whole human genome and contains 841 protein coding genes [65] out of over 20,000 genes in the genome. The Y chromosome is much smaller, spanning nearly 55 million base pairs and containing 63 genes [66].

X-inactivation or lyonization is a phenomenon that describes inactivation of one of the two X chromosomes in female mammals early in embryonic stem cell differentiation. The selection of which of the two X chromosomes will be permanently inactivated is random in placental mammals and takes place in all cells in the body besides the reproductive cells. Due to the randomness of the pattern in this process, some cells have the maternal X chromosome inactivated, whereas others have the paternal X chromosome inactivated. X-inactivation ensures females have one functional copy of the chromosome, as do males [67].

The X chromosome is silenced through a series of epigenetic modifications involving histone modifications and promoter methylation that result in a tightly packed and transcriptionally inactive structure called heterochromatin. The process is controlled by the X chromosome inactivation center (XIC). Transcription of the X-inactive specific transcript (*XIST*) on the X chromosome that will be inactivated initiates the process. *XIST* is a 17 kilobase, noncoding RNA molecule. While staying in the nucleus exclusively, *XIST* RNA coats the chromosome it was transcribed from and provides a template for epigenetic alterations that consist of sequential histone modifications [67]. Although, in the end, most of the X chromosome obtains the features of heterochromatin, some genes (~15%), mostly located on the p arm, escape X-inactivation. The majority of the genes that escape X-inactivation are also located on the Y chromosome in regions known as 'pseudo autosomal' regions. For the genes in the pseudo autosomal regions, both alleles are expressed and both men and women receive two copies of these genes [68].

Loss of X chromosome inactivation appears to play an important role in female carcinogenesis as shown in ovarian and breast cancers [69–71]. With the loss of X chromosome inactivation, the cancer cells acquire an extra active copy, which is thought to increase the expression of the oncogenes and TSGs on the chromosome [72]. Moreover, there is a debated interplay between *BRCA1* and the inactive X chromosome [73]. The loss of X chromosome inactivation is observed more commonly in basal like subtype and breast cancers with mutations in *BRCA1*, which have aggressive behavior [74–76].

In this chapter, I discuss the details of my examination of a potential link between X-inactivation driven imbalances observed in the X chromosome and several clinical outcomes by using RNA-seq data from tumor samples of females in the TCGA BRCA cohort. I start by briefly explaining the differences in the methods regarding the imputation and phasing of the genotypes on the X chromosome in comparison to the processing of autosomal chromosomes as described

in chapter 2, followed by identification of the X-inactivation driven imbalances through hapLOHseq. I conclude by comparing and contrasting the imbalance events on the X chromosome with that of DNA-based results to explore possible associations between these X-inactivation driven imbalances observed exclusively from RNA and various clinical features.

4.2 Study design

With this study, we present an extensive characterization of X-inactivation induced imbalances in the X chromosome using tumor samples from females. We use hapLOHseq, a sensitive, haplotype-based statistical approach to detect regions exhibiting allelic imbalance, which is defined as a deviation from the expected 1-to-1 ratio at germline heterozygous loci [1]. As our dataset, we utilize tumor bulk RNA-seq data from females in the TCGA's breast cancer cohort. We assess the chromosome level imbalances to identify X-inactivation induced imbalance patterns by comparing against a DNA - SNP array - based imbalance call set. Finally, we assess these patterns of chromosomal imbalance events that are exclusive to RNA and investigate whether an association between these chromosome level imbalances in the X chromosome and clinical features exist.

4.3 Materials and methods

4.3.1 Dataset

In the same manner described in chapter 2, bulk RNA-seq BAM files (hg38 genome build) and level 1 raw CEL files from the Affymetrix Genome-Wide Human SNP Array 6.0 (hg38 genome build) of 616 tumor samples from females in the TCGA breast cancer cohort were obtained along with the clinical information.

The level 1 raw CEL files of the matched-blood (hg19 genome build) samples were downloaded for the purpose of genotype imputation. As hg38 genome build was not an option for performing genotype imputation of the X chromosome by the Michigan Imputation Server, the hg19 genome build was chosen instead to impute the genotypes.

4.3.2 Processing of the tumor RNA-seq data

4.3.2.1 Genotyping and phasing

Our approach for SCNA detection depends on the allele specific signals at germline heterozygous sites and the germline genotypes can be obtained from (i) the tumor sample directly or from (ii) a matched-normal sample. In the study explained in chapter 2, we investigated the utility of using these two different sources for obtaining germline genotypes and demonstrated that the latter approach yields to an improved performance. Therefore, given the availability of the SNP DNA microarray data from matched-normal samples, the latter approach was preferred for investigating the detectible X inactivation patterns in the BRCA cohort.

We followed the same steps as for the autosomes (see chapter 2 for details) regarding genotype imputation and additionally, specifically for this study, included only the females, mapping the imputed genotypes from genome build hg19 to hg38, and phased the genotypes ourselves.

4.3.2.2 Detection of allelic imbalance from tumor samples

We used the same protocols described in chapter 2 for detecting allelic imbalance from tumor samples and postprocessing the event calls.

4.4 Results

Using 616 bulk RNA-seq tumor samples from the females in the TCGA BRCA cohort, we identified 629 regions that exhibit allelic imbalance. The smallest, median and largest event sizes

were 303.7 Kb, 143.7 Mb and 152.2 Mb, respectively. As summarized in Fig. 4.1, a large majority of the samples harbored at least one event, whereas only 25 of the samples had no region that exhibited allelic imbalance. Next, we assessed allelic imbalance at the chromosome level; Fig. 4.2 demonstrates the distribution of allelic imbalance burden per sample – defined as the fraction of the X chromosome that harbor allelic imbalance events – for all samples. Out of the 616 samples, burden was 50% or above for 567 and 75% and above for 495 with a median of 93%.

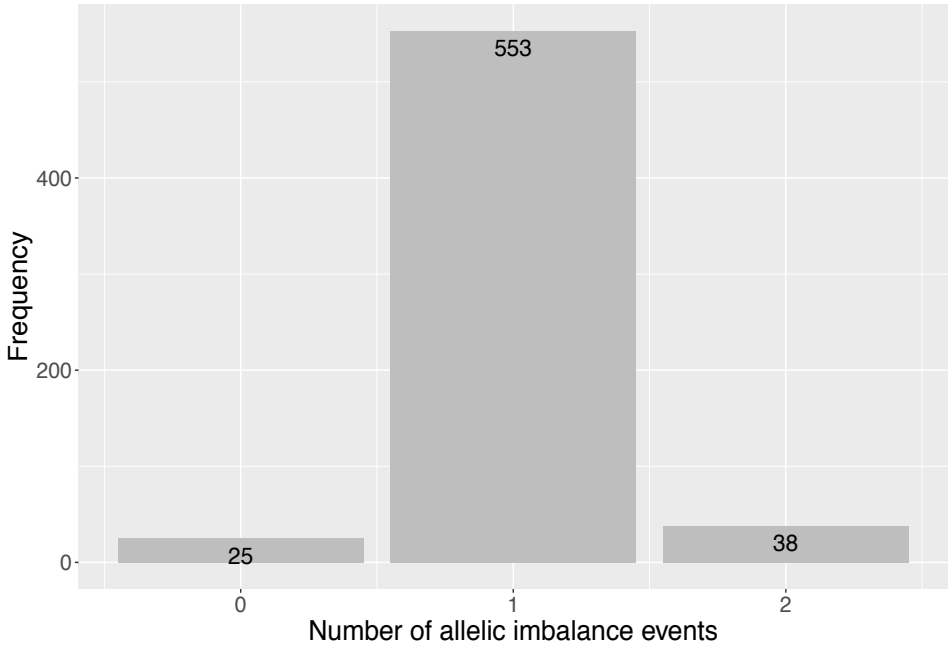


Fig. 4.1. Distribution of number of allelic imbalance events per sample. A barplot with number of allelic imbalance events on the X chromosome per sample across tumor samples from females in the TCGA BRCA cohort is shown.

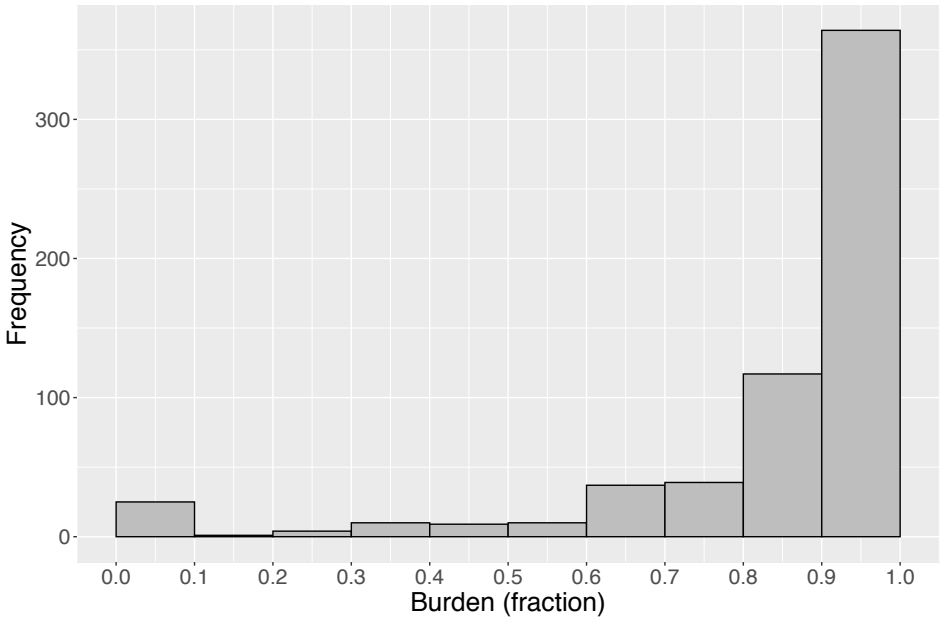


Fig. 4.2. Distribution of allelic imbalance burden per sample. A histogram with allelic imbalance burden per sample - defined as the fraction of the X chromosome spanned by allelic imbalance events - across females in the TCGA BRCA cohort is shown.

4.4.1 Comparison with an autosome for rates of detected imbalance

Following that, as a sanity check, we wanted to have a general conception of the profile of allelic imbalances observed in an autosome. In order to compare the X chromosome with an autosome, we selected chromosome 7, since, among the autosomes, it is the most similar in term of size (159.3 Mb) and the number of protein coding genes (984 genes) [77]. This within platform comparison aimed to eliminate possible technology/ RNA-seq related artifacts when making conclusions specific to the X chromosome.

Using the same set of samples utilized for evaluating the X chromosome (n = 616), we identified 627 allelic imbalance events on chromosome 7 with the median size of 22.8 Mb and the smallest and largest sizes of 10.1 Kb and 157.9 Mb, respectively. In comparison to the X chromosome, the size of the events on chromosome 7 differed noticeably. Despite the similarities between the chromosomes, the events on chromosome 7 tended to be smaller with only 267 out of 627 events larger than 50 Mb, while the events observed on the X chromosome were larger and mostly over 50 Mb (578 out of 629 events) (Fig. 4.3).

Additionally, the number of samples that do not exhibit allelic imbalance on chromosome 7 was noticeably higher with 203 samples (vs. 25 samples for the X chromosome) and a larger portion of the cohort with 158 samples harbored multiple events (Fig. 4.4) (vs. 38 samples for the X chromosome). At the chromosome level, allelic imbalance burden was significantly lower with a median of 26%. Out of 616 samples, 214 had at least 50% burden and 124 of these had at least 75% burden. Fig. 4.5 demonstrates allelic imbalance burden per sample distribution for chromosome 7. In conclusion, the comparative analysis between the X chromosome and chromosome 7 suggests that the allelic imbalance events exhibited by the X chromosome are largely driven by X-inactivation.

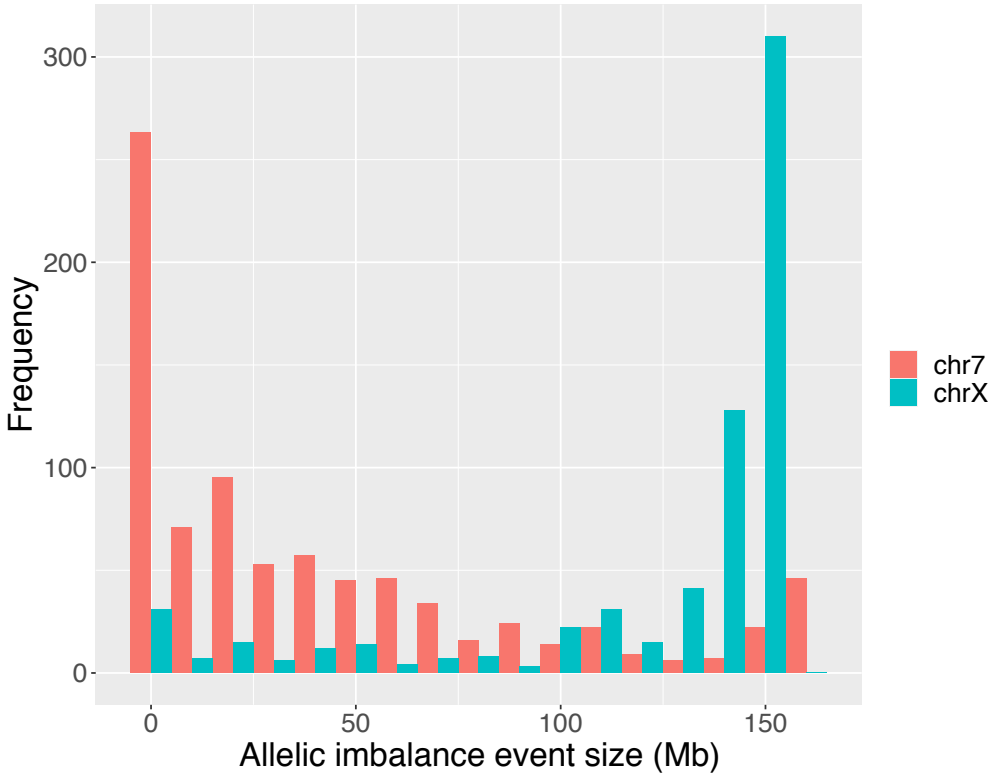


Fig. 4.3. Distribution of allelic imbalance event size. A histogram comparing the sizes of the allelic imbalance events detected on the X chromosome and chromosome 7 across females in the TCGA BRCA cohort is shown.

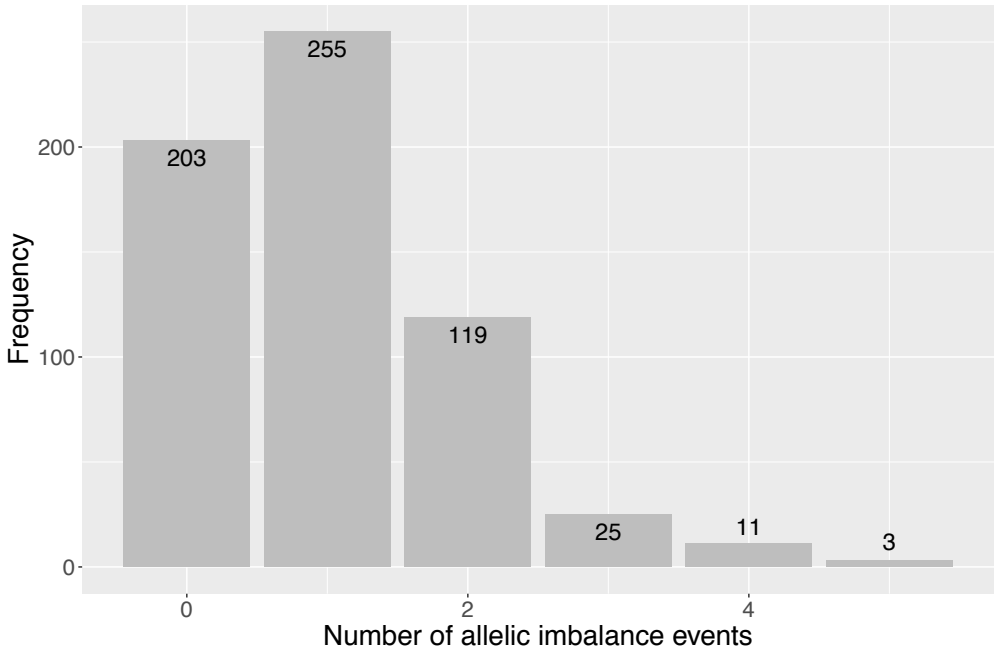


Fig. 4.4. Distribution of number of allelic imbalance events on chromosome 7 per sample. A barplot of the number of allelic imbalance events on chromosome 7 per sample across females in the TCGA BRCA cohort is shown.

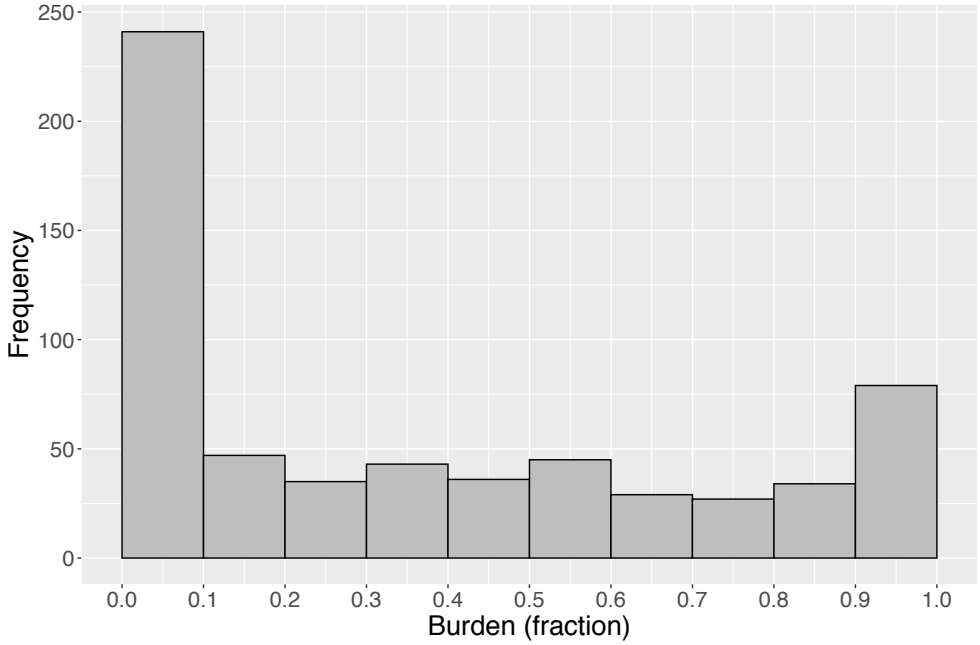


Fig. 4.5. Distribution of allelic imbalance burden per sample for chromosome 7. A histogram with allelic imbalance burden for chromosome 7 per sample - defined as the fraction of chromosome 7 spanned by allelic imbalance events - across females in the TCGA BRCA cohort is shown.

4.4.2 Comparison with SNP array allelic imbalance calls

As discussed more in depth in chapter 2, allelic imbalance events derived from a high density SNP microarray from matched-tumor DNA samples constituted our gold standard set when benchmarking the RNA-seq based allelic imbalance set of calls on the autosomes. Likewise, in this chapter, we used the SNP array based gold standard calls located on the X chromosome for the assessment of the RNA-seq derived calls.

Using SNP array data from the same set of tumor samples (n=616), we identified 682 allelic imbalance events. While 267 samples did not have an allelic imbalance event on the X chromosome, we detected 193 samples that exhibit a single allelic imbalance event and 156 samples with at least two events (per sample). The number of events per sample is demonstrated in Fig. 4.6.

The smallest, median, and largest events were 147.3 Kb, 8.3 Mb, and 155.7 Mb, respectively. By comparing the size of the events on the X chromosome among the SNP array and RNA-seq derived set of calls, we observed that, in general, the gold standard calls were smaller (Fig. 4.7), with 272 out of 682 events larger than 50 Mb (vs. 578 out of 629 events from RNA-seq call set) and 213 events above 75 Mb (vs. 558 events from RNA-seq call set). In terms of the distribution of the size of the calls, the calls on the X chromosome derived from SNP array appeared to be similar to the RNA-seq derived calls on chromosome 7, in contrast to the size of the RNA-seq derived calls on the X chromosome.

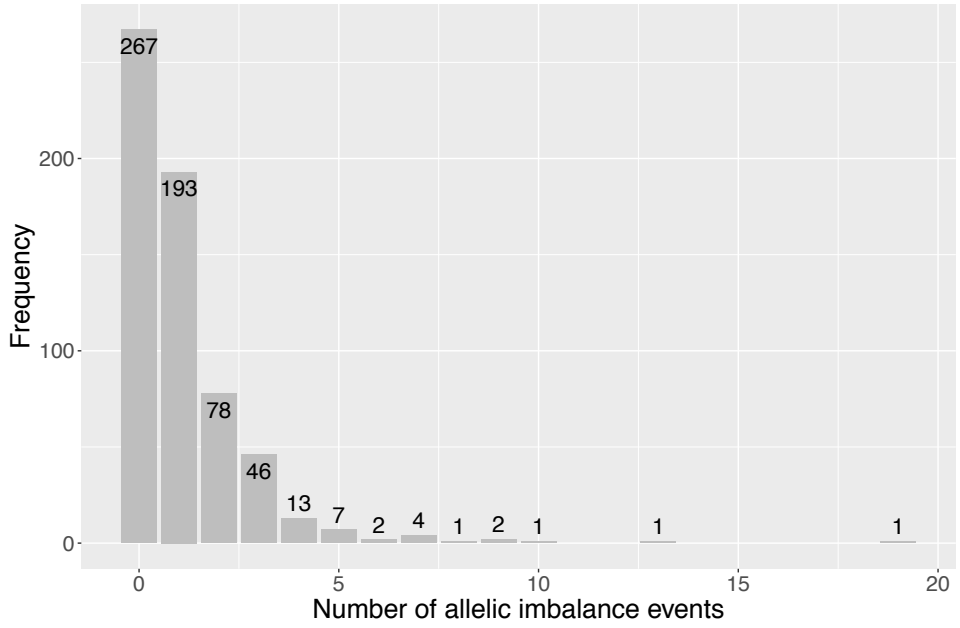


Fig. 4.6. Distribution of number of allelic imbalance events on the X chromosome per sample. A barplot of the number of allelic imbalance events on the X chromosome per sample from SNP array - DNA - derived gold standard call set across females in the TCGA BRCA cohort is shown.

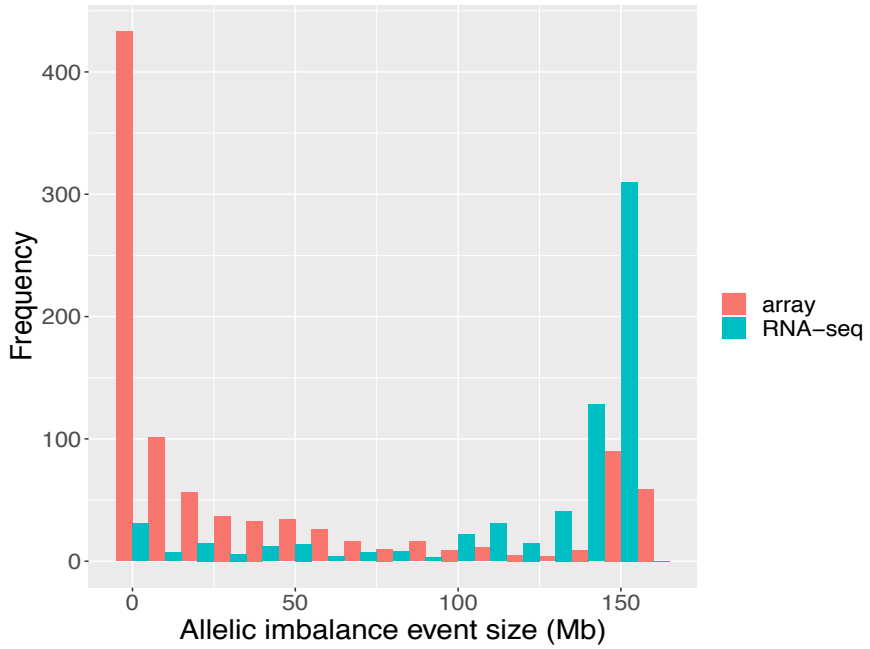


Fig. 4.7. Distribution of allelic imbalance event size. A histogram of the sizes of the allelic imbalance events detected on the X chromosome across females in the TCGA BRCA cohort contrasting RNA-seq and SNP array (gold standard) - DNA - derived call sets is shown.

At the chromosome level, SNP array derived burden exhibited a sharply bimodal distribution, with a peak centered around 5% and a second peak centered around 95% (Fig. 4.8). In comparison to RNA-seq, which had only few samples around 5% and most of the samples above 50% burden level, this distribution is distinct and drastically different. In fact, the distribution showed a higher similarity to that of RNA-seq derived burden of chromosome 7, with the exception of having a much higher peak at 95%.

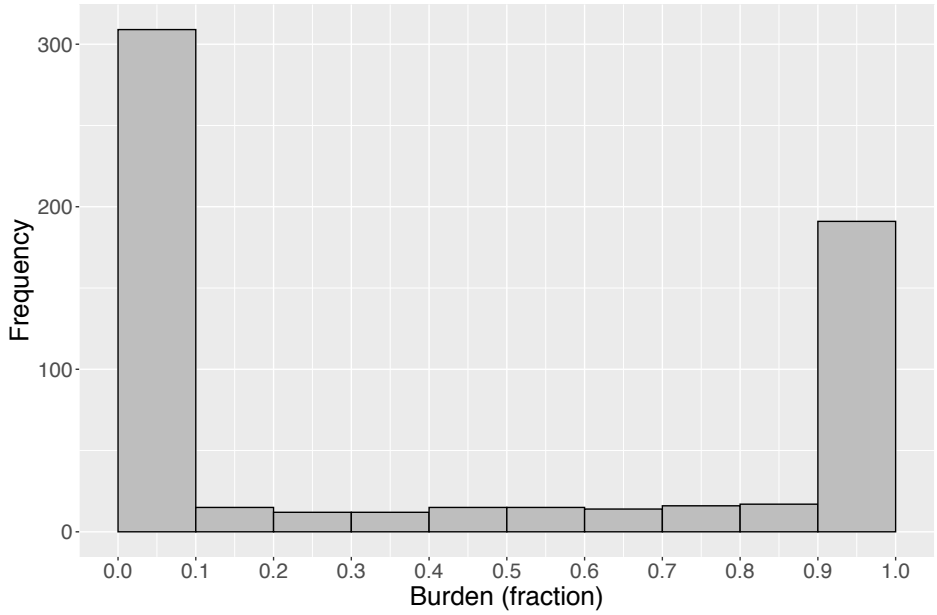


Fig. 4.8. Distribution of allelic imbalance burden per sample for the X chromosome. A histogram with allelic imbalance burden for the X chromosome per sample – defined as the fraction of the X chromosome spanned by allelic imbalance events – from SNP array (DNA) derived gold standard call set across females in the TCGA BRCA cohort is shown.

4.4.3 X-inactivation driven RNA-exclusive allelic imbalance

Subsequently, we investigated the 'RNA-exclusive' allelic imbalance events on the X chromosome. These events are the allelic imbalance events that were detected only from RNA and not from DNA (SNP array), probably due to the X-inactivation and epigenetic regulations. To define an X chromosome as RNA-exclusive, we used a cutoff to determine if majority of it had been spanned by allelic imbalance calls exclusive to the RNA and not by the DNA-based gold standard set of allelic imbalance calls (Appendix Fig. A2). Normally, the calls identified by RNA-

seq that were not detected in the gold standard were treated as putative FPs in our analyses (see chapter 2 for details). However, in theory, these calls might be true RNA-exclusive calls produced by mechanisms that can only be detected from RNA. After exploring different cutoffs, we decided to consider an X chromosome as RNA-exclusive if at least 80% of the chromosome had been spanned by RNA-seq derived allelic imbalance events and, at the same time, less than 20% of the chromosome had been spanned by the gold standard events. According to this definition, we classified 245 samples (out of 616) as having RNA-exclusive X chromosome driven by X-inactivation.

4.4.3.1 Association between the RNA-exclusive allelic imbalance and clinical features

There have been studies suggesting that loss of X chromosome inactivation plays a key role in carcinogenesis in females through increasing the expression of oncogenes - enabled via the extra active X chromosome copy, as well as through mechanisms that impact the course of the disease and become a factor in treatment options and outcome.

In order to probe the existence of possible clinical outcomes resulting from the varying degree of X-inactivation and expression of the genes that escaped X-inactivation, we assessed the association between having an RNA-exclusive X chromosome and several clinical features.

4.4.3.1.1 RNA-exclusive X chromosome and age

First, we investigated the link between age and the RNA-exclusive X chromosome. Taking the percent of new cases by age group in female breast cancer [78] into account, we categorized the samples into 3 groups. The number of samples with and without an RNA-exclusive X chromosome in each age category is summarized in Table 4.1. Although the proportion of samples with RNA-exclusive X chromosomes increased with age, when a Kruskal-Wallis rank sum test was performed to examine the differences between age groups based on the RNA-exclusive X

chromosome status, no statistically significant differences were found (chi-squared=3.47, p-value=0.18) among the 3 age categories (age < 45, 45 ≤ age < 75, and age ≥ 75).

RNA-exclusive chrX	Age < 45	45 ≤ Age < 75	Age ≥ 75
No	46 (68)	245 (60)	32 (52)
Yes	22 (32)	162 (40)	30 (48)
Total n	68 (100)	407 (100)	62 (100)

Table 4.1. RNA-exclusive chromosome X status and age. The number and percentage (in parenthesis) of samples that do and do not have an RNA-exclusive X chromosome distributed by age category is shown.

4.4.3.1.2 RNA-exclusive X chromosome and tumor stage

Second, we examined the differences between breast cancer stages in regard to the status of having an RNA-exclusive X chromosome. We placed the patients in the TCGA BRCA cohort in one of the 4 groups: stage I (stages I, IA, and IB), stage II (stages II, IIA, and IIB), stage III (stages III, IIIA, IIIB, and IIIC), and stage IV. The number of samples with and without an RNA-exclusive X chromosome for each cancer stage category is shown in Table 4.2. It appeared that as the cancer progresses, the samples with an RNA-exclusive X chromosome became less common. Indeed, a Kruskal-Wallis rank sum test showed that, the differences among the tumor stage categories (stages I, II, III, and IV) were statistically significant (chi-squared = 7.55, df = 3, p-value = 0.05). Moreover, stronger evidence suggesting a link between advancing tumor stage and negative RNA-exclusive chromosome X status was found when patients were grouped into early (stages I & II) and late (stages III & IV) stages with a p-value of 0.007 (chi-squared = 7.18, df = 1).

RNA-exclusive chrX	Stage I	Stage II	Stage III	Stage IV
No	52 (54)	170 (58)	90 (70)	5 (71)
Yes	44 (46)	125 (42)	39 (30)	2 (29)
Total n	96 (100)	295 (100)	129 (100)	7 (100)

Table 4.2. RNA-exclusive chromosome X status and tumor stage. The number and percentage (in parenthesis) of samples that do and do not have an RNA-exclusive X chromosome distributed by cancer stage category is shown.

Furthermore, we fit a logistic regression model including age (discrete variable: age < 45 as "1", 45 ≤ age < 75 as "2", and age ≥ 75 as "3") and tumor stage (stages I, II, III, and IV grouped numerically as discrete variables) as covariates. The results showed that, accounting for tumor stage, the odds of having an RNA-exclusive X chromosome increased by 43% (95% CI [0.95, 2.16]) for patients in older age categories (p-value = 0.08). Accounting for age, the odds of having an RNA-exclusive X chromosome decreased by 50% (95% CI [1.12, 2.01]) for patients with later tumor stages (p-value = 0.007).

4.4.3.1.3 RNA-exclusive X chromosome and overall survival

Third, we evaluated the impact of having an RNA-exclusive X chromosome on overall survival. We estimated the overall survival using the Kaplan Meier method and assessed the differences between curves with the log rank test. As the list of samples only included European females in the TCGA BRCA cohort already, there was no need to perform additional adjustment for the effects of possible confounding factors for race and gender. We did not observe a statistically significant difference between the groups that do and do not exhibit an RNA-exclusive X chromosome in terms of survival (Fig. 4.9; p-value=0.23).

Next, we used a multivariate Cox proportional hazards model for comparative survival analysis including RNA-exclusive chromosome X status as categorical and tumor stage and age as quantitative predictor variables (Fig. 4.10). The results indicated that tumor stage and age were

associated with an effect on prognosis with a significantly worse overall survival prognosis, whereas RNA-exclusive chromosome X status did not have a significant effect on prognosis. The p-value for RNA-exclusive chromosome X status was 0.49, with a hazard ratio (HR) of 0.78, indicating an insignificant relationship between the RNA-exclusive chromosome X status and decreased risk of death and association with a better survival. The p-value for tumor stage was < 0.001, with a HR = 2.74, which implies a significant association between tumor stage and increased risk of death and a poor survival. The p-value for age was 0.011, with a HR = 2.43, which suggests a significant link between age and increased risk of death and a poor survival.

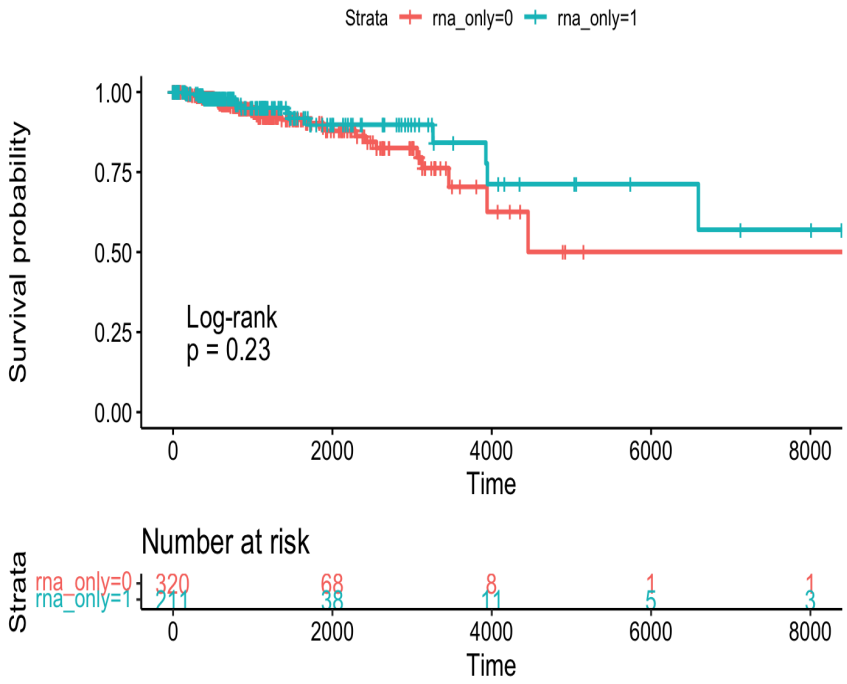


Fig. 4.9. The effect of RNA-exclusive chromosome X status on overall survival. Kaplan-Meier curve stratified according to RNA-exclusive chromosome X status is shown. RNA-exclusive = 1 indicates the status of having and RNA-exclusive = 0 indicates the status of not having an RNA-exclusive chromosome X.

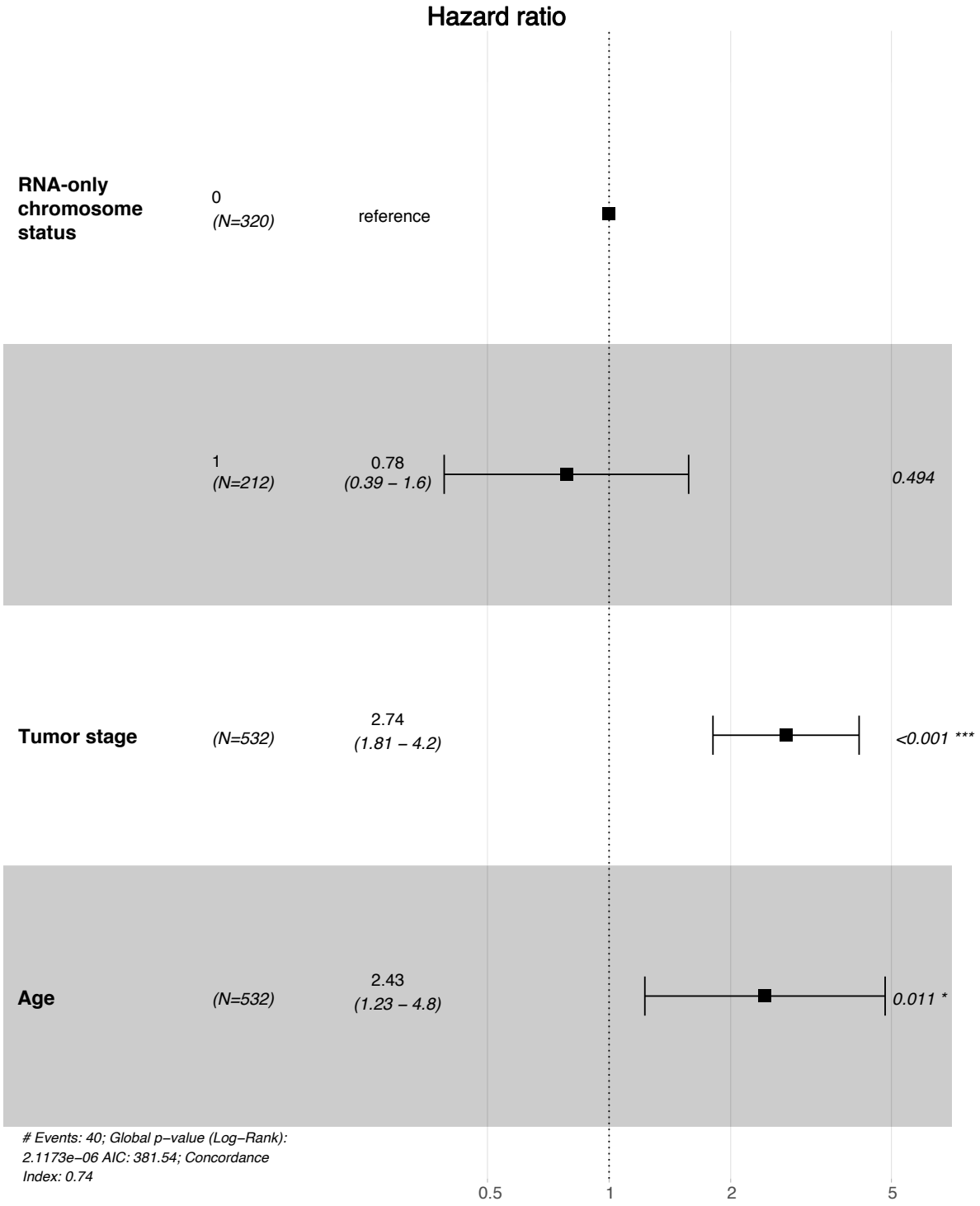


Fig. 4.10. Cox proportional hazards model. A forest plot of risk of death and respective hazard ratios derived from the Cox proportional hazards model for covariates RNA-exclusive chromosome status, tumor stage, and age category is shown.

4.4.3.1.4 RNA-exclusive X chromosome and BRCA molecular subtypes

Studies have suggested that the loss of X chromosome inactivation is more commonly observed in basal-like subtype of breast cancer [74–76]. We studied whether there is an association between the status of having an RNA-exclusive X chromosome and molecular subtypes. In regions where the loss of X chromosome inactivation is observed, two active copies of the X chromosome yield to a balanced expression of the parental alleles, leading to not having an RNA-exclusive X chromosome status. Therefore, we expected to observe RNA-exclusive X chromosome status to be less common among the samples coming from patients in the basal-like subtype category. To test this, we performed a chi-square test on the 251 samples for which we could obtain the molecular subtype information. After placing the samples in either basal-like or non-basal-like subtype categories, we observed that the samples that belong to the patients in the basal-like subtype category had the RNA-exclusive X chromosome status less frequently than the patients in the non-basal-like subtype category and the difference among the two groups is significant in the fraction of RNA-exclusive X chromosome status (chi-squared = 12.34, df = 1, p-value = 0.0004).

4.4.3.1.5 RNA-exclusive X chromosome and *BRCA1* carrier status

Studies propose that *BRCA1* plays an important role in X-inactivation through its influence on the concentration of *XIST* RNA [79,80] and its dysfunction may result in failure of the maintenance of X-inactivation by causing a disruption in the inactive X chromosome's association with partner proteins [81]. However, it remains unclear whether the impact of *BRCA1* loss on the inactivation would cause reactivation of the entire X chromosome or specific genes on the chromosome. We evaluated the existence of a link between the status of having an RNA-exclusive X chromosome and being a *BRCA1* carrier. Normally, in females, the epigenetic silencing of one of the X chromosomes through X-inactivation leads to an imbalance that is detectable through RNA

analysis and not DNA. Although X-inactivation causes an imbalance, the loss of it would lead to two active copies, which results in a balanced expression of the parental alleles, and thus, not having an RNA-exclusive X chromosome. As a consequence, we anticipated detecting fewer RNA-exclusive X chromosomes among the *BRCA1* carriers. In order to test this hypothesis, we dichotomized the samples into *BRCA1* carrier and non-carrier categories and placed only the samples carrying pathogenic variants into the carrier category. We performed Fisher's exact test on the 592 samples that the carrier status information was available on ClinVar [82] and found that *BRCA1* carriers tended to not have an RNA-exclusive X chromosome (9 out of 10 carriers), although it was not statistically significant (p-value of 0.09).

4.5 Discussion

The X chromosome inactivation in female mammals is an important epigenetic phenomenon that ensures dosage compensation between males and females. During the early development, one of the two X chromosomes is transcriptionally silenced arbitrarily, which results in the formation of the heterochromatin Barr body. Escapees, which make up 15% of the X-linked genes, are the genes that escape the transcriptional silencing and are expressed biallelically, thus have higher expression levels in females. In addition, another 10% of the X-lined genes escape X-inactivation at varying levels [83,84]. Although the importance of dosage compensation has been suggested, the effects the loss of X chromosome inactivation would have on human health have not been investigated sufficiently. In the literature, breast cancer cells, especially the aggressive forms, have been noted to miss the Barr body because of the loss of X chromosome inactivation leading to double the expected dose of genes on the X chromosome. Decompression of the Barr body's heterochromatin structure, succeeded by reactivation of the X-linked genes has been also proposed to explain the Barr body loss in cancer [85–88]. In addition, it has been suggested that

BRCA1 has a direct role in the loss of X chromosome inactivation through regulation of *XIST* expression and epigenetic relaxation of the inactive X chromosome [79]. The fact that defects in either genetic or epigenetic mechanisms may impact X chromosome inactivation necessitates conducting additional research on the topic to further understand the effect of loss of X chromosome inactivation in carcinogenesis.

In this study, we investigated the X chromosome inactivation's role in female cancers by utilizing the TCGA breast cancer data. Using 616 bulk RNA-seq tumor samples from females, we first identified allelic imbalance events on the X chromosome and then by contrasting these RNA-seq derived calls with the SNP array - DNA - derived allelic imbalance calls, we determined the X inactivation driven allelic imbalance events that are observed exclusively in RNA. In order to elaborate the effects of the loss of X inactivation in female tumorigenesis, we examined the association between certain clinical features and the binary status of having an RNA-exclusive X chromosome. For each sample, after calculating the fraction of the X chromosome affected by allelic imbalance events, we considered a sample as having an RNA-exclusive X chromosome if no more than the sample's X chromosome had been spanned by DNA derived calls and at least 80% of it exhibited allelic imbalance events, based on RNA.

We found that, out of 616 samples, only 25 samples had no allelic imbalance events detected from RNA. 567 samples had 50% and above and 495 samples had 75% and above allelic imbalance burden (considering only the X chromosome) with 93% median allelic imbalance burden. Next, we compared the findings to that of RNA derived chromosome 7, the autosome that is the most similar to the X chromosome in terms of size and number of genes it contains. The purpose of this within platform comparison was to identify the technology/ RNA-seq related artifacts, e.g., lack of coverage, if there are any when making conclusions specific to the X chromosome. We found that a much higher fraction of the samples (203 out of 616) did not have

any allelic imbalance events on chromosome 7 and the burden was noticeably lower with a median of 26%. While 214 samples had 50% or more burden, 124 samples had more than 75%. This comparison suggests that the rates of allelic imbalance burden are elevated in the X chromosome compared to the autosomes. We also evaluated the allelic imbalance profiles of the X chromosome detected via SNP array from matching tumor DNA. The findings were similar to that of chromosome 7 with 267 of the samples not exhibiting any allelic imbalance events and overall events were smaller in comparison to RNA. Comparisons at the chromosome level (allelic imbalance burden) showed that the SNP array derived burden had a bimodal distribution with a peak centered at 5% and a second peak centered at 95%, which is drastically different from the RNA-seq derived burden distribution with only few samples at 5% burden and most samples over 50%.

Subsequently, we probed the association between RNA-exclusive X chromosome status and physiological age to test if observations from peripheral blood studies that suggest age related increase in the levels of X chromosome inactivation skewing [89,90] holds in tissues besides blood. A Kruskal-Wallis rank sum test indicated no statistically significant difference among the age categories based on the RNA-exclusive X chromosome status. Next, we investigated the relationship between breast cancer stages in regard to the status of having an RNA-exclusive X chromosome. A Kruskal-Wallis rank sum test indicated a link between advancing tumor stage and negative RNA-exclusive chromosome X status when patients were placed in early (stages I & II) and late (stages III & IV) stages with a p-value of 0.007. These results agree with the findings in the literature and validate an association between the loss of X chromosome inactivation and more aggressive behavior in breast cancer. To examine this further, we also assessed the link between overall survival and the status of RNA-exclusive chromosome X. However, our results did not indicate a statistically significant association, although the patients in the negative RNA-

exclusive chromosome X status category were associated with worse overall survival. Additionally, we assessed the validity of the hypotheses that suggest the loss of X chromosome inactivation is more frequently observed in basal-like subtype and among *BRCA1* carriers. We recapitulated the former (chi-square test p-value = 0.0004), although we did not observe a statistically significant difference in the results regarding the latter (Fisher's exact test p-value = 0.09). In comparison to the anticipated rate of 3-5% for breast cancers associated with *BRCA1* germline pathogenic variant [91], the low rate of pathogenic *BRCA1* variants observed in the TCGA BRCA cohort was a limiting factor. Instead of using only ClinVar for annotation, other sources, such as the ARUP database [92] and the BRCA Exchange site [93], could be taken into consideration.

We note that, in this study, we were restricted to the samples from patients with European ancestry in order to use the HRC reference panel for genotype imputation, thus our results are not inclusive of all races. Another limitation of the study is, using the current cutoffs for defining RNA-exclusivity status, we cannot identify RNA-exclusive X chromosome status of a sample if greater than 20% of its X chromosome is spanned by allelic imbalance events detected through DNA. In order to enable flexible cutoffs when determining whether an X chromosome is RNA-exclusive or not, we used subtraction of DNA-derived chromosome X burden from that of RNA-derived and required it to be greater than 50%. This did not alter the results and conclusions drawn. We also assessed the clinical relationships after excluding those with high DNA burden where we could not assess whether RNA-exclusive changes occurred. After excluding the 292 high DNA burden samples, 245 sample did and 79 samples did not have an RNA-exclusive X chromosome. We still did not detect a statistically significant correlation between age and RNA-exclusive X chromosome status. After the exclusion, a Kruskal-Wallis rank sum test showed that the differences among the tumor stage categories were not statistically significant (p-val = 0.13)

although after accounting for age, the odds of having an RNA-exclusive X chromosome decreased by 9% (95% CI [0.69, 1.73]) for patients with later tumor stages with a marginally significant p-value of 0.07. We did not observe a statistically significant difference between the groups that do and do not exhibit an RNA-exclusive X chromosome in terms of survival (p-value=0.26). Lastly, the set was not amenable to molecular subtype and *BRCA1* carrier status association analyses because exclusion of the samples resulted in a small dataset which was not suitable, e.g., there were only 2 *BRCA1* carriers. Hence, when performing RNA-exclusive X chromosome association analysis on cohorts with high allelic imbalance levels from DNA, one needs to be aware of such limitations. Additionally, several studies reported the gain of an extra active X chromosome along with the lack of the inactive X chromosome in some breast cancers [94–96]. These findings suggest that although loss of X chromosome inactivation might have taken place, with the gain of an extra active X chromosome, allelic imbalance may still be observed. Therefore, using lack of imbalance might not always be enough and additional analysis, such as integration of phase concordance information, i.e., to estimate haplotype similarity, might be necessary to confirm loss of X chromosome inactivation. If it has taken place, the frequency based - excess - haplotype should bear resemblance to the germline haplotypes and not just reflect stochastic deviations.

As future directions, a natural extension of this study is to analyze the TCGA OV cohort to compare and contrast with the findings from the TCGA BRCA cohort. In addition, the association analysis performed for the *BRCA1* carriers could be repeated on the *BRCA2* carriers to test if RNA-exclusive X chromosome status is observed less frequently among the *BRCA2* carriers as well. Another natural extension to this work is performing scRNA-seq and epigenetics analyses from matched samples to validate and better contextualize the results.

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

As one of the hallmarks of cancer, genomic instability leads to alterations in DNA, including CNAs, which often affect extensive portions of the genome and span large numbers of genes. Allelic imbalance is a signature of genomic instability since most SCNAs cause a deviation from the expected 1:1 paternal-to-maternal allelic ratio. In addition to having an important role in carcinogenesis, allelic imbalance has also been linked to clinical features and outcomes. Thus, identification and comprehensive characterization of allelic imbalance may serve as a key step in shedding light on initiation and progression of cancer and may be used as a prognostic indicator. Despite the fact that RNA-seq may be the only available platform in certain settings, i.e., where tumor material or funding is limited, allelic imbalance detection from bulk RNA-seq data has not been well developed, perhaps due to the challenges caused by properties innate to RNA, such as non-uniform coverage of the genome which complicates distinguishing between allelic imbalance and dynamically varying gene expression.

5.1. Overall significance

In this dissertation, I discuss adaptation and application of hapLOHseq, a software originally developed by San Lucas et al. for detection of allelic imbalance events from exome sequencing, to bulk RNA-seq and also incorporation of genotypes imputed from a matched normal sample, into the hapLOHseq pipeline in order to increase the quality of estimated genotypes and haplotypes.

Adaptation of hapLOHseq to RNA-seq data makes the approach the first to leverage haplotypes estimates to infer large scale allelic imbalances from bulk RNA-seq data. Previous studies with SNP array data have demonstrated that utilizing haplotype information improves power for detection of allelic imbalance [1,3,53,54]. By utilizing haplotypes, hapLOHseq models signals at multiple markers jointly, thus requiring signals to be sustained beyond several mRNA transcripts, which may be regulated by factors innate to RNA and not caused by SCNAs.

Germline genotypes for input into hapLOHseq can be derived from: (i) tumor RNA-seq directly or (ii) SNP array or an alternative DNA sequencing platform with genome wide coverage ran on a matched normal sample. For the latter, easily collectible blood or buccal samples may be used to derive germline genotypes. SNP array data from these matching normals provides an economical approach to boost allelic imbalance detection from tumor samples. This option is especially feasible for existing large data repositories with well phenotyped cohorts, such as TCGA, TARGET, and ICGC. Subsequent to obtaining germline haplotypes, hapLOHseq quantifies evidence for allelic imbalance in a way that is robust to errors in phasing by exploiting switch consistency metric, i.e., switch consistencies are averaged to calculate phase concordance for a region. In addition to utilizing haplotype estimates for allelic imbalance inference and providing avenues for joint analysis of expression and CNAs, the approach is also novel in the sense that it uses imputed genotypes derived from SNP array data from a matched normal for obtaining highly accurate and phased genotypes, which hapLOHseq relies on.

In chapter 2, through application to 4,942 tumor samples across 28 cancer sites in the TCGA, I demonstrate that hapLOHseq is a highly effective method for detecting somatic megabase level copy number alterations from RNA in scenarios where data from germline DNA is available. To benchmark the method, we compared RNA derived events against those derived from SNP array (DNA), which were treated as a gold standard. In this study, the method achieved 85% median

sensitivity and 95% median specificity at the gene level. Evaluations at the chromosome arm and genome levels indicated performances similar to that of gene level. Furthermore, to show the method's utility in clinical settings, we successfully replicated subtype specific copy number alteration features associated with breast cancer. Comprehensive comparisons with other methods for detection of copy number alteration from bulk RNA-seq data, such as CaSPeR and SuperFreq, showed that hapLOHseq achieves significantly higher sensitivities and specificities at the gene level for the samples in the BRCA and GBM cohorts in comparison to CaSPeR. At the gene level, hapLOHseq performs similarly in comparison to SuperFreq for a subset of samples in the BRCA cohort that the sample sets overlap. As of note, performance characteristics heavily depend on the gold standard used, hence an absolute performance comparison entails improvements in the gold standard datasets. Overall, hapLOHseq enables robust identification of large allelic imbalances that represent gain, loss, or cn-LOH events, and a more in-depth tumor profiling when additional DNA analysis of the tumor sample is not feasible.

In chapter 4, I provide characterization of allelic imbalances harbored by the X chromosome specifically and investigate the role of X chromosome inactivation in female cancers through utilizing tumor samples from females in the BRCA cohort. A contrast of the X chromosome allelic imbalance profiles between RNA and DNA indicated that the rates of allelic imbalance burden are elevated for RNA when compared to the gold standard, presumably due to the epigenetic mechanisms underlying X-inactivation. To prevent artifacts that might have been inherent to RNA-seq from interfering with the results, we performed a within platform comparison by contrasting RNA derived calls exhibited by the X chromosome to that of an autosome (chromosome 7), which was chosen because of its similarity to the X chromosome in terms of size and the number of genes. Detection of high rates of allelic imbalance burden from the X chromosome (*vs.* chromosome 7) confirmed that the observations were not caused by the

sequencing technology or reasons inherent to RNA. Next, we investigated the link between varying degrees of X-inactivation, including loss of it, and clinical features. Fitting a logistic regression model to the data showed that after accounting for age, the odds of having an RNA-exclusive X chromosome (induced by X-inactivation) decreased by 50% (95% CI [1.12, 2.01]) for patients with later tumor stages (p-value = 0.007). In addition, compared to the samples from patients in non-basal-like subtype category, fewer RNA-exclusive X chromosomes have been observed from the samples in the basal-like subtype category (chi-squared = 12.34, df = 1, p-value = 0.0004). These findings are in line with studies that suggest more frequent observations of loss of X chromosome inactivation among basal-like subtype and more aggressive behavior in breast cancer. On the other hand, we did not observe a significant association between age, overall survival, and *BRCA1* carrier status and RNA-exclusive chromosome X status, although a marginally significant association (p-value=0.09) was detected for *BRCA1* carrier status. Investigations discussed in this chapter are an example of additional utilities that come naturally with RNA, e.g., to assess the role of epigenetically regulated mechanisms affecting large genomics regions in carcinogenesis, besides enabling a more comprehensive characterization of tumor samples through copy number inference when RNA is the only nucleic acid source for a tumor sample. Our future plans include analyzing samples in the TCGA ovarian cohort, repeating the association analysis performed for the *BRCA1* carriers on the *BRCA2* carriers, and utilizing scRNA-seq and epigenetics analyses from matched samples to validate some of our results and further explore the role of X-inactivation in female carcinogenesis.

5.2. Future directions

5.2.1. Combine hapLOHseq with other SCNA detection softwares

With the integration of imputation into the hapLOHseq pipeline to obtain germline genotypes, the approach provides high accuracy, phased germline genotypes and addresses several issues related to genotype calling from RNA. Other methods developed for detecting SCNAs from bulk RNA-seq, such as SuperFreq and CaSpER, could benefit from our germline heterozygote identification as the information utilized by these methods, i.e., read counts and BAF dispersions, is orthogonal to what hapLOHseq leverages. Therefore, applying hapLOHseq in combination with alternative SCNA detection methods or integrating them for joint analyses is an area of future study.

5.2.2. Determine copy number status of allelic imbalance segments

As of now, our approach outputs genomic locations of allelic imbalance events but does not inform about their copy number status, e.g., gain, loss, or cn-LOH. Determining copy number status of allelic imbalance segments is among future research directions. For this purpose, changes in read coverage observed in a region of allelic imbalance can be utilized to inform about the copy number status of that allelic imbalance event. Briefly, the underlying idea is that copy number gains result in increased read coverage, whereas copy number losses lead to decreased read coverage in comparison to baseline. To create a baseline, one may take advantage of normal samples, such as normal adjacent to tumor samples for the same tissue in the TCGA, and calculate the read depth of the regions corresponding to allelic imbalance segments from the tumor (subsetting heterozygous markers only). After comparing the read depths between tumor and normal, lower read depths (from tumor) would be categorized as loss and higher read depths would be categorized as gain. If the read depth is not deviated from the baseline but the region has a prominent BAF deviation, it would be categorized as cn-LOH.

Currently, our approach cannot identify balanced duplications, e.g., AABB, when there is no deviation from the expected 1-to-1 ratio. In addition to using coverage data to determine the copy number status of allelic imbalance segments, it may also be integrated into to the approach to handle this issue.

As of note, although all RNA-seq based approaches will be limited in regard to identifying alterations that do not span expressed genes, the limitations are mitigated over large genomic regions.

5.2.3. WES and RNA-seq joint analysis

Although not typically, both RNA-seq and WES data from the same tumor sample may occasionally be available. Under these circumstances, an integrative strategy may be employed to improve allelic imbalance calling performance, especially for genomic regions where heterozygous markers are sparsely distributed. Our analysis drew attention to certain scenarios where true allelic imbalance events were picked up by RNA or WES (mutually exclusive) and other scenarios where true events were subtly implied by both platforms but were not called by either because posterior probabilities did not reach the cutoff. In these settings, an alternative could simply be to combine allelic depths for reference and alternate alleles obtained from RNA-seq and exome sequencing at the marker level and to resume the regular hapLOHseq workflow thereafter.

5.2.4. Adaptation to single cell RNA-seq data

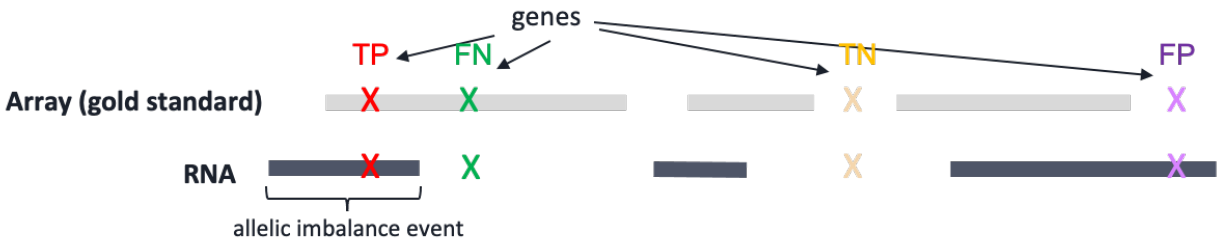
Single cell RNA-seq (scRNA-seq) technology has become increasingly popular and aims to quantify tumor heterogeneity and evolution while enabling detection of rare cell subpopulations with distinct mutations and informing about the evolutionary dynamics of cancer. There already is a bioinformatics tool named CHISEL [97] that utilizes haplotype information in single-cell DNA analysis and their approach shows that haplotype information can improve inference of

SCNAs at single cell level. Motivated by this and lack of methods for allelic imbalance detecting from scRNA-seq data, adapting our approach to scRNA-seq data is a natural area for further development.

5.2.5. Using more inclusive reference panels for increased diversity

In our studies, we restricted ourselves to samples from individuals of European ancestry to evaluate the approach's performance in a best case scenario with the resources at hand. However, recent efforts, i.e., construction of the TOPMed reference panel, enabled generation of genotype panels for individuals with more diverse genetic backgrounds. As part of our future research directions, we plan on conducting the same analysis on a subset of TCGA cohorts, this time not excluding individuals of non-European ancestry. More studies exploiting genotype panels with greater representation of those individuals with diverse genetic backgrounds will lead to more novel insights into tumor initiation and progression and result in improvements in clinical care and understanding of disparities in cancer.

APPENDIX A

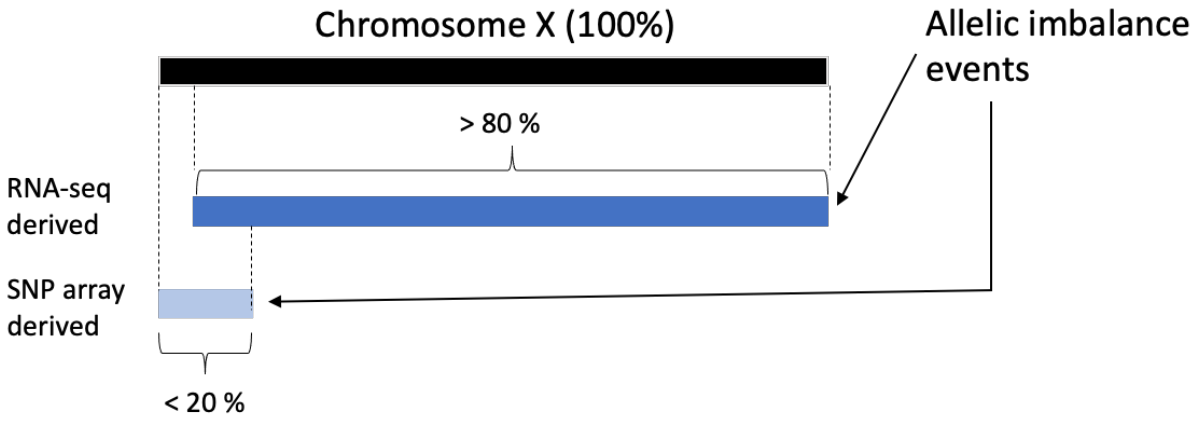


	Gold standard	RNA-seq
TP	1	1
FN	1	0
FP	0	1
TN	0	0

Sensitivity (TPR): TP/(TP+FN)

Specificity (TNR): TN/(TN+FP)

Appendix Fig. A1 Sensitivity and specificity calculation for the gene level assessment.



Appendix Fig. A2 RNA only chromosome X.

APPENDIX B

```
#################
# Event caller
#################
# min_boundary_prob: min post prob for event boundaries
# min_event_prob: min post prob that markers should reach in order to
call an AI event
# when classifying marker postprob:
# 0 if S1 < min_boundary_prob
# 1 if min_boundary_prob <= S1 < min_event_prob
# 2 if min_event_prob <= S1
# Run by: Rscript event_caller.R dir file_prefix min_boundary_prob
min_event_prob

suppressMessages(library(tidyverse))

ARGV <- commandArgs(trailingOnly=TRUE)

dir = ARGV[1]
file_prefix = ARGV[2] # e.g. 'TCGA-V5-A7RE-11A-11R-A354-31'
min_boundary_prob = as.numeric(ARGV[3])
min_event_prob = as.numeric(ARGV[4])

# read posterior.dat
sample_df = read_tsv(file.path(paste0(dir, '/posterior'),
paste0(file_prefix, '.posterior.dat')),
                     col_names = TRUE,
                     skip=0,
                     progress = FALSE,
                     show_col_types = FALSE)

# subset hets
# levels(factor(sample_df$GT))
sample_df_hets = sample_df %>%
  filter(GT == '0/1' | GT == '1/0' | GT == '0/0' | GT == '1/1')
#  filter(GT == '0/1' | GT == '1/0')

# filter post prob = NA
sample_df_hets = sample_df_hets %>%
  filter(!is.na(S1))

# classify markers
sample_df_hets = sample_df_hets %>%
  mutate(postprob.status = ifelse(S1 >= min_boundary_prob & S1 <
min_event_prob, 1,
                                                  ifelse(S1 >=
min_event_prob, 2, 0)))

call_events <- function(chromnum) {

  # subset chromosome and identify events
  sample_df_hets_chr = sample_df_hets %>%
```

```
    filter(CHR == paste0('chr', chromnum))

  first_het_on_chr = sample_df_hets_chr$POS[1]

  # index vectors for 0,1,2
  zero_ind = which(!is.na(match(sample_df_hets_chr$postprob.status, 0)))
  one_ind = which(!is.na(match(sample_df_hets_chr$postprob.status, 1)))
  two_ind = which(!is.na(match(sample_df_hets_chr$postprob.status, 2)))

  if(length(two_ind) != 0) { # if there's an event (2's) in the chr

    # call events
    i=1
    first_two_ind = two_ind[1]

    while (i <= nrow(sample_df_hets_chr)) {

      # first 0 to the left is event start
      max.less = max(zero_ind[zero_ind < first_two_ind])

      if(is.infinite(max.less)) { # if there is no 0 before 2

        event_start = first_het_on_chr # pos of the first het on chr

      } else {

        event_start = sample_df_hets_chr$POS[max.less + 1]
      }

      # first 0 to the right is event stop
      min.greater = min(zero_ind[first_two_ind < zero_ind]) # source of
warnings
      event_end = sample_df_hets_chr$POS[min.greater - 1]

      # if chr doesn't end w/ a 0, then use the last het marker's
position as event end pos
      if (is.na(event_end)) {

        event_end =
sample_df_hets_chr$POS[max(one_ind[length(one_ind)],two_ind[length(two_in
d)])]

        num_markers = which(sample_df_hets_chr$POS == event_end) -
which(sample_df_hets_chr$POS == event_start) + 1
        cat(paste(paste0('chr', chromnum), event_start, event_end,
num_markers, sep = '\t'), '\n')

        break

      } else {

        num_markers = which(sample_df_hets_chr$POS == event_end) -
which(sample_df_hets_chr$POS == event_start) + 1
        cat(paste(paste0('chr', chromnum), event_start, event_end,
num_markers, sep = '\t'), '\n')
```

```
      }

      # update i
      first_two_ind = min(two_ind[min.greater < two_ind])
      i = first_two_ind

    }

  }

}

# main
dir.create(file.path(dir, paste0('/events_boundary', min_boundary_prob,
'_peak', min_event_prob)), showWarnings = FALSE)
sink(paste0(dir, '/events_boundary', min_boundary_prob, '_peak',
min_event_prob, '/', file_prefix, '.bed'), append = TRUE)
cat(paste('#CHR', 'START', 'STOP', 'NUM_MARKERS', sep = '\t'), '\n')

for (j in 1:22) {
  call_events(j)
}

closeAllConnections()
```

BIBLIOGRAPHY

[1] S. Vattathil, P. Scheet, Haplotype-based profiling of subtle allelic imbalance with SNP arrays, Genome Res. 23 (2013) 152–158. https://doi.org/10.1101/gr.141374.112.

[2] Cancer Genome Atlas Network, Comprehensive molecular portraits of human breast tumours, Nature. 490 (2012) 61–70. https://doi.org/10.1038/nature11412.

[3] S. Sivakumar, F.A. San Lucas, Y.A. Jakubek, Z. Ozcan, J. Fowler, P. Scheet, Pan cancer patterns of allelic imbalance from chromosomal alterations in 33 tumor types, Genetics. 217 (2021) 1–12. https://doi.org/10.1093/genetics/iyaa021.

[4] A. Serin Harmanci, A.O. Harmanci, X. Zhou, CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data, Nat. Commun. 11 (2020) 89. https://doi.org/10.1038/s41467-019-13779-x.

[5] Cancer Facts & Figures 2022, (n.d.). https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2022.html (accessed March 7, 2022).

[6] B.D. Preston, T.M. Albertson, A.J. Herr, DNA replication fidelity and cancer, Semin. Cancer Biol. 20 (2010) 281–293. https://doi.org/10.1016/j.semcancer.2010.10.009.

[7] M.R. Stratton, P.J. Campbell, P.A. Futreal, The cancer genome, Nature. 458 (2009) 719–724. https://doi.org/10.1038/nature07943.

[8] S. Negrini, V.G. Gorgoulis, T.D. Halazonetis, Genomic instability--an evolving hallmark of cancer, Nat. Rev. Mol. Cell Biol. 11 (2010) 220–228. https://doi.org/10.1038/nrm2858.

[9] A.G. Knudson Jr, Mutation and cancer: statistical study of retinoblastoma, Proc. Natl. Acad. Sci. U. S. A. 68 (1971) 820–823. https://doi.org/10.1073/pnas.68.4.820.

[10] A.M. Taylor, J. Shih, G. Ha, G.F. Gao, X. Zhang, A.C. Berger, S.E. Schumacher, C. Wang, H. Hu, J. Liu, A.J. Lazar, Cancer Genome Atlas Research Network, A.D. Cherniack, R. Beroukhim, M. Meyerson, Genomic and Functional Approaches to Understanding Cancer Aneuploidy, Cancer Cell. 33 (2018) 676-689.e3. https://doi.org/10.1016/j.ccell.2018.03.007.

[11] R. Beroukhim, C.H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J.S. Boehm, J. Dobson, M. Urashima, K.T. Mc Henry, R.M. Pinchback, A.H. Ligon, Y.-J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M.S. Lawrence, B.A. Weir, K.E. Tanaka, D.Y. Chiang, A.J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F.J. Kaye, H. Sasaki, J.E. Tepper, J.A. Fletcher, J. Tabernero, J. Baselga, M.-S. Tsao, F. Demichelis, M.A. Rubin, P.A. Janne, M.J. Daly, C. Nucera, R.L. Levine, B.L. Ebert, S. Gabriel, A.K. Rustgi, C.R. Antonescu, M. Ladanyi, A. Letai, L.A. Garraway, M. Loda, D.G. Beer, L.D. True, A. Okamoto, S.L. Pomeroy, S. Singer, T.R. Golub, E.S. Lander, G. Getz, W.R. Sellers, M. Meyerson, The landscape of somatic copy-number alteration across human cancers, Nature. 463 (2010) 899–905. https://doi.org/10.1038/nature08822.

[12] T.I. Zack, S.E. Schumacher, S.L. Carter, A.D. Cherniack, G. Saksena, B. Tabak, M.S. Lawrence, C.-Z. Zhsng, J. Wala, C.H. Mermel, C. Sougnez, S.B. Gabriel, B. Hernandez, H. Shen, P.W. Laird, G. Getz, M. Meyerson, R. Beroukhim, Pan-cancer patterns of somatic

copy number alteration, Nat. Genet. 45 (2013) 1134–1140. https://doi.org/10.1038/ng.2760.

[13] K.A. Knouse, T. Davoli, S.J. Elledge, A. Amon, Aneuploidy in cancer: Seq-ing answers to old questions, Annu. Rev. Cancer Biol. 1 (2017) 335–354. https://doi.org/10.1146/annurev-cancerbio-042616-072231.

[14] U. Ben-David, G. Ha, Y.-Y. Tseng, N.F. Greenwald, C. Oh, J. Shih, J.M. McFarland, B. Wong, J.S. Boehm, R. Beroukhim, T.R. Golub, Patient-derived xenografts undergo mouse-specific tumor evolution, Nat. Genet. 49 (2017) 1567–1575. https://doi.org/10.1038/ng.3967.

[15] S.L. Carter, K. Cibulskis, E. Helman, A. McKenna, H. Shen, T. Zack, P.W. Laird, R.C. Onofrio, W. Winckler, B.A. Weir, R. Beroukhim, D. Pellman, D.A. Levine, E.S. Lander, M. Meyerson, G. Getz, Absolute quantification of somatic DNA alterations in human cancer, Nat. Biotechnol. 30 (2012) 413–421. https://doi.org/10.1038/nbt.2203.

[16] A. Taylor-Weiner, T. Zack, E. O'Donnell, J.L. Guerriero, B. Bernard, A. Reddy, G.C. Han, S. AlDubayan, A. Amin-Mansour, S.E. Schumacher, K. Litchfield, C. Turnbull, S. Gabriel, R. Beroukhim, G. Getz, S.L. Carter, M.S. Hirsch, A. Letai, C. Sweeney, E.M. Van Allen, Genomic evolution and chemoresistance in germ-cell tumours, Nature. 540 (2016) 114–118. https://doi.org/10.1038/nature20596.

[17] C.W. Brennan, R.G.W. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S.R. Salama, S. Zheng, D. Chakravarty, J.Z. Sanborn, S.H. Berman, R. Beroukhim, B. Bernard, C.-J. Wu, G. Genovese, I. Shmulevich, J. Barnholtz-Sloan, L. Zou, R. Vegesna, S.A. Shukla, G.

Ciriello, W.K. Yung, W. Zhang, C. Sougnez, T. Mikkelsen, K. Aldape, D.D. Bigner, E.G. Van Meir, M. Prados, A. Sloan, K.L. Black, J. Eschbacher, G. Finocchiaro, W. Friedman, D.W. Andrews, A. Guha, M. Iacocca, B.P. O'Neill, G. Foltz, J. Myers, D.J. Weisenberger, R. Penny, R. Kucherlapati, C.M. Perou, D.N. Hayes, R. Gibbs, M. Marra, G.B. Mills, E. Lander, P. Spellman, R. Wilson, C. Sander, J. Weinstein, M. Meyerson, S. Gabriel, P.W. Laird, D. Haussler, G. Getz, L. Chin, TCGA Research Network, The somatic genomic landscape of glioblastoma, Cell. 155 (2013) 462–477. https://doi.org/10.1016/j.cell.2013.09.034.

[18] A.B. Olshen, E.S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data, Biostatistics. 5 (2004) 557–572. https://doi.org/10.1093/biostatistics/kxh008.

[19] J. Staaf, D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Göransson, G. Juliusson, R. Rosenquist, M. Höglund, A. Borg, M. Ringnér, Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays, Genome Biol. 9 (2008) R136. https://doi.org/10.1186/gb-2008-9-9-r136.

[20] A. Li, Z. Liu, K. Lezon-Geyda, S. Sarkar, D. Lannin, V. Schulz, I. Krop, E. Winer, L. Harris, D. Tuck, GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays, Nucleic Acids Res. 39 (2011) 4928–4941. https://doi.org/10.1093/nar/gkr014.

[21] C. Yau, D. Mouradov, R.N. Jorissen, S. Colella, G. Mirza, G. Steers, A. Harris, J. Ragoussis, O. Sieber, C.C. Holmes, A statistical approach for detecting genomic

aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data, Genome Biol. 11 (2010) R92. https://doi.org/10.1186/gb-2010-11-9-r92.

[22] F.A. San Lucas, S. Sivakumar, S. Vattathil, J. Fowler, E. Vilar, P. Scheet, Rapid and powerful detection of subtle allelic imbalance from exome sequencing data with hapLOHseq, Bioinformatics. 32 (2016) 3015–3017. https://doi.org/10.1093/bioinformatics/btw340.

[23] R.L. Grossman, A.P. Heath, V. Ferretti, H.E. Varmus, D.R. Lowy, W.A. Kibbe, L.M. Staudt, Toward a Shared Vision for Cancer Genomic Data, N. Engl. J. Med. 375 (2016) 1109–1112. https://doi.org/10.1056/NEJMp1607591.

[24] The Cancer Genome Atlas Program, National Cancer Institute. (2018). https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga (accessed March 25, 2022).

[25] GenomeOC, Therapeutically Applicable Research to Generate Effective Treatments, (n.d.). https://ocg.cancer.gov/programs/target (accessed March 25, 2022).

[26] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L.D. Stein, V. Ferretti, The International Cancer Genome Consortium Data Portal, Nat. Biotechnol. 37 (2019) 367–369. https://doi.org/10.1038/s41587-019-0055-9.

[27] Z. Ozcan, F.A. San Lucas, J.W. Wong, K. Chang, K.H. Stopsack, J. Fowler, Y.A. Jakubek, P. Scheet, Chromosomal imbalances detected via RNA-sequencing in 28 cancers, Bioinformatics. (2022). https://doi.org/10.1093/bioinformatics/btab861.

[28] D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell. 144 (2011) 646–674. https://doi.org/10.1016/j.cell.2011.02.013.

[29] H. Hieronymus, R. Murali, A. Tin, K. Yadav, W. Abida, H. Moller, D. Berney, H. Scher, B. Carver, P. Scardino, N. Schultz, B. Taylor, A. Vickers, J. Cuzick, C.L. Sawyers, Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death, Elife. 7 (2018). https://doi.org/10.7554/eLife.37294.

[30] L. Liang, J.-Y. Fang, J. Xu, Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy, Oncogene. 35 (2016) 1475–1482. https://doi.org/10.1038/onc.2015.209.

[31] O. Nibourel, S. Guihard, C. Roumier, N. Pottier, C. Terre, A. Paquet, P. Peyrouze, S. Geffroy, S. Quentin, A. Alberdi, R.B. Abdelali, A. Renneville, C. Demay, K. Celli-Lebras, P. Barbry, B. Quesnel, S. Castaigne, H. Dombret, J. Soulier, C. Preudhomme, M.H. Cheok, Copy-number analysis identified new prognostic marker in acute myeloid leukemia, Leukemia. 31 (2017) 555–564. https://doi.org/10.1038/leu.2016.265.

[32] T. Ried, Y. Hu, M.J. Difilippantonio, B.M. Ghadimi, M. Grade, J. Camps, The consequences of chromosomal aneuploidy on the transcriptome of cancer cells, Biochim. Biophys. Acta. 1819 (2012) 784–793. https://doi.org/10.1016/j.bbagrm.2012.02.020.

[33] A. Shukla, T.H.M. Nguyen, S.B. Moka, J.J. Ellis, J.P. Grady, H. Oey, A.S. Cristino, K.K. Khanna, D.P. Kroese, L. Krause, E. Dray, J.L. Fink, P.H.G. Duijf, Chromosome arm aneuploidies shape tumour evolution and drug response, Nat. Commun. 11 (2020) 449. https://doi.org/10.1038/s41467-020-14286-0.

[34] H. Wang, L. Liang, J.-Y. Fang, J. Xu, Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers, Oncogene. 35 (2016) 2011–2019. https://doi.org/10.1038/onc.2015.304.

[35] T.B.K. Watkins, E.L. Lim, M. Petkovic, S. Elizalde, N.J. Birkbak, G.A. Wilson, D.A. Moore, E. Grönroos, A. Rowan, S.M. Dewhurst, J. Demeulemeester, S.C. Dentro, S. Horswell, L. Au, K. Haase, M. Escudero, R. Rosenthal, M.A. Bakir, H. Xu, K. Litchfield, W.T. Lu, T.P. Mourikis, M. Dietzen, L. Spain, G.D. Cresswell, D. Biswas, P. Lamy, I. Nordentoft, K. Harbst, F. Castro-Giner, L.R. Yates, F. Caramia, F. Jaulin, C. Vicier, I.P.M. Tomlinson, P.K. Brastianos, R.J. Cho, B.C. Bastian, L. Dyrskjøt, G.B. Jönsson, P. Savas, S. Loi, P.J. Campbell, F. Andre, N.M. Luscombe, N. Steeghs, V.C.G. Tjan-Heijnen, Z. Szallasi, S. Turajlic, M. Jamal-Hanjani, P. Van Loo, S.F. Bakhoum, R.F. Schwarz, N. McGranahan, C. Swanton, Pervasive chromosomal instability and karyotype order in tumour evolution, Nature. 587 (2020) 126–132. https://doi.org/10.1038/s41586-020-2698-6.

[36] C. Alkan, J.M. Kidd, T. Marques-Bonet, G. Aksay, F. Antonacci, F. Hormozdiari, J.O. Kitzman, C. Baker, M. Malig, O. Mutlu, S.C. Sahinalp, R.A. Gibbs, E.E. Eichler, Personalized copy number and segmental duplication maps using next-generation sequencing, Nat. Genet. 41 (2009) 1061–1067. https://doi.org/10.1038/ng.437.

[37] K.C. Amarasinghe, J. Li, S.M. Hunter, G.L. Ryland, P.A. Cowin, I.G. Campbell, S.K. Halgamuge, Inferring copy number and genotype in tumour exome data, BMC Genomics. 15 (2014) 732. https://doi.org/10.1186/1471-2164-15-732.

[38] A. Bouska, T.W. McKeithan, K.E. Deffenbacher, C. Lachel, G.W. Wright, J. Iqbal, L.M. Smith, W. Zhang, C. Kucuk, A. Rinaldi, F. Bertoni, J. Fitzgibbon, K. Fu, D.D. Weisenburger, T.C. Greiner, B.J. Dave, R.D. Gascoyne, A. Rosenwald, G. Ott, E. Campo, L.M. Rimsza, J. Delabie, E.S. Jaffe, R.M. Braziel, J.M. Connors, L.M. Staudt, W.-C. Chan, Genome-wide copy-number analyses reveal genomic abnormalities involved in transformation of follicular lymphoma, Blood. 123 (2014) 1681–1690. https://doi.org/10.1182/blood-2013-05-500595.

[39] G. Callagy, P. Pharoah, S.-F. Chin, T. Sangan, Y. Daigo, L. Jackson, C. Caldas, Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays, J. Pathol. 205 (2005) 388–396. https://doi.org/10.1002/path.1694.

[40] M.M. Weiss, E.J. Kuipers, C. Postma, A.M. Snijders, D. Pinkel, S.G.M. Meuwissen, D. Albertson, G.A. Meijer, Genomic alterations in primary gastric adenocarcinomas correlate with clinicopathological characteristics and survival, Cell. Oncol. 26 (2004) 307–317. https://doi.org/10.1155/2004/454238.

[41] L.B. Alexandrov, S. Nik-Zainal, D.C. Wedge, S.A.J.R. Aparicio, S. Behjati, A.V. Biankin, G.R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale, S. Boyault, B. Burkhardt, A.P. Butler, C. Caldas, H.R. Davies, C. Desmedt, R. Eils, J.E. Eyfjörd, J.A. Foekens, M. Greaves, F. Hosoda, B. Hutter, T. Ilicic, S. Imbeaud, M. Imielinski, N. Jäger, D.T.W. Jones, D. Jones, S. Knappskog, M. Kool, S.R. Lakhani, C. López-Otín, S. Martin, N.C. Munshi, H. Nakamura, P.A. Northcott, M. Pajic, E. Papaemmanuil, A. Paradiso, J.V. Pearson, X.S. Puente, K. Raine, M. Ramakrishna, A.L. Richardson, J. Richter, P.

Rosenstiel, M. Schlesner, T.N. Schumacher, P.N. Span, J.W. Teague, Y. Totoki, A.N.J. Tutt, R. Valdés-Mas, M.M. van Buuren, L. van 't Veer, A. Vincent-Salomon, N. Waddell, L.R. Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, J. Zucman-Rossi, P.A. Futreal, U. McDermott, P. Lichter, M. Meyerson, S.M. Grimmond, R. Siebert, E. Campo, T. Shibata, S.M. Pfister, P.J. Campbell, M.R. Stratton, Signatures of mutational processes in human cancer, Nature. 500 (2013) 415–421. https://doi.org/10.1038/nature12477.

[42] L. Peng, X.W. Bian, D.K. Li, C. Xu, G.M. Wang, Q.Y. Xia, Q. Xiong, Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types, Sci. Rep. 5 (2015) 13413. https://doi.org/10.1038/srep13413.

[43] A. Coudray, A.M. Battenhouse, P. Bucher, V.R. Iyer, Detection and benchmarking of somatic mutations in cancer genomes using RNA-seq data, PeerJ. 6 (2018) e5362. https://doi.org/10.7717/peerj.5362.

[44] M. Griffith, C.A. Miller, O.L. Griffith, K. Krysiak, Z.L. Skidmore, A. Ramu, J.R. Walker, H.X. Dang, L. Trani, D.E. Larson, R.T. Demeter, M.C. Wendl, J.F. McMichael, R.E. Austin, V. Magrini, S.D. McGrath, A. Ly, S. Kulkarni, M.G. Cordes, C.C. Fronick, R.S. Fulton, C.A. Maher, L. Ding, J.M. Klco, E.R. Mardis, T.J. Ley, R.K. Wilson, Optimizing cancer genome sequencing and analysis, Cell Syst. 1 (2015) 210–223. https://doi.org/10.1016/j.cels.2015.08.015.

[45] R. Kridel, B. Meissner, S. Rogic, M. Boyle, A. Telenius, B. Woolcock, J. Gunawardana, C. Jenkins, C. Cochrane, S. Ben-Neriah, K. Tan, R.D. Morin, S. Opat, L.H. Sehn, J.M. Connors, M.A. Marra, A.P. Weng, C. Steidl, R.D. Gascoyne, Whole transcriptome

sequencing reveals recurrent NOTCH1 mutations in mantle cell lymphoma, Blood. 119 (2012) 1963–1971. https://doi.org/10.1182/blood-2011-11-391474.

[46] S.P. Shah, R.D. Morin, J. Khattra, L. Prentice, T. Pugh, A. Burleigh, A. Delaney, K. Gelmon, R. Guliany, J. Senz, C. Steidl, R.A. Holt, S. Jones, M. Sun, G. Leung, R. Moore, T. Severson, G.A. Taylor, A.E. Teschendorff, K. Tse, G. Turashvili, R. Varhol, R.L. Warren, P. Watson, Y. Zhao, C. Caldas, D. Huntsman, M. Hirst, M.A. Marra, S. Aparicio, Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution, Nature. 461 (2009) 809–813. https://doi.org/10.1038/nature08489.

[47] K. Yizhak, F. Aguet, J. Kim, J.M. Hess, K. Kübler, J. Grimsby, R. Frazer, H. Zhang, N.J. Haradhvala, D. Rosebrock, D. Livitz, X. Li, E. Arich-Landkof, N. Shoresh, C. Stewart, A.V. Segrè, P.A. Branton, P. Polak, K.G. Ardlie, G. Getz, RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues, Science. 364 (2019). https://doi.org/10.1126/science.aaw0726.

[48] J. Fan, H.-O. Lee, S. Lee, D.-E. Ryu, S. Lee, C. Xue, S.J. Kim, K. Kim, N. Barkas, P.J. Park, W.-Y. Park, P.V. Kharchenko, Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data, Genome Res. 28 (2018) 1217–1227. https://doi.org/10.1101/gr.228080.117.

[49] R. Gao, S. Bai, Y.C. Henderson, Y. Lin, A. Schalck, Y. Yan, T. Kumar, M. Hu, E. Sei, A. Davis, F. Wang, S.F. Shaitelman, J.R. Wang, K. Chen, S. Moulder, S.Y. Lai, N.E. Navin, Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes, Nat. Biotechnol. (2021). https://doi.org/10.1038/s41587-020-00795-2.

[50] T. Tickle, I. Tirosh, C. Georgescu, M. Brown, B. Haas, inferCNV of the Trinity CTAT Project, Klarman Cell Observatory, Broad Institute of MIT and Harvard. (2019).

[51] C. Flensburg, A. Oshlack, I.J. Majewski, Detecting copy number alterations in RNA-Seq using SuperFreq, Bioinformatics. (2021). https://doi.org/10.1093/bioinformatics/btab440.

[52] K.H. Stopsack, C.A. Whittaker, T.A. Gerke, M. Loda, P.W. Kantoff, L.A. Mucci, A. Amon, Aneuploidy drives lethal progression in prostate cancer, Proc. Natl. Acad. Sci. U. S. A. 116 (2019) 11390–11395. https://doi.org/10.1073/pnas.1902645116.

[53] J.D. Baugher, B.D. Baugher, M.D. Shirley, J. Pevsner, Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method, BMC Genomics. 14 (2013) 367. https://doi.org/10.1186/1471-2164-14-367.

[54] P.-R. Loh, G. Genovese, R.E. Handsaker, H.K. Finucane, Y.A. Reshef, P.F. Palamara, B.M. Birmann, M.E. Talkowski, S.F. Bakhoum, S.A. McCarroll, A.L. Price, Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations, Nature. 559 (2018) 350–355. https://doi.org/10.1038/s41586-018-0321-x.

[55] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303. https://doi.org/10.1101/gr.107524.110.

[56] Y. Li, C.J. Willer, J. Ding, P. Scheet, G.R. Abecasis, MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes, Genet. Epidemiol. 34 (2010) 816–834. https://doi.org/10.1002/gepi.20533.

[57] J.M. Korn, F.G. Kuruvilla, S.A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P.J. Collins, K. Darvishi, C. Lee, M.M. Nizzari, S.B. Gabriel, S. Purcell, M.J. Daly, D. Altshuler, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs, Nat. Genet. 40 (2008) 1253–1260. https://doi.org/10.1038/ng.237.

[58] A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, D. Reich, Principal components analysis corrects for stratification in genome-wide association studies, Nat. Genet. 38 (2006) 904–909. https://doi.org/10.1038/ng1847.

[59] S. Das, L. Forer, S. Schönherr, C. Sidore, A.E. Locke, A. Kwong, S.I. Vrieze, E.Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W.G. Iacono, A. Swaroop, L.J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G.R. Abecasis, C. Fuchsberger, Next-generation genotype imputation service and methods, Nat. Genet. 48 (2016) 1284–1287. https://doi.org/10.1038/ng.3656.

[60] J. Fowler, F.A. San Lucas, P. Scheet, System for Quality-Assured Data Analysis: Flexible, reproducible scientific workflows, Genet. Epidemiol. 43 (2019) 227–237. https://doi.org/10.1002/gepi.22178.

[61] D.C. Koboldt, Q. Zhang, D.E. Larson, D. Shen, M.D. McLellan, L. Lin, C.A. Miller, E.R. Mardis, L. Ding, R.K. Wilson, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing, Genome Res. 22 (2012) 568–576. https://doi.org/10.1101/gr.129684.111.

[62] Y.A. Jakubek, K. Chang, S. Sivakumar, Y. Yu, M.R. Giordano, J. Fowler, C.D. Huff, H. Kadara, E. Vilar, P. Scheet, Large-scale analysis of acquired chromosomal alterations in non-tumor samples from patients with cancer, Nat. Biotechnol. 38 (2020) 90–96. https://doi.org/10.1038/s41587-019-0297-6.

[63] S. Vattathil, P. Scheet, Extensive Hidden Genomic Mosaicism Revealed in Normal Tissue, Am. J. Hum. Genet. 98 (2016) 571–578. https://doi.org/10.1016/j.ajhg.2016.02.003.

[64] D. Taliun, D.N. Harris, M.D. Kessler, J. Carlson, Z.A. Szpiech, R. Torres, S.A.G. Taliun, A. Corvelo, S.M. Gogarten, H.M. Kang, A.N. Pitsillides, J. LeFaive, S.-B. Lee, X. Tian, B.L. Browning, S. Das, A.-K. Emde, W.E. Clarke, D.P. Loesch, A.C. Shetty, T.W. Blackwell, A.V. Smith, Q. Wong, X. Liu, M.P. Conomos, D.M. Bobo, F. Aguet, C. Albert, A. Alonso, K.G. Ardlie, D.E. Arking, S. Aslibekyan, P.L. Auer, J. Barnard, R.G. Barr, L. Barwick, L.C. Becker, R.L. Beer, E.J. Benjamin, L.F. Bielak, J. Blangero, M. Boehnke, D.W. Bowden, J.A. Brody, E.G. Burchard, B.E. Cade, J.F. Casella, B. Chalazan, D.I. Chasman, Y.-D.I. Chen, M.H. Cho, S.H. Choi, M.K. Chung, C.B. Clish, A. Correa, J.E. Curran, B. Custer, D. Darbar, M. Daya, M. de Andrade, D.L. DeMeo, S.K. Dutcher, P.T. Ellinor, L.S. Emery, C. Eng, D. Fatkin, T. Fingerlin, L. Forer, M. Fornage, N. Franceschini, C. Fuchsberger, S.M. Fullerton, S. Germer, M.T. Gladwin, D.J. Gottlieb, X. Guo, M.E. Hall, J. He, N.L. Heard-Costa, S.R. Heckbert, M.R. Irvin, J.M. Johnsen, A.D. Johnson, R. Kaplan, S.L.R. Kardia, T. Kelly, S. Kelly, E.E. Kenny, D.P. Kiel, R. Klemmer, B.A. Konkle, C. Kooperberg, A. Köttgen, L.A. Lange, J. Lasky-Su, D. Levy, X. Lin, K.-H. Lin, C. Liu, R.J.F. Loos, L. Garman, R. Gerszten, S.A. Lubitz, K.L. Lunetta, A.C.Y. Mak, A. Manichaikul, A.K. Manning, R.A. Mathias, D.D. McManus, S.T. McGarvey, J.B. Meigs, D.A. Meyers, J.L. Mikulla, M.A. Minear, B.D. Mitchell, S. Mohanty, M.E.

Montasser, C. Montgomery, A.C. Morrison, J.M. Murabito, A. Natale, P. Natarajan, S.C. Nelson, K.E. North, J.R. O'Connell, N.D. Palmer, N. Pankratz, G.M. Peloso, P.A. Peyser, J. Pleiness, W.S. Post, B.M. Psaty, D.C. Rao, S. Redline, A.P. Reiner, D. Roden, J.I. Rotter, I. Ruczinski, C. Sarnowski, S. Schoenherr, D.A. Schwartz, J.-S. Seo, S. Seshadri, V.A. Sheehan, W.H. Sheu, M.B. Shoemaker, N.L. Smith, J.A. Smith, N. Sotoodehnia, A.M. Stilp, W. Tang, K.D. Taylor, M. Telen, T.A. Thornton, R.P. Tracy, D.J. Van Den Berg, R.S. Vasan, K.A. Viaud-Martinez, S. Vrieze, D.E. Weeks, B.S. Weir, S.T. Weiss, L.-C. Weng, C.J. Willer, Y. Zhang, X. Zhao, D.K. Arnett, A.E. Ashley-Koch, K.C. Barnes, E. Boerwinkle, S. Gabriel, R. Gibbs, K.M. Rice, S.S. Rich, E.K. Silverman, P. Qasba, W. Gan, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, G.J. Papanicolaou, D.A. Nickerson, S.R. Browning, M.C. Zody, S. Zöllner, J.G. Wilson, L.A. Cupples, C.C. Laurie, C.E. Jaquish, R.D. Hernandez, T.D. O'Connor, G.R. Abecasis, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program, Nature. 590 (2021) 290–299. https://doi.org/10.1038/s41586-021-03205-y.

[65] Chromosome X: Chromosome summary - Homo sapiens, Ensembl. (2017). http://mar2017.archive.ensembl.org/Homo_sapiens/Location/Chromosome?r=X (accessed January 26, 2022).

[66] Chromosome Y: Chromosome summary - Homo sapiens, Ensembl. (2017). http://mar2017.archive.ensembl.org/Homo_sapiens/Location/Chromosome?r=Y (accessed January 26, 2022).

[67] A. Spatz, C. Borg, J. Feunteun, X-chromosome genetics and human cancer, Nat. Rev. Cancer. 4 (2004) 617–629. https://doi.org/10.1038/nrc1413.

[68] M.J. Navarro-Cobos, B.P. Balaton, C.J. Brown, Genes that escape from X-chromosome inactivation: Potential contributors to Klinefelter syndrome, Am. J. Med. Genet. C Semin. Med. Genet. 184 (2020) 226–238. https://doi.org/10.1002/ajmg.c.31800.

[69] G.J. Pageau, L.L. Hall, S. Ganesan, D.M. Livingston, J.B. Lawrence, The disappearing Barr body in breast and ovarian cancers, Nat. Rev. Cancer. 7 (2007) 628–633. https://doi.org/10.1038/nrc2172.

[70] P.C. Cheng, J.A. Gosewehr, T.M. Kim, M. Velicescu, M. Wan, J. Zheng, J.C. Felix, K.F. Cofer, P. Luo, B.H. Biela, G. Godorov, L. Dubeau, Potential role of the inactivated X chromosome in ovarian epithelial tumor development, J. Natl. Cancer Inst. 88 (1996) 510–518. https://doi.org/10.1093/jnci/88.8.510.

[71] M.-H. Benoît, T.J. Hudson, G. Maire, J.A. Squire, S.L. Arcand, D. Provencher, A.-M. Mes-Masson, P.N. Tonin, Global analysis of chromosome X gene expression in primary cultures of normal ovarian surface epithelial cells and epithelial ovarian cancer cell lines, Int. J. Oncol. 30 (2007) 5–17. https://www.ncbi.nlm.nih.gov/pubmed/17143508.

[72] J. Kang, H.J. Lee, J. Kim, J.J. Lee, L.-S. Maeng, Dysregulation of X chromosome inactivation in high grade ovarian serous adenocarcinoma, PLoS One. 10 (2015) e0118927. https://doi.org/10.1371/journal.pone.0118927.

[73] D.P. Silver, S.D. Dimitrov, J. Feunteun, R. Gelman, R. Drapkin, S.D. Lu, E. Shestakova, S. Velmurugan, N. Denunzio, S. Dragomir, J. Mar, X. Liu, S. Rottenberg, J. Jonkers, S. Ganesan, D.M. Livingston, Further evidence for BRCA1 communication with the inactive X chromosome, Cell. 128 (2007) 991–1002. https://doi.org/10.1016/j.cell.2007.02.025.

[74] V. Borah, P.N. Shah, S.N. Ghosh, M.B. Sampat, D.J. Jussawalla, Further studies on the prognostic importance of Barr body frequency in human breast cancer: with discussion on its probable mechanism, J. Surg. Oncol. 13 (1980) 1–7. https://doi.org/10.1002/jso.2930130102.

[75] S. Ganesan, A.L. Richardson, Z.C. Wang, J.D. Iglehart, A. Miron, J. Feunteun, D. Silver, D.M. Livingston, Abnormalities of the inactive X chromosome are a common feature of BRCA1 mutant and sporadic basal-like breast cancer, Cold Spring Harb. Symp. Quant. Biol. 70 (2005) 93–97. https://doi.org/10.1101/sqb.2005.70.045.

[76] A.L. Richardson, Z.C. Wang, A. De Nicolo, X. Lu, M. Brown, A. Miron, X. Liao, J.D. Iglehart, D.M. Livingston, S. Ganesan, X chromosomal abnormalities in basal-like human breast cancer, Cancer Cell. 9 (2006) 121–132. https://doi.org/10.1016/j.ccr.2006.01.013.

[77] Chromosome 7: Chromosome summary - Homo sapiens, Ensembl. (2017). http://mar2017.archive.ensembl.org/Homo_sapiens/Location/Chromosome?r=7 (accessed January 26, 2022).

[78] Cancer of the Breast (Female) - Cancer Stat Facts, SEER. (n.d.). https://seer.cancer.gov/statfacts/html/breast.html (accessed January 26, 2022).

[79] S. Ganesan, D.P. Silver, R.A. Greenberg, D. Avni, R. Drapkin, A. Miron, S.C. Mok, V. Randrianarison, S. Brodie, J. Salstrom, T.P. Rasmussen, A. Klimke, C. Marrese, Y. Marahrens, C.X. Deng, J. Feunteun, D.M. Livingston, BRCA1 supports XIST RNA concentration on the inactive X chromosome, Cell. 111 (2002) 393–405. https://doi.org/10.1016/s0092-8674(02)01052-8.

[80] S. Ganesan, D.P. Silver, R. Drapkin, R. Greenberg, J. Feunteun, D.M. Livingston, Association of BRCA1 with the inactive X chromosome and XIST RNA, Philos. Trans. R. Soc. Lond. B Biol. Sci. 359 (2004) 123–128. https://doi.org/10.1098/rstb.2003.1371.

[81] S.M. Sirchia, L. Ramoscelli, F.R. Grati, F. Barbera, D. Coradini, F. Rossella, G. Porta, E. Lesma, A. Ruggeri, P. Radice, G. Simoni, M. Miozzo, Loss of the inactive X chromosome and replication of the active X in BRCA1-defective and wild-type breast cancer cells, Cancer Res. 65 (2005) 2139–2146. https://doi.org/10.1158/0008-5472.CAN-04-3465.

[82] M.J. Landrum, J.M. Lee, M. Benson, G.R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J.B. Holmes, B.L. Kattman, D.R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, Nucleic Acids Res. 46 (2018) D1062–D1067. https://doi.org/10.1093/nar/gkx1153.

[83] L. Carrel, H.F. Willard, X-inactivation profile reveals extensive variability in X-linked gene expression in females, Nature. 434 (2005) 400–404. https://doi.org/10.1038/nature03479.

[84] A.M. Cotton, B. Ge, N. Light, V. Adoue, T. Pastinen, C.J. Brown, Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome, Genome Biol. 14 (2013) R122. https://doi.org/10.1186/gb-2013-14-11-r122.

[85] R. Chaligné, E. Heard, X-chromosome inactivation in development and cancer, FEBS Lett. 588 (2014) 2514–2522. https://doi.org/10.1016/j.febslet.2014.06.023.

[86] D.M. Carone, J.B. Lawrence, Heterochromatin instability in cancer: from the Barr body to satellites and the nuclear periphery, Semin. Cancer Biol. 23 (2013) 99–108. https://doi.org/10.1016/j.semcancer.2012.06.008.

[87] T. Ohhata, A. Wutz, Reactivation of the inactive X chromosome in development and reprogramming, Cell. Mol. Life Sci. 70 (2013) 2443–2461. https://doi.org/10.1007/s00018-012-1174-3.

[88] T. Pollex, E. Heard, Recent advances in X-chromosome inactivation research, Curr. Opin. Cell Biol. 24 (2012) 825–832. https://doi.org/10.1016/j.ceb.2012.10.007.

[89] L. Busque, R. Mio, J. Mattioli, E. Brais, N. Blais, Y. Lalonde, M. Maragh, D.G. Gilliland, Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age, Blood. 88 (1996) 59–65. https://www.ncbi.nlm.nih.gov/pubmed/8704202.

[90] A. Zito, M.N. Davies, P.-C. Tsai, S. Roberts, R. Andres-Ejarque, S. Nardone, J.T. Bell, C.C.Y. Wong, K.S. Small, Heritability of skewed X-inactivation in female twins is tissue-specific and associated with age, Nat. Commun. 10 (2019) 5339. https://doi.org/10.1038/s41467-019-13340-w.

[91] I. Cusin, D. Teixeira, M. Zahn-Zabal, V. Rech de Laval, A. Gleizes, V. Viassolo, P.O. Chappuis, P. Hutter, A. Bairoch, P. Gaudet, A new bioinformatics tool to help assess the significance of BRCA1 variants, Hum. Genomics. 12 (2018) 36. https://doi.org/10.1186/s40246-018-0168-0.

[92] BRCA1 Database, (n.d.). https://arup.utah.edu/database/BRCA/Home/BRCA1_landing.php (accessed March 31, 2022).

[93] BRCA Exchange, (n.d.). https://brcaexchange.org/ (accessed March 31, 2022).

[94] M. Camargo, N. Wang, Cytogenetic evidence for the absence of an inactivated X chromosome in a human female (XX) breast carcinoma cell line, Hum. Genet. 55 (1980) 81–85. https://doi.org/10.1007/BF00329131.

[95] B. Dutrillaux, M. Muleris, M.G. Seureau, Imbalance of sex chromosomes, with gain of early-replicating X, in human solid tumors, Int. J. Cancer. 38 (1986) 475–479. https://doi.org/10.1002/ijc.2910380404.

[96] N. Wang, E. Cedrone, G.R. Skuse, R. Insel, J. Dry, Two identical active X chromosomes in human mammary carcinoma cells, Cancer Genet. Cytogenet. 46 (1990) 271–280. https://doi.org/10.1016/0165-4608(90)90112-n.

[97] S. Zaccaria, B.J. Raphael, Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL, Nat. Biotechnol. 39 (2021) 207–214. https://doi.org/10.1038/s41587-020-0661-6.

VITA

Zuhal Ozcan is the daughter of Hayriye and Ali Ozcan. Zuhal Ozcan was born in Turkey in 1989. She graduated from Vefa Anatolian High School and then studied at Sabanci University at Istanbul, earning Bachelor's Degree in Bioengineering and Biological Sciences in 2014. After spending two years conducting research at Swiss Federale Institute of Technology Lausanne (EPFL) and University of Pittsburgh, Pennsylvania, she began her graduate studies at the Graduate School of Biomedical Sciences in Summer of 2016. She will continue her training as a Laboratory Genetics and Genomics Fellow in the Department of Pathology at Henry Ford Health System, Michigan.