The Clinical Actionability and Evolution of Mutational Processes in Metastatic Cancer

Eric Yang Zhao

B. Sc. (Honours) The University of British Columbia 2013, May

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

The University of British Columbia

(Vancouver)

November 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

The Clinical Actionability and Evolution of Mutational Processes in Metastatic Cancer

submitted by Eric Yang Zhao in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Bioinformatics

Examining Committee:

Steven J. M. Jones, Medical Genetics
Supervisor

Christian Steidl, Pathology & Laboratory Medicine
Supervisory Committee Member

Cathie Garnis, Department of Surgery
University Examiner

Judy Wong, Medical Genetics
University Examiner

Additional Supervisory Committee Members:

Marco Marra, Medical Genetics
Supervisory Committee Member

Inanc Birol, Medical Genetics
Supervisory Committee Member

Stephen Yip, Pathology & Laboratory Medicine
Supervisory Committee Member

ABSTRACT

Cancers are characterized by somatic mutation arising from the interplay of mutagen exposure and deficient DNA repair. Whole genome sequencing of tumours reveals characteristic patterns of mutation, known as mutation signatures, which often correspond with specific processes such as cigarette smoke exposure or the loss of a DNA repair pathway. Quantifying DNA repair deficiency can have clinical implications. Cancer chemotherapies which induce DNA damage are known to be more effective against cancers with deficient DNA repair. However, it is not yet known whether mutation signatures can serve as reliable predictive biomarkers for response to these treatments. Furthermore, the current understanding of mutation signatures stems largely from studies of primary, untreated tumours, whereas metastasis underpins as much as 90% of cancer-related mortality. This thesis aims to (1) describe the association between mutation signatures and clinical response to DNA damaging chemotherapy, (2) enable accurate personalized assessment of mutation signatures and their evolution over time, and (3) characterize the evolution of mutational processes in metastatic cancers. To assess clinical actionability, we quantified signatures of single nucleotide variants, structural variants, copy number variants, and small deletions in 93 metastatic breast cancers, 33 of which received platinum-based chemotherapy. We found that patients with signatures of homologous recombination deficiency had improved responses and prolonged treatment durations on platinum-based chemotherapy. Next, we formulated a Bayesian model called SignIT, which improves the accuracy of individualized mutation signature analysis and infers

signature evolution over tumour subpopulations. We demonstrated SignIT's superior accuracy on both simulated data and somatic mutations from The Cancer Genome Atlas, and validated temporal dissection using whole genomes from 24 multiply-sequenced cancers. We highlighted a potential clinical application of mutation in a *BRCA1*-mutated pancreatic adenocarcinoma with low Homologous Recombination Deficiency (HRD) signature but exceptional response to platinum-containing chemotherapy. Finally, we deciphered mutation signatures from nearly 500 metastatic cancer whole genomes, revealing evolution of mutational processes associated with late metastasis and exposure to cytotoxic chemotherapy. Taken together, our findings demonstrate the complex interplay of factors shaping the metastatic cancer genome. We highlight both clinical opportunities of studying genomic instability and the additional insights available from understanding their temporal evolution.

LAY SUMMARY

Changes in DNA, known as mutations, happen for many reasons such as exposure to cigarette smoke and ultraviolet rays. Healthy human cells have tools to fix mutated DNA, but cancer cells often lose this ability. This might make tumours vulnerable to chemotherapy designed to damage DNA, because healthy cells can repair this damage but cancer cells cannot. Can we tailor treatments to exploit this vulnerability? To answer this question, we detected all the mutations in about 500 late-stage cancers of different types. The patterns of DNA mutation can reveal if a cancer is properly repairing DNA. Within breast cancers, we found that cancers unable to repair broken DNA were more sensitive to chemotherapies called cisplatin and carboplatin. This could help improve personalized treatment plans for some cancer patients. We also found mutation patterns caused by chemotherapy, showing that cancer treatments themselves can alter DNA.

PREFACE

All of the work contained within this thesis was conducted at the BC Cancer Agency Genome Sciences Centre under the auspices of the Personalized Oncogenomics Project (NCT 02155621), approved by the University of British Columbia Research Ethics Board (approval #H12-00137 and #H14-00681-A019). Written informed consent, including potential publication of findings, was obtained from patients prior to genomic profiling. Patient information was anonymized, and each was assigned an alphanumeric identification code. Whole-genome sequencing and RNA-seq data (.bam files) are deposited in the European Genome-Phenome Archive (www.ebi.ac.uk/ega/home) under the study accession number EGAS00001001159.

Chapters 1 and 5 are my original work, and contain brief excerpts from an invited book chapter

> Zhao EY, Jones MR, Jones SJM. "Whole Genome Sequencing in Cancer". *Next Generation Sequencing*. Ed. McCombie WR, Mardis ER, Knowles J, and McPherson JD. New York: Cold Spring Harbour Press, Accepted Feb 2018.

A version of chapter 2 has been published:

> Zhao EY, Shen Y, Pleasance E, Kasaian K, Leelakumari S, Jones M, Bose P, Ch'ng C, Reisle C, Eirew P, Corbett R, Mungall KL, Thiessen N, Ma Y, Schein JE, Mungall AJ, Zhao Y, Moore RA, Den Brok W, Wilson S, Villa D, Shenkier T, Lohrisch C, Chia S, Yip S, Gelmon K, Lim H, Renouf D,

Sun S, Schrader KA, Young S, Bosdet I, Karsan A, Laskin J, Marra MA, and Jones SJM. Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. Clin Cancer Res. 2017 Dec 15;23(24):7521-7530.

I conceptualized this project with my supervisor, Steven Jones. I was the primary researcher, leading the study design, clinical chart review, data analysis, interpretation of findings, and manuscript preparation. Yaoqing Shen, Erin Pleasance, Katayoon Kasaian, Sreeja Leelakumari, Martin Jones, Pinaki Bose, Carolyn Ch'ng, Caralyn Reisle, and Peter Eirew were involved in the analysis of specific patient cases and bioinformatics pipeline development. Richard Corbett, Karen Mungall, Nina Thiessen, and Yussanne Ma were responsible for developing and operating the bioinformatics pipelines. Jacqueline Schein, Andrew Mungall, Yongjun Zhao, and Richard Moore were responsible for developing and operating laboratory protocols for the handling and sequencing of samples. Wendie Den Brok, Sheridan Wilson, Diego Villa, Tamara Shenkier, Caroline Lohrisch, Stephen Chia, Karen Gelmon, and Sophie Sun were responsible for recruiting patients. Karen Gelmon, Howard Lim, Daniel Renouf, and Janessa Laskin provided guidance and supervision of the clinical chart review. Stephen Yip performed histopathology analysis. Yaoqing Shen, Howard Lim, Kasmintan Schrader, Sean Young, Ian Bosdet, and Aly Karsan assisted with analysis and interpretation of germline variants, as well as associated ethics discussions. Janessa Laskin and Marco Marra provided supervision over the Personalized Oncogenomics Project. All authors were involved in manuscript revision.

A version of chapter 3 has been submitted for publication:

Zhao EY, Pleasance ED, Jones MR, Shen Y, Reisle CR, Mungall AJ, Zhao Y, Moore RA, Laskin J, Marra MA, Jones SJM. SignIT: Inferring Muta-

tion Signatures and their Temporal Evolution in Individual Tumours. *In Review*.

In addition to chapter 3, aspects of the analysis submitted for publication also contributed to chapter 5. I was the primary researcher on this project and was responsible for conceptualization, methodology development, data analysis, and manuscript preparation. Erin Pleasance, Martin Jones, Yaoqing Shen, and Caralyn Reisle were involved in various aspects of data analysis and cohort assembly. Andrew Mungall, Yongjun Zhao, and Richard Moore were responsible for laboratory protocols, including the sequencing of samples. Janessa Laskin and Marco Marra supervised the Personalized Oncogenomics Project, which was the major source of data for the study. Steven Jones supervised the project and was involved in conceptualization and interpretation of findings throughout. All authors revised and approved the manuscript.

A version of chapter 4 has been submitted for publication as a case study, for which I was an equally contributing primary author:

Wong H, Zhao EY, Jones MR, Reisle CR, Eirew P, Pleasance ED, Grande BM, Karasinka JM, Kallogher SE, Lim HJ, Shen Y, Yip S, Morin RD, Laskin J, Marra MA, Jones SJM, Schrader KA, Schaeffer DF, and Renouf DJ. Temporal dynamics of genomic alterations in a BRCA1 germline mutated pancreatic cancer with low genomic instability burden but exceptional response to FOLFIRINOX. *In Review*.

Hui-li was responsible for detailing the clinical history, and performing retrospective review. I was principally responsible for the temporal analysis of *BRCA1* and other genomic events, as well as the analysis and timing of mutation signatures. Hui-li and I jointly conceptualized the project, interpreted the findings, and

wrote the manuscript. Martin Jones, Caralyn Reisle, Peter Eirew, and Erin Pleasance were responsible for initial genome and transcriptome characterization of the tumour as part of the Personalized Oncogenomics Project. Howard Lim, Yaoqing Shen, and Kasmintan Schrader were responsible for the analysis of germline variants. Bruno Grande and Ryan Morin were responsible for the clonality analysis copy number variants. Joanna Karasinka and Steve Kalloger were responsible for coordinating the study. Stephen Yip was responsible for histopathology analysis. Janessa Laskin and Marco Marra were responsible for supervising the Personalized Oncogenomics Project, the study for which this patient was recruited. Steven Jones, Kasmintan Schrader, David Schaeffer, and Daniel Renouf jointly supervised this project and were involved at various stages of conceptualization and interpretation of findings. All authors revised and approved the manuscript.

A version of chapter 5 is part of an in-draft manuscript which will detail the whole genome sequencing of the first 570 metastatic cancers in the Personalized Oncogenomics Project. This project is led by Erin Pleasance, Yaoqing Shen, and Martin Jones, and is supervised by Marco Marra. However, I am the lead researcher on the aspect of the project which involves mutation signatures, which is the only portion reported in this thesis. I was solely responsible for the analysis and interpretation of mutation signatures, with guidance from the study leads and Steven Jones. Erin Pleasance and Martin Jones were responsible for curating the final list of participants included in the study and grouping them into cancer type cohorts. Yussanne Ma oversees bioinformatics analyses for the project. Additionally, Laura Williamson was involved in the analysis of mismatch repair deficiency genes and microsatellite instability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADVI	automatic differentiation variational inference
BER	base excision repair
BTP	binary temporal partitioning
CCF	cancer cell fraction
CGH	comparative genomic hybridization
CI	clinical improvement
CNLOH	copy-neutral loss of heterozygosity
CNV	copy number variant
COSMIC	Catalogue of Somatic Mutations in Cancer
DDR	deficient DNA repair
EGA	European Genome-Phenome Archive
EM	expectation maximization
FFPE	formalin-fixed and paraffin embedded
Gb	gigabase
GDC	Genomic Data Commons
HMC	Hamiltonial Monte Carlo
HMM	hidden Markov model
HR	homologous recombination
HRD	homologous recombination deficiency
ICGC	International Cancer Genome Consortium
ITH	intratumour heterogeneity
kb	kilobase
LOH	loss of heterozygosity

LST	large-scale transition
MAF	mutation annotation format
Mb	megabase
MCMC	Markov-chain Monte Carlo
MCN	mutation copy number
MMR	mismatch repair
MMRD	mismatch repair deficiency
MSI	microsatellite instability
NER	nucleotide excision repair
NGS	Next-generation sequencing
NHEJ	non-homologous end joining
NMF	non-negative matrix factorization
NNLS	non-negative least squares
OCT	optimal cutting temperature
OS	overall survival
PARP	poly (ADP-ribose) polymerase
PCR	polymerase chain reaction
PD	progressive disease
PD-L1	Programmed death ligand 1
PDAC	pancreatic ductal adenocarcinoma
PNET	pancreatic neuroendocrine tumour
POG	Personalized Oncogenomics Project
QP	quadratic programming
ROS	reactive oxygen species
RPKM	reads per kilobase of transcript per million mapped reads
SA	simulated annealing

SD	stable disease
SNV	single nucleotide variant
SSE	sum of squared errors
SV	structural variant
TAI	telomeric allelic inbalance
TARGET	Therapeutically Applicable Research to Generate Effective Treatments
TCGA	The Cancer Genome Atlas
TDT	total duration on platinum-based therapy
TMB	total mutational burden
UV	ultraviolet
VAF	variant allele fraction
VUS	variants of uncertain significance
WAIC	Watanabe-Akaike information criterion
WES	whole exome sequencing
WGS	whole genome sequencing
WGTA	whole genome and transcriptome analysis
WTS	whole transcriptome sequencing

ACKNOWLEDGEMENTS

I am infinitely grateful to my supervisor Dr. Steven J. M. Jones for his mentorship. With his support, I have had the privilege of learning from and contributing to the fascinating research field of cancer genomics and precision oncology. In addition, I would like to thank the members of my thesis advisory committee, Dr. Marco Marra, Dr. Inanc Birol, Dr. Christian Steidl, and Dr. Stephen Yip. Their input at committee meetings has continually challenged me to critically consider the unifying questions and themes of my research, and to rise to the standards of a PhD.

My sincerest thanks go to the participants of the Personalized Oncogenomics (POG) project, who are contributing critical insights in the fight against cancer. Additionally, I thank the BC Cancer Foundation for providing generous funding support for POG.

Thank you to the POG directors Marco Marra and Janessa Laskin, as well as the POG coordinators and project managers, Robyn Roscoe, Alexandra Fok, Katherine Mui, Jessica Nelson, and Payal Sipahimalani, without whom this work would not be possible. Also, a big thank you to POG pathologists and oncologists who have at various times served as both research and clinical mentors to me, in particular Drs. Janessa Laskin, Daniel Renouf, Stephen Yip, Sophie Sun, Karen Gelmon, Howard Lim, Stephen Chia, Hui-li Wong, and Wendie Den Brok. I am also grateful for the administrative work of Louise Clarke, Cath Ennis, Leslie Alfaro, and Mhairi Sigrist.

I would like to thank all of the members of the Jones Lab, as well as the staff scientists, group leaders, project managers, and computational biologists involved in POG, who have been constant companions in this research endeavour. I am grate-

ful to some individuals in particular, who frequently made time to discuss my research and contributed significantly to the development of my skills: Erin Pleasance, Yaoqing Shen, Martin Jones, Jake Lever, Jasleen Grewal, Laura Williamson, Bruno Grande, Ryan Morin, Shaun Jackman, Katayoon Kasaian, Pinaki Bose, Darya Dargahi, Caroline Ch'ng, Caralyn Reisle, Daniel Paulino, Luka Culibrk, Richard Corbett, Martin Krzywinski, Karen Mungall, Andrew Mungall, and Yussanne Ma.

The results in this thesis are in whole or part based upon data generated by The Cancer Genome Atlas (TCGA) managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

I am grateful for funding support during my graduate studies from the BCCA-CIHR-UBC MD/PhD Studentship, a CIHR Vanier Canada Graduate Scholarship, a UBC 4 Year Doctoral Fellowship, a CIHR Canada Graduate Scholarship Master's Award, and a UBC Faculty of Medicine Graduate Award. I would also like to acknowledge financial support for travel and technology purchases from the BCCA-CIHR-UBC MD/PhD Studentship, the John Bosdet Travel Award, the UBC Graduate Studies Travel Award, and two travel awards from the UBC Bioinformatics Program.

My sincere thanks go to the UBC MD/PhD program director Lynn Raymond, associate director Torsten Nielson, and coordinator Jane Lee. I would also like to thank the UBC Bioinformatics program coordinator, Sharon Ruschkowski. Lastly, I thank my former research supervisors Dr. Marco Marra, Dr. Jan M. Friedman and Dr. Alex MacKay for welcoming me into their labs as an undergraduate student and empowering me to pursue further research endeavours.

Thank you to the members of the 2017 Class of Medicine at the University of British Columbia, who were my daily companions during the first two years of my

graduate program. Also, a tremendous thanks to my fellow MD/PhD colleagues, many of whom have shared the same timeline as myself: Andrea Jones, Amanda Dancsok, Philip Edgcumbe, Parker Jobin, Frank Lee, Victoria Baronas, Cynthia Min, Paulina Piesik, Adam Ramzy, Allen Zhang, David Twa, Jordan Squair, Daniel Woodsworth, Alexander Wright, William Guest, and many others.

My warmest thanks to my life partner, Kaylee Sohng, who (despite residing on the other end of the country) has seen me through all the ups and downs of this work with exceptional generosity. My deepest gratitude goes to my sister Sarah and my parents Liheng and Yongjun, who have ceaselessly supported me for nearly three decades.

DEDICATION

I dedicate this thesis to my parents Liheng Li and Yongjun Zhao, who have made so many sacrifices while instilling in me scientific curiosity and a passion for learning and service.

1

INTRODUCTION

1.1 RESEARCH AIMS

The objective of this thesis is to precisely characterize the evolution and clinical actionability of genomic instability in metastatic cancer. To address this objective, we begin by demonstrating the clinical applicability of mutation signatures in the treatment of metastatic breast cancers, using homologous recombination (HR) as a model system. Next, we develop a statistical model capable of accurately estimating the temporal evolution of mutational processes shaping individual tumour genomes. Last, we catalog the mutational processes shaping 484 metastatic cancers and relate them to potential clinical implications.

1.2 BACKGROUND

1.2.1 *Cancer is an Evolving Genetic Disease*

The first known work to probe the biological basis of heredity was published in 1866 by Gregor Mendel. In 1886, De Gouvea reported the first known case of inherited retinoblastoma, providing evidence that cancer, like other traits, is heritable. In the early 20th century, Boveri, Sutton, and Hunt demonstrated that genetic material was organized into chromosomes, and postulated that cellular processes

and chromosome damage triggered the onset of oncogenesis. Later, Wynder and Graham (1950) showed that cigarette smoking was associated with lung cancer in advance of the 1964 surgeon general report on the same topic (Surgeon General, 1964). However, it was only recently, with advances in genome sequencing, that the scale of DNA damage and mutation from tobacco smoke and other mutagenic sources was quantified across various types of cancer (Alexandrov et al., 2013a). This section will overview the recent literature highlighting the role of mutagenesis in tumour biology, and the opportunities and challenges it poses.

Somatic Mutations Arise from Exogenous and Endogenous Sources

Genome instability and mutagenesis are cancer hallmarks arising from a confluence of mutagenic exposures and deficient DNA repair (Hanahan and Weinberg, 2011). The total number of single nucleotide variants (SNVs) acquired by cancers varies by orders of magnitude both between and within tumour types, from as few as ten to as many as three million (Alexandrov et al., 2013a; Lawrence et al., 2013). Large scale structural variants (SVs) also exhibit heterogeneous patterns of occurrence, with substantial differences in mutation burden and distributions even between tumour subtypes (Nik-Zainal et al., 2016; Waddell et al., 2015). This reflects the tremendous diversity in the etiology of and genetic predispositions to individual cancers.

Exogenous mutational processes typically involve exposure to environmental carcinogens such as cigarette smoke or ultraviolet (UV) radiation. Mutations are also known to occur iatrogenically, through exposure to radiation therapy (Behjati et al., 2016) and chemotherapies. For example, the alkylating chemotherapy agent, temozolomide, commonly used to treat brain cancers, causes a hypermutating signature of cytosine-to-thymine (C$\rightarrow$T) transitions in cancers with specific

DNA repair deficiencies (Alexandrov et al., 2013a; Tomita-Mitchell et al., 2000; Yip et al., 2009). When exogenous mutagenesis occurs, it is often responsible for large numbers of mutations, accounting for the high mutation rates of melanomas, lung cancers, and gastrointestinal tract cancers, which arise in tissues frequently exposed to environmental stresses.

By contrast, *endogenous* mutational processes refer to the action of intracellular mechanisms which induce DNA damage or otherwise cause base changes. A frequently observed example is deamination, which causes a pattern of C→G and C→T mutations at TCN trinucleotide contexts. The deamination of 5-methylcytosine at CpG loci is thought to cause aging-related mutations across cell types (Alexandrov et al., 2015a). Another form of deamination, catalyzed by the APOBEC gene family, displays a more specific mutagenic profile (Nik-Zainal et al., 2012) and is often responsible for local hypermutation, known as "Kataegis" (Alexandrov et al., 2013a). Like exogenous processes, endogenous processes often accrue mutations in the context of DNA damage repair deficiencies. For example, metabolic damage due to reactive oxygen species (ROS) can cause the formation of DNA lesions, such as 8-oxoguanosine, which result in DNA mispairing. Excision of 8-oxoguanosine is performed by DNA glycosylase, encoded by the gene *MUTYH*. Consequently, mutations that inactivate *MUTYH* result in the accumulation of G→T/C→A mutations in the cancer genome (Pilati et al., 2017; Viel et al., 2017; Zehir et al., 2017).

Mutational processes each generate characteristic patterns of mutation known as mutation signatures (Alexandrov et al., 2013a). Others refer to them as "genomic scars" (Lord and Ashworth, 2012; Watkins et al., 2014), a term which emphasizes the lasting imprint that mutations leave on the cancer genome. Mutation signature analysis leverages the genome itself as a functional assay to study mu-

tagenesis and DNA repair. Many signatures have been associated with specific cancer types (Alexandrov et al., 2013a), histopathological subtypes (Wang et al., 2017), and response to chemotherapeutic agents (Rizvi et al., 2015; Telli et al., 2016).

Somatic mutation provides a basis for tumourigenesis. The stepwise occurrence of mutations in key oncogenes and tumour suppressors generates a diverse set of phenotypes amongst the cells of a tumour, upon which selective pressures can act (Nowell, 1976). This process is known to favour dedifferentiated cells with qualities concisely summarised as The Hallmarks of Cancer (Hanahan and Weinberg, 2011), including replicative immortality, loss of growth suppression, and evasion of immune detection among others.

The Role of Deficient DNA Repair

Cells are equipped with the molecular machinery to repair damage and errors in the genome. Molecular pathways that repair DNA point mutations include base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), and direct catalytic removal of lesions (i.e. the removal of 6-O-methylguanosine by the *MGMT* gene) (Weinberg, 2013). Mechanisms for repair of DNA strand breakages and cross-links include HR and non-homologous end joining (NHEJ) (Chang et al., 2017; Ranjha et al., 2018).

Like with mutagenic exposures, deficiencies in these DNA repair pathways can give rise to specific patterns of mutation. For example, MMR corrects erroneous base pairing, and a germline predisposition to mismatch repair deficiency (MMRD) gives rise to Lynch Syndrome, which carries a 50-80% lifetime risk of colorectal cancer (Kohlmann and Gruber, 1993). MMRD is commonly associated with large numbers of C→T mutations and microsatellite instability (MSI), the

genome-wide shrinkage or expansion of short repetitive sequences known as microsatellites (Thibodeau et al., 1993).

While DNA damage repair deficiency can drive somatic mutation, it is also hypothesized to produce vulnerability to DNA damaging chemotherapy. Overwhelming the compromised repair mechanism using platinum-based chemotherapy, alkylating agents, anthracyclines or other DNA damaging drugs is thought to stall replication and induce apoptosis (Helleday et al., 2008). A successful application of this strategy is the use of platinum-based chemotherapies in *BRCA1*/*BRCA2* mutated cancers (Farmer et al., 2005; Kennedy et al., 2004; Yang et al., 2011). Alternatively, inducing synthetic lethal inhibition of the poly (ADP-ribose) polymerase (PARP) gene family has recently been shown as a more targeted strategy of treating tumours with mutations in *BRCA1* or *BRCA2* (Engert et al., 2017; Helleday et al., 2008; Robson et al., 2017).

The Clinical Implications of Homologous Recombination Deficiency

Uncovering the role of homologous recombination deficiency (HRD) in cancer susceptibility was a major scientific breakthrough of the 1990s. Hall et al. (1990) and Narod et al. (1991) first reported that inheritance of chromosomal segment 17q12-23 was strongly associated with familial breast and ovarian cancers. A few years later, Miki et al. (1994) and Albertsen et al. (1994) pinpointed the exact locus of the *BRCA1* gene, and *BRCA2* was identified the following year (Wooster et al., 1995). Hereditary mutations in *BRCA1* and *BRCA2* confer an up to 85% lifetime risk of breast cancer and drive 5-10% of total cases (Canadian Cancer Society, 2014; National Cancer Institute, 2014). However, the mechanism by which *BRCA1*/*BRCA2* mutations resulted in cancer risk was unclear. In 1997, phosphorylation of *BRCA1* was found to occur in response to DNA damage, and *BRCA1*

was also found to complex *Rad51* at damaged sites (Scully et al., 1997a, 1997b). Gradually, the complementary roles of *BRCA1* and *BRCA2* in HR mediated DNA damage repair were uncovered (Roy et al., 2012; Yoshida and Miki, 2004).

HR refers to the exchange of similar or identical nucleotide sequences. It facilitates error-free repair of double strand breaks and inter-strand crosslinks, as well as error-free replication support and telomere maintenance (Li and Heyer, 2008). Additional studies have described cancer risk mutations in other genes of the HR pathway, including *ATM*, *MRN*, *MRE11*, *RAD50*, *NBS1*, *RAD51*, *XRCC2/3*, and the FANC family, as well as downregulation of *ATR* (Cerbinskaite et al., 2012).

Breast cancers with *BRCA1/BRCA2* mutations also display characteristic patterns of mutation. The first of these to be discovered were three large-scale copy number variant (CNV) patterns detected by hybrid-capture panels (Birkbak et al., 2012; Popova et al., 2012). This approach yielded three quantifiable scores (loss of heterozygosity: HRD-LOH, telomeric allelic imbalance: HRD-TAI, and large scale transition: HRD-LST) which correlated with BRCA1/2 mutation status (Timms et al., 2014). More recently, the whole genome analysis of breast cancers has identified characteristic SNV and SV signatures associated with *BRCA1/BRCA2* mutations, as well as short homologous stretches of DNA at deletion breakpoints known as microhomology (Alexandrov et al., 2013b; Davies et al., 2017; Nik-Zainal et al., 2016; Stephens et al., 2012).

HRD is a promising target for the administration of poly-ADP ribose polymerase (PARP) inhibitors (Gelmon et al., 2011; Kaufman et al., 2013) and platinum-based therapies such as cisplatin and carboplatin (Farmer et al., 2005; Kennedy et al., 2004; Yang et al., 2011). This is motivated by the substantial evidence of a link between germline *BRCA1* and *BRCA2* variants and sensitivity to platinum-based chemotherapy (Arun et al., 2011; Byrski et al., 2010; Tutt et al., 2015; Von Minck-

witz et al., 2014). Conversely, spontaneous reversion mutations which restore a functional copy of *BRCA1* or *BRCA2* have been observed in breast and ovarian cancers with acquired resistance to platinum-based chemotherapy (Afghahi et al., 2017; Patch et al., 2015; Swisher et al., 2008).

Telli et al. (2016) showed that the aforementioned HRD score was predictive of response to platinum-containing neoadjuvant chemotherapy in primary breast cancers. However, a phase III trial could not reproduce this finding in the advanced breast cancer setting (Tutt et al., 2015). What remains unclear is whether more precise quantification of HRD-associated DNA damage using whole genome sequencing (WGS) might predict response to platinum-based chemotherapy in advanced breast cancers where the HRD score on its own could not.

Tumour Heterogeneity and The Evolution of Mutational Processes

Nowell (1976) proposed a model of cancer clonal expansion, which asserts that cancers evolve over time by natural selection. Just as cancers change, so too do the processes which cause mutations in cancer. This surfaces a limitation of mutation signatures: they condense the life history of a tumour into a single time point. As Watkins et al. (2014) wrote, "[b]y chronicling the past but not documenting the present, genomic scar measures report whether or not a defect ... has been operative at some point in tumorigenesis and not whether it remains operative at the point of treatment." Mutations accumulated in the cancer genome do not disappear when the processes that created them grind to a halt. Therefore, observing a mutational pattern does not immediately reveal whether it arose long ago in tumour initiation or whether it represents an active and clinically relevant mutational process.

The lost temporal axis of cancer mutation can be partially restored using genomic features known to vary with time. Consider the analogy of the geological record. The fossilized remnants of past life have accumulated in the earth, but it can be inferred that fossils found in deeper strata likely represent earlier life forms than those found in shallow ones, even without the use of additional technology such as carbon dating. In the cancer genome, clonal heterogeneity and large scale chromosomal duplications are analogous to geological strata. Mutations occurring in a larger fraction of cancer cells likely occurred earlier than mutations occurring in few. Additionally, mutations present on multiple chromosomes in duplicated regions most likely occurred prior to the duplication event, and can also be inferred to be earlier-arising (Figure 1.1).

By inferring the relative timing of mutations, an analysis across publicly available cancer datasets in The Cancer Genome Atlas (TCGA) verified that mutational processes do indeed change over time (McGranahan et al., 2015). Those associated with aging, cigarette smoking, and UV radiation were prevalent among earlier-occurring mutations, whereas the impacts of DNA repair deficiencies were more likely to be split evenly between early and late mutations. While perhaps not as precise as taking multiple time-separated samples of a tumour to track its evolution, this approach enables the temporal dissection of mutations using a single biopsy, obviating the inconvenience, medical risk, and expense of sequencing multiple biopsies of a tumour.

Just as chromosomal duplications can temporally dissect point mutations, so too can point mutations provide a molecular clock for duplications. A late-arising duplication can be expected to reveal a region spanning many prior (and therefore now duplicated) point mutations. Conversely, an early-arising duplication will likely go on to acquire many late-arising (and therefore non-duplicated) point

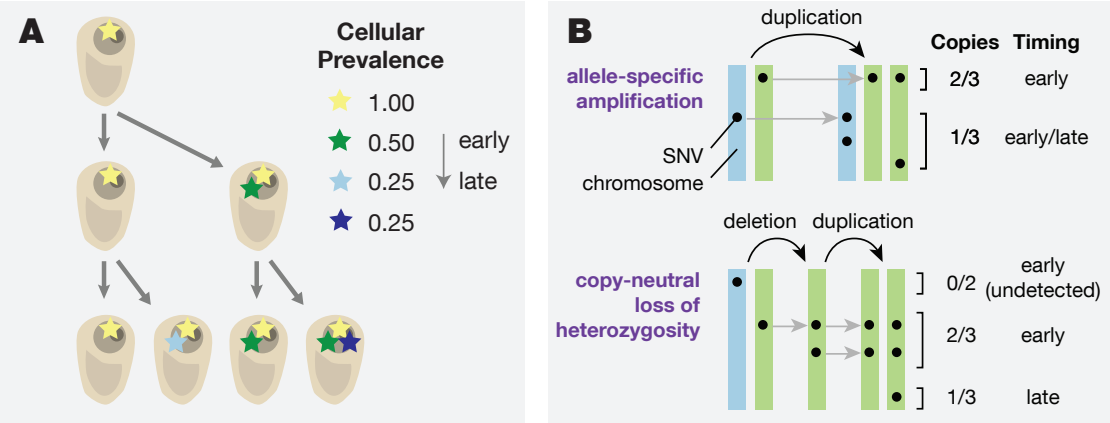

Figure 1.1: Mutational prevalence is related to time of mutation onset. Mutational prevalence is defined as the arithmetic product of cellular prevalence and mutational copy number, and has a direct impact on variant allele fraction. (A) For a given locus, the cellular prevalence is the fraction of cancer cells carrying a mutation at the locus. Higher cellular prevalence is associated with "trunk" mutations, which are more likely to be early arising amongst cancer cells. (B) Increased mutation copy number is associated with early arising mutations which occurred prior to duplication events.

mutations. Importantly, this model refers to "molecular time," which differs from true time by a factor proportional to the mutation rate, which can vary over time. Purdom et al. (2013) implemented a generative model which estimates the relative timing of chromosomal duplications. An important limitation is that relative timing of events in regions with more complex chromosomal abnormalities can only be inferred if the exact temporal ordering of events is known or can be deduced. This limitation stems from the inability of sequencing techniques to distinguish which homologous chromosome a DNA read originated from. The advent of long-read and linked-read (Zheng et al., 2016) sequencing technology may eventually help to address this limitation.

1.2.2 *Whole Genome Sequencing of Cancer*

The application of WGS to derive a more complete understanding of cancer has been a central goal of cancer researchers since before the human genome was first decoded in 2003 (Lander et al., 2001). It would take a further 5 years and a sea change in genome sequencing technology before the first application of next-generation whole genome sequencing to a cancer sample was described. (Ley et al., 2008) reported the analysis of a cytogenetically normal AML in 2008 only six months after the first human whole genome sequence by next-generation technologies was published (Wheeler et al., 2008). At the time, the bioinformatics tools and genomic resources to facilitate the in-depth analysis of whole genome data could be considered in their infancy compared to today's standards. Even so, the insights gained into both the approach taken to sequencing the tumour and the biology of the tumour itself were profound when compared to the targeted sequencing approaches commonly applied to cancer research at the time.

Today, the resources required for WGS analysis have decreased substantially. Alongside a steady reduction in the cost of WGS, there have been improvements in the technologies for generating and processing quality raw data as well as the tools and companion datasets that contextualize findings for biological and clinical interpretation. However, a majority of cancer genomics efforts remain focused around targeted deep sequencing and whole exome sequencing (WES) (Morris et al., 2017; Raphael et al., 2017).

Large-scale efforts using genome sequencing to characterize a wide variety of adult and paediatric cancers began in earnest as early as 2005. This included projects such as TCGA, the International Cancer Genome Consortium (ICGC), Catalogue of Somatic Mutations in Cancer (COSMIC), and Therapeutically Ap-

plicable Research to Generate Effective Treatments (TARGET), to name but a few. Not only have such efforts progressed our understanding of cancer as a genomic disease, they also provide the data needed for developing tools and resources that facilitate the rapid detection and analysis of potentially relevant genomic events (Cerami et al., 2012; Gao et al., 2013; Gonzalez-Perez et al., 2013; Rubio-Perez et al., 2015). However, since the bulk of the data produced is focused on coding regions of the genome, the available data are underpowered to inform how untranslated, intronic and intergenic regions might impact the molecular pathogenesis of disease. In many cases the data also lack comprehensive clinical annotation, which is required for linking genomic events to specific cancer types, prognoses, and treatment responses. Furthermore the majority of samples in these cohorts are from primary untreated disease and do not offer insight into how tumours respond to often complex and disparate treatment regimens (Robinson et al., 2017). Additional cancer cohorts of samples from multiple time-points and biopsy sites that include rich clinical information are therefore still required to better define tumour biology and the relationship to treatment history and response.

The number of tumour whole genome sequences that have been published and made publicly available has steadily increased over the past 10 years. These analyses have led to surprising insights into cancer biology, particularly from the analysis of SVs in tumour genomes (Chong et al., 2017). They range in scope from characterization of cancer cell lines (Pleasance et al., 2010b, 2010a) and n-of-one case reports with rich clinical detail (Ellis et al., 2012), to ultra-deep sequencing of a single tumour to uncover clonal heterogeneity (Griffith et al., 2015). Larger-scale efforts are also emerging, focussed both at the characterization of somatic mutations, including non-coding and SVs (Alexandrov et al., 2013b; Banerji et al., 2012; Nik-Zainal et al., 2016; Wang et al., 2014) and germline-specific analysis for the

discovery of predisposing factors (Foley et al., 2015). For example, Nik-Zainal et al. (2016) sequenced the whole genomes of 560 breast cancers, 260 supplemented with transcriptome sequencing. They demonstrated how such approaches fill gaps in our understanding of the genome between the exons and expand the known repertoire of biological mechanisms underlying tumorigenesis with potential clinical utility (Alexandrov and Stratton, 2014; Alexandrov et al., 2013b; Davies et al., 2017; Zolkind and Uppaluri, 2017).

WGS has many unique capabilities that together enable the complete cataloguing of somatic variation within a single experimental protocol (Figure 1.2). Whereas targeted sequencing has the advantage of being efficient and affordable while capturing much of the known actionable variation, WGS carries advantages in the analysis of genomic instability and SVs. The research within this thesis relies upon insights gleaned from recent advances in the analysis and interpretation of cancer WGS data. The following sections provide an overview of unique capabilities of WGS which are leveraged in later chapters.

High Resolution Structural Variant Analysis

Few images are more deeply enscribed in the history of human genetics than the karyotype, which emphasizes the functional importance of genomic organization. Over the years, a plethora of technologies have enabled inspection of SVs and CNVs in the genome. Some, such as fluorescence in-situ hybridization, are precisely targeted, while others, such as array-comparative genomic hybridization, are comprehensive at varying resolutions. WGS promises to deliver precise and comprehensive characterization of both SVs and CNVs. While many challenges stand in the way of achieving a complete "digital karyotype" of cancer, WGS has made dramatic advances towards this goal.

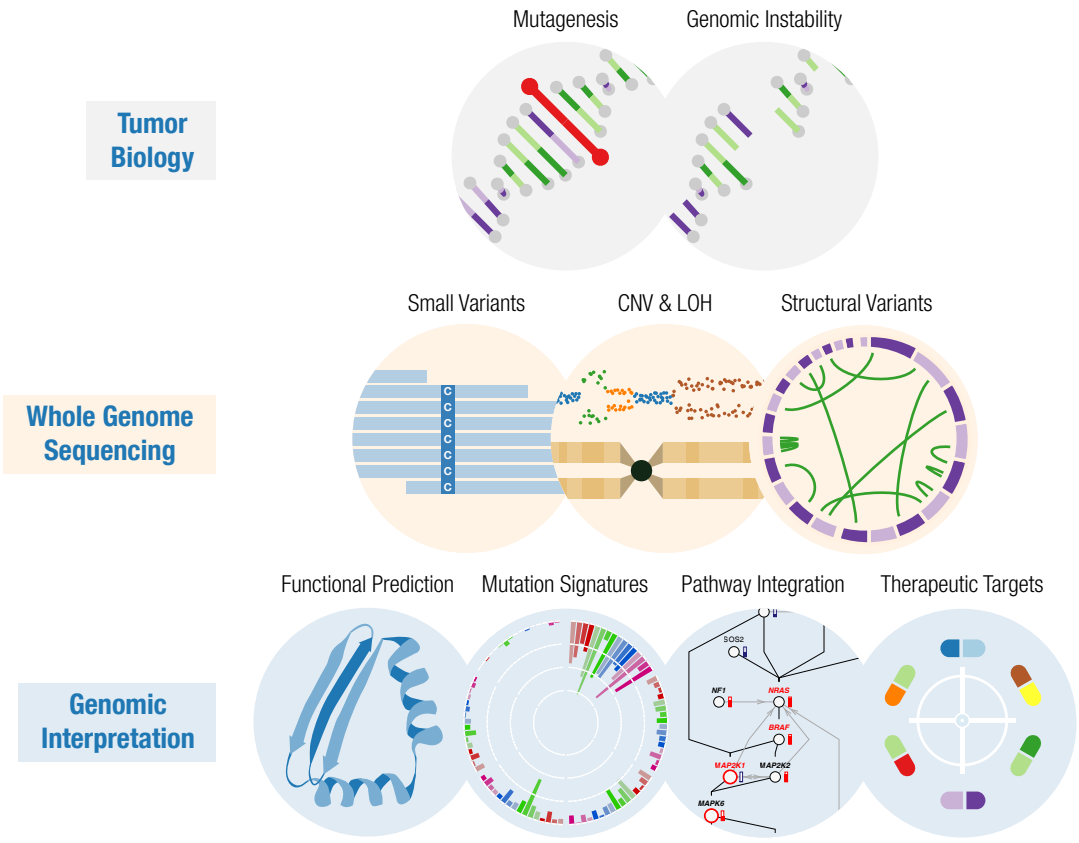

Figure 1.2: Whole genome sequencing data reveal diverse forms of genomic alteration. Tumour genomes exhibit frequent mutation and genomic instability which drive the hallmarks of cancer. Whole genome sequencing can catalog various forms of genomic alteration, enabling integrative analyses of tumour biology.

The cancer genome frequently features complex and interlocking patterns of somatic SVs, expanding the realm of possible cancer-driving alterations. Since the discovery of the Philadelphia chromosome (Nowell P., 1960), the characterization of oncogenic fusions has been central to cancer diagnosis and treatment. Aside from fusions, SVs also modulate gene regulation by rearranging the non-coding genome. Variants that impact the copy number or relative positioning of promoters and other regulatory elements can alter gene expression (Alaei-Mahabadi et al., 2017).

Paired-end WGS has become the standard for comprehensively and precisely cataloguing SVs and CNVs. While targeted arrays and WES can provide comparative read counts for CNV analysis, they lack the resolution to detect microamplifications and microdeletions and suffer from sequencing depth bias. These challenges, along with a need for computational methods to address them, limit the accuracy of CNV calls and result in high false discovery rates (Zare et al., 2017). While WES can detect gene fusions (Chmielecki et al., 2013), it misses fusions affecting splice sites, promoters, and other functionally critical loci.

SV analysis methods fall broadly into four categories: read density, split reads, paired end reads, and *de novo* assembly (Liu et al., 2015; Tattini et al., 2015). Recent methods combine these strategies to captalize on the unique strengths of each. For example, DELLY improves sensitivity by considering paired end reads while enabling base-pair breakpoint calling precision by examining split reads (Rausch et al., 2012). A major challenge in digital karyotype construction is poor concordance between SVs and CNV breakpoints (Alaei-Mahabadi et al., 2017), which suggests mismatched detection thresholds between these two modalities. This mismatch causes further difficulty in reconstructing complex rearrangements such as chromothripsis (Stephens et al., 2011) and chromoplexy (Baca et al., 2013).

SVs and CNVs play important roles in the somatic alteration of cancer genes, and are best comprehensively characterized by WGS. Evolving methodologies promise to move the field towards increasingly accurate reconstruction of the digital karyotype of cancer.

Methods for Analyzing Mutation Signatures

A mutation signature can be defined as a set of somatic mutation types which occur at specific relative frequencies. For example, a simple mutation signature could describe the relative frequencies of each base change amongst SNVs. To avoid redundancy, it is typical to collapse complementary bases, resulting in six mutation classes: C→A, C→G, C→T, T→A, T→C, T→G. A widely-used extension of this classification includes the 3′ and 5′ trinucleotide context. For example, a C→T mutation in an Ap**C**pG context would be considered one mutation class. Because there are four possible 3′ and four possible 5′ bases, this parameterization yields a total of 96 SNV classes (Alexandrov et al., 2013b).

This approach accounts for bias in the frequencies of mutations observed in specific trinucleotide contexts. For example, deamination of methyl-cytosine is a common cause of C→T mutations, and often occurs at CpG sites, which are frequently methylated. As a result, signatures of deamination often feature high rates of C→T mutations in N**C**G trinucleotide contexts (Nik-Zainal et al., 2012).

The analysis of mutation signatures involves two steps: (1) generating mutation count vectors, and (2) inferring signatures and exposures. Using the chosen parameterization, the number of mutations of each class is determined to yield a mutation count vector, also known as a mutation catalog. The inference of signatures and exposures was first performed using non-negative matrix factorization (NMF), which determines a dimensionally reduced set of mutation signatures and

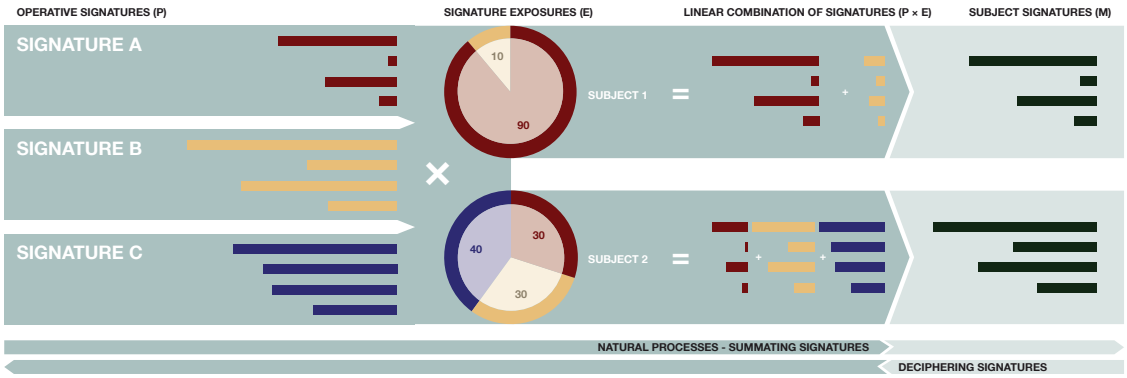

Figure 1.3: The analysis of mutation signatures by non-negative matrix factorization. Mutational occurrence probabilities are modeled as a linear combination of signatures. This defines a generative process whereby the genomes of individual cancers differ in the relative contributions of each mutation signature. In reality, the mutation count matrix (M) is known, and the signatures (P) and exposures (E) are unknown. Non-negative matrix factorization is an unsupervised learning method which infers a stable factorization of the mutation count matrix, which is often referred to as "deciphering mutation signatures *de novo*."

their relative contributions to each sample's genome (Alexandrov et al., 2013b). Thus, the mutation counts of a given genome are modeled as a linear combination of the signatures, which is consistent with the notion of multiple overlapping mutational processes each exerting additive effects (Figure 1.3). While this method can be used to derive mutation signatures both from genomes and exomes, the number of mutations sampled by WES often insufficient to detect all but the most hypermutating signatures, unless many hundreds of cancers are sequenced.

There are many variations on the analysis of mutation signatures which involve modifying (1) the parameterization of mutation types, or (2) the dimensionality reduction algorithm. An example of varying mutation types is the inference of SV mutation signatures in breast cancer (Nik-Zainal et al., 2016), wherein mutations were classified by size, type (deletions, duplications, inversions, and translocations), and whether or not their breakpoints are clustered. Alternative dimension-

ality reduction algorithms include principal components analysis (Gehring et al., 2015), expectation-maximization (Fischer et al., 2013), empirical Bayes (Rosales et al., 2017), and Bayesian NMF (Kim et al., 2016). Additionally, mutation signatures can be determined in subsets of the total mutation set to address specific biological hypotheses. For example, Supek and Lehner (2017) partitioned mutations into clustered and non-clustered sets to determine which mutational processes generate local hypermutation, and McGranahan et al. (2015) partitioned mutations into inferred early and late sets to examine mutation signature evolution.

The first integrative analysis of mutation signatures across cancer types aggregated 21 mutation signatures across 7,042 cancers, mostly from publicly available sequencing data (Alexandrov et al., 2013a). Later, this set of "consensus" mutation signatures was expanded to 30 to incorporate the continued discovery of novel signatures (for example Poon et al. (2015)). The 30-signature reference set is described in detail at http://cancer.sanger.ac.uk/cosmic/signatures.

Lastly, methods for determining the contributions of known mutation signatures to individual cancer genomes have also emerged (Huang et al., 2017; Rosenthal et al., 2016). The previously described mutation signature analysis methods can be considered *de novo* signature analysis, because they simultaneously infer mutation signatures and their contributions to individual cancers from scratch. By contrast, n-of-1 mutation signature analysis requires determining the best fit of an individual mutation profile against a set of known signatures. In the personalized cancer genomics paradigm, it is important to be able to reproducibly analyze the signatures of individual genomes to identify patterns of mutation with potential therapeutic implications.

Methods for Analyzing Tumour Heterogeneity

The analysis of intratumour heterogeneity in cancer is an area of active research. Significant advances in single-cell genome sequencing are enabling the directly sampling of genomic heterogeneity across tumour cells. While single-cell sequencing technologies are rapidly improving, resource limitations make them prohibitive at this time for use in clinical research, where large cancer cohorts are often required to distinguish significant effects. Another strategy for understanding tumour evolution involves the statistical inference of cancer cell subclones via the analysis of digital next-generation sequencing (NGS) read counts. This approach can be made more computationally feasible with the collection and comparison of multiple cancer sequencing timepoints.

NGS by Illumina-based protocols involves the capture, sequencing, and alignment of fragmented DNA reads. This process yields digital read depth counts at each locus, which, at sufficient sequencing depth, enables genome-wide statistical inference of DNA copy number. Additionally, somatic mutation loci can be queried for the fraction of total reads supporting the variant allele, also known as the variant allele fraction (VAF). The number of variant reads is related to (1) the tumour content or purity of the sample, (2) the tumour and normal cell copy number at the mutated locus, (3) the number of DNA copies carrying the mutation, also known as the mutation copy number (MCN), and (4) the fraction of cancer cells carrying the mutation, also known as cancer cell fraction (CCF).

If the tumour content and copy number are known, the MCN and CCF of a given mutation can be estimated from the VAF. The random sampling of fragmented reads from the bulk tumour sample also introduces noise. With sufficiently deep sequencing, discrete clusters representing mutations of varying MCN, and sometimes different CCF, can be observed. At lower sequencing depths,

it may be challenging to deconvolute MCN and CCF. A common goal for tumour heterogeneity inference methods is to infer the number of tumour subclones present as well as their relative CCF (Miller et al., 2014; Roth et al., 2014). These techniques are often designed to work on deep sequencing data, wherein sites of known somatic mutation are sequenced to a depth of hundreds or thousands of reads.

A simplified approach to CCF estimation has been applied to the temporal analysis of mutation signatures from WES or WGS data. This involves partitioning cancer somatic mutations according to their MCN and CCF (McGranahan et al., 2015) and has yielded promising findings regarding mutation signature evolution in lung cancer (Bruin et al., 2014), breast cancer (Yates et al., 2017), and liver cancer (Letouzé et al., 2017).

Complementarity with the Transcriptome

Alongside advancements in WGS, RNA sequencing technologies including whole transcriptome sequencing (WTS) can provide complementary insights in a personalized medicine setting. WTS enables genome-wide quantification of gene expression, which substantially expands the capture of potentially actionable molecular aberrations. Substantial efforts exist aiming to use tumour expression profiles to refine cancer diagnoses and subtyping. Translation of these efforts into interpretable and clinically actionable parameters such as the widely used PAM50 gene set for subtyping, stratification, and prognostication of breast cancers (Chia et al., 2012) can further advance these aims. The integration of gene expression data into functional clusters or pathways using statistical methods or visualizations can help identify dysregulated cancer pathways, including DNA repair pathways (Mulligan et al., 2014). This can provide rationale for guiding targeted cancer

treatment, including the use of experimental therapeutics (Tomasetti et al., 2017). When treatments fail, gene expression analysis can also elucidate resistance mechanisms and potentially suggest follow-up targets (Jones et al., 2010).

The existence and functional impacts of many events observed in the genome can be further analyzed by WTS. Amplified and deleted genes can be assessed for differential expression. The presence of oncogenic or deleterious mutations on the expressed transcript can be confirmed. Exon skipping and intron retention can be identified and potentially linked to splice site variants. Transcriptome assembly can facilitate detection of potentially oncogenic alternative transcripts. The presence of oncogenic gene fusions can be confirmed, and their expression verified. The effects of promoter and enhancer mutations on gene expression can be quantified. Tumour suppressors such as *BRCA1* and *MLH1* can be assessed for potential epigenetic silencing.

1.2.3 *From Cancer Genome to Personalized Oncogenomics*

The scope of sequencing and its applications has also broadened into the clinic. For specific cancer types, the use of targeted genomic panels for both germline susceptibility and known 'actionable' somatic mutations is becoming routine in many cancer centres (Bosdet et al., 2013; De Leeneer et al., 2011). The development and application of larger-scale gene panels is also seeing routine use on a large number of patient samples (Zehir et al., 2017). These high-throughput approaches drive discovery in the clinical context and quantify the frequency and prevalence of both well-characterized and novel variants in cancer-related genes. However, they capture a tiny fraction of the genomic complexity that can exist in an individual tumour.

Jones et al. (2010) published the first attempt to characterise a whole cancer genome for a clinical application. This analysis involved sequencing an adenocarcinoma of the tongue and identifying genomic amplification and concurrent abundant expression of the RET oncogene as a potential driver of the disease. This discovery led to a personalised treatment approach using kinase inhibitors to target the RET protein. Subsequent analysis of a post-treatment sample after disease progression provided comprehensive insight into how the tumour evolved to circumvent the treatment regimen in a way that a targeted approach could not have achieved. The success of this study led to a pilot for the Personalized Oncogenomics Project (POG) study at the BC Cancer Agency, now in its 5th year, which aims to leverage whole genome analysis with the intent to treat based on the genomic information. Sequencing of the initial 100 patients on this trial required development of pipelines and comprehensive interpretation tools (Laskin et al., 2015) and promised to restratify cancers by molecular features rather than by site of origin.

The inherent genomic complexity of cancers gives rise to a range of genomic events and signatures that are becoming increasingly relevant in patient treatment stratification. However, it is far from certain that an individual's tumour will harbour previously described and functionally characterized genomic events. The successful clinical application of personalized genomic medicine therefore must rely on broad screening approaches, a conclusion that was also reached in a study comparing WES and WGS in gastric cancer (Wang et al., 2014). A whole genome approach is currently the most efficient way to build a comprehensive picture of the genomic variation in a tumour without resorting to multiple technical platforms. That being said, there are still significant challenges that must be overcome before approaches such as whole genome and transcriptome analysis

(WGTA) can be universally adopted. However, on the assumption that sequencing costs will follow the historical downward trend, a more gradual uptake of WGTA for more refined stratification and subtyping of rare tumours may be achievable in the short term. Furthermore, as unanticipated clinical successes from WGTA continue to permeate the field and the infrastructure to support such approaches mature, a transition towards ever more comprehensive sequencing in the clinical setting is expected.

Outstanding Challenges in the Clinical Translation of Mutation Signatures

The study of mutational processes in cancer is yielding promising predictive biomarkers (Davies et al., 2017; Rizvi et al., 2015). However, many challenges stand in the way of clinical translation. While HRD mutation signatures have been proposed as targets for DNA damaging chemotherapy (Alexandrov et al., 2015b), their association with therapeutic outcomes has not yet been established.

Additional technical challenges limit biomarker studies of mutation signatures. The individualized analysis of SNV mutation signatures is possible, but existing approaches are susceptible to bleed of signal between like signatures, which can lead to false positive signature identification. Moreover, methods for individualized temporal dissection of mutation signatures provide point estimates without confidence intervals which are challenging to interpret in the clinical setting. Additionally, while SNV mutation signatures can be captured by WES, the accuracy of assessing processes with low to medium mutagenicity, such as HRD, is poor. WGS not only addresses this, but also simultaneously enables the detection of SV signatures (Nik-Zainal et al., 2016). However, unlike with SNV signatures, there is not yet a consensus set of SV signatures applicable across cancer types.

Lastly, the understanding of how mutation signatures evolve in the metastatic setting is limited, as nearly all studies of mutation signatures have taken place in primary, untreated cancers. Metastasis underlies as much as 90% of cancer-related mortality, (Chaffer and Weinberg, 2011), and the clinical features of metastatic dissemination and overall survival can be highly variable, even amongst subtypes of a cancer (Kennecke et al., 2010). Only recently have genomic studies investigated actionable cancer genes (Robinson et al., 2017) and the tumour evolution (Yates et al., 2017) of metastatic cancers. Understanding the factors which shape mutagenesis in metastatic cancers will help to guide the clinical application of mutation signatures.

1.3 THESIS OBJECTIVES AND CHAPTER OVERVIEW

The overarching goal of this this thesis is to elucidate the clinical actionability of mutation signatures and their temporal evolution. The HR pathway was an ideal model system to study clinical actionability of mutation signatures, as it relates to the commonly mutated genes *BRCA1* and *BRCA2* and to readily available platinum-based chemotherapy. I hypothesized that mutation signatures of HRD are independently associated with response and resistance to platinum-based chemotherapy, even in cancers lacking *BRCA1*/*BRCA2* mutations. I further hypothesized that in cancers with low or moderate HRD mutation signatures, observing the continued activity of HRD mutagenesis could be similarly associated with response to platinum-based chemotherapy. The corollary to this hypothesis is that past exposure to platinum-based chemotherapy would be associated with the suppression of HRD mutation signatures in late mutations.

The aim of chapter 2 is to assess the predictive value of mutation signatures arising from HRD. In this chapter, I describe an observational study of 93 advanced-stage breast cancers, 33 of which were treated with platinum-based chemotherapy. We made use of the recently-developed HRDetect metric, which combines six distinct signatures of HRD to form a more robust model that has been shown to predict mutations in *BRCA1/BRCA2* with high accuracy (Davies et al., 2017). We found that patients with increased HRDetect scores had improved outcomes on platinum-based chemotherapy, and that HRDetect was a superior predictor than any of the six signatures alone. This work was published in Clinical Cancer Research (Zhao et al., 2017).

Examining temporal shifts in mutation signatures required the development of novel mutation analysis software. In chapter 3, I present SignIT: mutation signature inference in individual tumours. SignIT is a Bayesian hierarchical model which provides improved accuracy and clinical interpretability of individualized signature analysis. A natural extension of the SignIT model enabled the inference of mutation signatures across tumour subpopulations with inferred relative timing.

Having formulated SignIT, I demonstrated two applications of mutation signature timing. In chapter 4, I analyzed the evolution of HRD in a pancreatic adenocarcinoma to address the hypothesis that ongoing HRD activity is associated with treatment outcome. In chapter 5, I report mutation signatures and their temporal evolution across 484 metastatic cancers. This analysis revealed an association between prior platinum-based chemotherapy exposure and the suppression of HRD in late mutations. I also uncovered signatures arising from chemotherapy-associated mutagenesis.

2

THE CLINICAL ACTIONABILITY OF HOMOLOGOUS RECOMBINATION DEFICIENCY IN ADVANCED BREAST CANCER

2.1 INTRODUCTION

Genomic instability and mutagenesis are hallmarks of human cancers that can arise from deficient DNA repair processes. One such process, HR, involves strand invasion by homologous sequences to facilitate error-free repair of double strand breaks and inter-strand crosslinks (Li and Heyer, 2008). Mutations in genes responsible for HR are prevalent among human cancers. The *BRCA1* and *BRCA2* genes are centrally involved in HR, DNA damage repair, end resection, and checkpoint signaling (Joosse, 2012). Inherited mutations in *BRCA1* and *BRCA2* account for 5-10% of all breast cancers, conferring an up to 85% lifetime risk (Canadian Cancer Society, 2014; National Cancer Institute, 2014). There is also emerging evidence suggesting that germline *BRCA1* and *BRCA2* mutated cancers are associated with sensitivity to platinum-based chemotherapy (Arun et al., 2011; Byrski et al., 2010; Tutt et al., 2015; Von Minckwitz et al., 2014) and PARP inhibitors (Robson et al., 2017). This is further supported by resistance to platinum-based agents arising from secondary mutations that cause somatic reversion of germline *BRCA1/2* variants (Norquist et al., 2011).

HRD is complex, and its myriad causes are not fully understood. However, examining characteristic patterns of mutation, collectively known as mutation signatures or genomic scars, can provide an aggregate functional metric of pathway function. For example, *BRCA1* and *BRCA2* are associated with characteristic CNV patterns (Timms et al., 2014), which have been suggested to independently predict platinum sensitivity in primary breast cancer (Telli et al., 2016). However, a clinical trial in advanced stage triple negative breast cancer did not verify this association (Tutt et al., 2015). Meanwhile, new genomic correlates have refined the detection of HRD. Large-scale genome profiling across thousands of cancers has revealed characteristic patterns of mutation giving rise to millions of somatic SNVs (Alexandrov et al., 2013a) and SVs (Nik-Zainal et al., 2016). Recent efforts aggregated six HRD-associated signatures into a single score called HRDetect to accurately classify breast cancers by their *BRCA1* and *BRCA2* status (Davies et al., 2017).

With this improved capability to quantify "BRCA-ness," there is substantial interest in its therapeutic implications in breast cancer (Alexandrov et al., 2013a; Davies et al., 2017; Jacot et al., 2015; Lips et al., 2013; Stecklein and Sharma, 2014). Importantly, these measures may be able to identify *BRCA1*- and *BRCA2*-intact but HR-deficient tumours to guide eligibility for HRD-targeted clinical trials and treatment decision-making. However, there is not yet direct evidence that aggregated genomic scar metrics predict platinum sensitivity. In this observational biomarker study, we perform WGS to identify HRD mutation signatures in a cohort of 93 patients with advanced stage breast cancers and associate them with molecular, pathologic, and clinical features. Using HRDetect, we aggregate HRD signatures and demonstrate their association with clinical benefit on platinum-based chemotherapy.

2.2.1 *Somatic Mutation Signatures*

Using a published framework (Alexandrov et al., 2013b), we deciphered the mutation signatures of 1,182,840 somatic SNVs and 11,393 SVs from the whole genomes of 93 advanced-stage breast cancers.

Of the nine resulting SNV signatures, numbered V1-V9 (Fig. 2.1A), six closely matched previously described mutation signatures available from COSMIC. V9 (Signature 3) and V6 (Signature 8) are associated with HRD (Alexandrov et al., 2013a; Davies et al., 2017; Nik-Zainal et al., 2016). V4 (Signature 1) is associated with aging (Alexandrov et al., 2015a). V1 (Signature 2) and V2 (Signature 13) are associated with APOBEC deaminase activity. V3 (Signature 17) has been observed across many cancer types, but its etiology is unclear.

The three remaining signatures, V5, V7, and V8, represent novel breast cancer mutational signatures. V5 predominantly displays C→T mutations in CpCpY contexts (see Table A.1 for nomenclature) and was present in only three cancers. V7 is characterized by high pyrimidine transition rate with enrichment in NpYpG contexts and was observed across many tumours spanning histological and molecular subtypes. V8 demonstrated moderate enrichment of all base substitution types when flanked by T and A bases, and was present at low levels across many tumours. These novel signatures may reflect the advanced, recurrent, and drug-treated nature of our cohort, whereas previous mutation signatures have been derived from primary untreated cancers. Further study is necessary to verify etiology. Potential etiologies of signatures V3 and V5 will be discussed in more depth in chapter 5.

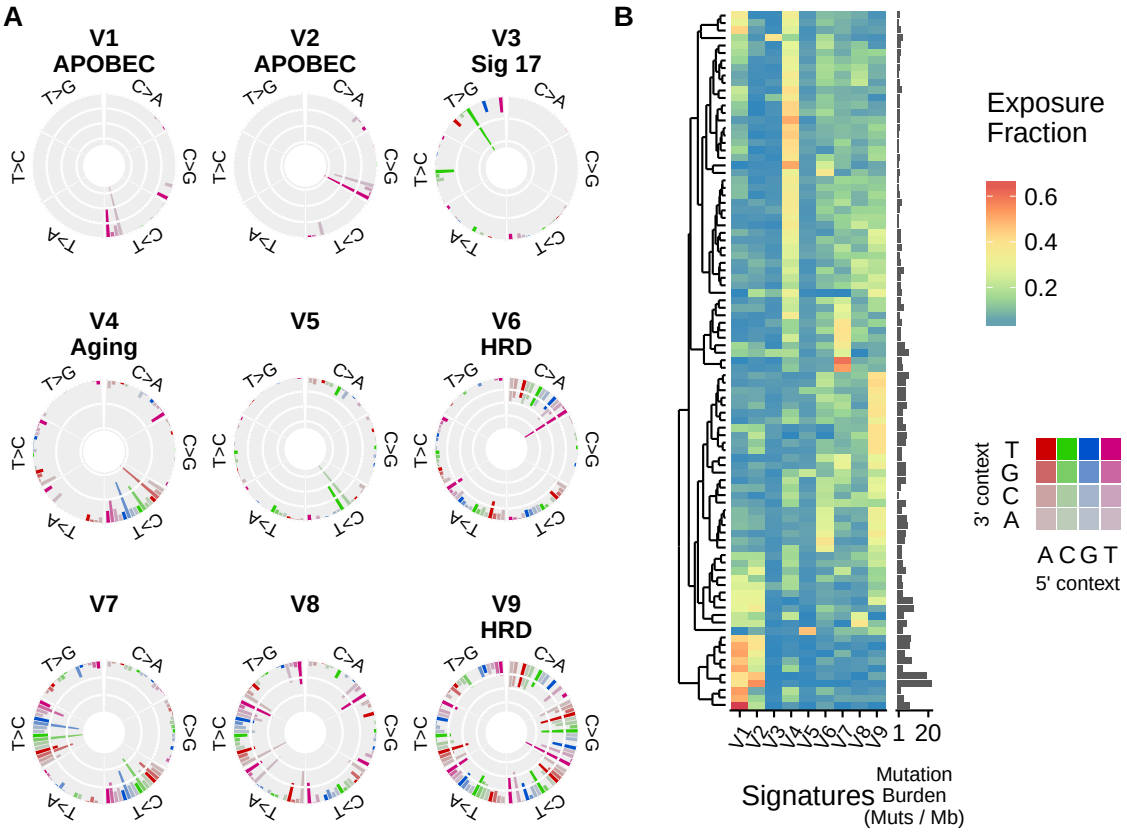

Figure 2.1: Nine signatures of single nucleotide variation deciphered from 93 breast cancer whole genomes. (A) Signatures are visualized according to relative frequencies of mutations grouped by base change and 3′/5′ context. Six of nine signatures match previously published mutation signatures (cosine similarity > 0.9), five of which are associated with hypothetical etiologies. (B) Fractional exposures and mutation burdens across the patient cohort, ordered by hierarchical clustering, reveals groups defined by aging (top cluster with dominant V4), homologous recombination deficiency (middle cluster with dominant V9), and APOBEC deamination (lower cluster with dominant V1 and V2).

Hierarchical clustering revealed that most cases of high SNV burden were driven by APOBEC or HRD associated processes (Fig. 2.1B), which together were dominant in 46 (49%) of the 93 sequenced breast cancers. The aging mutation signature was ubiquitous across cancers, and was the dominant signature in 31 (33%) cases, all of which had low mutation burden (< 5 SNVs per Mb).

92 samples were classified into intrinsic subtypes based on expression profiles of PAM50 (Chia et al., 2012) genes. Non-parametric analysis demonstrated significant differences in signatures V2, V3, V8, and V9 across subtypes (Appendix Table A.2). Post-hoc pairwise Dunn tests revealed elevated V3, V8, and V9 within basal-like cancers (Figure 2.2), suggesting that diverse mutagenic etiologies, including HRD, underlie this subtype. Elevated signature V9 was also most common among triple-negative tumours.

The six deciphered SV signatures (Figure 2.3), numbered R1-R6, closely resembled the six previously described breast cancer signatures reported by Nik-Zainal et al. (2016). R1-R4 and R6 uniquely matched previously described signatures. By visual inspection, R5 matches previously described rearrangement signature 5 albeit with more non-clustered translocations.

2.2.2 *Genomic Findings Associated with HRD*

Alongside these four SNV and SV mutation signatures, we measured two additional HRD-associated patterns of somatic mutation, the HRD index, and microhomology at deletion breakpoints. The HRD index measures the frequency of large scale loss of heterozygosity (LOH), telomeric allelic inbalance (TAI), and large-scale transition (LST) events (Timms et al., 2014), and was computed using allelic copy number ratios inferred from read alignment frequencies. The proportion of

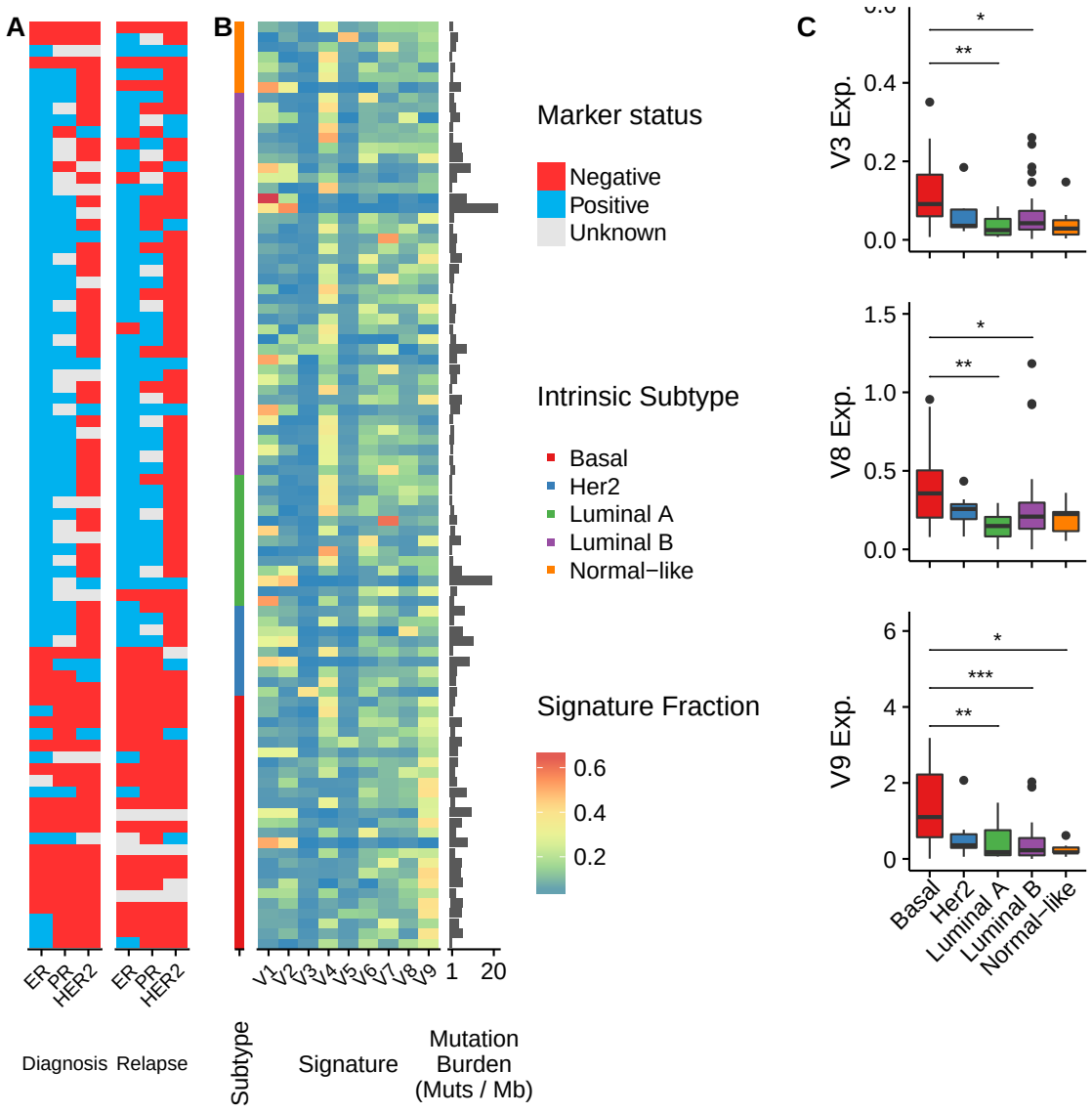

Figure 2.2: Breast cancer signatures across subtypes. Comparison of SNV mutation signatures across (A) histological and (B) molecular subtypes shows more frequent signature V9 (homologous recombination deficiency) exposure in triple-negative and basal-like breast cancers. (C) Four signatures (V2, V3, V8, V9) exhibited statistically significant differences across molecular subtypes (Kruskal-Wallis test with adjusted p-values). Subtype-specific signature exposures are shown here, with pairwise statistical significance testing performed by the Dunn non-parametric test of multiple comparisons.

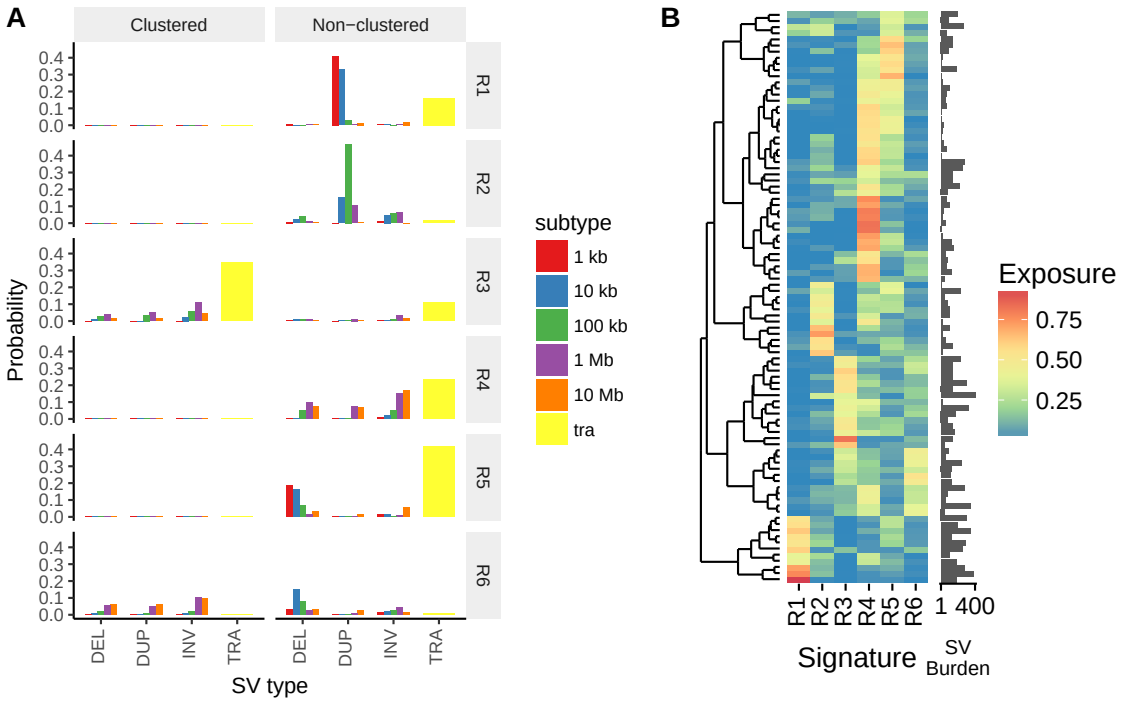

Figure 2.3: Breast cancer structural variant signatures. (A) Six structural variant (SV) mutation signatures deciphered from breast cancer whole genomes. SVs were classified according to breakpoint clustering, mutation type, and size. (B) Signature exposures were normalized to sum to 1 for each sample, then arranged according to hierarchical clustering alongside SV mutation burden.

small deletions associated with microhomology was determined by comparing sequences flanking deletion breakpoints. As per a published method (Davies et al., 2017), all six scores were log transformed, normalized, and combined into a single HRDetect predictor. This was performed using a logistic function with the same coefficients as those reported by Davies et al. (2017) to ensure consistency with the previously model.

19 breast cancers had high HRDetect scores (> 0.7), 37 had moderate scores (0.005-0.7), and 37 had low scores (< 0.005). All cancers underwent genome-wide characterization of germline and somatic point mutations, insertions and deletions, and copy loss in gene regions and splice sites.

Across the 93 breast cancers, HRDetect predicted pathogenic germline and somatic variants in BRCA1 and BRCA2 with high accuracy and an optimal differentiating threshold of 0.74 (Fig. 2.4B). These findings closely agree with the previously established threshold of 0.70 (Davies et al., 2017). Because variants of uncertain significance (VUS) have previously not been associated with increased HRDetect (Davies et al., 2017), we classified VUS as non-pathogenic mutations for the purposes of this analysis. Elevated HRDetect scores were observed in all tumours with observed *BRCA1/BRCA2* frame shifts, nonsense mutations, homozygous deletions, or splice variants identified as likely pathogenic in ClinVar (Fig. 2.5). There were 11 cases with germline missense VUS. The most common of these was *BRCA2* T1915M, which had a global minor allele frequency of 1.14% and has conflicting reports of both reducing (Serrano-Fernández et al., 2009) and increasing (Johnson et al., 2007) breast cancer risk. In our study, seven breast cancers (BR004, BR027, BR032, BR036, BR064, BR074, BR086) harboured germline *BRCA2* T1915M, of which three (BR004, BR036, BR086) were homozygous in the tumour and displayed a wide range of HRDetect scores (0, 0.04, and 0.62 respectively).

However, BR086 exhibited coincident homozygous deletion of *RAD51* which may account for the elevated score. These data therefore do not provide clear evidence for pathogenicity of *BRCA2* T1915M.

A number of other genes involved in HR demonstrated tentative associations with HRDetect scores. Elevated HRDetect was observed in three cases with homozygous deletion of *PTEN* as well as one case with two coincident *PTEN* missense mutations (F278L and P38S). However, one case with homozygous *PTEN* A126D somatic mutation was associated with a low HRDetect score. Homozygous deletions in *RAD50*, *RAD51*, and *MCPH1* were observed in some tumours with moderate or high HRDetect scores. *MCPH1* is a potential cancer susceptibility gene (Mantere et al., 2016) whose deletion may be a poor prognostic marker (Tsuneizumi et al., 2002). Although recurrently deleted in our cohort, its link to HRD signatures was inconsistent.

High HRDetect scores were also associated with triple negative and basal-like breast cancers (Table 2.1). Of 19 samples with high HRDetect, 11 (58%) were classified as basal-like. Among low HRDetect samples, only 2 (5%) were basal like. Luminal B and normal-like tumours were more likely to have low HRDetect scores, whereas most (7/9) HER2-like tumours displayed moderate HRDetect. Receptor status was assessed by immunohistochemistry and retrieved from pathology records, which were available for 79 tumours at primary and 76 at relapse (Figure 2.2). High HRDetect was inversely associated with positive receptor status in all three receptors. 50% of high HRDetect tumours were triple negative, compared to only 6% of primary and 15% of metastatic low HRDetect tumours.

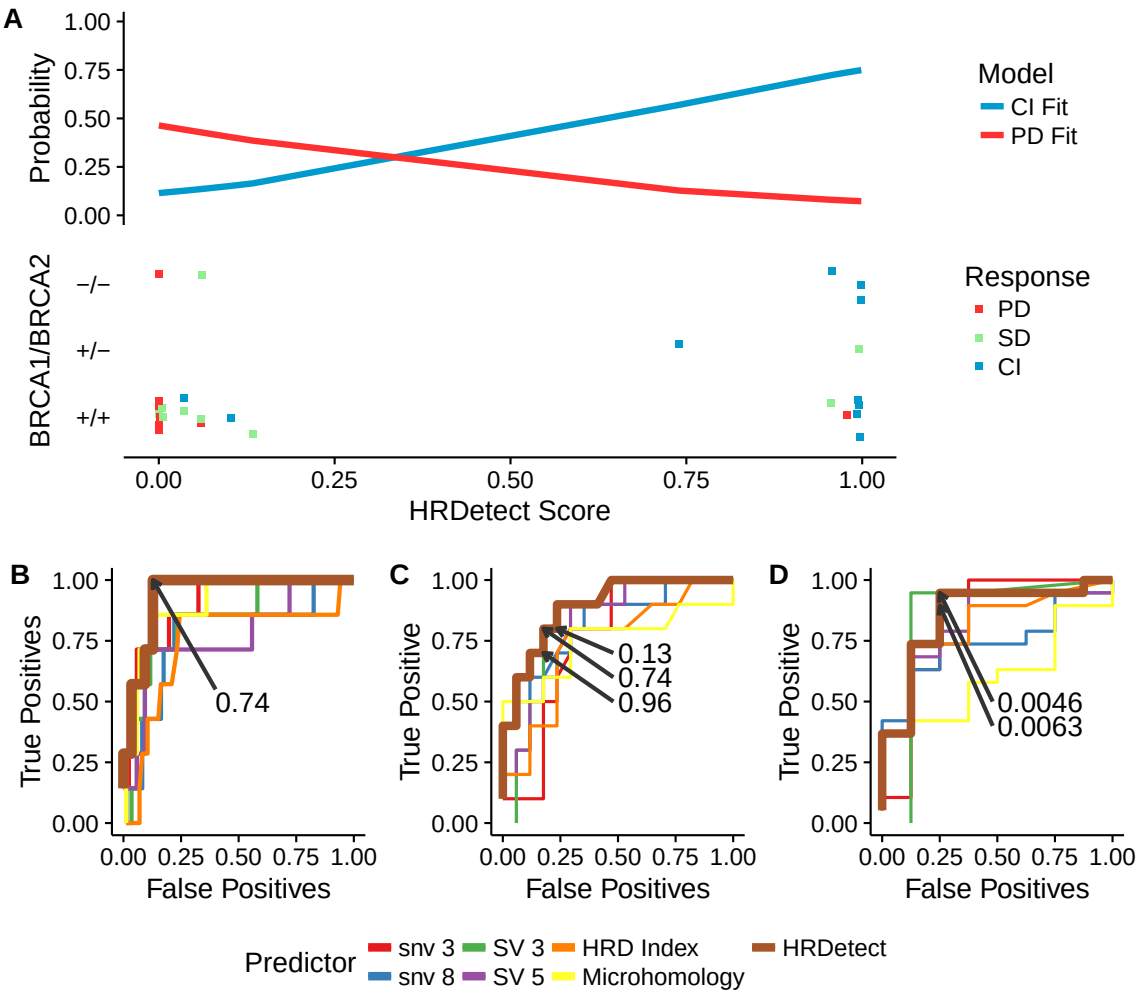

Figure 2.4: Association of platinum-based treatment outcomes with HRDetect, an aggregate of six homologous recombination deficiency (HRD) mutation signatures. (A) The HRDetect score is significantly associated with clinical improvement (CI) on platinum-based chemotherapy (logistic regression, adjusted for *BRCA1/2* status and treatment timing, p = 0.006). There was also a trend between low HRDetect and progressive disease (PD; p = 0.112). Moreover, of 8 *BRCA1/2*-intact cases with elevated HRDetect score, 5 responded favorably to platinum-based chemotherapy. Receiver-operator characteristic for (B) BRCA status and (C,D) therapeutic outcomes on platinum-based chemotherapy (C: CI; D: stable disease, SD). These suggest optimal HRDetect thresholds of 0.7 and 0.005 for CI and SD respectively. Specific near-threshold HRDetect values are labelled. In all three ROC curves, HRDetect had a superior area under the curve than its six constituent mutation signatures.

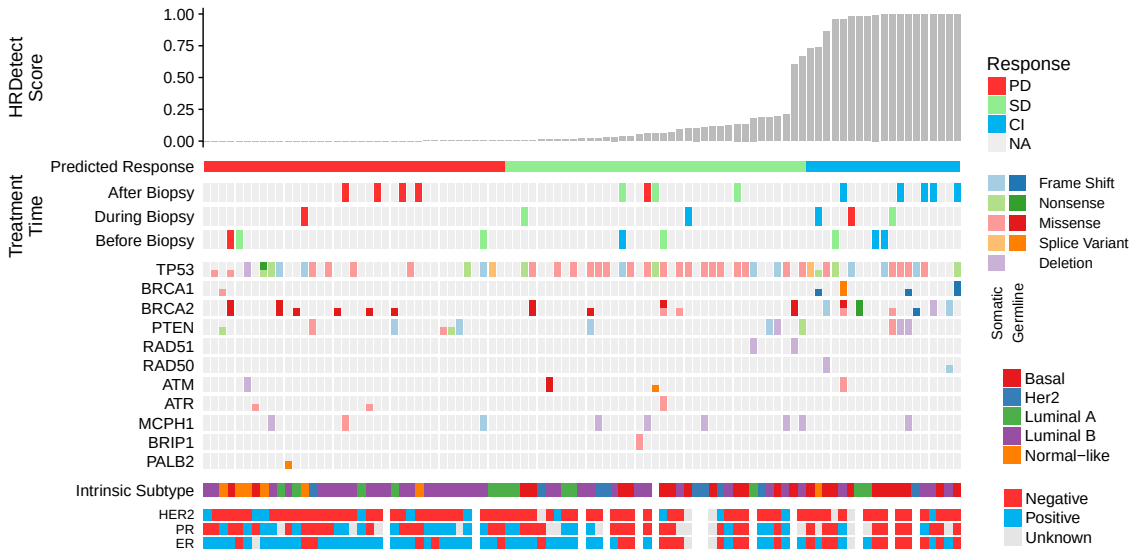

Figure 2.5: HRDetect scores, mutations in key homologous recombination genes, and outcomes on platinum-based therapy. Six distinct mutation signatures associated with homologous recombination deficiency (HRD) were deciphered from 93 breast cancer whole genomes and aggregated into a single HRDetect score. Radiology reports during and after treatment regimens involving platinum-based chemotherapy were reviewed for evidence of clinical improvement (CI), stable disease (SD), or progressive disease (PD). Analysis of receiver-operator characteristic curves suggested HRDetect thresholds of 0.7 for CI and 0.005 for SD, indicated here by a colourbar.

Table 2.1: Summary of patient molecular and clinical characteristics by HRDetect status.

HRDetect status	Low (<0.005)	Moderate	High (>0.7)	Total
Sample counts				
Total Count	37	37	19	93
Treated Count	9	13	11	33
Treated and Imaged	8	7	11	26
Pathogenic *BRCA1/2* Variant	0	0	7	7
Response to Platinum				
CI	0	2	8	10
SD	2	4	2	8
PD	6	1	1	8
Median TDT (days)	56 (n=9)	71 (n=13)	143 (n=11)	
Median OS (days)	122 (n=6)	160 (n=8)	384 (n=5)	
Intrinsic Subtype				
Basal	2	12	11	24
HER2 Amplified	1	7	1	9
Luminal A	6	5	2	13
Luminal B	22	12	4	38
Normal-like	6	0	1	7
Primary receptor status				
ER (positive/negative)	31 / 3	20 / 9	7 / 8	58 / 20
PR (positive/negative)	18 / 4	11 / 10	4 / 12	33 / 26
HER2 (positive/negative)	4 / 23	4 / 22	0 / 14	8 / 59
Triple negative	2 (6%)	8 (28%)	8 (50%)	18
Metastatic receptor status				
ER (positive/negative)	27 / 6	17 / 10	5 / 10	49 / 26
PR (positive/negative)	15 / 15	9 / 13	2 / 10	26 / 38
HER2 (positive/negative)	6 / 28	4 / 22	1 / 13	11 / 63
Triple negative	5 (15%)	6 (21%)	8 (50%)	19

2.2.3 *HRD Mutation Signatures are Associated with Platinum Outcomes*

High HRDetect scores were significantly associated with clinical improvement on platinum-based chemotherapy, even after adjusting for *BRCA1*/*BRCA2* status and treatment timing (p = 0.006, n = 26; Table 2.4). HRDetect demonstrated areas under the ROC curve of 0.89 for clinical improvement (CI) and 0.86 for stable disease (SD), which exceeded those of its component signatures (Fig. 2.4B, C; Table 2.3). Optimal thresholds of 0.005 for predicting SD and 0.7 for predicting CI were chosen (Fig. 2.4B, C). Sensitivity, specificity, positive predictive value, and negative predictive value were computed for both thresholds and are reported in Table 2.2.

Biallelic loss of *BRCA1* or *BRCA2* was also associated with clinical improvement on platinum-based chemotherapy (Fig. 2.4A) but was observed in only three of 26 treated patients with available imaging. By comparison, 11 patients demonstrated HRDetect scores above 0.7, of whom 8 experienced CI, 2 experienced SD, and 1 had disease progression. Therefore, HRDetect scores correctly identified five additional patients without biallelic loss of *BRCA1* or *BRCA2* who benefited from platinum-based therapy. In a joint logistic model, *BRCA1* and *BRCA2* status did not contribute significantly to the predictive value of HRDetect (Table 2.4).

Table 2.2: Test metrics of HRDetect predictions computed using specified thresholds. Elevated HRDetect scores computed from whole genome sequencing of a breast cancer cohort were associated with improve response to platinum-based chemotherapy. Receiver-operator characteristic (ROC) curves suggested thresholds of 0.005 for stable disease (SD) and 0.7 for clinical improvement (CI). Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were computed based on true/false positive/negative rates for both thresholds.

Response	threshold	accuracy	sensitivity	specificity	PPV	NPV
SD or CI	0.005	0.85	0.89	0.75	0.89	0.75
CI	0.7	0.81	0.8	0.82	0.73	0.88

Table 2.3: Area under the curve of homologous recombination deficiency (HRD) signatures in platinum response prediction. Six distinct HRD-associated mutation signatures were computed using whole genome sequencing data from 93 advanced-stage breast cancers. The six signatures were normalized and aggregated using a logistic regression model with coefficients trained in a previous study (Davies et al., 2017). Retrospective clinical review was performed to classify best reported radiographic response to platinum-based chemotherapy into three categories: clinical improvement (CI), stable disease (SD), and progressive disease (PD). Receiver-operator characteristics (ROC) were computed for the aggregated metric, called HRDetect, as well as the six original signatures. Treatment success groups were defined as either CI or the union of CI and SD response groups. For both success metrics, the area under the curve of each ROC curve is reported here.

Predictor	*BRCA1/2* status AUC	CI AUC	CI & SD AUC
snv 3	0.897	0.756	0.836
snv 8	0.777	0.826	0.743
SV 3	0.832	0.838	0.845
SV 5	0.769	0.821	0.822
HRD Index	0.743	0.741	0.812
Microhomology	0.899	0.75	0.605
HRDetect	0.94	0.891	0.855

Table 2.4: Logistic regression model odds ratios of clinical improvement (CI) on platinum-based chemotherapy. HRDetect scores were computed using six mutation signatures derived from whole genome sequencing of a breast cancer cohort. Germline and somatic assessment of mutation status, deletions, and loss of heterozygosity of *BRCA1* and *BRCA2* were assessed to determine mono-allelic and bi-allelic loss of function. Retrospective clinical review classified responses to platinum-based chemotherapy, which was modelled using logistic regression with HRDetect scores, *BRCA1* & *BRCA2* status, and treatment timing as predictors. HRDetect was significantly associated with platinum response with a log odds ratio of 3.2 (odds ratio = 16, p = 0.006).

	z	p	Log Odds Ratio	Lower CI	Upper CI
Intercept	-1.9	0.061	-2.1	-4.6	-0.12
HRDetect	2.8	0.0057	3.2	1.1	5.7
BRCA+/-	-0.22	0.83	-0.46	-4.8	3.9
BRCA-/-	0.54	0.59	0.73	-2.1	3.7
Tx During Biopsy	-0.12	0.9	-0.21	-4.1	3.1
Tx After Biopsy	-0.023	0.98	-0.028	-2.5	2.5

2.2.4 *Effects of HRDetect on Overall Survival and Treatment Duration*

Of patients treated post-biopsy with platinum-based chemotherapy, there was a statistically significant difference in overall survival (OS) depending upon HRDetect (p = 0.04, n = 33). 5 patients with predicted CI (HRDetect > 0.7) demonstrated a median survival of 384 days, 8 with predicted SD (0.7 > HRDetect > 0.005) had a median survival of 160 days, and 6 patients with predicted progressive disease (PD) (HRDetect < 0.005) had a median survival of 122 days. This difference should be interpreted with caution due to small sample size and other treatments received besides platinum, but represents a promising trend which warrants further study.

In addition to OS, total duration on platinum-based therapy (TDT) was used as a surrogate for clinical response. In practice, platinum-based chemotherapy is typically continued in responding patients until disease progression or significant toxicity. Figure 2.8 verifies that, in 26 patients with available imaging, patients with reported radiographic response were more likely to undergo a longer duration of treatment. HRDetect scores were significantly associated with extended TDT with a hazard ratio of 0.24 (0.081 - 0.95; p = 0.01, n = 33), after adjusting for *BRCA1* and *BRCA2* mutation status, timing of treatment, and patient age (Fig. 2.6B). Tumours were classified based on HRDetect scores into predicted treatment response categories. There was a significant difference in TDT ($p < 0.001$, n = 33; Fig. 2.6A) between patients with predicted CI (median 143 days), SD (median 71 days), and PD (median 56 days). This amounts to an estimated three-month difference in median TDT between high HRD and low HRD cases.

2.2.5 *Feasibility of HRD Analysis in Personalized Medicine*

The development of precision oncology initiatives (Laskin et al., 2015; Meric-Bernstam et al., 2013; Mestan et al., 2011; Zehir et al., 2017) has necessitated genome analysis pipelines compatible with "N of 1" cases. One challenge of mutation signature analysis by NMF is the reliance upon large cohorts of sequenced tumours. This has led to techniques to determine the most likely composition of signatures for a single isolated sample (Rosenthal et al., 2016). HRD analysis provides a promising target for personalized treatment decision-making. Thus, in addition to cohort-based *de novo* signature discovery, we also computed individual-tumour best fit signature exposure profiles for HRD-associated SNV signatures 3 (V9) and 8 (V6) and SV signatures 3 (R1) and 5 (R5) using non-negative least

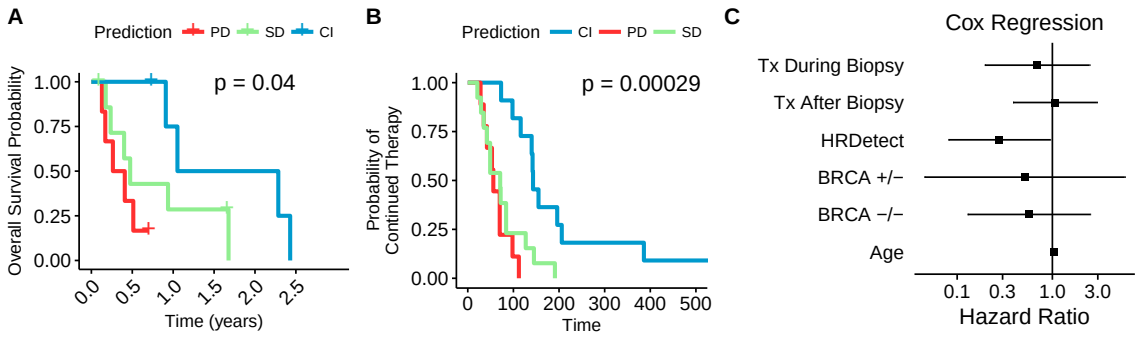

Figure 2.6: Homologous recombination deficiency is associated with extended overall survival (OS) and total duration on platinum-based therapy (TDT). (A) Among patients treated after the sequencing biopsy (n = 19), OS was computed as the duration between first post-biopsy treatment and death. There was a statistically significant (p = 0.04) difference between patients predicted to be CI (HRDetect > 0.7), SD (0.7 > HRDetect > 0.005), and PD (HRDetect < 0.005). (B) Platinum-treated patients (n = 33) with different predicted treatment outcomes also experienced significantly different TDT as part of standard care for advanced breast cancer. (C) Multivariate Cox survival model demonstrated a significant association between HRDetect and TDT independently of *BRCA1/2* mutation status. 95% confidence intervals are shown for the hazard ratio.

squares (NNLS) - details in Methods. We then recomputed HRDetect scores using these individualized NNLS signature exposures to assess accuracy.

HRDetect scores and all four HRD-associated SNV and SV signatures demonstrated high concordance between NMF and NNLS approaches based on Pearson linear regression ($r > 0.9$; Fig. 2.7). Employing the selected thresholds of 0.005 for SD and 0.7 for CI, 86 out of 93 cancers were concordantly classified by NNLS and NMF, including all cases predicted to experience CI. NNLS reclassified 4 cancers from PD to SD, and 3 from SD to PD.

These findings demonstrate that NNLS-based N of 1 computation of mutation signature exposures provides robust HRD estimates concordant with a cohort-based NMF approach. This is promising for the application of HRD biomarkers in sequencing-driven treatment guidance. However, this approach may not translate to WES data or similarly targeted sequencing approaches due to the lower numbers of sampled mutations.

2.3 DISCUSSION

In this retrospective study, HRD mutation signatures were associated with clinical benefit on platinum based chemotherapy in advanced stage breast cancer. Specifically, we demonstrated that HRDetect, the same model independently trained to predict *BRCA1* and *BRCA2* status with high sensitivity and specificity (Davies et al., 2017), was also significantly associated with favorable response to platinum chemotherapy response and longer TDT. Moreover, we identified an optimal HRDetect threshold of 0.7, which agrees with the previously established cut-off for *BRCA1*/*BRCA2* status (Davies et al., 2017). Therefore, our findings both inde-

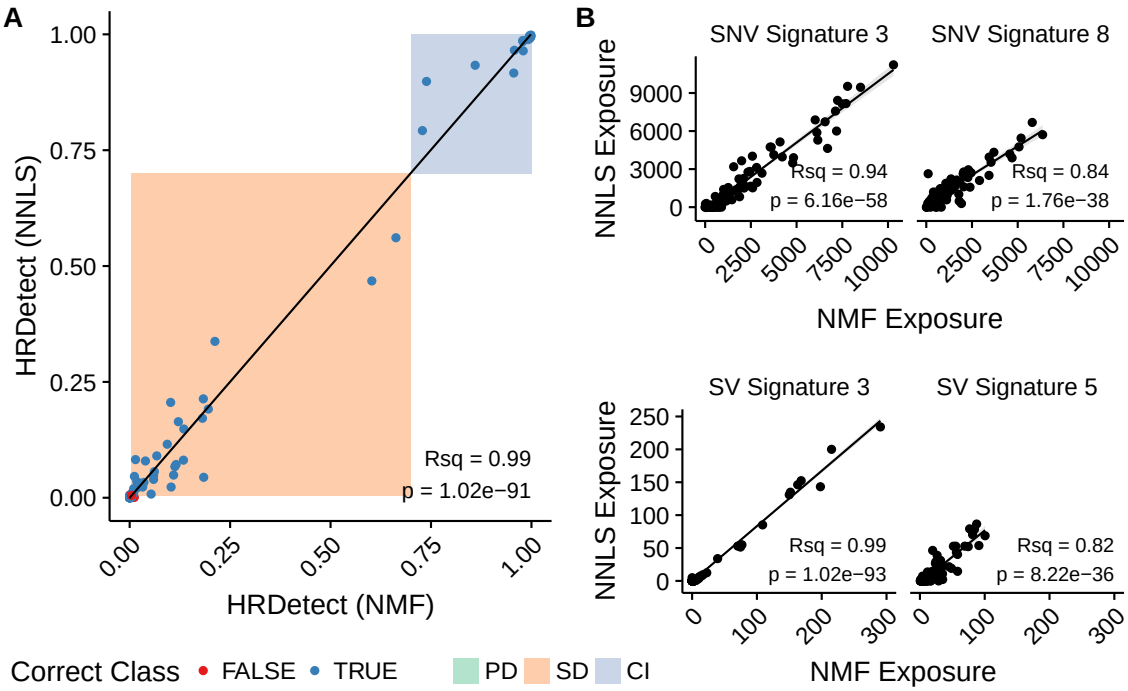

Figure 2.7: N of 1 signatures by non-negative matrix factorization (NNLS) accurately reproduce HRDetect scores. (A) HRDetect scores were computed using component signatures derived from both NNLS and non-negative matrix factorization (NMF). Scores obtained by the two approaches were strongly correlated (Pearson's R squared = 0.99) and demonstrated high classification concordance based on selected thresholds. (B) Individual HRD-associated mutation signatures were concordant between the two approaches (Pearson's R squared > 0.82 for all signatures).

pendently validate the HRDetect model and provide promising evidence for its clinical relevance.

A key limitation of this study is the ability to establish causation. As this was an observational cohort of advanced-stage breast cancers undergoing standard chemotherapy treatments, some patients were sequenced during or after courses of platinum-based chemotherapy. To mitigate the impacts of tumour evolution, we limited analyses to patients sequenced within two years of treatment. Another significant challenge when studying treated tumours is that platinum-associated mutagenesis may impact the mutation signature profile, especially in cancers biopsied after treatment. A few factors help to mitigate this challenge, but cannot entirely rule out platinum-induced mutagenesis. First, we adjusted for the treatment timing in statistical analyses of the association between HRDetect and clinical outcomes. Second, there has been reproducible evidence of HRD-associated signatures in cohorts of predominantly primary tumours (Alexandrov et al., 2013a; Davies et al., 2017; Nik-Zainal et al., 2012, 2016; Timms et al., 2014), which are a close match to the signatures we found. Lastly, the aggregation of six distinct signatures into a more robust metric should help mitigate the impact of platinum-induced mutagenesis affecting any one signature in particular. Notably, the investigation of advanced stage breast cancers is an important feature of this study. Whereas a previous trial did not find that the HRD index alone was predictive in advanced breast cancer (Tutt et al., 2015), our findings renew promise for aggregated metrics such as HRDetect. However, studying advanced stage tumours inevitably introduces potential confounders such as variable treatment histories. Therefore, well-designed prospective clinical trials are needed to further validate HRDetect as a predictive biomarker.

Another caveat is the threshold selection for predicting CI. A threshold of 0.7 was chosen because it both agrees with the model trained by Davies et al. (2017) and optimally separated responders from non-responders in our cohort. However, there was a sharp decline in HRDetect scores below 0.7, with no cases falling between 0.25 and 0.5, and no cases with treatment response data between 0.5 and 0.7. This suggests that a superior threshold may exist between 0.25 and 0.7, and a study with greater sample size may be require to pinpoint it.

HRD is common among breast cancers. Based on our HRDetect predictive thresholds, 19 cases (20%) showed potentially targetable high HRD status (HRDetect > 0.70). An additional 37 cancers (40%) showed moderate HRD status consistent with stable disease on platinum-based chemotherapy (HRDetect > 0.005). By comparison, biallelic germline and somatic mutations were detected in only 11 cases, and known pathogenic variants in only 7. Similarly, a previous analysis of 560 breast cancer genomes, which additionally examined promoter hypermethylation, estimated the frequency of BRCA-null breast cancers at 14% (Nik-Zainal et al., 2016). The analysis of HRD signatures may identify patients who could benefit from platinum-based therapy otherwise undetected on *BRCA1/2* screening. These signatures may also have implications for sensitivity to PARP inhibitors, which exploit a synthetic lethal interaction between *PARP-1* and the HR pathway. Germline mutations in *BRCA1* and *BRCA2* are associated with improved response to PARP inhibitors (Robson et al., 2017). Additional translational research incorporating WGS is necessary to reveal whether HRD mutation signatures are similarly associated with PARP inhibitor response independently of *BRCA1/2* status.

Clinical translation of HRD mutation signatures requires sufficient capture of somatic SNVs and SVs to infer the processes underlying mutagenesis. While HRDetect improves upon the accuracy of the clinically employed LOH, TAI, and

LST metrics, it requires WGS, which currently poses technical and financial challenges for clinical use. Further research to develop predictive models that exclude SV signatures may enable application on cancer exomes or other targeted sequencing methods, which can capture sufficient somatic mutations for SNV signature but not SV signature analysis. Additionally, orthogonal HRD assays, for example employing gene set expression profiling (Mulligan et al., 2014), may also serve as lower cost parameters for treatment prediction. Nevertheless, as sequencing costs fall, WGS provides unique opportunities to integrate diverse markers of genomic instability and mutagenesis within a single protocol. Moreover, we demonstrated that NNLS mutation signature analysis enables accurate N of 1 HRD signature investigation for genome-driven personalized medicine initiatives.

Quantifying HRD signatures supplements existing knowledge and paradigms of cancer detection and stratification. HRDetect scores were associated not only with *BRCA1* and *BRCA2*, but also potentially with other genes such as *PTEN*. This approach provides a functional indicator for mutations whose impact on gene function is uncertain, potentially expanding the repertoire of known causative variants which comprise hereditary cancer screening (Polak et al., 2017). Additionally, we observed that HRD signatures were more common in, but not exclusive to, triple-negative and basal-like breast cancers. This agrees with previous work (Nik-Zainal et al., 2016) and helps to situate HRD in the context of other widely-used breast cancer markers. A topic for future investigation is the value of screening basal-like and triple negative breast cancers for signatures of HRD.

Breast cancer remains the most common cancer diagnosis in women worldwide. It is evident that a substantial proportion are driven in some part by HRD. Here, we have quantified the relationship between aggregated HRD signatures and measures of sensitivity to platinum-based chemotherapy, providing the basis

for further investigation of this putative predictive biomarker in prospective trials. In doing so, this study demonstrates the potential for mutation signatures to guide clinical therapy in a precision oncology setting.

2.4 METHODS

2.4.1 *Patient Samples, Ethics, and Data Policy*

93 study participants with advanced stage breast cancer underwent tumour biopsies at the BC Cancer Agency and collaborating hospitals as part of the POG project, the first 100 cases of which were described in an earlier publication (Laskin et al., 2015). This study includes data from the first 93 verified breast cancer cases which underwent whole genome characterization and met quality assurance standards.

2.4.2 *Sample Collection, Preparation, and Sequencing*

Biopsy samples were embedded in optimal cutting temperature (OCT) compound and sectioned. Pathology review was completed for each specimen, including assessment of tumour content. Genome libraries from tumor and peripheral blood (normal control) as well as transcriptome libraries from tumour were constructed using Illumina protocols. Whole genome and transcriptome sequencing was performed on an Illumina HiSeq2000 or HiSeq2500 sequencer. The details of library construction and sequencing have been previously described (Bose et al., 2015; Sheffield et al., 2015).

2.4.3 *Bioinformatic Analysis*

Sequencing reads were aligned to the human reference genome (GSCh37) by the BWA aligner (v0.5.7) (Li and Durbin, 2009, 2010). Somatic SNVs and small insertions/deletions were processed using samtools (Li et al., 2009) and Strelka (v0.4.6.2) (Saunders et al., 2012). CNVs were called using CNASeq (v0.0.6) as described in (Jones et al., 2010) and LOH by APOLLOH (v0.1.1) (Ha et al., 2012). The matched normal genome was used to subtract germline variants and to report cancer risk variants in 98 select actionable genes, pre-approved by an ethics committee. Germline variant pathogenicity was estimated according to established ACMG guidelines (Richards et al., 2015) using a local curated variant database and custom-built risk calculator established by the BC Cancer Agency Cancer Genetics Laboratory. Transcriptomes were repositioned using JAGuaR (version 2.0.3) (Butterfield et al., 2014). Differential expression analysis was performed by comparing RPKM expression levels against a compendium of 16 normal tissues from the Illumina BodyMap 2.0 project (available from ArrayExpress, queryID: E-MTAB-513) as described in (Jones et al., 2010). Intrinsic subtypes were determined by performing Spearman rank-order correlations on the expression of genes in the PAM50 gene set (Chia et al., 2012) for each breast cancer subtype between sequenced samples and 823 breast cancers derived from The Cancer Genome Atlas (The Cancer Genome Atlas, 2012). For each sample, the subtype with the greatest correlation coefficient was taken as the intrinsic subtype (Figure 2.2). One tumour sample did not pass quality control for RNA-seq and was excluded from analyses involving intrinsic subtypes.

2.4.4 *Determining HRDetect Scores*

HRDetect scores were computed by aggregating six mutation signatures associated with HRD: (1) SNV signature 3/V9, (2) SNV signature 8/V6, (3) SV signature 3/R1, (4) SV signature 5/R5, (5) the HRD index, and (6) the fraction of deletions with microhomology. All signatures were normalized and log transformed as previously described (Davies et al., 2017), and HRDetect scores were computed using a logistic model with the same intercept and coefficients as those reported in the previously trained model, without any retraining or adjustment (Davies et al., 2017). The intercept was -3.364 and the coefficients were 1.611, 0.091, 1.153, 0.847, 0.667, and 2.398 respectively for the six HRD signatures. The sections that follow detail the computation of the six component signatures. A complete pipeline for computing HRDetect scores is available at github.com/eyzhao/hrdetect-pipeline.

2.4.5 *Single Nucleotide Variant Mutation Signatures*

Somatic SNVs called by Strelka were used for mutation signature calculation. SNVs were categorized based on 6 variant types and 16 trinucleotide context subtypes to yield a total of 96 mutation classes. Mutation signatures were deciphered using a published framework (Alexandrov et al., 2013b), which employs NMF to infer both the operative signatures prevalent across the 93-genome cohort and the relative exposure of each signature to each genome. Exposures are modeled as the number of mutations contributed by a mutation signature. Fractional exposure was defined as the proportion of a genome's total mutation burden contributed by a particular signature. Signature stability estimates were obtained by bootstrap re-sampling with 1 008 iterations (84 iterations over 12 cores). The similarity of

signatures to thirty previously described mutational signatures (available from cancer.sanger.ac.uk/cosmic/signatures) was quantified using the cosine similarity metric. Solutions with a 7 to 10 signature model were found to best maximize signature stability and minimize Frobenius reconstruction error. Among these, a 9-signature model was selected as it yielded one signature with maximal cosine similarity to the previously described HRD-associated Signature 3.

2.4.6 *Structural Variant Mutation Signatures*

Large scale somatic SVs were reconstructed by *de novo* assembly of tumor and normal reads using ABySS and Trans-ABySS (Robertson et al., 2010). Candidate SVs were re-aligned to the reference genome to resolve breakpoints. Additionally, we used DELLY (v0.6.1) to obtain an independent SV set by reference-based analysis of split and paired end reads (Rausch et al., 2012). Germline events were filtered out by subtracting SVs found in the matched normal genome. SVs detected by the two methods were merged to yield a high quality consensus set, containing an intersection of variants called by both methods where matching breakpoint loci were separated by no more than 20 base pairs.

32-parameter SV mutation catalog vectors were computed by binning variants based on breakpoint clustering, SV type, and SV length (Nik-Zainal et al., 2016), yielding a 32 by 93 catalogue matrix. This matrix was decomposed by NMF (like with SNV signatures) using a 6-signature model, which was chosen to maximize signature stability and minimize Frobenius reconstruction error. Pairwise comparisons of newly deciphered mutation signatures to six previously described signatures was performed by cosine similarity metric.

2.4.7 *Calculation of the HRD Index*

For each cancer genome, the HRD index was computed as the arithmetic sum of LOH, TAI, and LST scores. CNV and LOH analysis pipelines yielded coordinates segmenting whole genomes by allele-specific copy number ratios. We created an R package called HRDtools which computes LOH, TAI, and LST scores based on the genome-wide CNV profile (available from github.com/eyzhao/hrdtools). Because the HRD index relies upon large-scale events, HRDtools first filters out small events occurring within contiguous events at least 100 times larger. The three scores are then determined based on published guidelines (Timms et al., 2014)

2.4.8 *Analysis of Deletion Microhomology*

Somatic deletions were detected based on sequence alignment using Strelka. Sequences flanking deletion breakpoints were obtained. The microhomology fraction was determined as the proportion of deletions which were larger than three base pairs and demonstrated overlapping microhomology at the breakpoints.

2.4.9 *Review of Clinical Case Data*

Retrospective chart review was performed to obtain treatment history and clinical response to chemotherapy regimens. We queried a province-wide registry of oncology therapeutic records (Wu et al., 2013) to obtain dates of (1) birth, (2) death if applicable, (3) most recent cancer diagnosis, and (4) start and end dates of all platinum-based chemotherapy regimens administered to treat the most recent

cancer diagnosis along with therapies used in combination. Treatment timelines and clinical response are presented in Figure 2.8. All patients were treated as part of standard cancer care either prior to, during, or after the sequencing biopsy. Platinum-treated patients were given standard doses of cisplatin (30 mg/m^2 on days 1 and 8 of a 21 day cycle) or carboplatin (calculated in milligrams as glomerular filtration rate + 25, multiplied by 6 for monotherapy or 5 in combination regimens).

To assess therapeutic benefit, three outcomes were chosen: OS, TDT, and clinical response based on imaging. OS was assessed in patients treated after sequencing (n = 19) and was computed as the duration from first post-biopsy dose of platinum-based chemotherapy to death. TDT was examined as a surrogate for therapy effectiveness. To improve relevance to the present diagnosis, TDT included only treatment regimens occurring within 2 years of sequencing biopsy (n = 33; Figure 2.8).

Clinical imaging reports were reviewed to evaluate platinum response including fludeoxyglucose positron emission tomography and computed tomography obtained during or within two months after the period of platinum-based therapy, compared to pre-treatment scans. Treatment response was classified as follows: (1) CI, any tumor shrinkage of one or more lesions with no evidence of growth or new lesions; (2) SD, either no change in lesions or decreased size of some lesions with growth of others; or (3) PD, disease progression with no associated tumor shrinkage. The best observed response per regimen was recorded.

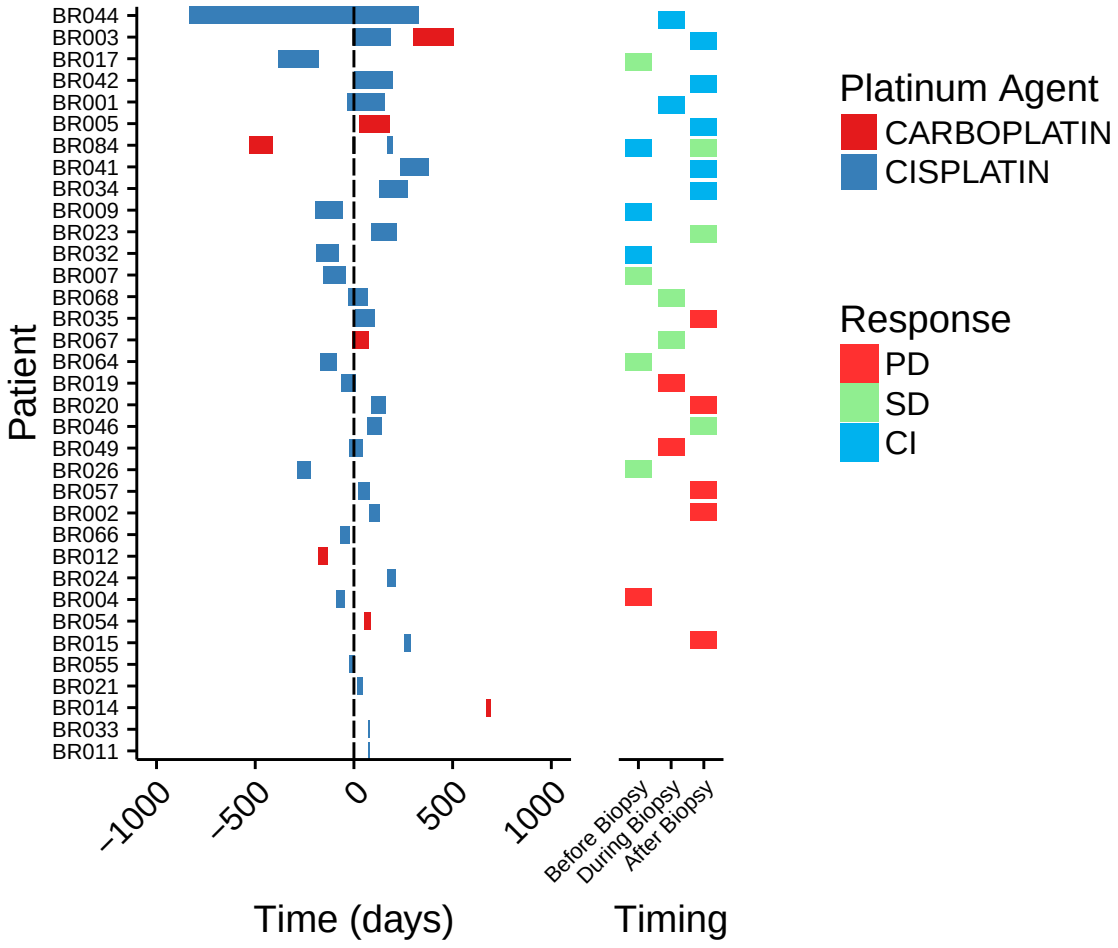

Figure 2.8: Treatment timelines and radiographic outcomes on platinum-based chemotherapy arranged by total duration on platinum-based chemotherapy. The time axis is aligned to the biopsy date, which is centred at time zero. Treatment timelines were obtained from the Outcomes and Surveillance Integration System (OaSIS) of the BC Cancer Agency, which aggregates cancer therapy data across provincial registries. Radiographic outcomes were obtained from a retrospective review of radiologist reports specific to periods of platinum-based treatment.

3

SIGNIT: INFERRING MUTATION SIGNATURES AND THEIR TEMPORAL EVOLUTION IN INDIVIDUAL TUMOURS

3.1 INTRODUCTION

Mutagenic processes in cancer leave characteristic patterns of somatic SNVs (Alexandrov et al., 2013a) and SVs (Nik-Zainal et al., 2016). These mutation signatures reveal exposures such as tobacco smoke (Alexandrov et al., 2016) and ultraviolet radiation, as well as DNA repair deficiencies (Polak et al., 2017). They have also been shown to correlate with the etiology, biology, and pathology of tumours (Schulze et al., 2015; Wang et al., 2017). Recent studies have also revealed therapeutic implications of mutation signatures, suggesting opportunities for predictive biomarker clinical trials (Alexandrov et al., 2015b; Le et al., 2015; Rizvi et al., 2015; Zhao et al., 2017). Increasingly, high throughput sequencing is being investigated for its potential to guide cancer precision therapy (Kumar-Sinha and Chinnaiyan, 2018; Zehir et al., 2017). The use of mutation signatures as biomarkers for personalized medicine will require accurate, robust, and interpretable mutation signature analysis in individual tumours.

Most mutation signature methods focus on detecting signatures *de novo*, which requires large cancer genome cohorts (Alexandrov et al., 2013b; Baez-Ortega and Gori, 2017; Fischer et al., 2013; Shiraishi et al., 2015). Currently, two methods exist for n-of-1 mutation signature decomposition by fitting to consensus refer-

ence signatures: deconstructSigs (Rosenthal et al., 2016) and SignatureEstimation (Huang et al., 2017). deconstructSigs performs signature selection and point estimation of signature exposures, while SignatureEstimation additionally reports credible intervals. However, a significant challenge when fitting samples to reference signatures is multicollinearity: correlated features between signatures can cause bleed of signal between them. This can cause overfitting, resulting in over- or underestimation of clinically relevant mutation signatures.

The temporal evolution of mutation signatures is also crucial to their biological and clinical interpretation. Temporal dissection of cancer mutation sets has shown that exogenous and aging-related mutagenic processes act earlier than endogenous mutagens and DNA repair deficiencies (Bruin et al., 2014; McGranahan et al., 2015; Rosenthal et al., 2016). Temporal shifts in mutagenesis may also reveal shifts in therapeutic targets.

One approach for tracking changing mutation signatures is serial sequencing. However, this strategy is costly and impractical in the clinical setting as it requires rebiopsy. An alternative strategy is to use digital NGS read counts to infer the cellular prevalence (also known as cancer cell fraction) and number of chromosomal copies carrying each mutation. Both are directly related to mutation timing: somatic mutations present on multiple copies likely occurred *before* duplication, and mutations with high cellular prevalence likely occurred *before* subclone development (Figure 1.1). Previous approaches have partitioned mutations *a priori* into "early" and "late" categories (Bruin et al., 2014; McGranahan et al., 2015). Henceforth, we will refer to this approach as binary temporal partitioning (BTP). This method is limited because it makes hard assumptions about the underlying tumour clonal architecture, uses arbitrary thresholds to define "clonal" mutations,

and treats each variant independently rather than inferring shared parameters from the complete data.

In recent years, Bayesian probabilistic models have generated substantial advances in the inference of heterogeneous tumour subpopulations from both somatic SNVs and read depth data (Fischer et al., 2014; Ha et al., 2014; Miller et al., 2014; Roth et al., 2014). These methods define a hierarchical data-generating probabilistic model, then infer model parameters (such as population prevalences and signature exposures) to best reflect the data. Bayesian inference is often more robust to noise and provides full posterior distributions over parameter estimates.

Here, we present SignIT, an R package featuring a Bayesian hierarchical model for accurate and robust mutation signature analysis of individual tumours. Full posterior estimates of signature exposures juxtaposed with comprehensive mutation signature bleed mapping can significantly enhance interpretability. SignIT also includes an extended model which enables joint inference of mutation signatures and temporally distinct tumour subpopulations, exposing signature evolution. We assess SignIT's n-of-1 signature accuracy against deconstructSigs and SignatureEstimation using both simulated mutation count vector and somatic mutation data from TCGA. We validated SignIT's temporal analysis using WGS data from 24 serially sequenced primary-metastasis tumour pairs. Lastly, we apply SignIT to the analysis of 543 metastatic whole genomes, the first ever temporal analysis of mutation signatures in metastatic cancer.

3.2.1 *SignIT Reports Credible Intervals and Signature Bleed*

We begin with an example to illustrate SignIT's output. P10 is a patient who underwent whole genome sequencing of her metastatic breast cancer, revealing 10,068 somatic SNVs. Mutation signature analysis by SignIT revealed elevated signatures 3 and 8, both associated with HRD (Davies et al., 2017), as well as slight involvement of signatures 2, 9, and 17 (Figure 3.1A). SignIT reports full posterior probability distributions which reflect the stochasticity of somatic mutation as well as uncertainties due to signature bleed. Moreover, 2D projections of the posterior distribution provide a pairwise map of signature bleed (Figure 3.1B), which is visualized below signature exposures as a non-directed graph. Signature bleed presents as anti-correlation in the posterior samples because the existence of reference signatures with similar profiles will produce mutually exclusive solutions. Signatures 3 and 8, for example, have correlated mutation spectra, with a cosine similarity of 0.76.

SV signatures can also be analyzed. P10 possessed 146 somatic SVs, which were fitted against six SV signatures previously identified in breast cancer (Nik-Zainal et al., 2016). This revealed involvement of rearrangement signatures 2 and 5, with signature bleed between them (Figure 3.1C).

3.2.2 *Resilience to Complexity and Noise*

To assess the accuracy of signature exposures, we created a mutation signature simulation R package called msimR (github.com/eyzhao/msimR). Mutation

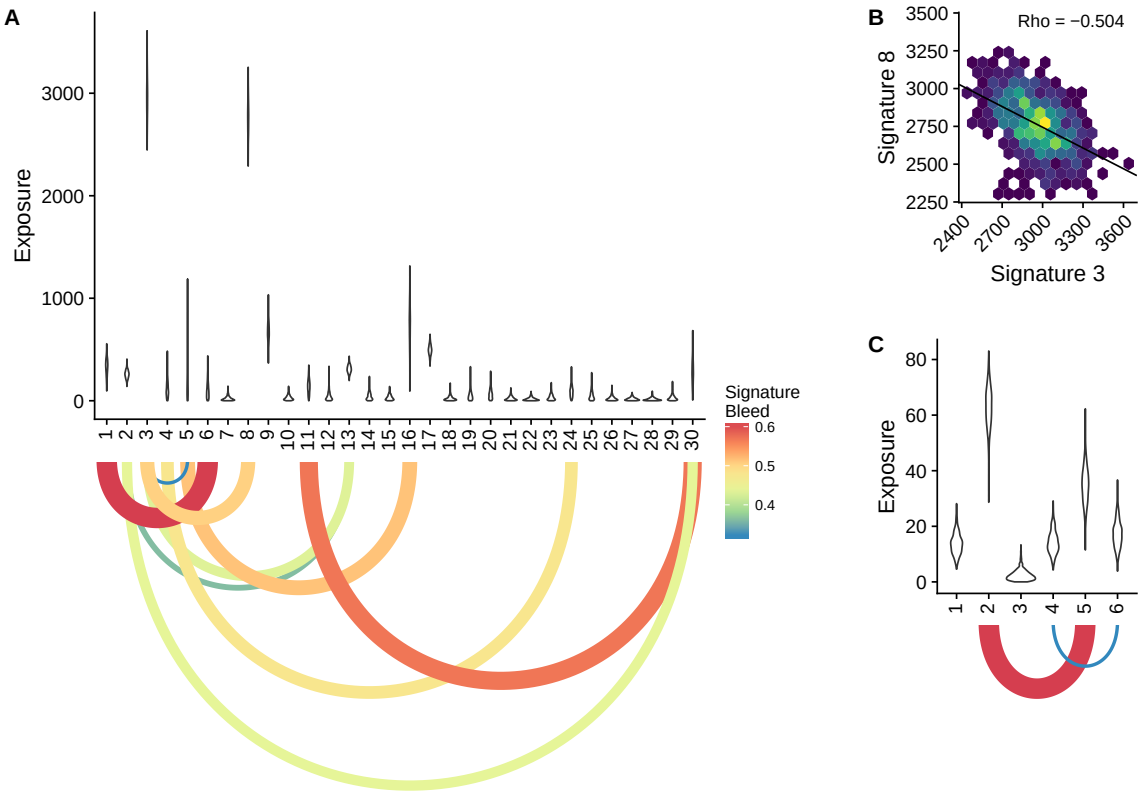

Figure 3.1: SignIT reports complete posterior distributions along with signal bleed between signatures. (A) Complete Bayesian inference over mutation counts determines posterior parameter estimates for each mutation signature exposure. Signature bleed, quantified as the negative Spearman correlation coefficient per 2D slice of the posterior distribution, is quantified pairwise between signatures and plotted as a graph below posteriors. (B) An example of anticorrelated exposure posteriors between Signature 3 and Signature 8 is shown as a 2D density plot. (C) Mutation signature exposures can also be computed for structural variant mutation signatures, using a 6-signature reference set.

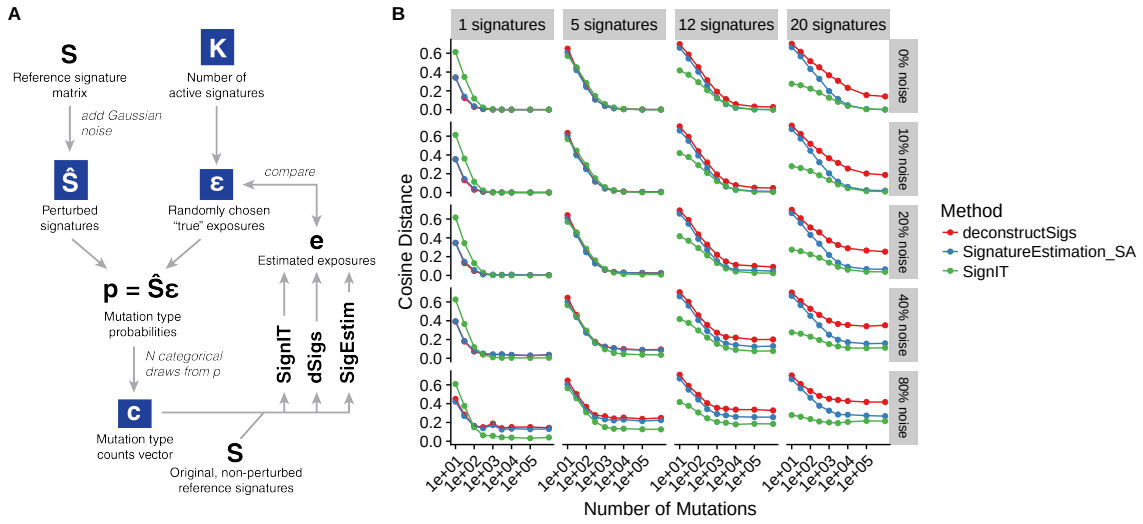

Figure 3.2: SignIT improves signature estimation for complex models with noisy data. (A) Mutation count vectors were simulated by varying the mutation burden, number of active signatures, and amount of noise introduced into the reference signature matrix. Exposures were estimated using three signature decomposition methods and compared against true exposures. (B) 500 count vectors were simulated at each condition and the similarity of estimated to true exposures was computed using the cosine distance (lower values indicate better accuracy). The mean cosine similarity per condition is shown.

count vectors were generated from known simulated signature exposures with varying mutation burden and model complexity (number of active signatures). Additionally, there can be uncertainty in the mutational profile of processes which generated somatic mutation, as no reference set can be expected to capture all the possible biological variability of mutagenesis. To emulate this biological variability, random Gaussian perturbation of reference signatures was introduced (Figure 3.2A). It is expected that performance improves with increasing mutation burden (sample size) and declines with increasing model complexity (the number of signatures, or dimensionality) and reference signature noise.

Simulated genomes with higher mutation burden yielded more accurate signature exposures across all conditions and methods, as demonstrated by decreased

cosine distance (Figure 3.2B), which is defined as 1 − cosinesimilarity. However, accuracy declined in more complex genomes with larger numbers of active signatures. SignIT was either equally accurate or more accurate than other methods in all settings with the exception of low-complexity, low-mutation genomes. SignIT was superior in genomes with many active processes and was also substantially more robust to perturbation of the underlying reference signatures.

Error rates were quantified per signature to identify over- and underestimated signature exposures. Both deconstructSigs and SignatureEstimation frequently underestimated signatures, which may result in the loss of actionable information (Appendix Figure A.1). Particularly difficult to resolve were signatures most similar to other signatures and are therefore most likely to exhibit signature bleed, especially signature 5. Conversely, absent signatures are frequently overestimated by all methods (Appendix Figure A.2). However, where SignIT inflated exposures, it did so with a lower relative error. This robustness against dramatic over- or underestimation of signature exposures is necessary for confident clinical interpretation.

In most settings, SignIT takes longer to run than other methods, but scales to realistic mutation burdens with practical runtimes (Figure 3.3). Using default settings (8 chains in parallel with 200 burn-in iterations and 200 sampling iterations each), SignIT ran in 20 seconds on tumours with 100 mutations and 154 seconds on tumours with 1,000,000 mutations.

3.2.3 *SignIT Better Reproduces Signatures in Cancer Data*

To evaluate SignIT on real cancer genome mutation data, we analyzed whole exomes from nine cohorts of TCGA. Mutation signatures in each cohort were

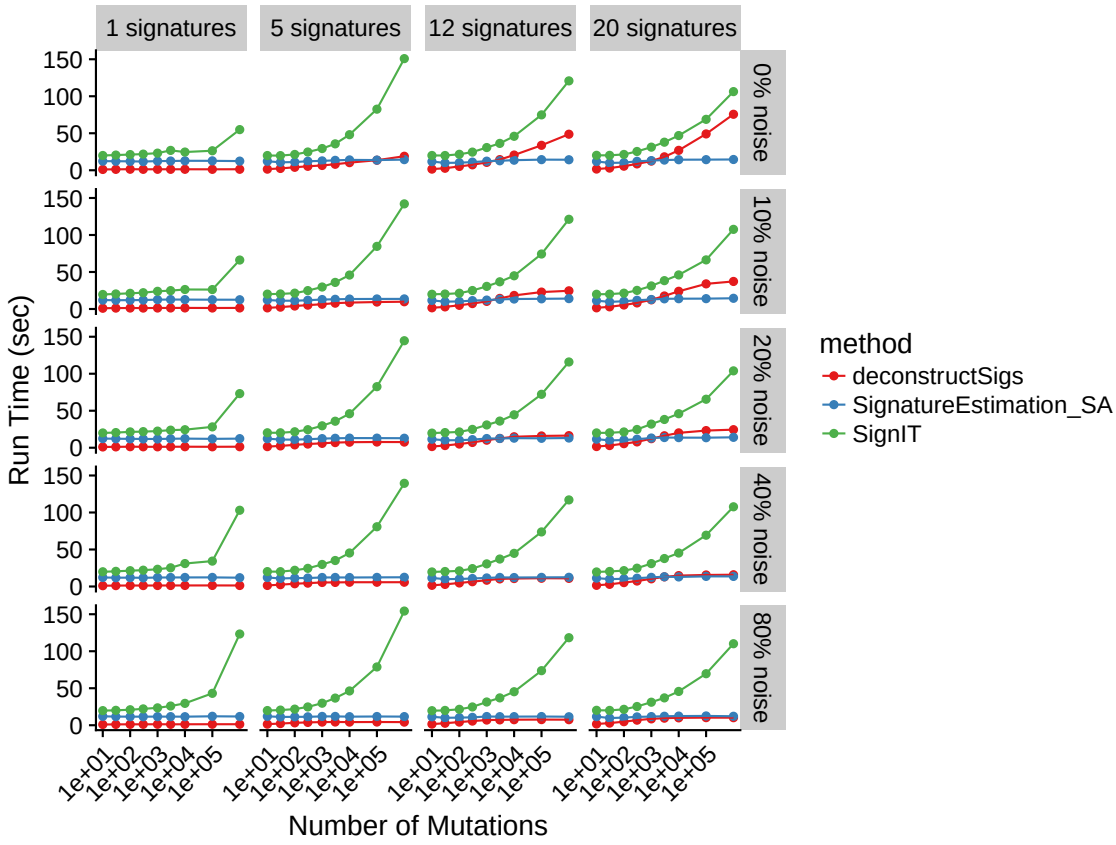

Figure 3.3: Runtimes of n-of-1 mutation signature decomposition tools. Simulated mutation catalogs were generated under various conditions and their exposures were re-estimated using deconstructSigs, SignatureEstimation, and SignIT. The number of mutations varied from 10 to 1,000,000, the number of active signatures varied from 1 to 20, and random perturbation of reference signatures varied from 0 to 80 percent. Runtimes were captured across 500 trials under each set of conditions. Mean runtimes are presented here in seconds.

deciphered by NMF (Alexandrov et al., 2013b), as well as by SignIT, deconstructSigs, and SignatureEstimation. NMF signatures were compared against the full COSMIC 30-signature reference set to determine the best match for each *de novo* signature. To best emulate a clinical sequencing scenario, n-of-1 signature analysis was rendered entirely blind to NMF results. Exposures were computed against the entire COSMIC 30-signature reference set. For each COSMIC signature matched by NMF, exposures were compared to those of each n-of-1 method by Spearman correlation, which was chosen because of its robustness to outlier (hypermutating) signatures.

SignIT exposures demonstrated greater concordance with *de novo* NMF methods across all signatures and cohorts than deconstructSigs or SignatureEstimation (Figure 3.4A). While the methods were comparable for hypermutating signatures such as Signatures 2, 4, 7, and 13, SignIT substantially improved concordance with NMF for lower-exposure signatures (Figure 3.4B).

3.2.4 *SignIT Infers the Temporal Evolution of Signatures*

Returning to patient P10, we next undertake the temporal dissection of signatures across tumour subpopulations (Figure 3.5A). SignIT identified two mutational subpopulations with prevalences of 1.0 and 0.31 accounting for 80.5% and 19.5% of total mutation burden respectively. The two populations display different mutational profiles, with the more prevalent (earlier) population being enriched for Signature 3 and the less prevalent (later) population for Signatures 16, 17, and 30. Dividing mutations into early and late sets using BTP (McGranahan et al., 2015) agreed closely with results from SignIT (Figure 3.5B).

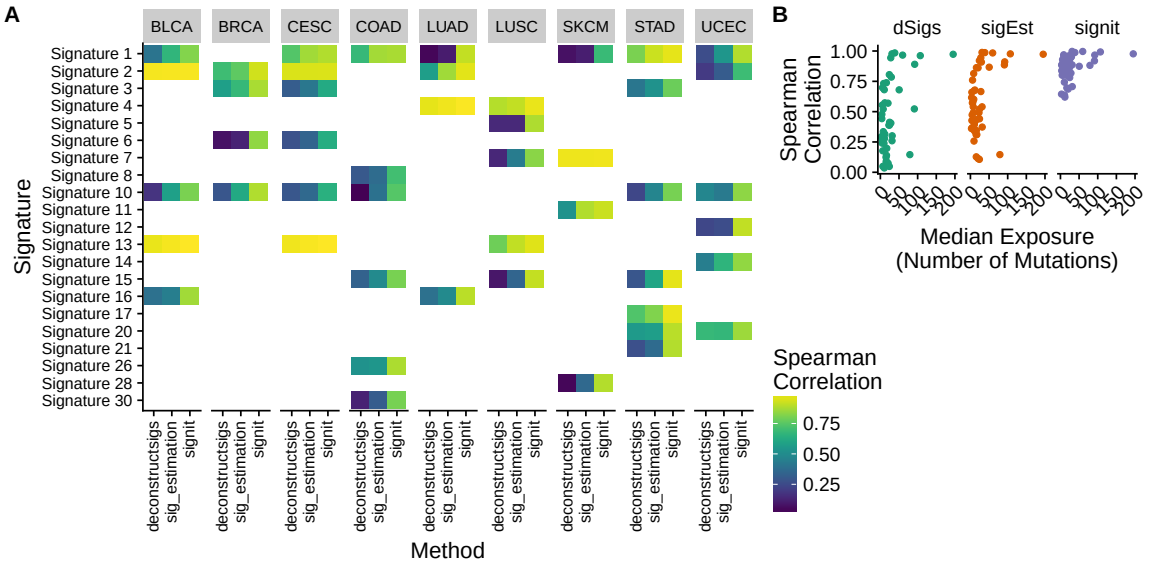

Figure 3.4: Comparison of NMF and n-of-1 methods across nine cancer exome cohorts. (A) Mutation signatures were deciphered *de novo* in 9 cohorts from The Cancer Genome Atlas (TCGA) and matched to the most similar corresponding reference signature. N of 1 mutation signature exposures were estimated using three methods and compared against *de novo* signature exposures using the Spearman correlation coefficient. (B) SignIT demonstrated improved accuracy, providing significant improvement in resolving signal from less mutagenic signatures.

Early-arising, population 1 signatures were concordant with both primary (cosine similarity = 0.988; Figure 3.5C) and metastatic (cosine similarity = 0.976; Figure 3.1A) exposure profiles, both of which showed elevated signatures 3 and 8. Population 2 therefore may provide insights into mutational processes later in metastasis.

3.2.5 *Metastatic Tumours Demonstrate Divergence of Mutational Processes*

To assess the temporal dissection of mutation signatures, we performed WGS of 24 metastatic tumours with paired sequencing of primaries. 20 of those primaries were sequenced from formalin-fixed and paraffin embedded (FFPE) tissue, and 4 from OCT tissue. The intersection of primary and metastatic SNVs was used to derive signatures of the primary tumour. The rationale for using the intersection is to focus on mutations present in the primary which persisted in the metastasis and to filter out false positives introduced by FFPE. Subpopulation-specific signature exposures computed by SignIT were compared to primary tumour signatures. The divergence away from the primary was quantified for the signatures of each subpopulation in the metastatic sample using the cosine distance.

As expected, all cases demonstrated mutation signature divergence from the primary tumour in the least prevalent (latest) subpopulation (Figure 3.6). In most cases, prevalent early subpopulations were similar to the primary tumour (cosine distance < 0.2), even when signatures from the bulk metastatic tumour differed from the primary. This suggests that signature timing can reveal the early mutational processes of tumorigenesis using a later metastatic sample.

Early mutations derived from the BTP method similarly matched primary tumour signatures except in three cases (P07, P08, and P15). All three were charac-

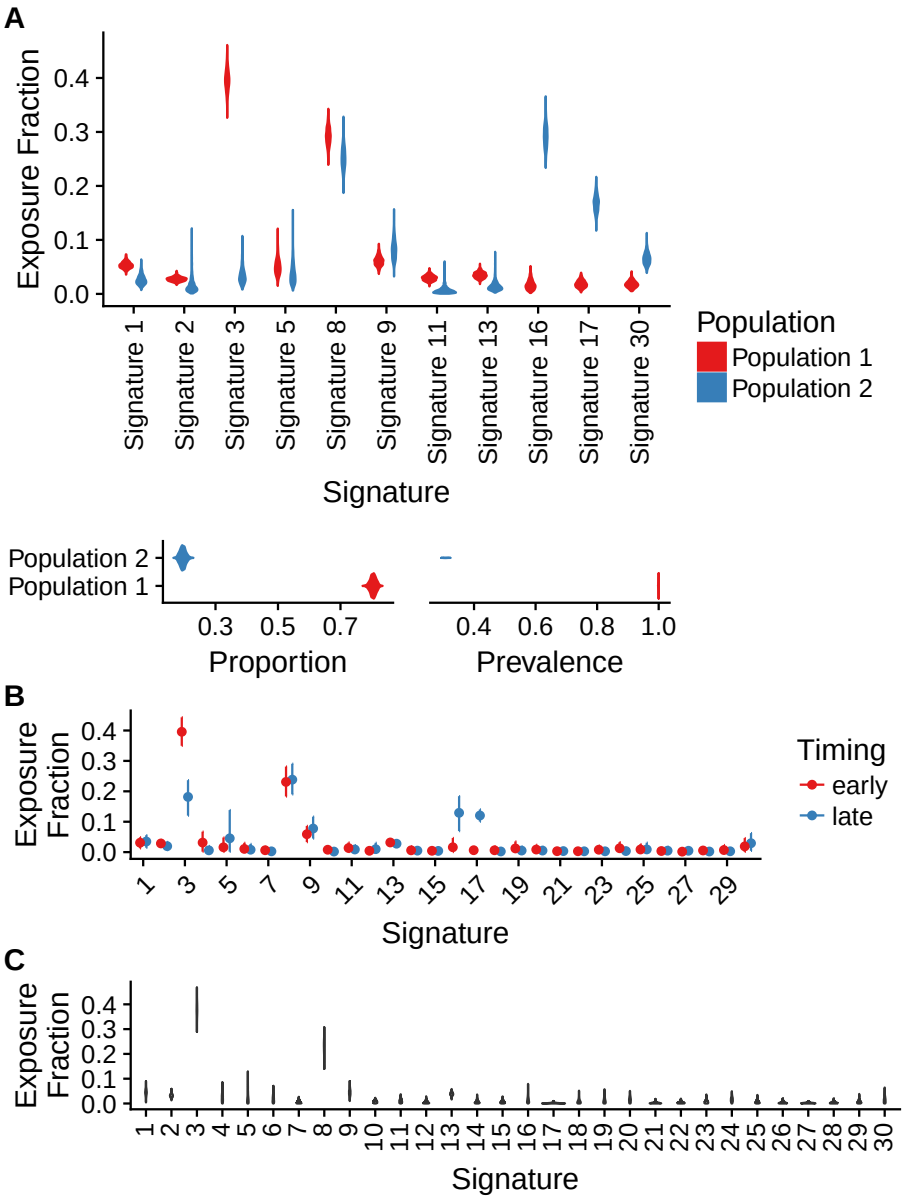

Figure 3.5: Subpopulation-specific mutation signatures in a somatic cancer whole genome. (A) SignIT was used to infer subpopulation-specific mutation signatures in a breast cancer. This revealed two temporally distinct subpopulations with mutational prevalences of 1.0 and 0.31 giving rise to 80.5% and 19.5% of mutations respectively. (B) Temporally dissected mutation signatures were similar to those deciphered by binary temporal partitioning. (C) Signatures deciphered from mutations shared with the sequenced primary tumour also agreed with early subpopulation mutation signatures.

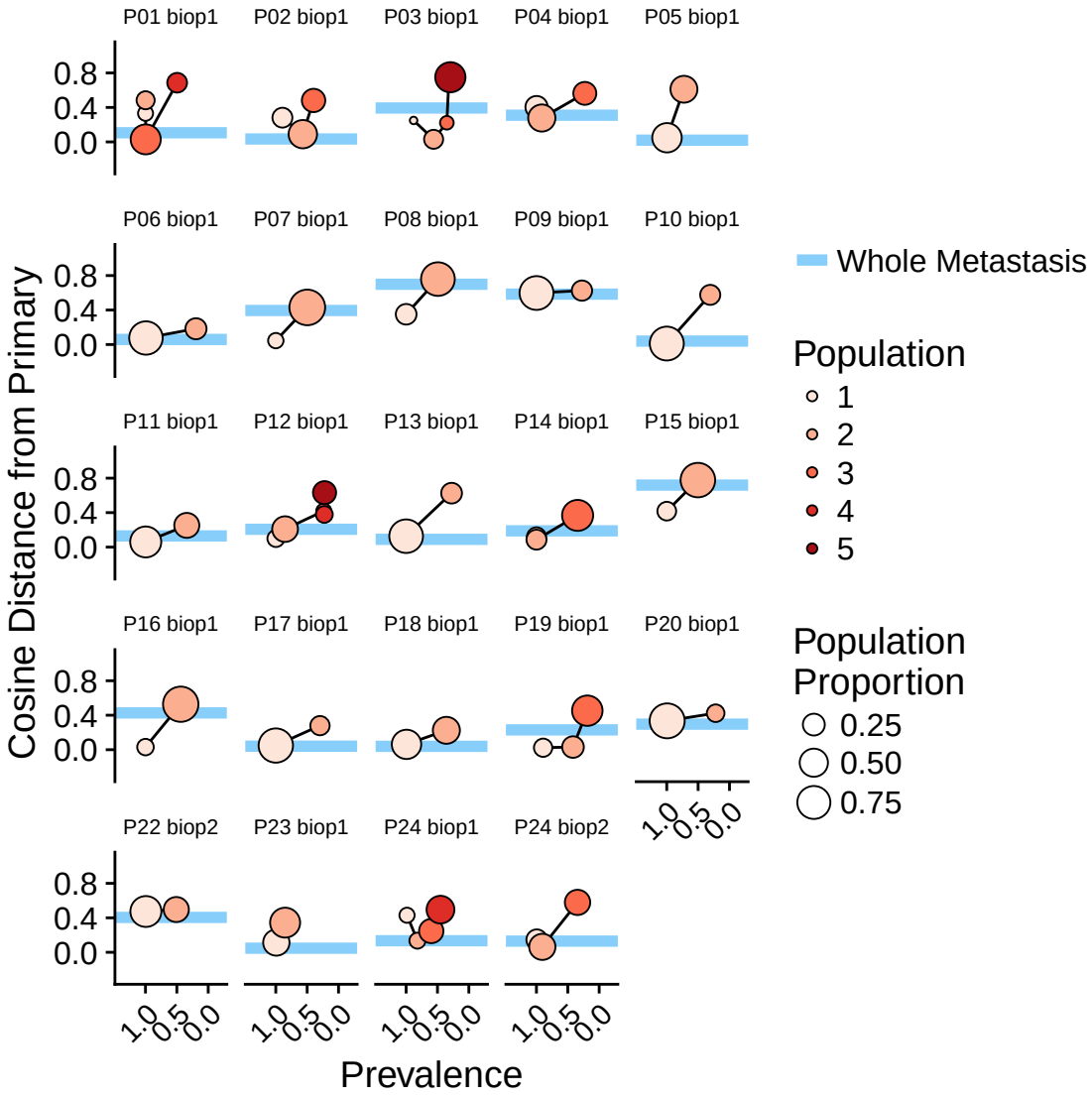

Figure 3.6: Mutation signatures in serially sequenced metastatic tumours demonstrate time-dependent divergence from the primary. To validate SignIT and explore signature evolution in metastatic tumours, SignIT was used to decipher population-specific signatures in metastatic tumours with whole genome sequencing of paired primaries. Cosine distance was used to determine the similarity of mutation signature exposures in each subpopulation to those of mutations shared with the primary tumour. More prevalent populations typically demonstrated similarity with the primary, even when bulk metastasis signatures differed greatly. Signatures in lower-prevalence populations diverged over time.

terized by highly mutagenic late (lower-prevalence) subpopulations (Figure 3.6). Examining P07 as an example (Figure 3.7), SignIT early (population 1) mutation signatures were a closer prediction of primary tumour signatures than the early mutations from BTP, which overestimated APOBEC-related signatures 2 and 13. Whereas SignIT found that 91% of mutations originated from the lower-prevalence, late subpopulation, BTP identified 80% of mutations as late-arising. This suggests that BTP may have suffered from contamination of the early mutation pool with late-arising APOBEC-associated mutations.

SignIT also improves temporal dissection when tumours harbour subpopulations with prevalence values near 1.0. For example, P05 had a subpopulation with a relatively high prevalence of 0.73. This population demonstrated a dramatic drop in signature 1 and rise in signature 8, which was not resolved as clearly by BTP (Figure 3.8). SignIT mitigates these limitations of BTP by fitting the cancer's subpopulation structure.

3.3 DISCUSSION

Mutation signatures and genomic instability are an emerging part of the ever-growing scientific literature focussed on clinically actionable cancer biomarkers. SignIT constitutes a substantial advance in mutation signature analysis and interpretation in individual tumours. Along with providing novel insights into mutation signature bleed and tumour subpopulation structure, our findings demonstrate SignIT's accuracy and its robustness against model complexity and noise. SignIT's inference of subpopulation-specific signatures improves upon previous approaches because it directly models the underlying clonal structure. Analysis of tumours sequenced at multiple time points revealed frequent divergence of

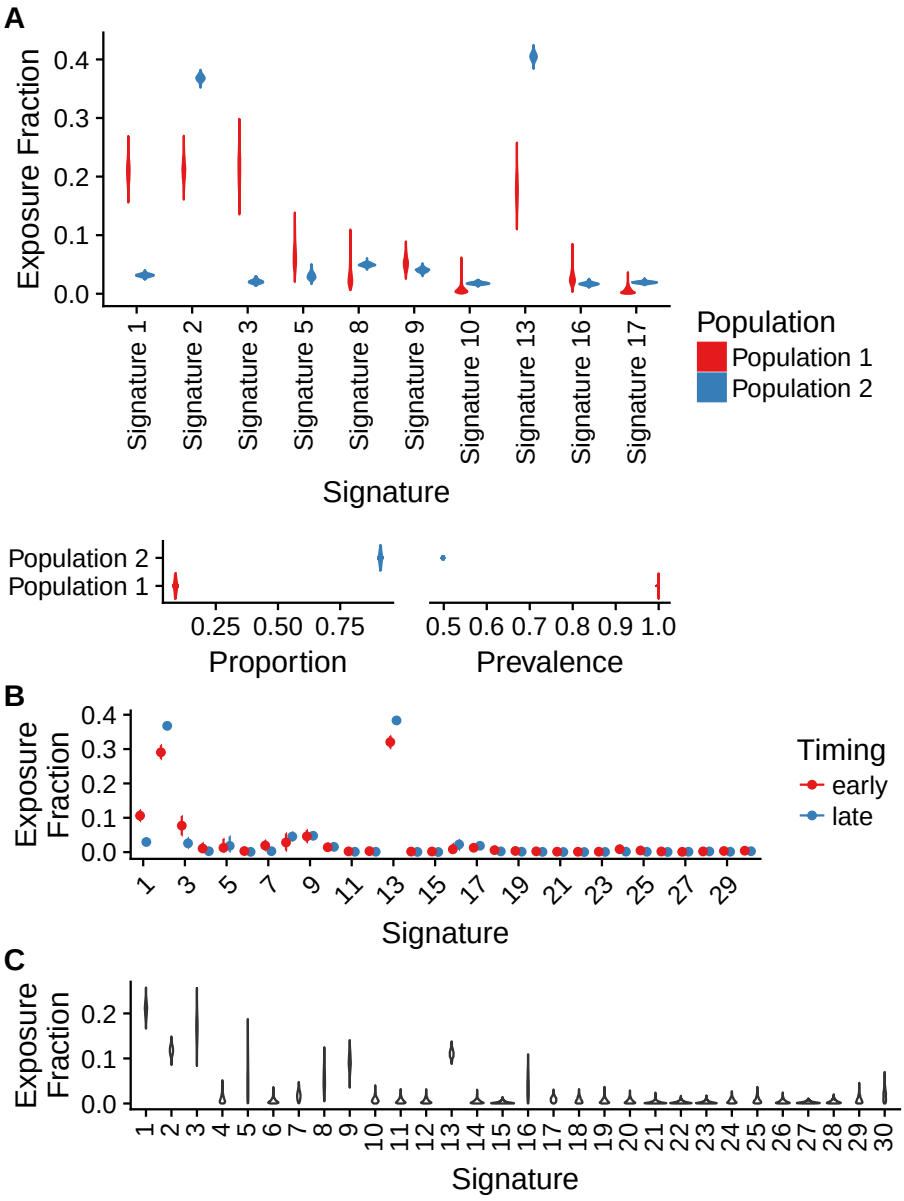

Figure 3.7: SignIT improves upon binary temporal partitioning (BTP) by modeling the tumour subpopulation structure. When the tumour subpopulation structure disagrees with the strict assumptions of BTP, temporally dissected mutation signatures can be inaccurate. (A) In this temporal analysis of a lung adenocarcinoma, the lower prevalence population 2 was highly mutagenic. (B) This feature may have resulted in contamination of the smaller "early" mutation pool, resulting in poor temporal separation. (C) SignIT's early signatures better match those of the archival sample.

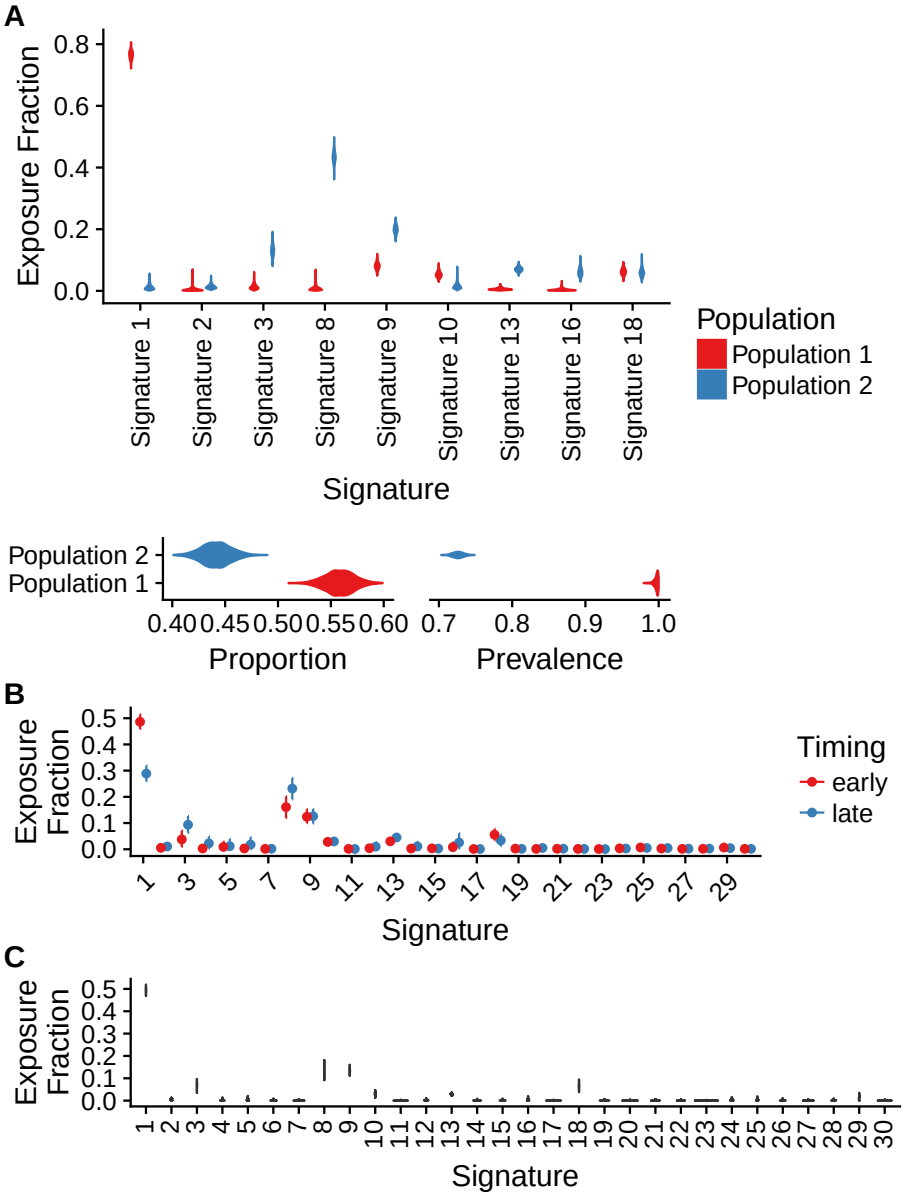

Figure 3.8: A colorectal cancer demonstrates errors in binary partitioning resulting from unusually high mutational prevalence of population 2. (B) This resulted in erroneous estimates of signatures 1 and 8. (A) By contrast, SignIT jointly models the population structure and signatures, enabling clearer delineation of signatures between subpopulations. (C) SignIT's results more accurately reproduce mutation signatures from the matched archival sample.

mutation signatures, highlighting the need to track the evolution of mutagenic processes in metastasis.

SignIT offers the ability to resolve the mutational history of individual tumours at greater resolution than previously possible. The temporal dissection of mutation signatures can inform many questions of biological interest. For example, understanding the earliest mutational processes in cancer may help inform tumour prevention and early detection. Temporally resolving mutation signatures could also improve the understanding of mechanisms underlying known and novel signatures. Lastly, signature timing can isolate mutagenic processes characteristic of key disease phases such as metastasis.

Successful mutation signature analysis necessitates some technical trade-offs. For instance, SignIT infers mutational prevalence as a surrogate for mutation timing without deconvolving the influences of variant copy number and cellular prevalence. It is technically challenging to determine both these factors independently without targeted deep sequencing, which is frequently used to estimate clonal composition (Roth et al., 2014). However, the analysis of mutation signatures requires the broad capture of large numbers of variants, which is most economical by lower-depth WGS or WES. A potential compromise, which SignIT supports, involves using targeted deep sequencing to predetermine fixed prevalence parameters in SignIT inference.

Fulfilling the promise of cancer precision medicine will require rapid integration of orthogonal genomic biomarkers into research and clinical practice. SignIT provides a novel, easy to use, and methodologically rigorous approach to mutation signature analysis to improve interpretation of n-of-1 signature analysis. It also enables temporal dissection of mutation signatures, which is applicable to emerging biological and clinical questions in cancer genomics.

3.4 METHODS

SignIT is available from github.com/eyzhao/SignIT. Version v1.0.1 was used for all analyses described in this thesis.

3.4.1 *The SignIT Generative Model*

Bayesian inference involves the definition of a generative model, then learning the parameters of that model which provide the best fit to data. Upon convergence, Hamiltonial Monte Carlo (HMC) enables probabilistically proportionate sampling from the complete posterior distribution over the parameters. Here, we describe the generative model used in SignIT to perform Bayesian inference over signature exposures.

Mathematics of De Novo *Mutation Signature Analysis*

The early work on mutation signatures aimed to identify recurrent patterns of somatic mutation which could explain the mutational processes frequently observed in cancer. These approaches utilize NMF or similar dimensionality reduction methods to reduce a mutation count vector denoting the frequencies of a set of mutation types into a smaller set of signatures. We refer to these approaches, in aggregate, as *de novo* mutation signature analysis, because they identify novel mutation signatures using unsupervised learning methods.

First, let V be a set of mutation classes parametrized for N mutations arising from K mutation signatures in G genomes. Let $\mathbf{M}$ be a $V \times G$ mutation counts matrix, $\mathbf{S}$ a $V \times K$ mutation signature matrix, and $\mathbf{E}$ a $K \times G$ exposures matrix. In

the mutation signature model, $\mathbf{M} = \mathbf{SE}$. Determining an optimal solution for $\mathbf{S}$ and $\mathbf{E}$ given a mutation count matrix $\mathbf{M}$ can be performed by NMF.

N of 1 Mutation Signature Analysis

In the n-of-1 case, let $G = 1$, yielding $\mathbf{c} = \mathbf{Se}$, where $\mathbf{c} \in \mathbb{N}^V$ is a V-dimensional non-negative integer mutation count vector and $\mathbf{e} \in \mathbb{R}^{+K}$ is a K-dimensional non-negative exposures vector. In the common parametrization of SNVs based on base change and $3'/5'$ context, $V = 96$. Let $\mathbf{S}$; $\left(0 < S_{ij} < 1 \forall i = 1 \ldots V, j = 1 \ldots K\right)$ be a $V \times K$ matrix of known signatures, where K is the number of known reference signatures.

At the time of writing, the most commonly used reference signature set for SNVs is a 30-signature matrix available from COSMIC at cancer.sanger.ac.uk/cosmic/signatures. However, any set of reference mutation signatures may be used. The mutational spectrum of a tumour is modeled as a linear combination of contributing signatures, with coefficients $\mathbf{e}; e_1, e_2, \ldots, e_K > 0$.

Given known values for $\mathbf{m}$ and $\mathbf{S}$, the goal is to determine the best fit value of $\mathbf{e}$. The most common cost function is the sum of squared errors (SSE), such that $\mathbf{e}$ is chosen to minimize $|\mathbf{Se} - \mathbf{m}|^2$. Given that $\mathbf{e}$ must be non-negative in all dimensions (there cannot be a negative number of mutations), this problem is known as non-negative least squares. NNLS is well-studied and readily implemented using quadratic programming (QP), which rapidly converges to an optimal solution.

Limitations of NNLS

There are three sources of error in the QP solution to NNLS which limit its accuracy and utility to clinical cancer sequencing. The first is sampling error in mutagenesis. The accrual of mutations can be viewed as a random process where,

for the i^{th} mutation, the responsible mutational process z_i is drawn from the categorical distribution, $z_i \sim \text{Categorical}(\mathbf{e})$, and the mutational class x_i is subsequently drawn from $x_i \sim \text{Categorical}(\mathbf{S}_{:,z_i})$. This results in a categorical noise profile, which QP is likely to overfit to, especially for small mutation counts. On the other hand, SignIT models the data-generating process underlying mutation counts (known as a categorical mixture model) and can therefore account for noise rather than overfitting to it. This allows SignIT to yield estimates of solution uncertainty, rather than point estimates.

The second source of error arises due to multicollinearity of the reference signature matrix. Signatures in the reference set often exhibit correlation with each other, due to similarities in their mutation profiles. This can result in spurious mutation signature elevation when similar signatures are present, a phenomenon known as "signature bleed". Multicollinearity poses an inherent mathematical limitation on the ability to call mutation signatures with certainty. SignIT addresses signature bleed by mapping mutual exclusivities in signature activation. In MCMC, posterior sampling density eventually converges on the true posterior distribution; any two signatures which bleed with one another will have anticorrelated MCMC samples. This reflects the fact that a fixed portion of mutations can be explained by various linear combinations of the two similar signatures.

The last source of error arises from the reference signature matrix, $\mathbf{S}$. The "true" signatures giving rise to a cancer genome may differ slightly from those catalogued in $\mathbf{S}$, resulting in the mis-estimation of exposures. This signature bias is more difficult to account for without prior knowledge of uncertainties in the reference signature matrix. However, N of 1 signature decomposition methods can be tested for robustness against reference signature bias by observing the accu-

racy and stability of results when the reference signatures are perturbed with a pre-determined noise profile.

Bayesian N of 1 Signature Analysis

SignIT models the acquisition of mutations as a categorical mixture of K mutational processes, where K is the number of reference mutation signatures. The reference mutation matrix, $\mathbf{S}$, has dimensions $V \times K$, where V is the number of mutation classes (most commonly $V = 96$) and variant classes are defined by the base change and $3'/5'$ mutation context (Alexandrov et al., 2013b). Every column of $\mathbf{S}$ is a probabilistic simplex denoting the probability distribution of mutation classes associated with a single reference signature. Let $\mathbf{e}$ be a K-dimensional simplex denoting signature proportions, also known as exposures. A signature's exposure is the probability of a random mutation occurring as a result of that signature. Note the analogy to a topic model, where signatures represent topics, mutation classes constitute the vocabulary, and mutations serve as words. To generate a dataset of N mutations, first select each mutation's signature, u_i, by drawing from a categorical distribution $u_i \sim \text{Categorical}(\mathbf{e})$. Next, determine the mutation class, $v_i \sim \text{Categorical}(\mathbf{S}_{:,i})$. Repeating this process for $i = 1, 2 \ldots, N$ yields a set of N mutations, each belonging to a specific class. SignIT vectorizes this by encoding mutation count vectors rather than individual mutations, which provides significantly better performance.

Vectorization

For efficiency, SignIT's implementation vectorizes the categorical mixture model presented in the main text, yielding an equivalent but simpler generative model. We begin by recognizing that the product $\mathbf{Se}$ yields an V-dimensional probability

simplex denoting the probability of observing each mutation class. We next define the V-dimensional mutation count vector $\mathbf{c}$, where c_i indicates the number of mutations belonging to the ith mutation class. In this scheme, the previously presented categorical model can be equivalently presented as $\mathbf{c} \sim \text{Multinomial}(n = N, p = \mathbf{Se})$. This vectorized parametrization yields significantly faster gradient calculation, sampling, and likelihood calculation.

The Temporal Subpopulation Model

SignIT includes a mathematically consistent extension of the aforementioned mixture model to jointly infer signature exposures and temporally separated tumour subpopulations. The complete hierarchical model (Figure 3.9) generates both the VAF and mutation class of each variant using coinciding beta-binomial and categorical finite mixtures respectively. The categorical mixture component is identical to the previously described model, except that exposures are replaced by mixing components with $K \cdot L$ elements, where L is the number of subpopulations being modelled. Subpopulations are distinguished based on prevalence, $\mu' = Fm$, defined as the product of clonal cellular prevalence, F, and the number of allelic copies carrying the mutation m. Higher prevalence subpopulations are associated with earlier-arising mutations.

We assume the presence of L latent temporally distinct subpopulations which give rise to varying mutant allele counts. To account for overdispersion, which is commonly observed amongst NGS reads, we model variant allele counts using a beta-binomial finite mixture model. Note that the beta-binomial distribution is parameterized with mean (μ) and concentration (κ), which relate to the shape parameters α and β by the relations $\mu = \dfrac{\alpha}{\alpha + \beta}$ and $\kappa = \alpha + \beta$.

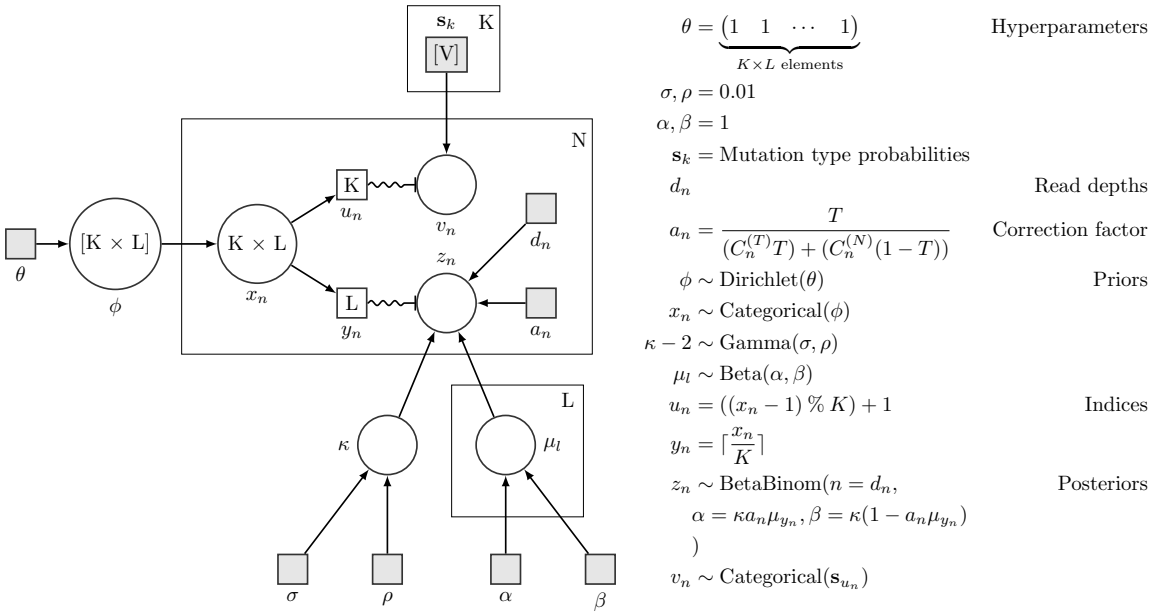

Figure 3.9: Complete SignIT joint population-signature model. For each mutation, mutation type (v_n) and variant read depth (z_n) are jointly drawn based on the selected signature (u_n) and population (y_n) respectively. These responsibility terms are deterministically mapped from x_n, which is chosen from the mixing probabilities ϕ. The signature probabilities (s_k) are user-determined. The beta-binomial population prevalences (μ_l) and shared concentration term (κ) determine read depths. The correction factor (a_n) is computed from total read depth (d_n), tumour copy number ($C^{(T)}$), normal copy number ($C^{(N)}$), and tumour content (T).

The SignIT population model assumes that the probability of sampling a variant allele at a mutated locus is

$$\frac{\mu T}{(C_n^{(T)} T) + (C_n^{(N)}(1-T))},$$

where $C^{(T)}$ is the positive integer tumour copy number, $C^{(N)}$ is the normal copy number, and T is the tumour content ($0 < T < 1$). We define μ as a joint "mutational prevalence," which is the arithmetic product of cellular prevalence (the fraction of cancer cells possessing a mutation at the locus) and mutation copy number (the number of copies per cell which carry the mutant allele).

Let ψ be a probability simplex with L elements. For the nth mutation, draw the subpopulation index $z_n \sim \text{Categorical}(\boldsymbol{\phi})$. This index selects the subpopulation prevalence μ_{y_n}. Letting κ be the concentration parameter which determines the degree of beta-binomial overdispersion, we draw the variant allele depth

$$z_n \sim \text{BetaBinom}(n = d_n, \alpha = \kappa a_n \mu_{y_n}, \beta = \kappa(1 - a_n \mu_{y_n})),$$

where $a_n = \dfrac{T}{(C_n^{(T)} T) + (C_n^{(N)}(1-T))}$.

This model makes a few assumptions. First, the "infinite sites" assumption that no locus undergoes multiple independent somatic mutations. Second, tumour copy number states are assumed to be clonal across the genome. Third, genomic read counts are overdispersed for all sites with a common concentration coefficient, κ.

The Complete Subpopulation Signature Model

The complete SignIT model unites the signature and population models 3.9. To facilitate this, each population-signature combination requires its own model coef-

ficient. To replace ψ from the population model, let ϕ be a simplex of length $K \times L$. The mixing index drawn from ϕ is $x_n \sim \text{Categorical}(\phi)$, and simultaneously encodes a population index y_n and a signature index u_n. The u^{th} signature and y^{th} population correspond to position $x = K(y-1) + u$ in ϕ, where K is the number of populations. The deterministic inverse mappings are $u_n = ((x_n - 1)\,\%\,K) + 1$ and $y_n = \lceil \frac{x_n}{K} \rceil$, where % is the modulo (remainder) operator and $\lceil ... \rceil$ is the ceiling (upwards rounding) operator.

Upon assignment of u_n and y_n, the remainder of the generative model proceeds as previously described. u_n selects the mutation signature, which determines the probability vector, $\mathbf{s}_{u_n}$, across mutation types. The mutation type is drawn from $v_n \sim \text{Categorical}(\mathbf{s}_{u_n})$. y_n selects the population mean μ_l, and the variant allele depth is drawn from a beta-binomial distribution.

This combination of signature and population models defines a joint categorical-beta-binomial mixture model, which allows likelihoods to be computed simultaneously across both mutation types and variant allele counts. In other words, this generative model simultaneously samples variant allele depth and mutation type at each mutated locus. Using Bayesian inference methods, we can then estimate posterior distributions over the parameters $\boldsymbol{\phi}, \boldsymbol{\mu}$, and κ based on the provided genomic data.

Avoiding Degeneracies in the Beta-Binomial Mixture Model

Finite mixture models with degenerate components often give rise to multimodal posterior distributions. This is due to non-identifiability of the mixture components (Betancourt, 2017). To remedy this, we enforced ordering over the subpopulation prevalences $\boldsymbol{\mu}$. Moreover, in order to sample $\boldsymbol{\mu}$ from $\text{Beta}(1,1)$ while impos-

ing natural limitations on its allowable interval, we sampled $\mu_l' \sim \text{Logistic}(0,1)$ and applied the inverse logit transform $\mu_l = \frac{\exp(\mu_l')}{1+\exp(\mu_l')}$.

Implementing Bayesian Inference

The SignIT hierarchical models are encoded using the Stan (2.17.0) probabilistic programming language (Carpenter et al., 2017). Stan is cross-platform and provides robust Bayesian samplers with a unified modeling language, as well as visual diagnostics for chain convergence via ShinyStan (Stan Development Team, 2017). Posteriors can be sampled by HMC for basic mutation signatures and by either HMC or automatic differentiation variational inference (ADVI) for population-specific signatures. SignIT was implemented and tested in R (version 3.4.1) using the RStan package. Analyses included in this manuscript were run on an x86_64 CentOS6 Linux cluster.

The SignIT signature model posteriors are sampled by Markov-chain Monte Carlo (MCMC), by default employing four chains each traversing 200 burn-in iterations and 200 sampling iterations with no thinning. The complete SignIT population-signature model by default employs ten chains with 200 burn-in iterations and 300 sampling iterations each. These parameters can be tuned, however, during testing these values have yielded consistent convergence with sufficient sampling density and effective sample sizes over the posterior with acceptable autocorrelation. The output contains summary statistics along with an attached Stan model output, which can be easily run through diagnostics in ShinyStan (Stan Development Team, 2017).

Because the SignIT population-signature model can be slow to sample by MCMC, especially in hypermutated cases, we also enable estimation of posteriors by variational inference. SignIT leverages Stan's ADVI module, which auto-

maticaly selects a variational family and performs optimization to minimize the Kullback-Liebler divergence. Upon convergence, 1000 iterations are drawn from the posterior distribution via importance sampling. SignIT treats these 1000 iterations equivalently to MCMC iterations in subsequent analyses.

The relevant version of all dependencies installed for this analysis are part of an Anaconda virtual environment which can be installed and executed on a Unix-based terminal using the following commands.

```
git clone https://github.com/eyzhao/bio-pipeline-dependencies.git
git checkout tags/SignIT-paper-dependencies
make
source miniconda3/bin/activate dependencies
```

Selecting the Number of Subpopulations

To select the number of subpopulations, we recommend performing inference over models ranging from 1 to at least 5 subpopulations and computing the Watanabe-Akaike information criterion (WAIC) on each. SignIT provides a function for automatically computing the WAIC on sampler output. The model which minimizes WAIC should be preferred.

If there is insufficient time or computational resources to attempt a range of models, SignIT can also provide a rough estimate of the optimal population count. Maximum *a posteriori* parameter estimates are obtained using the population model 3.9, excluding the portion for inferring signatures, for models ranging from 1 to 5 subpopulations. Bayesian information criteria (BIC) are computed for each model's parameter estimates and the model with minimum BIC is chosen.

3.4.2 *Simulated Genomes*

To evaluate mutation signature decomposition accuracy against known "true" exposures, we devised a mutation signature simulation package called msimR, available at github.com/eyzhao/msimR. Somatic mutation count vectors were simulated by drawing mutations of 96 distinct SNV classes from a multinomial distribution, $v_i \sim \text{Multinomial}(N, \hat{\mathbf{S}}\boldsymbol{\epsilon})$, where $\boldsymbol{\epsilon}$ represents a theoretical set of "true" exposures (Figure 3.2A). Active mutation signatures were randomly selected and all non-contributing signatures had their exposures set to zero. For each simulated mutation set, contributing signatures were selected at random from a uniform distribution $\text{Uniform}(0, 1)$ and the resulting $\boldsymbol{\epsilon}$ vector was normalized to sum to 1.

Aside from varying signature exposures, an additional source of variability may arise from differences between the reference signatures and the "true" biological mutational processes driving a tumour. It is unlikely that any static set of reference signatures can accurately reflect all possible mutational processes. Therefore, to simulate inaccuracies in the underlying reference signature set, a reference signature perturbation was performed by introducing Gaussian noise into $\mathbf{S}$. Given a perturbation factor p between 0 and 100 and "true" reference signature matrix $\mathbf{S}$, a perturbed signature matrix $\hat{\mathbf{S}}$ was randomly computed where $\hat{S}_{ij} \sim \text{Normal}(\mu = S_{ij}, \sigma = \frac{p}{100} * S_{ij}); i = 1, 2, \ldots V; j = 1, 2, \ldots K$, with the restriction that $\hat{S}_{ij} >= 0$.

In addition to SignIT, we implemented deconstructSigs v1.8.0 (Rosenthal et al., 2016) and SignatureEstimation v1.0.0 (Huang et al., 2017). All mutation signatures were deciphered "blindly," using only the mutation count vector $\mathbf{v}$ and the complete consensus (non-perturbed) matrix of reference signatures $\mathbf{S}$ (not $\hat{\mathbf{S}}$). This

scenario best reflects the real-world problem where only the observed mutations and reference signature matrix are known (see Figure 3.2A).

Simulated count vectors were generated with combinations of three parameters. (1) The number of mutations, N, was varied from 10 to 10^6; (2) The number of contributing signatures was varied from 1 to 20; and (3) the amount of reference signature perturbation was varied from 0% to 80%. For each combination of parameters, 500 random mutation count vectors were generated for a total of 90,000. Each count vector was decomposed into exposures using each of the three methods. Deviation between calculated exposures and true exposures was reported using the cosine distance ($1 - \frac{\boldsymbol{\epsilon} \cdot \mathbf{e}}{|\boldsymbol{\epsilon}||\mathbf{e}|}$, Figure 3.2B).

3.4.3 *Publicly Available Cancer Mutation Data*

2,748,760 cancer somatic SNVs from 4,563 exomes called by four mutation callers (MuSE, MuTect2, VarScan2, and SomaticSniper) via a harmonized pipeline were obtained as mutation annotation format (MAF) files from TCGA using the Genomic Data Commons (GDC) portal (gdc.cancer.gov). This study included data from 9 TCGA cancer cohorts; BLCA, BRCA, CESC, COAD, LUAD, LUSC, SKCM, STAD, and UCEC were chosen because they have the most cumulative somatic mutations and are thus more likely to yield reproducible, high quality mutation signatures by NMF. 459,552 SNVs called by only one of four callers were filtered out, leaving a total of 2,289,208 somatic SNVs (Table 3.1). Mutation signatures we deciphered *de novo* using NMF from each cohort using the WTSI framework (Alexandrov et al., 2013b), then the best-matching reference signature was chosen based on cosine similarity for comparison with n-of-1 methods. Where multiple *de novo* signatures best matched one reference signature, only the top match was cho-

sen for comparison. N of 1 mutation signatures were fitted using SignIT, deconstructSigs, and SignatureEstimation using the complete set of 30 SNV mutation signatures as a reference matrix, and blinded to the NMF analysis. Comparison between NMF and each n-of-1 method was performed by Spearman correlation of sample exposures between matching signatures (Figure 3.4).

Table 3.1: The numbers of samples and variants in each TCGA cohort analyzed.

Cohort	Sample Count	Total SNVs	Excluded SNVs	SNVs for NMF
BLCA	411	145,980	22,681	123,299
BRCA	983	130,254	34,701	95,553
CESC	289	112,735	18,573	94,162
COAD	399	287,035	66,401	220,634
LUAD	562	224,075	39,886	184,189
LUSC	491	195,436	28,051	167,385
SKCM	466	431,179	35,731	395,448
STAD	433	225,913	57,429	168,484
UCEC	529	996,153	156,099	840,054

3.4.4 *De Novo Signature Analysis*

SNVs were categorized based on 6 variant types and 16 trinucleotide context subtypes to yield a total of 96 mutation classes. Within each cohort, mutation signatures were deciphered using a published framework (Alexandrov et al., 2013b), which employs NMF to infer both the operative signatures prevalent across samples and the relative exposure of each signature to each sample. Signature stability estimates were obtained by Monte Carlo simulation with 1000 iterations (10 iterations over 100 cores). In each cohort, signature models involving 2 to 8 signatures were attempted and the solution which maximized signature stability and min-

imized reconstruction error was selected (Appendix Figure A.3). The similarity of signatures to thirty previously described mutational signatures (available from cancer.sanger.ac.uk/cosmic/signatures) was quantified using the cosine similarity metric and the most similar corresponding signature was selected in each case.

3.4.5 *Structural Variant Mutation Signatures*

For the SV signatures in Figure 3.1C, SVs were categorized as per Nik-Zainal et al. (2016) based on the mutation type (deletion, duplication, inversion, or translocation), the length of SV (except for translocations), and whether the SV breakpoints were clustered. Clustered breakpoints were in segments with breakpoint density at least 10 times greater than average, and segments were determined using a piecewise linear fitting with smoothness parameter $\gamma = 25$ and minimum breakpoints per segment $k_{\text{min}} = 10$, implemented using the copynumber package (v1.18.0) in R. This yielded a 32-class parameterization. SignIT analysis was performed on the resulting mutation count vector using six previously published SV signatures (Nik-Zainal et al., 2016) as the reference matrix.

3.4.6 *Whole Genome Sequencing of Metastatic Cancers*

Study participants with advanced stage cancers underwent tumour biopsies as part of the POG Project (Laskin et al., 2015). The study was approved by the University of British Columbia Research Ethics Board (REB# H12-00137 and H14-00681). Written informed consent, including potential publication of findings, was obtained from patients prior to genomic profiling. Patient information was anonymized, and each was assigned an alphanumeric identification code.

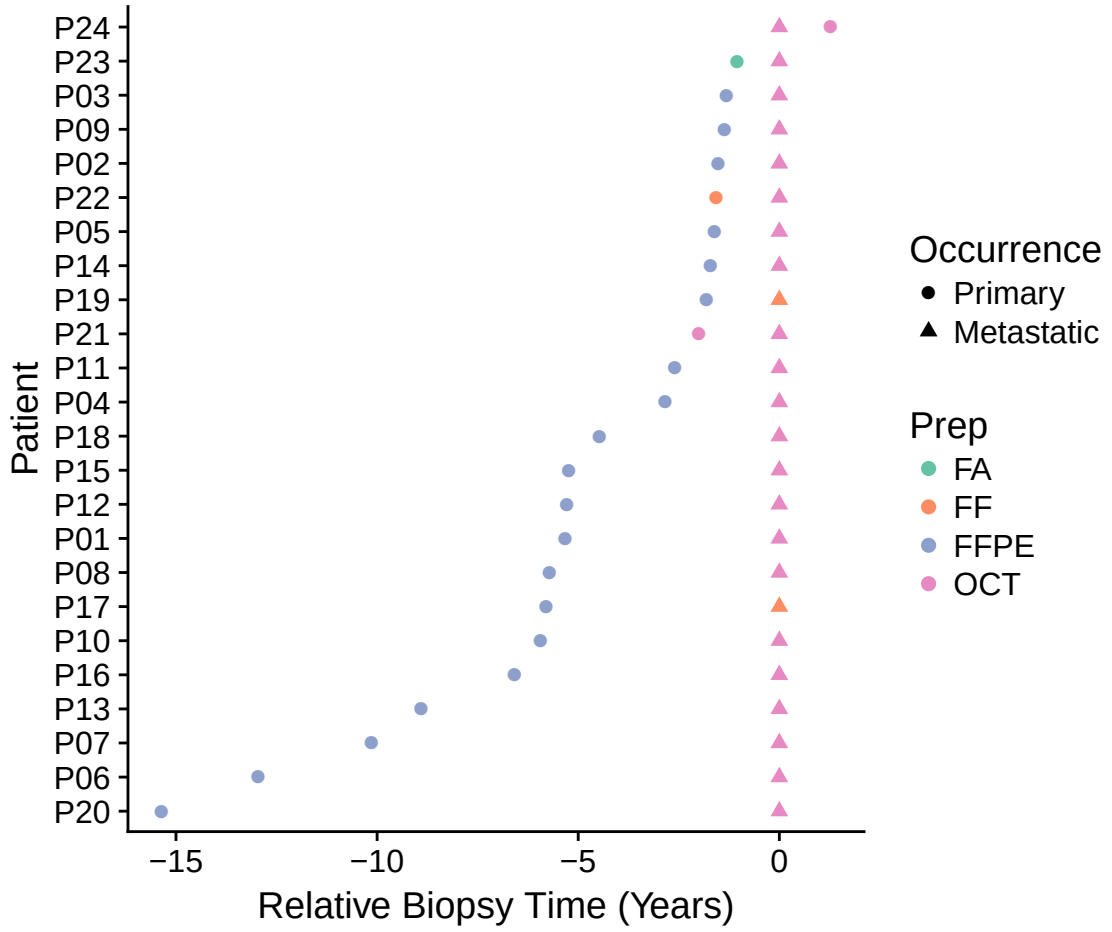

Figure 3.10: The time of sample collection for multiply sequenced tumours. Whole genome analysis of multiply sequenced tumours was performed in 24 patients. Timing of primary tumour sample collection relative to the metastatic biopsy ranged from -15 to +1 years.

24 patients underwent whole genome sequencing of multiple temporally and/or spatially distinct tumours (Appendix Table A.3). Of these, the primary tumour was biopsied before the metastatic tumour in all except one case (Figure 3.10). Whole-genome sequencing data (.bam files) have been submitted to the European Genome-Phenome Archive (EGA) (www.ebi.ac.uk/ega/home) under the study accession number EGAS00001001159.

The details of library construction, sequencing, and bioinformatics of metastatic samples have been previously described (Jones et al., 2010). Briefly, biopsy samples were embedded in OCT compound and sectioned. Pathology review was performed to select sections for sequencing. Genome libraries were constructed from tumor and peripheral blood (normal control) and sequenced using Illumina protocols and on a HiSeq sequencer. Reads were aligned to hg19 by the BWA aligner (v0.5.7) (Li and Durbin, 2009, 2010). Somatic SNVs and small insertions/deletions were processed using samtools (Li et al., 2009) and Strelka (v0.4.6.2) (Saunders et al., 2012). CNVs were called using CNASeq (v0.0.6). LOH was called by APOLLOH (v0.1.1) (Ha et al., 2012). SVs called both from ABySS *de novo* assembly (Jackman et al., 2017) and by DELLY (Rausch et al., 2012) were intersected based on events with breakpoints less than 20 base pairs apart.

3.4.7 *Ploidy-correction of Copy Number Variants*

SignIT relies upon accurate calling of CNVs in order to correct for variant allele probabilities. For every metastatic cancer which underwent whole genome sequencing, a most likely ploidy model was determined by manual review using CNV and LOH calls and informed by tumour content estimates both from pathology assessment and bioinformatic analysis. In order to correct for ploidy, the absolute copy numbers of segments called by the CNASeq hidden Markov model (HMM) were adjusted. The tumour-normal depth ratios, $\overline{R}$, used as input for segmentation are computed as

$$\overline{R} = \frac{C^{(T)}T + C^{(N)}(1-T)}{TP + C^{(N)}(1-T)},$$

where $\overline{R}$ is the mean tumour-to-normal read depth ratio across the segment, T is the tumour content, and P is the ploidy. $C^{(T)}$ is the estimated absolute tumour copy number of the segment, and $C^{(N)}$ is the normal copy number, assumed to be 2. The ploidy model is chosen manually by inspection of allele-specific read depths. The numerator is proportional the relative abundance of reads from the tumour sample, which is a mixture of tumour and normal cells. The denominator is the relative abundance for a region with no copy number abberation. Rearranging yields

$$C^{(T)} = \frac{(\overline{R})(TP + C^{(N)}(1-T)) - C^{(N)}(1-T)}{T}.$$

3.4.8 *Whole Genome Sequencing of Primary Tumours*

4 primary samples were sequenced as frozen or OCT samples as described in the previous section. 20 primary samples were sequenced from FFPE material.

Whole genome libraries from primary tumour were constructed as previously described (Chong et al., 2016) with modifications. The input amount of FFPE DNA samples varied from 100 ng to 2 μg depending on availability. To improve library quality, the sheared genomic DNA was either size-selected by polyacrylamide gel electrophoresis or by solid phase reversible immobilization bead-based size selection to remove smaller DNA fragments from highly degraded strands. Libraries were sequenced on the Illumina HiSeq2500 sequencer using paired-end sequencing with read lengths of 100 or 125 bp.

4

CLINICAL APPLICATION OF MUTATION TIMING IN A BRCA1-MUTATED PANCREATIC ADENOCARCINOMA

4.1 INTRODUCTION

HR facilitates error-free repair of double-strand DNA breaks and interstrand crosslinks (Li and Heyer, 2008). Mutations in *BRCA1*, *BRCA2*, and other genes responsible for HR are prevalent among human cancers, causing HRD and genomic instability (Scully and Livingston, 2000).

WGS efforts have identified mutational and structural rearrangement signatures linked to *BRCA1* and *BRCA2* mutations in breast and other cancers (Lord and Ashworth, 2016), which may predict response to platinum-based chemotherapy and PARP inhibitors. However, the role of signature timing on treatment response has not been elucidated, but could help distinguish currently active, actionable mutational processes from historically active ones. In chapter 2, we demonstrated an association between signatures of HRD and response to platinum-based chemotherapy. In chapter 3, we developed a method to perform n-of-1 analysis of mutation signatures and their temporal evolution.

Here, we present the first clinical application of HRD dynamics across spatially and temporally distinct biopsies of a pancreatic ductal adenocarcinoma (PDAC). This approach helped to reconcile paradoxical findings: genomic stability and low HRD mutation signature despite a germline *BRCA1* mutation and exceptional re-

sponse to FOLFIRINOX. Our findings highlight the potential value of considering timing in the clinical interpretation of mutation signatures.

4.2 CASE REPORT

As part of an ongoing study exploring the use of comprehensive molecular analysis to inform treatment decision-making (NCT 02155621) (Laskin et al., 2015), a 67-year-old male with metastatic PDAC and the germline founder mutation *BRCA1* c.68_69delAG (185delAG) consented to undergo biopsy of a liver metastasis for molecular analysis. The primary tumor had been resected previously, followed by 6 months of adjuvant cisplatin/gemcitabine chemotherapy, before detection of liver metastases 12 months after surgery and 6 months after discontinuing cisplatin/gemcitabine. Liver biopsy was performed before commencement of palliative FOLFIRINOX chemotherapy (5-fluorouracil, oxaliplatin and irinotecan). He had an excellent response to treatment, with CA19-9 halving in 2 months and complete PET response within 4 months (Figure 4.1A). Oxaliplatin was held after 16 cycles due to peripheral neuropathy and the patient continued to have disease control on first-line chemotherapy at last follow up, 18 months later. This represents an exceptional response, as median overall survival for metastatic pancreatic cancer is less than 6 months, or 11 months with FOLFIRINOX treatment (Conroy et al., 2011).

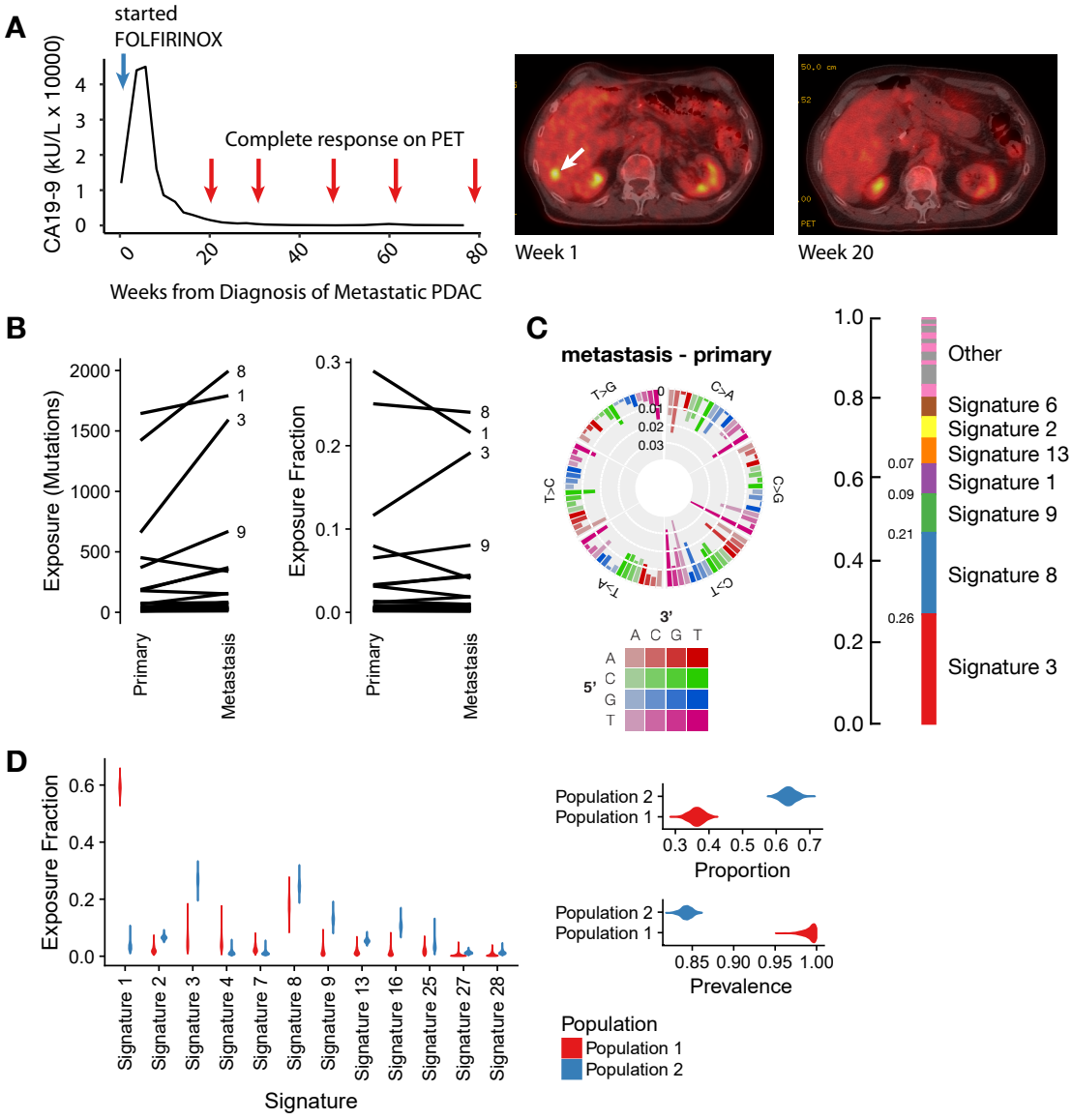

Figure 4.1: Evolution of single nucleotide variant (SNV) mutation signatures in a pancreatic adenocarcinoma with exceptional response to FOLFIRINOX. (A) At 20 weeks of treatment with FOLFIRINOX, the patient exhibited a complete response, which was maintained for over 18 months. (B) Mutation signature exposures in the primary tumor and metastasis reveal a substantial rise in the homologous recombination deficiency (HRD) signature. (C) The complete catalogue of new somatic SNVs (present in metastasis but absent in primary tumor), with SNVs categorized into 96 classes based on variant type and 3′/5′ context was matched against 30 pre-defined mutation signatures. This revealed dominant involvement of the HRD signatures (Signature 3 and 8). (D) Temporal dissection by SignIT revealed two tumour subpopulations, which revealed a drop in signature 1 and rise in signatures 3 and 8.

4.3 RESULTS

4.3.1 *BRCA1 Loss in the Primary and Metastasis*

Both the primary tumor and metastasis demonstrated genomically stable structural variant profiles based on previous characterization of the pancreatic cancer genome landscape (Waddell et al., 2015). The *BRCA1* c.68_69delAG frameshift variant, heterozygous in the germline, was homozygous in both tumors as demonstrated by copy-neutral loss of heterozygosity (CNLOH) spanning most of chromosome 17 and detailed analysis of aligned reads (Figure 4.2). Analysis of the BRCA1 transcripts showed the presence of the mutation in all expressed transcripts.

4.3.2 *Timing of the BRCA1 Loss*

Based on analysis with cancerTiming (Purdom et al., 2013), the CNLOH event on chromosome 17 resulting in homozygosity of *BRCA1* was the 9^{th} earliest of 30 events ($\pi_0 = 0.15$) in the metastasis and 17^{th} of 41 events ($\pi_0 = 0.40$) in the primary, suggesting that it was not among the earliest tumor-initiating events (Figures 4.4C, D). *TP53* loss of function was also observed and was likely a simultaneous occurrence due to the same CNLOH event. Cellular prevalence estimation using TITAN (Ha et al., 2014) converged on a 4-subclone model but suggests that the chr17 LOH event was clonal in the metastasis (Figure 4.3). The clonality of this event minimizes the risk of platinum resistance by selection of a *BRCA1* wild-type subclone. Tumor content in the pancreatic primary was insufficient to estimate clonality.

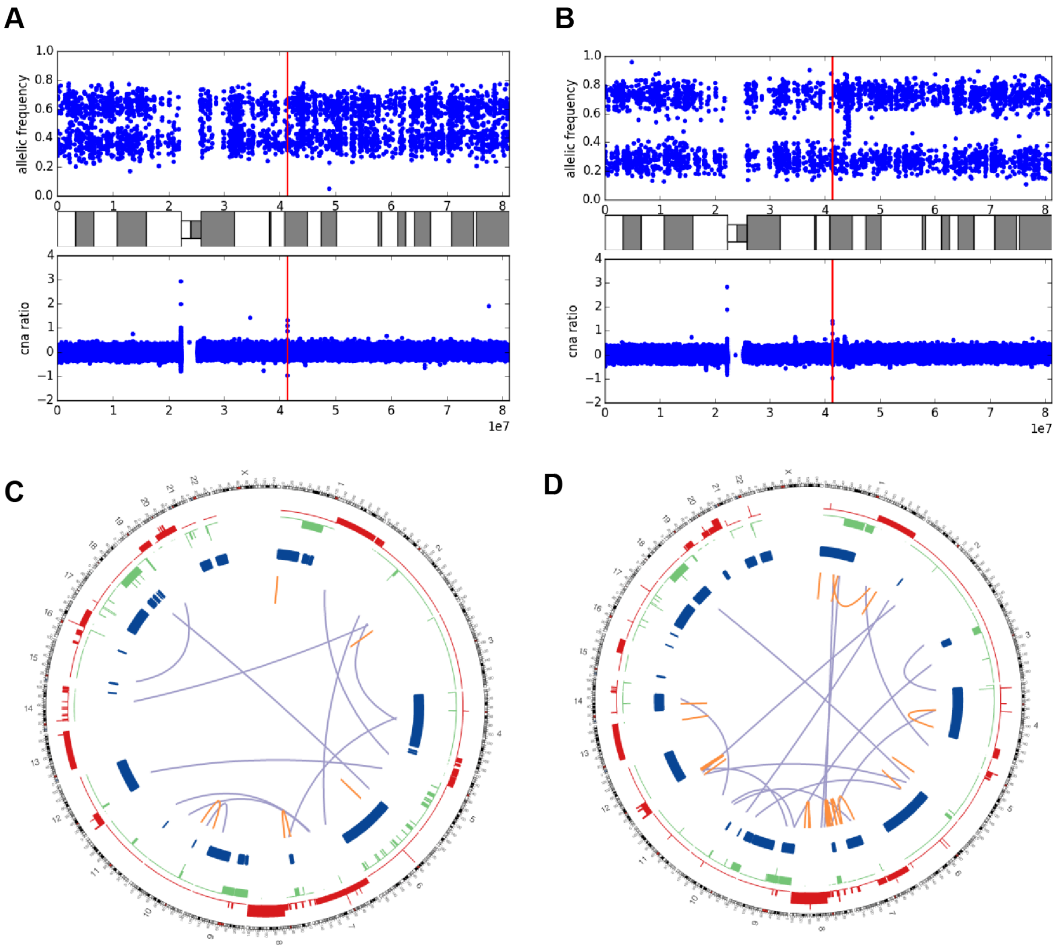

Figure 4.2: Genomic analysis and clinical evolution of a germline *BRCA1* c.68_69delAG-associated pancreatic ductal adenocarcinoma (PDAC) primary tumor (left) and metastasis (right). **(A, B)** Copy-neutral loss of heterozygosity (CNLOH) of chromosome 17. The allelic ratio is shown in the top-left and top-right plots. The copy number variant (CNV) ratios are shown in the bottom-left and bottom-right plots. LOH regions were called using APOLLOH. The position of *BRCA1* is shown as a red vertical line in all plots. **(C, D)** Structural variants, LOH and CNV events are depicted from the centre outwards. The green and red bars represent copy loss and copy gain respectively. Estimated tumor content by sequencing was 25% in the primary tumor and 49% in the metastasis. **(E)** Development of liver metastases with rising CA19-9 that peaked at 45,000 kU/L. Within 4 weeks of commencing FOLFIRINOX, CA19-9 decreased by over 50%; positron emission tomography (PET) complete response was seen at 20 weeks. At the time of writing, the patient has an ongoing PET complete response and suppression of CA19-9, 79 weeks after commencing treatment.

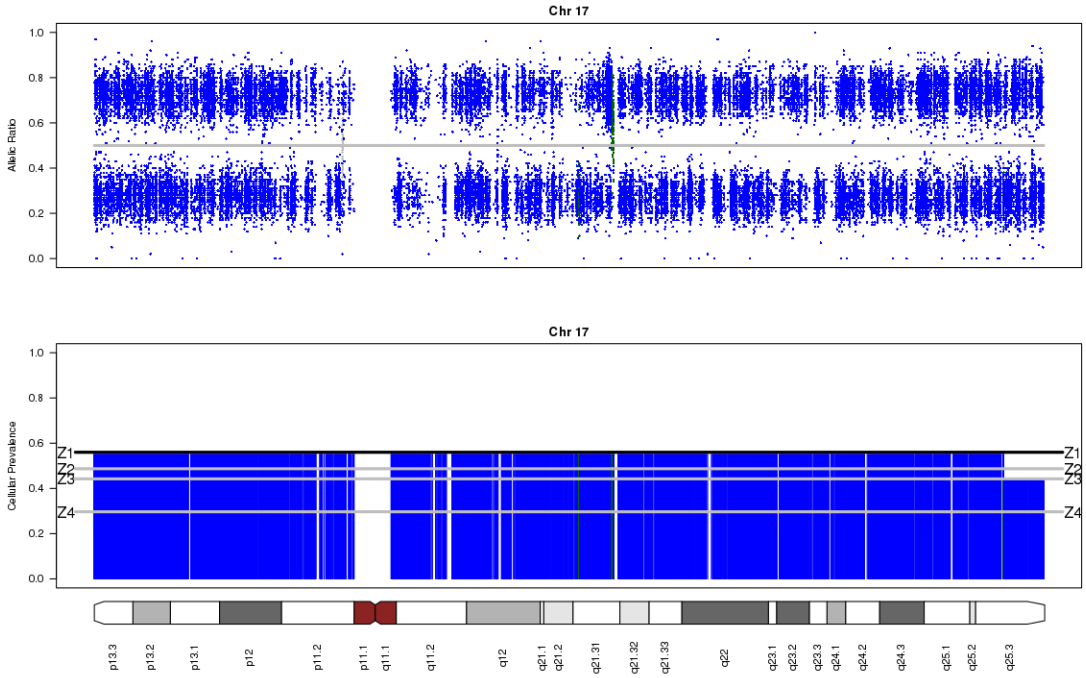

Figure 4.3: Joint calling of CNV, LOH, and clonal status performed across the metastatic genome using TITAN. The 4-clone model yielded an optimal fit to the data. Investigation of chromosome 17 revealed that the CNLOH event affecting the chr17 genes *TP53* and *BRCA1* was clonal, with high cellular prevalence.

Inferred timing of shared genomic events was significantly correlated (r = 0.7, $p = 4.8 \times 10^{-6}$) between primary and metastasis samples (Figure 4.4A), with events in the metastasis consistently inferred to be "earlier" (Figure 4.4B). This is expected, and reflects the "aging" of shared genomic events during the approximately one-year gap between sequencing of the two samples. To our knowledge, this is the first biological validation of a CNV timing inference model using multiple sequencing time points.

4.3.3 *Evolution of Mutation Signatures from Primary to Metastasis*

The relative contributions of 30 previously described mutation signatures (Alexandrov et al., 2013a) were determined from 5683 SNVs in the primary and 8315 in the metastasis. Signature 3 and, to a lesser extent, Signature 8 have been associated with HRD (Davies et al., 2017; Nik-Zainal et al., 2012). Mutations associated with signature 3 rose by 1593 in the metastasis, and signature 8 rose by 1421, more than any other signature (Figure 4.1A). Further, of new somatic mutations (present in the metastasis but absent in the primary tumor), 26% were associated with the HRD signature (Figure 4.1B), suggesting major involvement of HRD-associated mutagenesis in the evolution of this PDAC. Strong signature bleed was observed between signatures 3 and 8 (Figure 4.5), but there was little bleed between HRD and other signatures.

4.3.4 *Evolution of Orthogonal HRD-associated Mutational Signatures*

The presence of recently described genomic signatures associated with HRD (Davies et al., 2017) was investigated in the primary and metastasis. Rearrange-

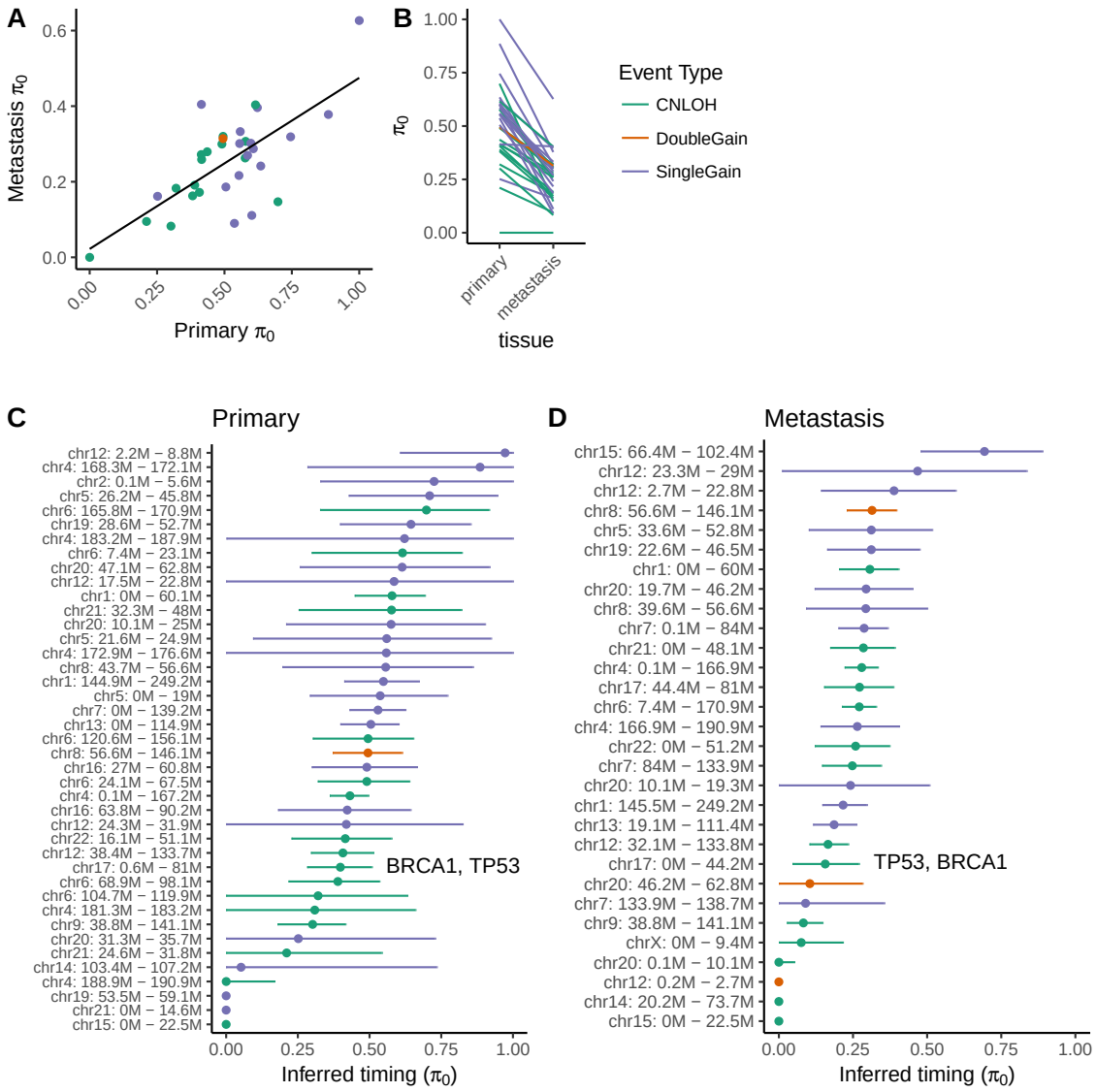

Figure 4.4: Comparison of inferred timing for events shared between pancreatic primary tumor and metastasis. (A) Inferred timing of copy-change events in the primary tumor and metastasis were strongly correlated, with a slope of 0.45, intercept of 0.02, r = 0.7, and $p = 4.8 \times 10^{-6}$. Only shared genomic events in overlapping regions are shown. **(B)** Events in the metastasis sample were inferred to have arisen "earlier" in the course of tumorigenesis, consistent with the timing of sample collection. **(C, D)** Inferred timing with 95% confidence intervals in the primary and metastasis. Event coordinates are labelled along the y-axis. In case of two-copy gain, only the timing of the first copy gain is shown. The loss of heterozygosity (LOH) event encompassing both *BRCA1* and *TP53* has been highlighted, and is the 17th earliest of 41 inferred events in the primary and 9th of 30 events in the metastasis.

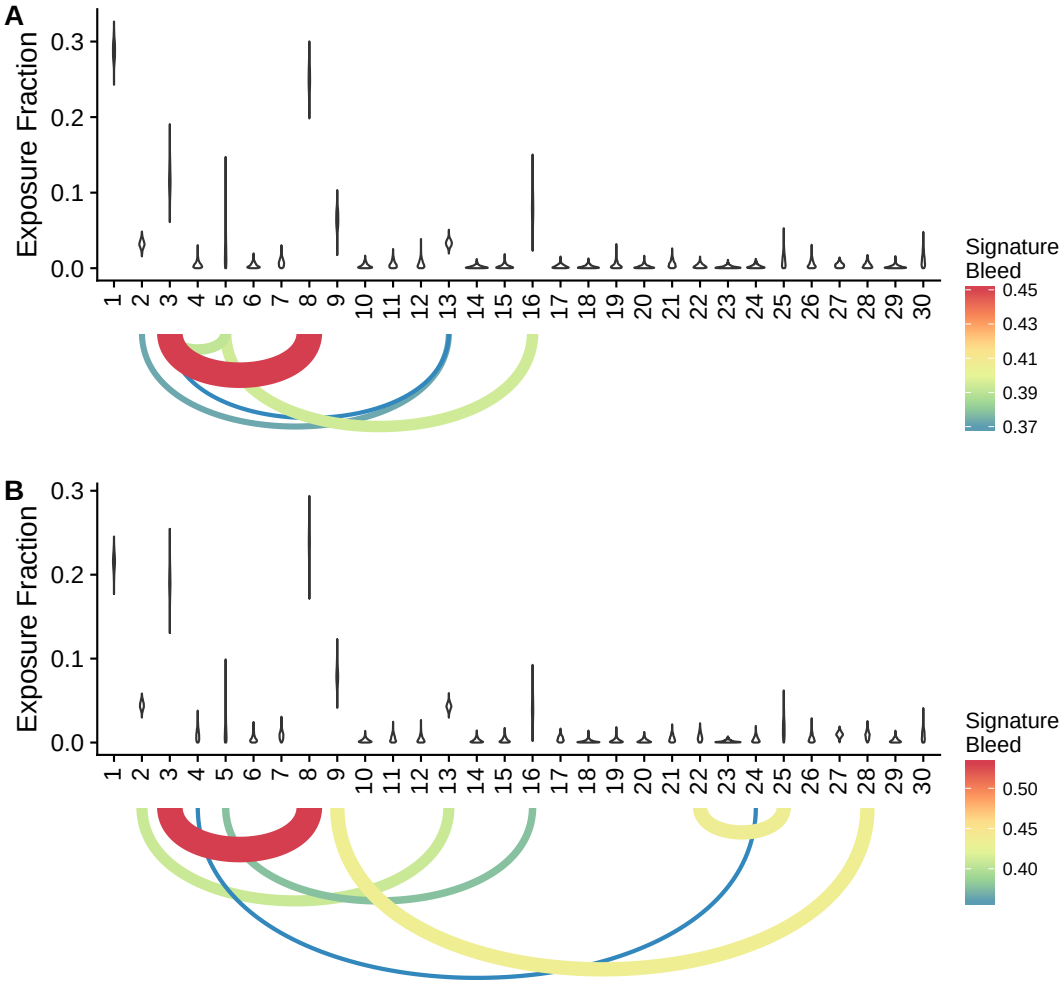

Figure 4.5: Mutation signature bleed between signatures 3 and 8. Using SignIT, mutation signature posteriors were estimated from a full Bayesian solution to a categorical mixture model using Hamiltonial Monte Carlo. Signature bleed was collected based on anticorrelation in 2D projections of the posterior probability distribution. Mutation signature exposures from (A) the primary sample and (B) the metastatic sample demonstrate increasing involvement of Signature 3, with bleed between Signatures 3 and 8.

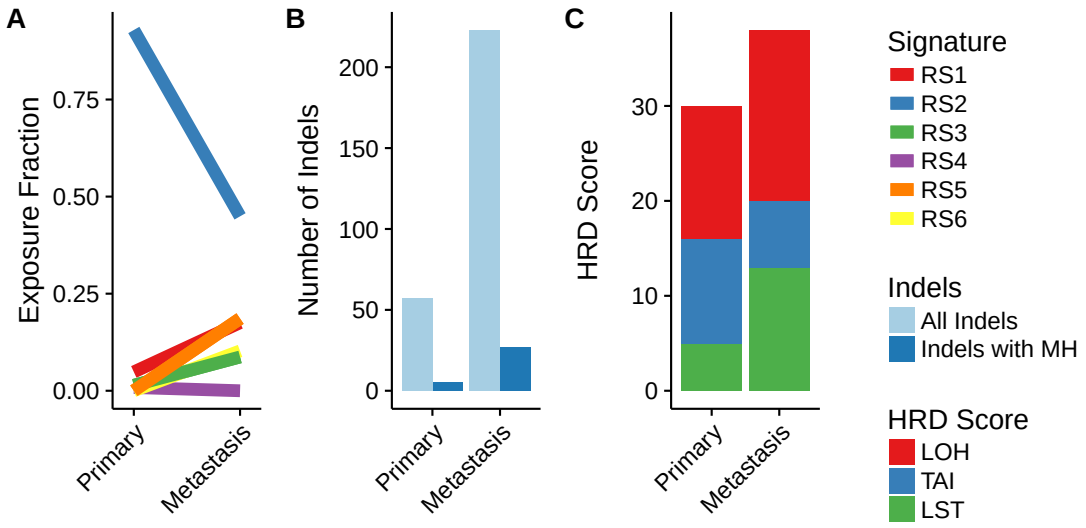

Figure 4.6: Evolution of structural variation alterations between the pancreatic primary and metastasis. (A) Rearrangement signatures 3 and 5, associated with HRD, were low but rose between the primary and metastasis. (B) The fraction of indels with microhomology rose from 8% to 12%. (C) The total HRD score rose from 30 to 38, largely driven by a rise in loss of heterozygosity (LOH) and large-scale transitions (LST). RS: Rearrangement signature; MH: microhomology; TAI: telomeric allelic imbalance.

ment signatures 3 and 5 (Figure 4.6A) and the fraction of indels with microhomology (Figure 4.6B) were low, but rose between the primary and the metastasis. The HRD composite score, combining LOH, TAI, and LST, increased from 30 in the primary to 38 in the metastasis (Figure 4.6C). A caveat is that the low tumour content of the primary tumour may impact the accuracy of SV and CNV calling. This was the highest observed HRD score among the first 25 PDAC cases in our study cohort.

4.3.5 *Mutation Signature Timing*

Analysis with SignIT revealed two temporally distinct tumour subpopulations with mutational prevalences of 0.99 and 0.84. The higher-prevalence population reflects signatures from clonal mutations (present in every tumour cell) or mutations present on multiple copies. Both are associated with earlier-arising mutations: clonal mutations occur before subclone branching, and multi-copy mutations occur before replication of the associated segment 1.1. These early mutations account for 36% of mutations, while later mutations make up 64%.

The early subpopulation was dominated by Signature 1, which accounted for 59% of mutations, whereas Signatures 3, 8, 9, and 16 were active in the later subpopulation. These findings agree with the observed increase in signatures 3, 8, and 9 from the primary to the metastatic sample. In particular, Signature 3 exposure rose from 6% of mutations in population 1 to 27% in population 2. These findings provide additional evidence that HRD remains an active mutation-causing process in the metastatic tumour, despite the overall low SV burden and moderate HRD signature exposure.

4.4 DISCUSSION

Previous studies have reported platinum sensitivity in PDACs with *BRCA1/2* mutations, rampant SVs, and strong mutation signature (Waddell et al., 2015). Here, we explored the genomic evolution of a *BRCA1* germline mutated PDAC with a paradoxically low HRD mutation signature and genomic instability burden. Based on observations involving the temporal dynamics of mutational signatures, we postulate that HRD onset in this case may have occurred too recently to produce

a heavy burden of genomic instability, thus resulting in the absence of an unstable rearrangement signature. However, rising HRD signature exposure suggests that HRD remains a "currently active" process, which may explain the patient's excellent and sustained response to FOLFIRINOX, a platinum-containing chemotherapy.

This analysis has some limitations. The archival primary sample had low tumour content (25%), which is known to limit the accuracy of mutation calling, and thus required analytical validation of major findings. Mitigating factors include the findings that timing of CNV and LOH events was concordant between primary and metastasis and temporal analysis of the metastasis corroborated mutation signature evolution patterns. Moreover, SNVs were called by Strelka, which is designed to operate under low cellularity (Saunders et al., 2012). Another important limitation is that FOLFIRINOX contains agents other than oxaliplatin, namely fluorouracil, leucovorin, and irinotecan, which may have contributed to the durable treatment response. Notably, the patient did not exhibit such a dramatic response to adjuvant cisplatin/gemcitabine therapy in the primary setting. While this could be explained by low HRD mutagenic activity in the primary tumor, the action of non-oxaliplatin agents cannot be neglected. Lastly, it is possible that exposure to platinum-based therapy may have driven HRD-associated mutagenesis in the metastatic tumour. However, other studies have discovered HRD-associated signatures in treatment-naive primaries (Alexandrov et al., 2015b; Nik-Zainal et al., 2016), and they differ from recently-discovered platinum-associated signatures (Boot et al., 2017; Szikriszt et al., 2016). Despite these caveats, we believe this case raises important educational questions on whether temporal evolution of HRD signature activity may help refine prediction of therapy response.

Although HRD is commonly considered an early tumor-initiating event, a recent study suggests that the *BRCA1* and *BRCA2* mutation signature is also prevalent in late-arising mutations (McGranahan et al., 2015). The exploration of mutation timing has not yet been widely adopted, and presents with numerous technical challenges (Purdom et al., 2013). This case was an opportune candidate for timing analysis due to the availability of primary tissue and the large chr17 CN-LOH event spanning *BRCA1* and *TP53*. Consequently, these findings raise several questions beyond the scope of this brief report. Do temporal dynamics vary across pathogenic *BRCA1*, *BRCA2*, or other HRD-associated gene variants, and is "late-onset" HRD a common phenomenon?

With a growing body of evidence supporting the role of HRD as a predictive marker of response to platinum-based therapy and PARP inhibitors across various tumor types, there is increasing interest in new approaches to identify genomic scars associated with HRD. Although WGS techniques provide a cross-sectional snapshot of the cancer genome at a fixed moment in time, they can also be used to infer the relative timing of genomic events. We hope that ongoing comprehensive molecular analysis with high quality prospective treatment and outcome information will facilitate a deeper understanding of the nuances in HRD-related mutational processes, resulting in improved clinically predictive accuracy of HRD assessment.

4.5 METHODS

4.5.1 *Tissue Collection, Processing, and Storage*

Following informed consent, patients underwent image-guided metastatic biopsies as part of the Personalized OncoGenomics program of British Columbia (NCT 02155621, University of British Columbia Clinical Research Ethics Board approval no. H12-00137). Up to 5 biopsy cores were obtained using 18-22G biopsy needles and embedded in optimal cutting temperature (OCT) compound. Tumor sections were reviewed by a pathologist to confirm the diagnosis, evaluate tumor content and cellularity and to select areas most suitable for DNA and RNA extraction. Peripheral venous blood samples were obtained at the time of biopsy and leukocytes isolated for use as a germline DNA reference. DNA and RNA were extracted for genomic and transcriptomic library construction, which have been previously described in detail (Sheffield et al., 2015).

Tissue from the primary pancreatic tumor and liver metastasis were sequenced, with leukocytes isolated from blood samples used as a germline DNA reference. Tumour content was estimated at 49% for the metastatic sample and 25% for the primary. The low tumour content of the primary sample necessitates careful interpretation of variant calls along with orthogonal validation of key findings.

The primary pancreatic tumor sample was obtained from the previously resected specimen that had been snap-frozen at the time of surgery and stored at the BC Gastrointestinal Biobank at -80°C for approximately 18 months prior to extraction and analysis. All samples were handled under sterile conditions and transported in dry ice.

4.5.2 *Sequencing and Bioinformatics*

Paired-end reads were generated on an Illumina HiSeq2500 sequencer and aligned to the human reference genome GSCh37 by the BWA aligner (Li et al., 2009) (v0.5.7). Somatic SNVs and small insertions/deletions were processed using SAMtools (Li and Durbin, 2010) and Strelka (Saunders et al., 2012) (v0.4.6.2). Regions of CNV were determined using CNASeq (v0.0.6) and LOH by APOL-LOH (Ha et al., 2012) (v0.1.1). Tumor content and ploidy models were estimated from sequencing data through analysis of the CNA ratios and allelic frequencies of each chromosome. This was then compared to theoretical models (Ha et al., 2012) for diploid, triploid, tetraploid, and pentaploid genomes at various tumor contents (10% intervals from initial lab estimate). The resulting analysis was a diploid model at 25% tumor content in the PDAC primary and 49% tumor content in the PDAC metastasis. Structural variation was detected by de novo assembly of tumor reads using ABySS and Trans-ABySS (Robertson et al., 2010), followed by variant discovery using DELLY (Rausch et al., 2012).

CNV and LOH Analysis with TITAN

Joint detection of CNV and LOH was performed on the metastatic sample using TITAN (Ha et al., 2014). The TitanCNA Bioconductor package (version 1.12.0) and its dependencies were installed in R (version 3.3.2). The germline heterozygous mutations required by TITAN were called using MutationSeq (version 4.3.8) (Ding et al., 2012) installed in Python (version 2.7.13). The germline variants were filtered for those present in dbSNP (release 138, common_all). For more information, see https://github.com/MO-BCCRC/titan_workflow. The cellular prevalence of each event was estimated according to a 4-subclone model, which yielded

the best Bayesian information criteria fit to the SNV read counts. TITAN provided an estimated tumour content of 0.56, which is similar to the tumour content estimate obtained by manual review (0.49). TITAN also provided an average tumour ploidy of 2.05. These findings were used to assess the clonal status of the LOH event on chromosome 17 spanning BRCA1 and TP53.

4.5.3 *Mutation Timing Analysis*

The relative temporal ordering of large scale genomic events can be performed by leveraging SNV burden as a "molecular clock". Note that this method can only infer the timing of events for which the precise history is known with reasonable confidence. As a result, only regions with CNLOH or allele-specific amplification with 1-copy or 2-copy gain have inferrable timing. Thus, Figure 2 shows the inferred timing for the subset of events which fit this criterion.

This analysis was performed using the cancerTiming module of the R programming language (Purdom et al., 2013). Because larger genomic events yield more accurate timing, small events which interrupt adjacent larger ones were automatically filtered out 2. This filtering step dramatically improved the assocation in timing between the primary and metastatic samples. For each segment, cancerTiming computes π_0, the probability of a random SNV within the affected loci occurring prior to the event. Greater π_0 values suggest later occurrence of the CNV and/or LOH. Bootstrap distributions were computed non-parametrically using 1000 iterations 95% confidence intervals were determined by reporting the 25th and 975th ordered values from the resulting distribution.

4.5.4 *Mutation Signature and Signature Timing Analysis*

Patterns of somatic SVs, SNVs, and CNVs were interrogated to determine the contribution of HRD to mutagenesis. SNV and SV count vectors were computed as previously described (Alexandrov et al., 2013b; Nik-Zainal et al., 2016). Using the 30 consensus signatures from COSMIC as a reference set (cancer.sanger.ac.uk/cosmic/signatures), signature exposures were computed using SignIT.

4.5.5 *Additional HRD Metrics: Deletion Microhomology and HRD Score*

The proportion of deletions exhibiting overlapping microhomology was determined by identifying matching sequences between deleted ends and flanking regions. HRD scores were computed as the arithmetic sum of LOH, TAI, and LST metrics, which in turn were determined using CNV and LOH patterns (Figure 4.7) based on previously described guidelines (Timms et al., 2014).

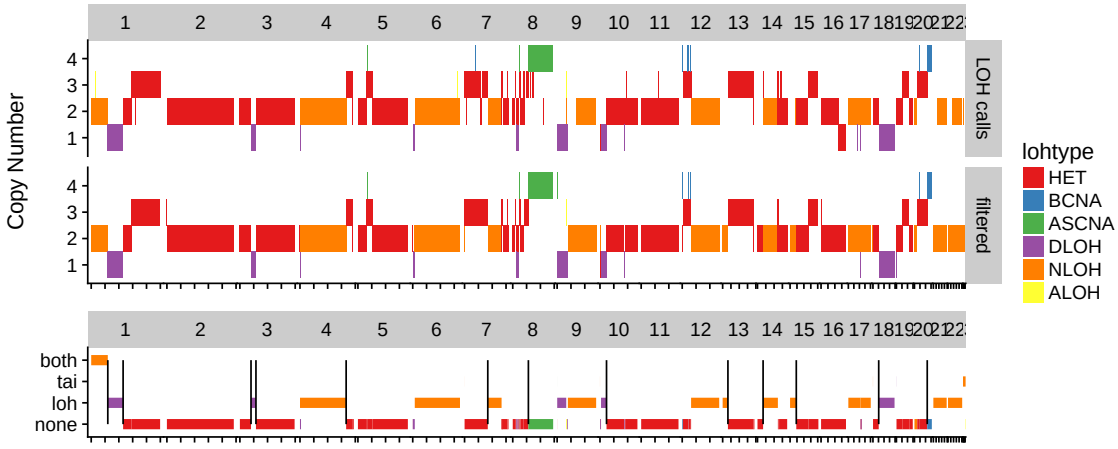

Figure 4.7: Filtering of small segments for mutation timing and HRD scores pre-processing. The purpose of filtering is to (1) fill in gaps where no calls exist with a normal, heterozygous segment, and (2) remove tiny segments adjacent to two equivalent larger segments, as these are likely to represent later or spurious events. HRD scores are composed of telomeric allelic imbalance (TAI), large loss of heterozygosity (LOH), and large-scale transitions (LST). LST junctions are shown by the vertical black lines in the lower figure.

5

THE EVOLUTION OF MUTATIONAL PROCESSES IN METASTATIC CANCER

5.1 INTRODUCTION

Although metastasis underlies up to 90% of cancer-related mortality (Seyfried and Huysentruyt, 2013), genomic instability and mutation signatures are mostly studied in primary tumours. Mutation signatures are recurrent patterns of somatic mutation frequently associated with specific mutational mechanisms such as tobacco and UV exposure (Alexandrov et al., 2016), endogenous mutagenic processes such as deamination (Roberts et al., 2013), and DNA repair deficiencies (Nik-Zainal et al., 2012; Polak et al., 2017). The analysis of mutation signatures in primary cancer sequencing data has catalogued over 30 known signatures (Alexandrov et al., 2013b; Letouzé et al., 2017; Nik-Zainal et al., 2016). Analysis of digital NGS read counts has also revealed that the activity of mutational processes changes over time (McGranahan et al., 2015). This additional insight can help characterize the ordering of mutagenic impacts throughout carcinogenesis and progression.

Recent studies suggest that certain mutation signatures may predict chemotherapy response. Hypermutating tobacco, UV radiation, and MMR have been associated with increased neoantigen burdens and sensitivity to immunotherapies in lung, gastrointestinal, urothelial, and skin cancers (Iyer et al., 2017; Lauss et al., 2017; Le et al., 2015; Rizvi et al., 2015). Signatures of HRD have been associated

with distinct cancer subtypes (Wang et al., 2017) and sensitivity to platinum-based chemotherapies (Telli et al., 2016; Zhao et al., 2017). Understanding mutational processes in metastatic cancers could uncover actionable targets and refine models of progression and drug resistance.

Also of interest in metastatic cancers are mutational spectra manifest by exposure to cytotoxic chemotherapies. For example, a specific hypermutation signature of C→T transitions pervades the genomes of *MGMT*-methylated, MMR-deficient glioblastomas treated with temozolomide alkylator chemotherapy (Alexandrov et al., 2013a; Yip et al., 2009). Despite various efforts to catalogue the mutations induced in model organisms by chemotherapy exposures (Meier et al., 2014; Segovia et al., 2015; Szikriszt et al., 2016), few matching signatures have been observed in sequenced patient samples. However, a signature recently discovered in cisplatin treated human cell lines was also found in 8 hepatocellular and 2 esophageal cancers, all with histories of cisplatin exposure (Boot et al., 2017).

To catalogue the mutational signatures of metastatic cancer, whole genome and transcriptome analysis of 571 advanced cancers was performed as part of the BC Cancer Agency Personalized Oncogenomics Project. Additionally, we performed temporal dissection of mutation signatures to map their evolutionary trajectories through cancer progression. This is the largest study to date of metastatic cancer whole genomes, revealing novel mutation signatures and chemotherapy-associated evolution of mutational processes. Our findings highlight the complex interplay of factors shaping the somatic genomes of metastatic cancers.

De novo inference of mutation signatures was successful in 12 cohorts out of 23, containing a total of 484 patients with 9,646,146 somatic SNVs. Primary site of origin varied (Table 5.1), with the largest cohorts being breast (n = 144), colorectal (n = 87), and lung (n = 68). Hierarchical clustering over signatures deciphered independently from each cohort yielded 20 independent mutation signatures. Signatures were compared against the current 30-signature COSMIC reference set using the cosine similarity metric. 11 signatures closely matched with at least one previously observed signature from the COSMIC set (Appendix Figure A.4).

Table 5.1: The number of patients belonging to each cancer type specific cohort.

Cohort	Number of Participants	Primary Site / Cancer Type
BRCA	144	Breast
COLO	87	Colorectal
LUNG	68	Lung
SARC	50	Sarcoma
MISC	45	Miscellaneous (i.e. unknown primary)
PAAD	42	Pancreatic
OV	28	Ovarian
CHOL	14	Cholangiocarcinoma
SECR	12	Secretory gland tumors
SKCM	12	Skin
LYMP	11	Lymphoma
STAD	11	Stomach
ESCA	10	Esophageal
HNSC	6	Head & neck
UVM	6	Uveal melanoma
KDNY	5	Kidney
ACC	4	Adenoid cystic
THCA	4	Thyroid

Cohort	Number of Participants	Primary Site / Cancer Type
THYM	4	Thymoma
PRAD	3	Prostate
GBM	2	Glioblastoma
HCC	2	Hepatocellular
BLCA	1	Bladder

We inferred the temporal evolution of mutation signatures to map the progression trajectory of genomic instability. Signature evolution has only been previously investigated in primary, untreated cancers (Letouzé et al., 2017; McGranahan et al., 2015). The temporal dissection of novel metastatic signatures may help distinguish markers of metastasis.

We have numbered novel metastatic signatures Signatures M1 to M9. Diagrams of all novel signatures are provided in Figure 5.1. The mean timing of mutation signatures across every cancer type is summarized in Figure 5.2. The similarity of signatures to COSMIC reference signatures is shown in Figure 5.3. Signature exposures for every cancer sample are provided in Appendix Fig. A.5.

5.2.1 *Aging-related Mutation Signatures*

Of the 9 novel signatures (Figure 5.1), some were variations on known signatures. We identified signatures 1 and 5, known to be associated with aging (Alexandrov et al., 2015a). Signature M1 was characterized by C→T mutations in CpG contexts, and matched the aging-related signature 1B previously found in many primary tumours (Alexandrov et al., 2013a) but left out of COSMIC. Aging-related signatures were not observed in skin and lung cancer cohorts (SKCM and LUNG). In skin cancers, this is likely due to the small sample size and strong presence of the UV signature. In lung cancers, signatures 3, M4, and M6 were correlated

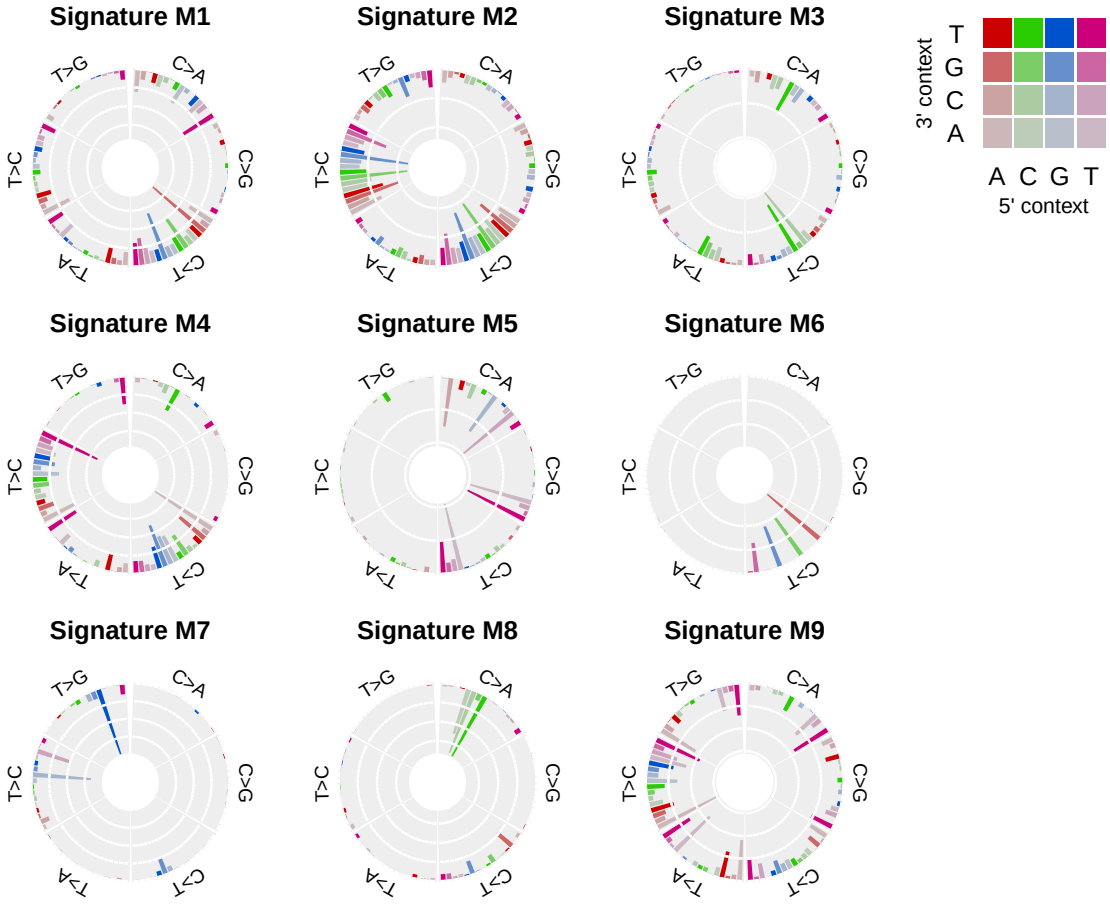

Figure 5.1: Novel metastatic signatures not catalogued in COSMIC. The analysis of recurrent mutation signatures across 12 metastatic cancer cohorts identified 9 signatures not found in the COSMIC signature catalog. These included signatures of aging (M1), cisplatin exposure (M3), mismatch repair deficiency (M4), and APOBEC deamination (M5). The etiology of the remaining signatures remains unclear.

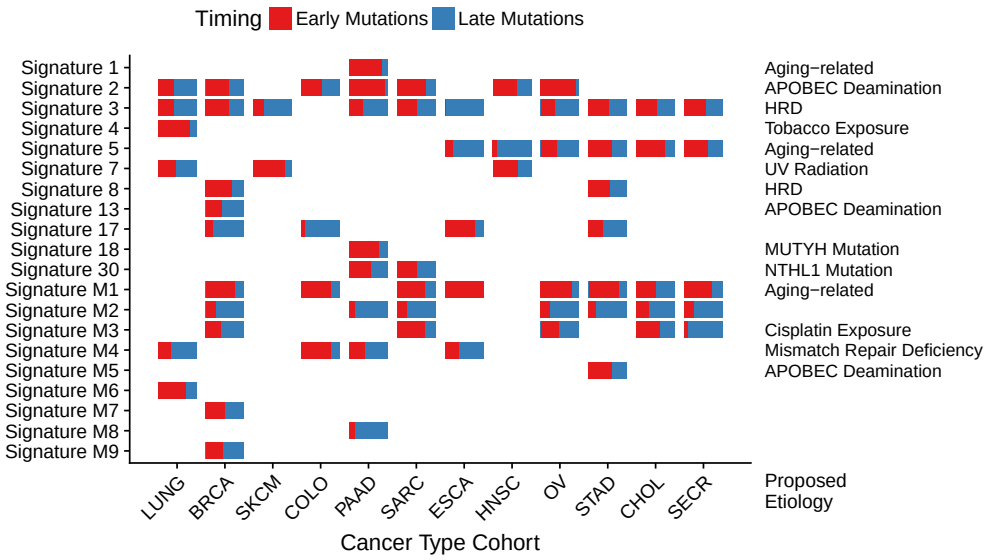

Figure 5.2: Mutation signatures and their temporal dissection in metastatic cancer. 20 *de novo* signatures deciphered from metastatic cancer whole genomes were found recurrently across tumours of 12 cancertype-specific cohorts. Temporal dissection revealed signatures biased towards early- or late-arising mutations.

and their linear combination matched signature 1, suggesting that presence of these the three signatures together obviated the need for a separate aging-related signature. A similar previous analysis suggested that signature 1B may often be composed of signature 1 together with signature 5 (Alexandrov et al., 2016).

Like in primary tumours, the aging-related signatures 1 and M1 were early-arising mutational processes across cancer types. Aging signatures were particularly elevated in cancers of rapidly proliferating epithelial cells, such as colorectal cancer, which agrees with previous findings (Alexandrov et al., 2015a). Despite previous evidence that signature 5 is also aging-related, we found that elevated signature 5 occurred primarily in late-arising mutations.

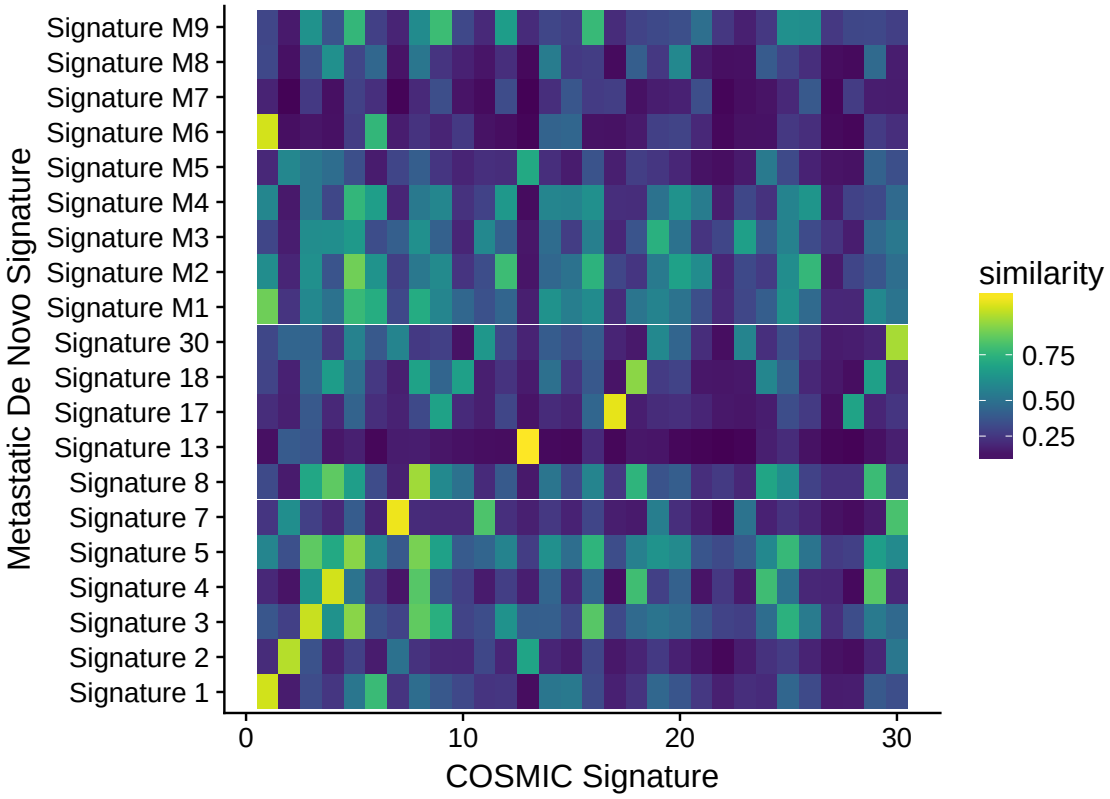

Figure 5.3: ***De novo*** **mutation signatures deciphered from metastatic cancers. Of these, 11 matched known signatures from COSMIC and 9 were novel.** To compare *de novo* signatures against those in COSMIC, The cosine similarity metric was computed pairwise between each metastatic signature and each COSMIC signature.

5.2.2 *Signatures of Exogenous Mutation*

Signature 4, associated with tobacco exposure (Alexandrov et al., 2016), was a specific indicator of smoking history as expected (Figure 5.4A). Signature 4 was early-arising in all but one of the participants with a known smoking history.

Signature 7, associated with UV radiation, was found in skin cancers and head and neck cancers as well as one hypermutated cancer of the lung. These cases had, on average, approximately 100 times the mean mutation burden. 4 out of 12 UV hypermutated cases displayed a bias towards early mutations, and 5 fit a single-population model (Figure 5.4B). Subsequent review was conducted into the UV-positive lung tumour. Clinical review showed that this patient had multiple prior skin cancers, and assessment of pathology and gene expression profiles suggested that this cancer was a sarcomatoid lung tumour which likely originated from a spindle cell carcinoma of the scalp.

5.2.3 *Signatures of Endogenous Mutation and DNA Repair Deficiency*

Mutation signatures arising from APOBEC deamination, signatures 2 and 13 (Roberts et al., 2013), were common across cancers of various types. Additionally, we identified a similar signature (M5) in stomach adenocarcinoma, which likely shares a similar mutational mechanism. These three signatures were observed across both early and late mutations. This disagrees with a previous finding that signature 2 was late-arising in bladder, head & neck, and lung cancers and signature 13 was early-arising in bladder cancers (McGranahan et al., 2015).

Signatures 3 and 8, associated with HRD (Davies et al., 2017; Nik-Zainal et al., 2012), were observed in both early and late mutations. Higher signature expo-

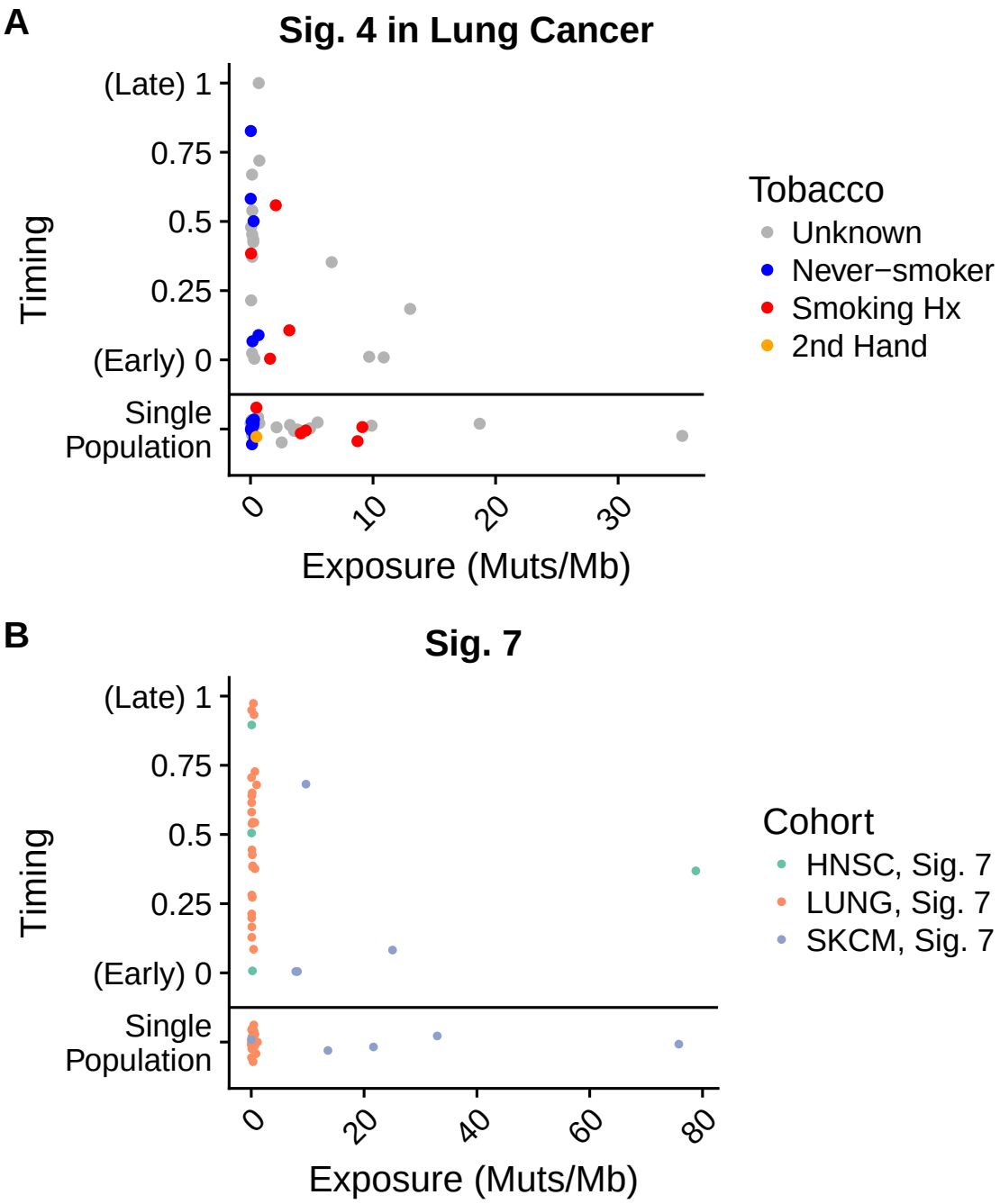

Figure 5.4: The hypermutating signatures of tobacco smoking and ultraviolet radiation. (A) Mutation signature 4, associated with tobacco exposure, was early-arising and elevated in patients with a known history of cigarette smoking. The signature was absent from never-smokers, as well as one patient who reported frequent exposure to second-hand smoke. (B) Mutation signature 7, associated with ultraviolet radiation was prevalent among skin cancers and one head & neck cancer, resulting over 10-100 times the mutation burden in exposed cases.

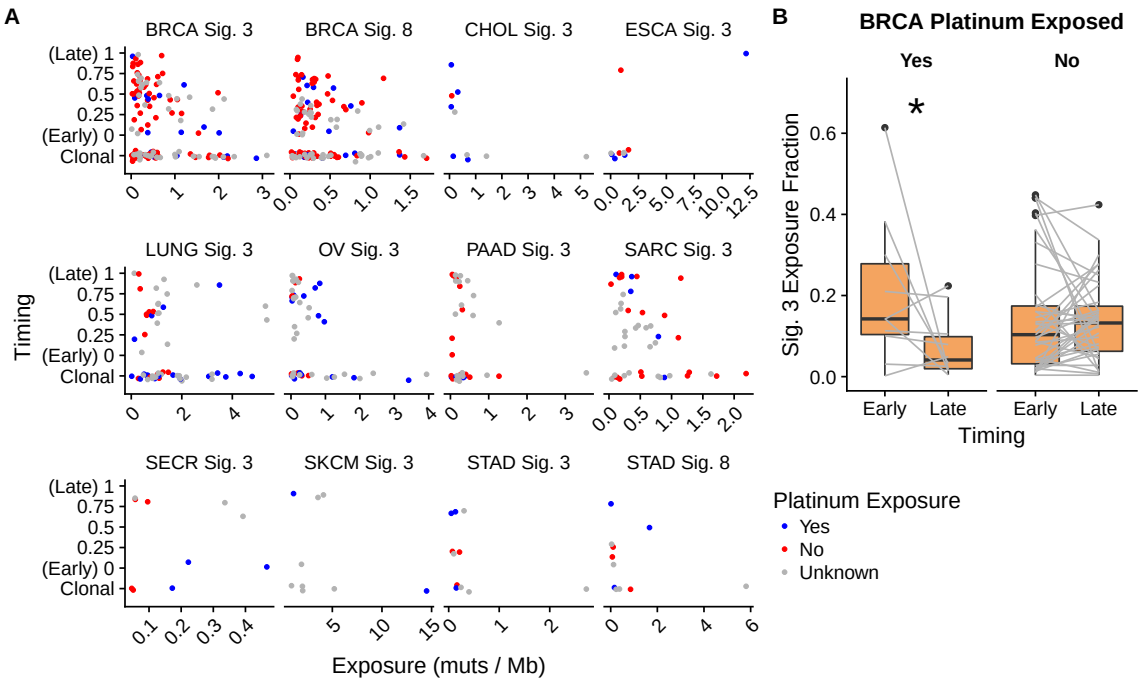

Figure 5.5: Platinum-exposure is associated with temporal evolution of homologous recombination deficiency (HRD) associated mutation signatures. (A) Mutation signatures 3 and 8, associated with HRD, were found in many cancer types. They were early-biased in high-exposure breast cancers and sarcomas. (B) In breast cancer, prior exposure to platinum-based chemotherapy was associated with a decrease in signature 3 from early to late mutations compared to non-platinum exposed tumours (p = 0.033).

sures were associated with early-arising mutations in breast cancer and sarcoma. Signature 3 was also observed in ovarian, pancreatic, and stomach cancers, as previously described (Alexandrov et al., 2015b). In breast cancer, prior exposure to platinum-based chemotherapy was associated with a decrease in the late mutation activity of signature 3 (p = 0.033, Figure 5.5). As discussed in the previous section, the elevation of signature 3 in lung cancers is likely artifactual.

Signature 30 was observed in two highly mutated cases, an undifferentiated round cell sarcoma and a pancreatic neuroendocrine tumour (PNET). Signature 30 was recently induced in cancer organoid models by the mutation of *NTHL1*

(Drost et al., 2017), a DNA glycosylase which participates in BER. In our cohort, both cases with elevated signature 30 carried deleterious mutations in the *NTHL1* gene. The *NTHL1* mutation in the PNET was a germline fusion event with the nearby genes *TRAF7* and *TSC2* and was previously described in detail (Wong et al., 2018).

Signature M4 was a driver of hypermutation in MMR-deficient tumours, and was associated with elevated MSI scores (Figure 5.6A). The signature profile was characterized by C→T and T→C transitions, and of the COSMIC signatures associated with MMR, it was closest to signature 26 (Figure 5.3). Timing bias of MMR hypermutation varied (Figure 5.6B).

Aside from hypermutated cases, MMR signatures demonstrated temporal variability across tumour types. In particular, colorectal cancers carried increased signature M4 exposure in early-biased or single-population tumours. Aside from mutation of genes responsible for MMR, a common etiology of MMR deficiency is hypermethylation of the *MLH1* promoter (Kuismanen et al., 2000; Li et al., 2013) which is associated with decreased *MLH1* expression. Although we did not directly measure methylation, the expression of *MLH1* was estimated from transcriptome data. Excluding cases with MSI or carrying germline mutations in an MMR gene, signature M4 exposure was negatively correlated with the expression of *MLH1* ($p = 0.0029$, Figure 5.6C,D).

Signature M7 was a signature of unknown etiology and accounted for 27,547 mutations (0.72% of mutation burden) in a single breast cancer sample. It has a specific profile of GCG→GTG, GTC→GCC, TTC→TCC, and GTT→GCT mutations.

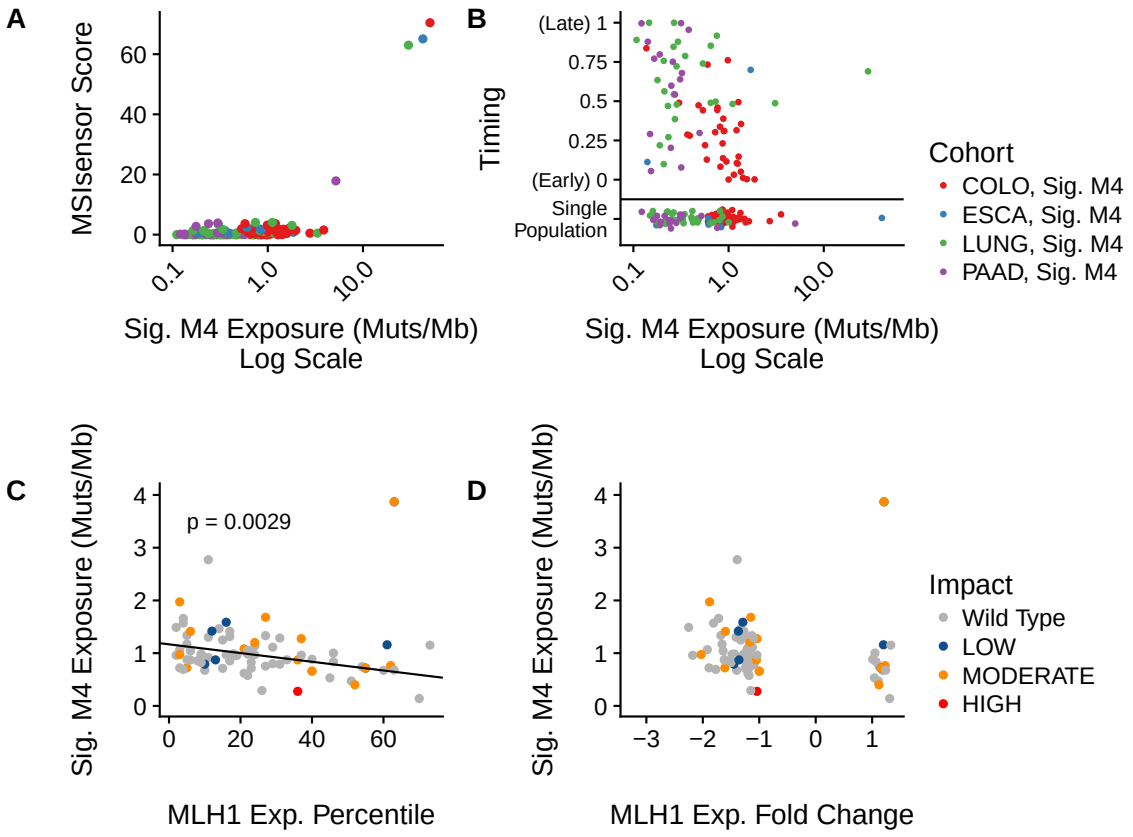

Figure 5.6: A novel signature of mismatch repair (MMR) deficiency is associated with microsatellite instability and underexpression of MLH1. (A) Mutation signature M4 was the only signature associated with microsatellite instability. (B) Temporal dissection of signature M4 revealed a distinct cluster of colorectal cancers with elevated early-arising signature exposure. (C,D) Cases with germline mutations in an MMR gene (*MSH2*, *MSH3*, *MSH6*, *PMS1*, *PMS2*, and *MLH1*) are shown according to their SNPeff-predicted mutation impact (low, moderate, or high). Excluding cases with germline MMR gene mutations, signature M4 was significantly correlated with *MLH1* expression.

5.2.4 *The Late-arising Signatures of Metastases and Chemotherapy Exposure*

Two signatures of unknown etiology, signatures 17 and M2 were biased towards late-arising mutations. This was observed across cancer types (Appendix Figure A.7) and biopsy sites (Appendix Figure A.8), suggesting that these signatures may relate to shared mutational processes occurring in progression or treatment. Signature 17 has been previously found in a small number of cancers of the liver (Letouzé et al., 2017) and breast (Nik-Zainal et al., 2016).

To explore exposures to common DNA-damaging chemotherapies as a potential etiology for mutation signatures, we examined drug-signature associations in 7 common chemotherapy agents with known DNA damaging properties. The signatures included in this analysis were those of late-arising or unknown etiology: signatures 5, 17, M2, M3, and M8.

Three drug-signature pairs displayed statistically significant differences in signature exposure between drug-exposed and non-exposed groups. Signature 17 was elevated in cancers exposed to oxaliplatin (p = 3.2e-07, median of 1685 vs. 228 mutations) and fluorouracil (p = 2.8e-06, median of 828 vs. 199 mutations), which are commonly administered in combination as part of FOLFOX regimens to treat gastrointestinal and pancreatic cancers (André et al., 2004; Conroy et al., 2011). The trend was observed both in cancers fitting single-population models and multi-population models, with the latter displaying a clear late signature bias in oxaliplatin-treated cases (Figure 5.7C).

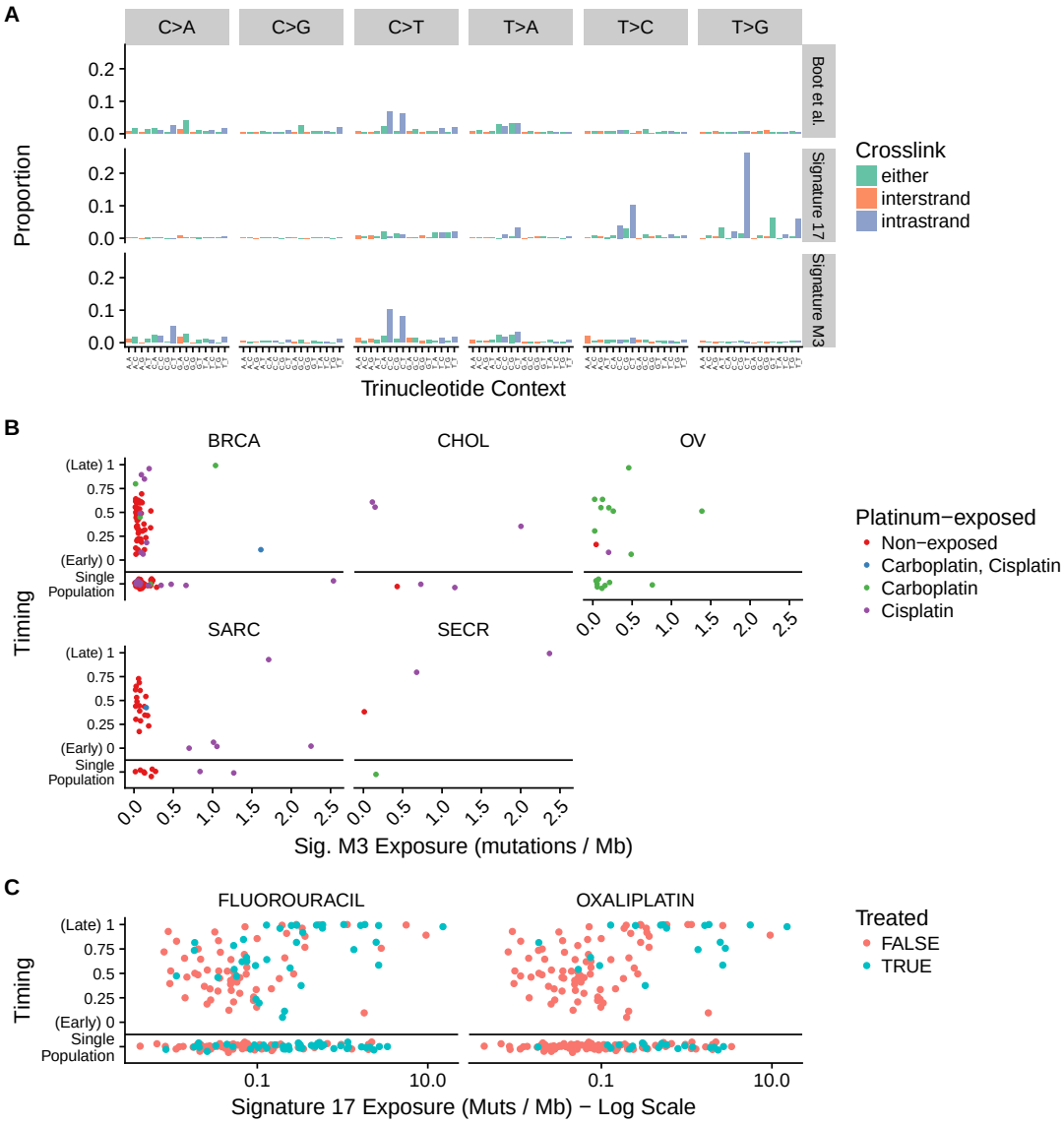

Figure 5.7: Screening of drug-signature interactions reveals statistically significant associations with cisplatin, oxaliplatin, and fluorouracil. (A) Mutation signatures associated with platinum-based chemotherapies demonstrated features consistent with intrastrand crosslink formation. Signature M3 strongly resembled a signature induced by Boot et al. (2017) in cell lines by treatment with cisplatin. Our findings provide independent discovery of this mutation signature in clinical samples. (B) Signature M3 was deciphered during *de novo* mutation signature analysis in five cancer types. Elevation of signature M3 was observed in association with 16 cancers of various types previously treated with cisplatin, as well as many carboplatin-treated ovarian cancers. (C) In addition, the elevation of signature exposure in association with oxaliplatin treatment was biased towards late-arising mutations.

5.2.5 *Signature M3 Results from Exposure to Cisplatin*

Signature M3 was elevated in patients exposed to cisplatin (p = 8e-07, median of 232 vs. 653 mutations), which is commonly administered with gemcitabine to treat cancers of the bladder, lung, breast, liver, and bile duct. Elevated signature M3 was associated with platinum exposure in four out of five osteosarcomas, as well as three other sarcomas, one breast cancer, three colangiocarcinomas, and two salivary duct carcinomas (Figure 5.7B).

A recently posted pre-print article in the bioRxiv (Boot et al., 2017) independently discovered a signature nearly identical to M3 (cosine similarity = 0.94) by treating human cell lines with cisplatin. Both signatures M3 and 17 display high rates of mutations consistent with intrastrand crosslink formation.

Whereas signature M3 was enriched for C→T mutations in CpCpY contexts, signature M8 exhibited C→A mutations in the same context. Signature M8 was found in only a single pancreatic adenocarcinoma, to which it contributed 13648 mutations, accounting for 90% of mutation burden. The only chemotherapy to which the pancreatic cancer had been previously exposed was gemcitabine. However, definitive conclusions cannot be made from this single observation.

5.3 DISCUSSION

Mutation signatures represent an emerging tumour biomarker orthogonal to existing clinical modalities. Understanding the evolution of mutation signatures from carcinogenesis to progression would undoubtedly inform research on their effective clinical translation. Here, we have performed the largest investigation of mutation signatures across metastatic cancer whole genomes. In addition to dis-

covering novel signatures in the metastatic setting, we inferred their temporal evolution, which can help guide the association of signatures with potential etiologic factors, such as exposure to DNA-damaging chemotherapies.

Our analysis also uncovered known and novel trends of temporal mutation signature bias. Signatures of aging were early arising, as previously observed; as were signatures associated with cigarette smoke and UV radiation. Additionally, temporal dissection revealed that non-hypermutating, early involvement of MMR-deficient signatures correlated with underexpression of *MLH1* specifically in colorectal cancers. Discrepancies in temporal bias compared with previous analyses (McGranahan et al., 2015), such as observed with APOBEC signatures, may result from numerous differences between the studies. McGranahan et al. (2015) analyzed WES data, which is less likely to yield stable signature solutions than WGS, but allowed for a greater number of cases per cancer type. Additionally, the previous study employed binary temporal partitioning, which we demonstrated previously can result in lower integrity temporal dissection than SignIT depending on the underlying clonal structure. Importantly, McGranahan et al. (2015) studied primary cancers, which may lack certain mutational processes specific to metastasis.

An example of signatures specific to advanced, chemotherapy-treated tumours are those arising from genotoxic chemotherapy exposures. Signature M3 was associated with cisplatin exposure across diverse cancer types. Boot et al. (2017) found an identical signature *in vitro*, as well as in 8 hepatocellular carcinomas and 2 esophageal cancers, which together with our findings provides strong evidence of a direct link to drug exposure. We further described signature M3 cisplatin treated breast cancers, cholangiocarcinomas, sarcomas, and salivary gland tumours.

In addition to signature M3, we also found an association between signature 17 and treatment with oxaliplatin and fluorouracil. In this case, the signature was late-biased, suggesting later onset of drug exposure during the course of the disease. Despite having been observed in other cancers, the etiology of signature 17 remains unknown, and no *in vitro* studies have definitively linked it to a mutagenic agent. However, mutational spectra arising from oxaliplatin and fluorouracil exposure have not yet been studied using *in vitro* methods. However, examination of signatures M3 and 17 revealed shared enrichment of mutations consistent with intrastrand crosslink formation. This form of DNA damage is typically repaired by NER (Huang and Li, 2013). Deficiencies in NER or related pathways such as translesion synthesis may explain why some platinum-treated tumours display these signatures while others do not.

The temporal analysis of signatures in association with drug exposures also enables the study of hypothetical drug resistance mechanisms. Tumours with mutations in the HR-associated genes *BRCA1* and *BRCA2* are more sensitive to platinum-based chemotherapies (Arun et al., 2011; Byrski et al., 2010; Tutt et al., 2015; Von Minckwitz et al., 2014). We showed in chapter 2 that breast cancers with signatures of HRD, including signature 3, are also associated with improved outcomes on platinum-based chemotherapy, independent of *BRCA1* and *BRCA2* mutations. We observed that prior exposure to platinum-based chemotherapy was associated with depression of signature 3 exposure in late-arising mutations. A hypothetical resistance mechanism to platinum-based chemotherapy is the reversion (or back-mutation) of *BRCA1* and *BRCA2*, restoring function to the mutant gene (Dhillon et al., 2011; Norquist et al., 2011; Swisher et al., 2008). This finding suggests that reversion of HRD, and therefore a drop in signature 3, may be asso-

ciated with platinum resistance even in the absence of detected *BRCA1* or *BRCA2* reversion mutations.

The analysis of signature timing provides evidence to suggest signature evolution in response to chemotherapy exposures. However, inferring timing from a single biopsy alone is limited in its ability to definitely attribute mutation signatures to chemotherapy exposures. The availability of sequencing data from multiple time points would be helpful in this regard, but can be costly and technically infeasible to obtain. A further limitation of this study is the lack of available clinical data regarding prior radiotherapy at the time of analysis. As a result, we could not investigate the association of radiation exposure with mutation signatures. However, a previous analysis of radiation-treated second malignancies predominantly identified signatures of SVs, insertions, and deletions rather than SNVs (Behjati et al., 2016).

This analysis also highlights continued technical challenges in the application and interpretation of mutation signatures. There were many disagreements in signature evolution between our analysis and previous work (McGranahan et al., 2015), such as discrepant timing of signatures 1B, 2, and 13. It is likely that multicollinearity between signatures, resulting in mutation signature bleed, plays a significant role in these discrepancies. For example, in chapter 3 we demonstrated that mutation signature 5 is similar to many other signatures, and thus may bleed signal with them. Additionally, the aging signature 1B (or M1) can be formulated from a linear combination of signatures 1 and 5 (Nik-Zainal and Morganella, 2017). This may reconcile why signature 5 was correlated with age-of-onset in a previous study (Alexandrov et al., 2015a) yet is biased towards late mutations in our study and others (McGranahan et al., 2015). This suggests that signature 5 itself may in-

dependently arise from both an aging-related process, and a different late-arising mechanism.

Over the past decade, mutation signature analysis has emerged as a valuable tool for the precise delineation of genomic instability and mutation in cancers. The applicability of this approach across tumour types makes it an attractive option for biomarker discovery in personalized cancer analysis and treatment. By investigating the mutation signatures of advanced cancers, we have aimed to shed light on the diverse mutagenic influences at play during invasion and metastasis.

5.4 METHODS

5.4.1 *Whole Genome Sequencing of Metastatic Cancers*

Study participants underwent tumour biopsies as part of the POG Project (Laskin et al., 2015). The study was approved by the University of British Columbia Research Ethics Board (REB# H12-00137 and H14-00681). Written informed consent, including potential publication of findings, was obtained from patients prior to genomic profiling. Whole-genome sequencing data (.bam files) have been submitted to the European Genome-Phenome Archive (EGA) (www.ebi.ac.uk/ega/home) under the study accession number EGAS00001001159.

The details of library construction, sequencing, and bioinformatics of metastatic samples have been previously described (Jones et al., 2010). Briefly, biopsy samples were embedded in OCT compound and sectioned. Pathology review was performed to select sections for sequencing. Genome libraries were constructed from tumor and peripheral blood (normal control) and sequenced using Illumina protocols on a HiSeq sequencer.

5.4.2 *Mutation Calling*

Reads were aligned to hg19 by the BWA aligner (v0.5.7) (Li and Durbin, 2009, 2010). Somatic SNVs and small insertions/deletions were processed using samtools (Li et al., 2009) and Strelka (v0.4.6.2) (Saunders et al., 2012). CNVs were called using CNASeq (v0.0.6).

5.4.3 *Mutation Signature Analysis*

Mutation signature analysis was performed on 571 cancer whole genomes from 23 cancer type cohorts. Somatic SNVs called by Strelka were categorized based on 6 variant types and 16 trinucleotide context subtypes to yield a total of 96 mutation classes. Mutation signatures were deciphered using a published framework (Alexandrov et al., 2013b) for non-negative matrix factorization (NMF) of the mutation catalog matrix into *de novo* mutation signatures and the relative exposure of each signature to each cancer genome. Fractional exposure was defined as the proportion of a genome's total mutation burden contributed by a particular signature.

Signature stability estimates were obtained by bootstrap re-sampling with 1,000 iterations (10 iterations over 100 cores). The solution which best maximizes signature stability and minimizes Frobenius reconstruction error, n_{opt} was chosen for each cohort with the formula

$$n_{\text{opt}} = \text{argmin}_n \left(\frac{R_n - \min(\mathbf{R})}{\max(\mathbf{R}) - \min(\mathbf{R})} - \frac{S_n - \min(\mathbf{S})}{\max(\mathbf{S}) - \min(\mathbf{S})} \right),$$

where S_n and R_n are the signature stability and reconstruction error values for the n-signature solution and **S** and **R** are vectors containing stability and reconstruction error values for all values of n. The model selection in each cohort is shown in Appendix Fig. A.6

Mutation signature analysis of a total of 23 cancer type cohorts was attempted (Table 5.1). All but 12 cohorts failed mutation signature analysis because of (1) too few samples, (2) too few SNVs, or (3) excessive heterogeneity in mutation signatures (as was the case in the MISC cohort). An analysis was marked failed if every sample had its own private mutation signature (meaning dimensionality reduction did not take place) or if the stability and reconstruction error estimates were poor across all attempted models.

5.4.4 *Temporal Analysis of Mutation Signatures*

Temporal analysis of mutation signatures based on mutation types and NGS variant allele counts was performed using SignIT. Cases which fit models described by greater than one subpopulation can be subject to mutation signature timing analysis. SignIT requires the annotation of SNV calls with tumour and normal copy number. Prior to annotation, CNVs from CNAseq were first corrected for ploidy using the following formula

$$C^{(T)} = \frac{(\overline{R}+1)(TP + C^{(N)}(1-T)) - C^{(N)}(1-T)}{T}.$$

Where $\overline{R}$ is the mean GC-corrected tumour-to-normal read depth ratio across the segment, T is the tumour content, and P is the ploidy. $C^{(T)}$ is the estimated absolute tumour copy number of the segment and was rounded to the nearest whole number, and $C^{(N)}$ is the normal copy number, assumed to be 2. SNVs

in regions with greater than 5 copies were filtered out, as precise copy number estimation becomes difficult.

262 cases best fit a model with one subpopulation, while 215 fit multiple temporally distinct subpopulations thus enabling signature timing. Mean early and late mutation signature exposures were computed by fitting a weighted linear model of exposure fraction versus subpopulation prevalence. The timing bias was computed as the fraction of late-arising mutations,

$$\frac{\text{late exposure}}{\text{late exposure} + \text{early exposure}},$$

and could vary from 0 for early mutation signatures to 1 for late mutation signatures. To generate Figure 5.2, results were grouped by cohort and signature, and the total number of early- and late-arising mutations across all samples was computed.

5.4.5 *Microsatellite Instability Scores*

Microsatellite instability was quantified from paired tumour-normal whole genome sequencing using the previously described tool, MSIsensor (Niu et al., 2014). Microsatellites and homopolymers were identified in the hg19 reference genome using default parameters (homopolymers ≥ 5, microsatellites repeat unit length ≤ 5) followed by subsampling to 50,000 sites randomly distributed across the genome. Somatic status of sites with sufficient coverage (20 spanning reads in both normal and tumour samples) was determined using default settings (median 1369 sites tested per sample). The percentage of tested sites that were unstable in the tumour sample compared to the normal sample was used as a measure of MSI. General classification of MSI status were as follows: microsatellite stable ($< 10\%$),

MSI-low (10-30%) and MSI-high (> 30% of somatic sites unstable). Four out of six cases that had conventional immunohistochemical testing for MMR deficiency and an MSI score of $\geq$ 10% (MSI-low or MSI-high) also tested positive for MMR deficiency by immunohistochemistry, supporting the accuracy of MSIsensor analysis.

5.4.6 *Quantifying Gene Expression from Transcriptomes*

Transcriptomes were repositioned using JAGuaR (version 2.0.3) (Butterfield et al., 2014). Differential expression analysis was performed by comparing reads per kilobase of transcript per million mapped reads (RPKM) values against a compendium of 16 normal tissues from the Illumina BodyMap 2.0 project (available from ArrayExpress, queryID: E-MTAB-513) as previously described (Jones et al., 2010). For every sample, the expression percentile of each gene was computed against expression data for that gene across all samples from TCGA.

5.4.7 *Retrospective Clinical Review*

A retrospective review of chemotherapy exposures including treatment start and end dates was performed, aided by a provincial clinical cancer database (Wu et al., 2013). Additionally, relevant patient demographics such as age at diagnosis and tobacco smoking history were obtained.

5.4.8 *Analysis of Drug-Signature Associations*

Chemotherapy exposure data were available in 408 out of 484 patients, who altogether had been exposed to 119 distinct chemotherapy drug types. Among the 20 most commonly used chemotherapy agents, 7 with known DNA damaging qualities were chosen for investigation: cyclophosphamide, cisplatin, fluorouracil, doxorubicin, capecitabine, carboplatin, and oxaliplatin. Late-arising signatures and those of unknown etiology (5, 17, M2, M3, M8) were each assessed for differences in exposure between therapy-exposed and non-exposed patients by the Wilcoxon signed-rank test. Resulting p-values were adjusted for multiple hypothesis testing using the Bonferroni-Holm method.

6

CONCLUSION

6.1 SUMMARY OF MAJOR FINDINGS

This thesis had three major aims. First, to assess the association between DNA damage repair mutation signatures and response to DNA-damaging chemotherapy. Second, to enable accurate individualized mutation signature decomposition and temporal dissection. Last, to characterize the evolution of mutation signatures in metastatic cancers.

To assess the clinical actionability of mutation signatures, we studied HR as a model system. This allowed us to build upon knowledge that cancers with *BRCA1* and *BRCA2* mutations are more sensitive to platinum-based chemotherapy. Moreover, past efforts to quantify genomic scars as predictors of *BRCA1*/*BRCA2* mutation (Timms et al., 2014) and platinum response (Telli et al., 2016) already demonstrated promising findings in primary breast cancers, but did not replicate in the metastatic setting (Tutt et al., 2015). In chapter 2, we used WGS of advanced breast cancers to demonstrate that response to platinum-based chemotherapies was associated with mutation signatures of HRD. The novel aspect of this work was the integration of multiple independent signatures of various mutation types, whereas previous studies had relied only upon signatures of CNV and LOH. Our findings suggest that genome-scale analysis of mutational processes can more accurately inform the clinical management of cancers with HRD.

The use of mutation signatures in clinical guidance calls for accurate individualized decomposition of signature exposures. However, the majority of mutation signature methods perform *de novo* inference of signatures from large cancer datasets. As part of our breast cancer study, we found that signature inference by NNLS using the 30 signature reference set from COSMIC yielded accurate n-of-1 HRD predictions. We built upon this in chapter 3 by proposing SignIT, a hierarchical Bayesian model which performs accurate, robust, and interpretable individualized inference of mutation signatures. Using simulated data and WES mutation calls from TCGA, we demonstrated SignIT's superiority over alternative approaches.

A challenge in the interpretation of mutation signatures or genomic scars is that they represent the aggregate mutational history of a tumour rather than the relevant mutational processes still active at the time of treatment. Serial sequencing at multiple timepoints could map out the mutational trajectory, but would be inconvenient, costly, and impose additional medical procedures. Instead, we extended SignIT to integrate genomic read depth data in order to infer the presence of temporally distinct mutational subpopulations. This enables the tracking of mutation signatures from a single sequencing timepoint. Using data from multiply sequenced metastatic tumours, we demonstrated that early prevalent subpopulations demonstrated signatures similar to those of the primary, whereas later-arising subpopulations diverged. By directly inferring tumour subpopulations rather than partitioning mutations using hard assumptions, SignIT improved upon previous attempts at reconstructing mutation signature timing.

In chapter 4, we demonstrated clinical implications of mutation signature timing in a pancreatic adenocarcinoma with a germline *BRCA1* variant but paradoxically low SV burden and HRD signature exposure. We synthesized the timing

of chr17 LOH and of the HRD signature using both computational techniques and comparison to the primary. The findings suggested later than expected somatic LOH of *BRCA1* and onset of HRD, which may reconcile the *BRCA1* variant, the low HRD signature, and the cancer's exceptional response to FOLFIRINOX therapy.

The development of SignIT allowed us to address the final aim. To date, mutational processes have been studied almost exclusively in primary, treatment-naive cancers. Successful application of mutation signatures to the analysis of advanced cancers will require an understanding of the unique forces shaping somatic mutation in metastases. In chapter 5, we deciphered mutation signatures from the whole genomes of nearly 500 metastatic, treated cancers. This uncovered novel signatures, one of which (signature M3) has been shown to arise *in vitro* from cells treated with cisplatin (Boot et al., 2017). Additionally, the HRD-associated signature 3 was suppressed in the late-arising mutations of cancers previously exposed to platinum-based chemotherapy, which hints at potential resistance mechanisms. These findings confirm that metastatic cancers are characterized by shifts in mutagenesis borne of selective pressures and exposure to DNA damaging therapies.

6.2 THE CLINICAL IMPLICATIONS OF GENOMIC INSTABILITY

Mutation signatures blur the line between the genotype and phenotype. While mutagenesis shapes the cancer's genotype, it also reveals the integrity of its DNA repair processes. By leveraging the whole genome as a functional assay, mutation signatures permit biologists to directly view the effects of DNA repair deficiencies whether or not their root cause is identifiable.

DNA repair mechanisms, such as HR, are complex and not fully understood. For example, not every HRD tumour is explained by mutation or hypermethylation of *BRCA1/BRCA2*. Whereas *BRCA1/BRCA2* underlie 5-10% of breast cancer, as many as 22% of primary breast cancers carry signatures of HRD (Davies et al., 2017). Furthermore, of the observed mutations, many are VUS, without clear evidence linking to breast cancer risk. Polak et al. (2017) recently demonstrated that mutation signatures can help to delineate the functional relevance of mutations in DNA repair genes. This has ramifications both for guidance of treatment and for the screening of hereditary cancer risk. The mutation signature of a targetable pathway can serve as an indicator of that pathway's function. However, even in objective responders to platinum-based chemotherapy, recurrence rates are high (Dent et al., 2007; Nagourney et al., 2000; Sirohi et al., 2008). Follow-up analysis by WGS offers the potential to probe the origins of acquired drug resistance (Jones et al., 2010). For instance, HRD tumours have been observed to acquire secondary mutations which restore the reading frame of *BRCA1/BRCA2* (Patch et al., 2015; Swisher et al., 2008). Here, the temporal dissection of mutation signatures may come into play. We found that breast cancers with prior exposure to platinum-based chemotherapy exhibited a decrease in HRD signature activity in late mutations. Again, the analysis of mutational processes may obviate the need to identify the specific somatic event giving rise to resistance, so long as a late shift in the signature is observed.

In contrast to HRD suppression, the pancreatic case study in chapter 4 demonstrated the principle of late-onset HRD. As a common source of cancer susceptibility, HRD is thought to be an early cancer driver. However, the analysis of mutation signature timing in primary cancers by McGranahan et al. (2015) and our own temporal analysis in metastatic cancers (chapter 5) suggests that HRD

may require time to accrue a notable mutation burden. In the clinical setting, this could result in the discounting of a low but active HRD signature and the missed opportunity for targeted treatment. Therefore, it is preferable to delineate currently active processes from historically active ones in order to appreciate the full timeline of actionable mutagenesis.

6.3 THE MUTATIONAL PROCESSES OF METASTATIC CANCERS

Studying the association of mutation signatures with drug exposures was made possible by the availability of clinical treatment data. This infrastructure allowed the screening and joint modeling of drug-signature interactions. The association between signature M3 and cisplatin exposure emerged directly from this analysis. Aside from the temozolomide signature, this is the first verified signature of a chemotherapy exposure independently discovered in patient samples by NMF.

Signature M3 appears to be a specific, but not sensitive marker of platinum exposure as many platinum-exposed samples do not display the signature. It is not yet clear whether its accrual depends on the loss of specific DNA repair processes as with signature 11 in temozolomide treatment. It is also unclear whether signature M3 could be a marker of acquired drug sensitivity or resistance. The signature's unbiased temporal occurrence suggests that it may be a remnant of DNA damage retained through the proliferation of a resistant clone. The lack of late-bias also suggests that, in a small fraction of tumours, cisplatin treatment may induce as many as 2 mutations per megabase and could shape early tumour cells which seed to metastatic sites. This finding calls for further study the clinical consequences of cisplatin-dependent mutagenesis, especially in metastatic cancer types where signature M3 is frequently observed.

The association of signature 17 with oxaliplatin and fluorouracil exposure stands in contrast to signature M3 because of its bias towards late-arising mutations. Therefore, we posit that signature 17 may result from direct drug exposure of the metastatic cancer rather than the expansion of a resistant clone. However, independent validation of this signature in fluorouracil/oxaliplatin treated cells is necessary, as no studies have yet examined the mutagenic profile of oxaliplatin.

Prior to the successful induction of signature M3 in human cell lines by cisplatin treatment (Boot et al., 2017), mutagenic profiles were also accumulated *in vitro* via cisplatin treatment of chicken DT40 lymphoblast cells (Szikriszt et al., 2016) and *caenorhabditis elegans* (Meier et al., 2014). However, these signatures have not subsequently been identified in sequenced human samples. The distinction suggests that signatures generated experimentally in model organisms may not be as applicable to human cancers as those generated in human cells. This further implies that mutation signatures stem from a delicate interplay between mutagens and repair pathways, and that subtle variation between species can dramatically alter mutagenesis.

6.4 LIMITATIONS

The study of platinum response in breast cancer was part of a larger study with the goal to guide individual treatment decisions using WGTA. The population studied was selected for inclusion, and may not reflect the full population of metastatic breast cancers. Additionally, this project occurred in two phases, the first of which included the first 100 cases (Laskin et al., 2015). Sequencing protocols differed slightly between the two phases, which could introduce batch effects.

However, bioinformatics pipelines are standardized across samples and earlier cases are occasionally re-run with updated analysis tools.

The clinical data on treatment durations were derived from a provincial database of pharmacy records at cancer centres (Wu et al., 2013). Some treatments are missing from this data, especially those delivered in different jurisdictions or health authorities, or which were part of certain clinical trials. However, data on standard treatments such as the platinum-based chemotherapies discussed here, were near complete. Additionally, for the breast cancers studied in chapter 2, missing treatment dates were reintroduced during retrospective clinical review.

Another major limitation is the lack of standard timelines for the assessment of treatment response. This precluded the use of standard objective response criteria such as RECIST (Eisenhauer et al., 2009) and necessitated the creation of a custom response scale. Instead, treatment duration was available as a secondary measure of patient outcome, and was found to correlate with rated treatment responses.

A limitation of SignIT is the assumption that CNVs are clonal, meaning they are identical across all subclones of a tumour. Some methods such as TITAN (Ha et al., 2014) can estimate the cellular prevalence of CNVs. Future iterations of SignIT could provide an option to include such estimates within the model to more precisely adjust the expected variant allele counts.

Lastly, in the temporal analysis of mutation signatures by SignIT, slightly more than half of cancers fit a single-population model. This suggests that there was insufficient clonal diversity and/or too few SNVs in regions of copy number variation to accurately estimate the timing of signatures. Deeper sequencing may be necessary to uncover identifiable tumour subpopulations in these cases. It is not certain, however, whether cancers which tend to fit a single-population model also systematically differ in the timing of mutational processes. If so, then the removal

of these cases from analysis could confound the findings in chapter 5 and harm generalizability.

6.5 A ROLE FOR MUTATION SIGNATURES IN PRECISION ONCOLOGY

Since 2014, the POG project has incorporated mutation signatures in the treatment-focused personalized analysis of cancer genomes. HRD scores were introduced in late 2016. Within the first 139 breast cancers, over 25% of cases were deemed to have an actionable target based on mutation signature, HRD score, or mutation burden. A major limitation is that distinct actionable targets from mutation signatures remain limited to HRD (for platinum/PARP inhibitors) and hypermutation (for immunotherapy). In addition to POG, other personalized cancer sequencing initiatives have also incorporated some element of mutation signature analysis (Tuxen et al., 2016; Zehir et al., 2017).

Cost remains a barrier to the integration of mutation signatures at scale into clinical care. The accuracy of mutation signature decomposition improves with increased sampling of mutations. For example, the signature 3 exposure of a typical metastatic breast cancer varied from 0 to 10,000 mutations, and total mutation burdens from 0 to 60,000. WES yields approximately 100 times fewer mutations, and targeted panels fewer still. When partitioning mutations across 96 classes, it can thus be challenging to identify clinically relevant signatures with any confidence. Using a large targeted panel, Zehir et al. (2017) could quantify signatures only of hypermutating processes (POLE, MMR, tobacco, UV, and temozolomide), even with 10,000 samples. Worse still, SV signatures are infeasible at the scale of exomes. Low-depth sequencing, such as employed by Nik-Zainal et al. (2016), is a potential cost-saving solution for mutation signature analysis. However, this

would render temporal dissection of mutational processes challenging or impossible without follow-up targeted sequencing of mutated loci.

The integration of mutation signatures into clinical practice will likely depend on the feasibility of WGS itself. The value added by WGS rises with continued characterization of the cancer genome and the proliferation of datasets supporting the contextualization of clinically relevant findings. Meanwhile, sequencing costs continue to fall, but the cost of genome analysis has not followed suit. Weymann et al. (2017) showed that the cost of WGTA within POG was $34,886 per patient from 2012 to 2015 with a downwards trajectory driven primarily by decreasing sequencing cost. The construction of automated analysis pipelines to surface known and hypothetical actionable targets needs to be a continued focus to realize the goal of scalable precision oncogenomics.

6.6 FUTURE RESEARCH DIRECTIONS

In the meantime, there remain many opportunities for research into the actionability of DNA repair deficiency. Within the field of HRD, a well-designed prospective trial leveraging HRDetect or a similar aggregate measure of HRD is necessary. Such a trial could compare the response to cisplatin/gemcitabine treatment between HRD and non-HRD breast cancers. Moreover, the recent approval of olaparib for use in germline *BRCA1*/*BRCA2* mutated breast cancers (Center for Drug Evaluation and Research, 2018) raises the possibility of a PARP inhibitor trial. Also promising is the recent development of a DNA G-quadruplex stabilizer (Xu et al., 2017), which is now in phase I clinical trial (Canadian trial NCT02719977). G-quadruplex structures are sites of frequent DNA damage which requires repair

by HR. Inducing or stabilizing G-quadruplex structures in the context of HRD may facilitate tumour-targeted cell death.

Another priority is better characterization of HRD's actionability in cancer types other than breast cancer. HRD has also been observed in ovarian, pancreatic, and gastric cancers (Alexandrov et al., 2015b; Davies et al., 2017), as well as osteosarcoma (Kovac et al., 2015), wherein the PARP inhibitor talazoparib has been effective *in vitro* (Engert et al., 2017).

Regarding platinum resistance, we hypothesized in chapter 5 that HR restoration could indicate acquired resistance to platinum-based chemotherapy. This finding may eventually inform development of biomarkers to monitor for drug resistance, similar to how HRD onset could be a marker of drug sensitivity. However, there has not yet been evidence showing that *BRCA1*/*BRCA2* reversion mutations lead to a drop in HRD mutation signature activity. Unfortunately, no *BRCA1*/*BRCA2* reversions have been confirmed in our metastatic cancer cohort because paired primary cancers were sequenced in only select cases. However, a WGS study of 92 chemoresistant ovarian cancers confirmed five cases of *BRCA1*/*BRCA2* reversion. Data from this study are available via European Genome-Phenome Archive (EGA), and could be used to assess the evolution of HRD mutation signatures in association with *BRCA1*/*BRCA2* reversion. Another promising approach is the sequencing of circulating tumour DNA to monitor for reversion mutations in *BRCA1*/*BRCA2* (Christie et al., 2017; Mayor et al., 2017; Weigelt et al., 2017). Exome-scale capture of circulating tumour DNA could eventually enable non-invasive monitoring of mutation signatures, which would dramatically improve feasibility of clinical application.

While this thesis has focused on the timing of SNV mutation signatures, the timing of SV signatures remains a challenge. Graph-based approaches have been

proposed to reconstruct genome-wide rearrangement histories (Greenman et al., 2012), but outstanding challenges exist relating both to analytical difficulties and the accuracy of CNV and SV callers (Maciejowski and Imielinski, 2016). Likewise, the timing of CNV-based signatures such as the HRD score could be achieved by methods such as those described by Purdom et al. (2013) and applied in chapter 4, but these in turn depend upon knowledge of rearrangement history. Despite these challenges, early attempts to chart the timing of common cancer driver events have already been made (Gerstung et al., 2017).

6.7 Looking forward: biomarker discovery in the era of genomic data

If nothing else, I hope that this thesis has conveyed the importance and complexity of relating a novel WGS biomarker, the mutation signature, to its therapeutic potential. There are fundamental biological questions to consider, such as the confluence of factors which shape and alter mutation signatures, and whether mutational processes evolve over time. There are also technical details to unmask, such as signature bleed. Most importantly, there is the feedback loop which enables the hypothesis, discovery, and follow-up of relevant clinical associations.

For some forward-thinking jurisdictions, collating the complete genomic information of tumours coupled with extensive clinical information will provide an unprecedented research platform to understand the mechanisms underlying therapeutic response, acquired resistance, and failure. Furthermore, the serial application of WGTA, undertaken many times during the course of the disease could provide a real-time view of cancer progression and treatment response. This feed-

back loop will be invaluable for the study of cancers where the goal is to improve disease stratification and therapeutic intervention.

I have been privileged to glimpse the earliest of genomics applications aimed at guiding cancer treatment decision-making. I hope that, in the coming era of genomic medicine, these efforts expand and continue to generate insights, supported by the clinical infrastructure necessary to render them actionable.

BIBLIOGRAPHY

Afghahi, A., Timms, K.M., Vinayak, S., Jensen, K.C., Kurian, A.W., Carlson, R.W., Chang, P.-J., Schackmann, E., Hartman, A.-R., Ford, J.M., et al. (2017). Tumor BRCA1 Reversion Mutation Arising during Neoadjuvant Platinum-Based Chemotherapy in Triple-Negative Breast Cancer Is Associated with Therapy Resistance. Clinical Cancer Research *23*.

Alaei-Mahabadi, B., Bhadury, J., Karlsson, J.W., Nilsson, J.A., and Larsson, E. (2017). Global analysis of somatic structural genomic alterations and their impact on gene expression in diverse human cancers. PNAS *113*, 13768–13773.

Albertsen, H., Smith, S.A., Mazoyer, S., Fujimoto, E., Stevens, J., Williams, B., Rodriguez, P., Cropp, C.S., Slijepcevic, P., Carlson, M., et al. (1994). A physical map and candidate genes in the BRCA1 region on chromosome 17q12–21. Nature Genetics *7*, 472–479.

Alexandrov, L.B., and Stratton, M.R. (2014). Mutational signatures: The patterns of somatic mutations hidden in cancer genomes.

Alexandrov, L., Jones, P., Wedge, D., Sale, J., Campbell, P., Nik-Zainal, S., and Stratton, M. (2015a). Clock-like mutational processes in human somatic cells. Nat Commun.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S. a J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013a). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013b). Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Reports *3*, 246–259.

Alexandrov, L.B., Nik-Zainal, S., Siu, H.C., Leung, S.Y., and Stratton, M.R. (2015b). A mutational signature in gastric cancer suggests therapeutic strategies. Nature Communications *6*.

Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutation signatures associated with tobacco smoking in human cancer. Science *354*, 618–622.

André, T., Boni, C., Mounedji-Boudiaf, L., Navarro, M., Tabernero, J., Hickish, T., Topham, C., Zaninelli, M., Clingan, P., Bridgewater, J., et al. (2004). Oxaliplatin, Fluorouracil, and Leucovorin as Adjuvant Treatment for Colon Cancer. New England Journal of Medicine *350*, 2343–2351.

Arun, B., Bayraktar, S., Liu, D.D., Barrera, A.M.G., Atchley, D., Pusztai, L., Litton, J.K., Valero, V., Meric-Bernstam, F., and Hortobagyi, G.N. (2011). Response to neoadjuvant systemic therapy for breast cancer in BRCA mutation carriers and noncarriers: a single-institution experience. Journal of Clinical Oncology *29*, 3739–3746.

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated Evolution of Prostate Cancer Genomes. Cell *153*, 666–677.

Baez-Ortega, A., and Gori, K. (2017). Computational approaches for discovery of mutational signatures in cancer. Briefings in Bioinformatics *69*, 26–33.

Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L., et al. (2012).

Sequence analysis of mutations and translocations across breast cancer subtypes. Nature *486*, 405–409.

Behjati, S., Gundem, G., Wedge, D., Roberts, N., Tarpey, P., Cooke, S., Van, L., Alexandrov, L., Ramakrishna, M., Davies, H., et al. (2016). Mutational signatures of ionizing radiation in second malignancies. Nature Communications 12605.

Betancourt, M. (2017). Identifying Bayesian Mixture Models.

Birkbak, N.F., Wang, Z.C., Kim, J...Y., Eklund, A.C., Li, Q., Tian, R., Bowman-Colin, C., Li, Y., Greene-Colozzi, A., Iglehard, J.D., et al. (2012). Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. Cancer Discovery *2*.

Boot, A., Huang, M., Ng, A.W.T., Kawakami, Y., Chayama, K., Teh, B.T., Nakagawa, H., and Rozen, S.G. (2017). In-depth characterization of the cisplatin mutational signature in a human cell line and in esophageal and liver tumors. bioRxiv 189233.

Bosdet, I.E., Docking, T.R., Butterfield, Y.S., Mungall, A.J., Zeng, T., Coope, R.J., Yorida, E., Chow, K., Bala, M., Young, S.S., et al. (2013). A Clinically Validated Diagnostic Second-Generation Sequencing Assay for Detection of Hereditary BRCA1 and BRCA2 Mutations. The Journal of Molecular Diagnostics *15*, 796–809.

Bose, P., Pleasance, E.D., Jones, M., Shen, Y., Ch'ng, C., Reisle, C., Schein, J.E., Mungall, A.J., Moore, R., and Ma, Y. (2015). Integrative genomic analysis of ghost cell odontogenic carcinoma. Oral Oncology *51*, e71–e75.

Bruin, E.C. de, McGranahan, N., Mitter, R., Salm, M., Wedge, D.C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A.J., et al. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science (New York, N.Y.) *346*, 251–256.

Butterfield, Y.S., Kreitzman, M., Thiessen, N., Corbett, R.D., Li, Y., Pang, J., Ma, Y.P., Jones, S.J.M., and Birol, İ. (2014). JAGuaR: junction alignments to genome for RNA-seq reads. PloS One *9*, e102398.

Byrski, T., Gronwald, J., Huzarski, T., Grzybowska, E., Budryk, M., Stawicka, M., Mierzwa, T., Szwiec, M., Wiśniowski, R., and Siolek, M. (2010). Pathologic complete response rates in young women with BRCA1-positive breast cancers after neoadjuvant chemotherapy. Journal of Clinical Oncology *28*, 375–379.

Canadian Cancer Society (2014). BRCA gene mutations.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. Journal of Statistical Software; Vol 1, Issue 1 (2017).

Center for Drug Evaluation and Research (2018). Approved Drugs - FDA approves olaparib for germline BRCA-mutated metastatic breast cancer.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery *2*, 401–404.

Cerbinskaite, A., Mukhopadhyay, A., Plummer, E.R., Curtin, N.J., and Edmondson, R.J. (2012). Defective homologous recombination in human cancers. Cancer Treatment Reviews *38*, 89–100.

Chaffer, C.L., and Weinberg, R.A. (2011). A Perspective on Cancer Cell Metastasis. Science *331*, 1559–1564.

Chang, H.H.Y., Pannunzio, N.R., Adachi, N., and Lieber, M.R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nature Reviews Molecular Cell Biology *18*, 495–506.

Chia, S.K., Bramwell, V.H., Tu, D., Shepherd, L.E., Jiang, S., Vickery, T., Mardis, E., Leung, S., Ung, K., Pritchard, K.I., et al. (2012). A 50-Gene Intrinsic Subtype Classifier for Prognosis and Prediction of Benefit from Adjuvant Tamoxifen. Clinical Cancer Research 1–8.

Chmielecki, J., Crago, A.M., Rosenberg, M., O'Connor, R., Walker, S.R., Ambrogio, L., Auclair, D., McKenna, A., Heinrich, M.C., Frank, D.A., et al. (2013). Whole-exome sequencing identifies a recurrent NAB2-STAT6 fusion in solitary fibrous tumors. Nature Genetics *45*, 131–132.

Chong, L.C., Twa, D.D.W., Mottok, A., Ben-Neriah, S., Woolcock, B.W., Zhao, Y., Savage, K.J., Marra, M.A., Scott, D.W., Gascoyne, R.D., et al. (2016). Comprehensive characterization of programmed death ligand structural rearrangements in B-cell non-Hodgkin lymphomas. Blood *128*, 1206–1213.

Chong, Z., Ruan, J., Gao, M., Zhou, W., Chen, T., Fan, X., Ding, L., Lee, A.Y., Boutros, P., Chen, J., et al. (2017). novoBreak: local assembly for breakpoint detection in cancer genomes. Nature Methods *14*, 65–67.

Christie, E.L., Fereday, S., Doig, K., Pattnaik, S., Dawson, S.-J., and Bowtell, D.D. (2017). Reversion of BRCA1/2 Germline Mutations Detected in Circulating Tumor DNA From Patients With High-Grade Serous Ovarian Cancer. Journal of Clinical Oncology *35*, 1274–1280.

Conroy, T., Desseigne, F., Ychou, M., Bouché, O., Guimbaud, R., Bécouarn, Y., Adenis, A., Raoul, J.-L., Gourgou-Bourgade, S., Fouchardière, C. de la, et al. (2011). FOLFIRINOX versus Gemcitabine for Metastatic Pancreatic Cancer. New England Journal of Medicine *364*, 1817–1825.

Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., Zou, X., Ramakrishna, M., Martin, S., Boyault, S., and Sieuwerts, A.M. (2017). HRDetect is a pre-

dictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nature Medicine.

De Leeneer, K., Hellemans, J., De Schrijver, J., Baetens, M., Poppe, B., Van Criekinge, W., De Paepe, A., Coucke, P., and Claes, K. (2011). Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations. Human Mutation *32*, 335–344.

Dent, R., Trudeau, M., Pritchard, K.I., Hanna, W.M., Kahn, H.K., Sawka, C.A., Lickley, L.A., Rawlinson, E., Sun, P., and Narod, S.A. (2007). Triple-Negative Breast Cancer: Clinical Features and Patterns of Recurrence. Clinical Cancer Research *13*, 4429–4434.

Dhillon, K.K., Swisher, E.M., and Taniguchi, T. (2011). Secondary mutations of BRCA1/2 and drug resistance. Cancer Science *102*, 663–669.

Ding, J., Bashashati, A., Roth, A., Oloumi, A., Tse, K., Zeng, T., Haffari, G., Hirst, M., Marra, M.A., Condon, A., et al. (2012). Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics *28*, 167–175.

Drost, J., Boxtel, R. van, Blokzijl, F., Mizutani, T., Sasaki, N., Sasselli, V., Ligt, J. de, Behjati, S., Grolleman, J.E., Wezel, T. van, et al. (2017). Use of CRISPR-modified human stem cell organoids to study the origin of mutational signatures in cancer. Science (New York, N.Y.) *358*, 234–238.

Eisenhauer, E., Therasse, P., Bogaerts, J., Schwartz, L., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). European Journal of Cancer *45*, 228–247.

Ellis, M.J., Ding, L., Shen, D., Luo, J., Suman, V.J., Wallis, J.W., Van Tine, B.A., Hoog, J., Goiffon, R.J., Goldstein, T.C., et al. (2012). Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature *486*, 353–360.

Engert, F., Kovac, M., Baumhoer, D., Nathrath, M., and Fulda, S. (2017). Osteosarcoma cells with genetic signatures of BRCAness are susceptible to the PARP inhibitor talazoparib alone or in combination with chemotherapeutics. Oncotarget *8*, 48794–48806.

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., et al. (2005). Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature *434*, 917–921.

Fischer, A., Illingworth, C.J., Campbell, P.J., and Mustonen, V. (2013). EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biology *14*, R39.

Fischer, A., Vázquez-García, I., Illingworth, C.J.R., and Mustonen, V. (2014). High-definition reconstruction of clonal composition in cancer. Cell Reports *7*, 1740–1752.

Foley, S.B., Rios, J.J., Mgbemena, V.E., Robinson, L.S., Hampel, H.L., Toland, A.E., Durham, L., and Ross, T.S. (2015). Use of Whole Genome Sequencing for Diagnosis and Discovery in the Cancer Genetics Clinic. EBioMedicine *2*, 74–81.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Science Signaling *6*, pl1–pl1.

Gehring, J.S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics btv408.

Gelmon, K.A., Tischkowitz, M., Mackay, H., Swenerton, K., Robidoux, A., Tonkin, K., Hirte, H., Huntsman, D., Clemons, M., Gilks, B., et al. (2011). Olaparib in patients with recurrent high-grade serous or poorly differentiated ovarian carcinoma or triple-negative breast cancer: a phase 2, multicentre, open-label, non-randomised study. The Lancet Oncology *12*, 852–861.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Mitchell, T.J., Rubanova, Y., Anur, P., Rosebrock, D., Yu, K., et al. (2017). The evolutionary history of 2,658 cancers. bioRxiv 161562.

Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Tamborero, D., Schroeder, M.P., Jene-Sanz, A., Santos, A., and Lopez-Bigas, N. (2013). IntOGen-mutations identifies cancer drivers across tumor types. Nature Methods *10*, 1081–1082.

Greenman, C., Pleasance, E., Newman, S., Yang, F., Fu, B., Nik-Zainal, S., Jones, D., Lau, K., Carter, N., Edwards, P., et al. (2012). Estimation of rearrangement phylogeny for cancer genomes. Genome Research *22*, 346–361.

Griffith, M., Miller, C.A., Griffith, O.L., Krysiak, K., Skidmore, Z.L., Ramu, A., Walker, J.R., Dang, H.X., Trani, L., Larson, D.E., et al. (2015). Optimizing Cancer Genome Sequencing and Analysis. Cell Systems *1*, 210–223.

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., Giuliany, R., Rosner, J., Oloumi, A., and Shumansky, K. (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. Genome Research *22*, 1995–2007.

Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L.M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., et al. (2014). TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Research *24*, 1881–1893.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., and King, M.C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. Science (New York, N.Y.) *250*, 1684–1689.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Helleday, T., Petermann, E., Lundin, C., Hodgson, B., and Sharma, R.A. (2008). DNA repair pathways as targets for cancer therapy. Nature Reviews Cancer *8*, 193–204.

Huang, Y., and Li, L. (2013). DNA crosslinking damage and cancer - a tale of friend and foe. Translational Cancer Research *2*, 144–154.

Huang, X., Wojtowicz, D., Przytycka, T.M., and Curtis, C. (2017). Detecting presence of mutational signatures in cancer with confidence. Bioinformatics *10.1093*.

Iyer, G., Audenet, F., Middha, S., Carlo, M.I., Regazzi, A.M., Funt, S., Al-Ahmadie, H., Solit, D.B., Rosenberg, J.E., and Bajorin, D.F. (2017). Mismatch repair (MMR) detection in urothelial carcinoma (UC) and correlation with immune checkpoint blockade (ICB) response. J Clin Oncol *35*, suppl; abstr 4511.

Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S.A., Jahesh, G., Khan, H., Coombe, L., Warren, R.L., et al. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Research *27*, 768–777.

Jacot, W., Theillet, C., Guiu, S., and Lamy, P.-J. (2015). Targeting triple-negative breast cancer and high-grade ovarian carcinoma: refining BRCAness beyond BRCA1/2 mutations? Future Oncology *11*, 557–559.

Johnson, N., Fletcher, O., Palles, C., Rudd, M., Webb, E., Sellick, G., dos Santos Silva, I., McCormack, V., Gibson, L., Fraser, A., et al. (2007). Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility. Human Molecular Genetics *16*, 1051–1057.

Jones, S.J.M., Laskin, J., Li, Y.Y., Griffith, O.L., An, J., Bilenky, M., Butterfield, Y.S., Cezard, T., Chuah, E., and Corbett, R. (2010). Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. Genome Biology *11*, 1–12.

Joosse, S.A. (2012). BRCA1 and BRCA2: a common pathway of genome protection but different breast cancer subtypes. Nature Reviews Cancer *12*, 372–372.

Kaufman, B., Shapira-Frommer, R., Schmutzler, R.K., Audeh, M.W., Friedlander, M., Balmana, J., Mitchell, G., Fried, G., Bowen, K., and Fielding, A. (2013). Olaparib monotherapy in patients with advanced cancer and a germ-line BRCA1/2 mutation: An open-label phase II study. In ASCO Annual Meeting Proceedings, p. 11024.

Kennecke, H., Yerushalmi, R., Woods, R., Cheang, M.C.U., Voduc, D., Speers, C.H., Nielsen, T.O., and Gelmon, K. (2010). Metastatic Behavior of Breast Cancer Subtypes. Journal of Clinical Oncology *28*, 3271–3277.

Kennedy, R.D., Quinn, J.E., Mullan, P.B., Johnston, P.G., and Harkin, D.P. (2004). The role of BRCA1 in the cellular response to chemotherapy. Journal of the National Cancer Institute *96*, 1659–1668.

Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., Tiao, G., Kwiatkowski, D.J., Rosenberg, J.E., Van Allen, E.M., D'Andrea, A.D., et al. (2016).

Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nature Genetics *48*, 600–606.

Kohlmann, W., and Gruber, S.B. (1993). Lynch Syndrome (University of Washington, Seattle).

Kovac, M., Blattmann, C., Ribi, S., Smida, J., Mueller, N.S., Engert, F., Castro-Giner, F., Weischenfeldt, J., Kovacova, M., and Krieg, A. (2015). Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. Nature Communications *6*.

Kuismanen, S.A., Holmberg, M.T., Salovaara, R., Chapelle, A. de la, and Peltomäki, P. (2000). Genetic and Epigenetic Modification of MLH1 Accounts for a Major Share of Microsatellite-Unstable Colorectal Cancers. The American Journal of Pathology *156*, 1773–1779.

Kumar-Sinha, C., and Chinnaiyan, A.M. (2018). Precision oncology in the age of integrative genomics. Nature Biotechnology *36*, 46–60.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Laskin, J., Jones, S., Aparicio, S., Chia, S., Ch'ng, C., Deyell, R., Eirew, P., Fok, A., Gelmon, K., Ho, C., et al. (2015). Lessons learned from the application of whole-genome analysis to the treatment of patients with advanced cancers. Molecular Case Studies *1*.

Lauss, M., Donia, M., Harbst, K., Andersen, R., Mitra, S., Rosengren, F., Salim, M., Vallon-Christersson, J., Törngren, T., Kvist, A., et al. (2017). Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. Nature Communications *8*, 1738.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Le, D.T., Uram, J.N., Wang, H., Bartlett, B.R., Kemberling, H., Eyring, A.D., Skora, A.D., Luber, B.S., Azad, N.S., Laheru, D., et al. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. New England Journal of Medicine *372*, 2509–2520.

Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.-F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., et al. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nature Communications *8*, 1315.

Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M., et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature *456*, 66–72.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

Li, X., and Heyer, W.-D. (2008). Homologous recombination in DNA repair and DNA damage tolerance. Cell Research *18*, 99–113.

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

Li, X., Yao, X., Wang, Y., Hu, F., Wang, F., Jiang, L., Liu, Y., Wang, D., Sun, G., and Zhao, Y. (2013). MLH1 promoter methylation frequency in colorectal cancer patients and related clinicopathological and molecular features. PloS One *8*, e59064.

Lips, E.H., Mulder, L., Oonk, A., Kolk, L.E. van der, Hogervorst, F.B.L., Imholz, A.L.T., Wesseling, J., Rodenhuis, S., and Nederlof, P.M. (2013). Triple-negative breast cancer: BRCAness and concordance of clinical features with BRCA1-mutation carriers. British Journal of Cancer *108*, 2172–2177.

Liu, B., Conroy, J.M., Morrison, C.D., Odunsi, A.O., Qin, M., Wei, L., Trump, D.L., Johnson, C.S., Liu, S., and Wang, J. (2015). Structural variation discovery in the cancer genome using next generation sequencing: computational solutions and perspectives. Oncotarget *6*, 5477–5489.

Lord, C.J., and Ashworth, A. (2012). The DNA damage response and cancer therapy. Nature *481*, 287–294.

Lord, C.J., and Ashworth, A. (2016). BRCAness revisited. Nature Reviews Cancer *16*, 110–120.

Maciejowski, J., and Imielinski, M. (2016). Modeling cancer rearrangement landscapes: from pattern to mechanism, and back. Current Opinion in Systems Biology *1*, 54–61.

Mantere, T., Winqvist, R., Kauppila, S., Grip, M., Jukkola-Vuorinen, A., Tervasmäki, A., Rapakko, K., and Pylkäs, K. (2016). Targeted Next-Generation Sequencing Identifies a Recurrent Mutation in MCPH1 Associating with Hereditary Breast Cancer Susceptibility. PLOS Genetics *12*, e1005816.

Mayor, P., Gay, L.M., Lele, S., and Elvin, J.A. (2017). BRCA1 reversion mutation acquired after treatment identified by liquid biopsy. Gynecologic Oncology Reports *21*, 57–60.

McGranahan, N., Favero, F., Bruin, E.C. de, Birkbak, N.J., Szallasi, Z., and Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. Science Translational Medicine *7*.

Meier, B., Cooke, S.L., Weiss, J., Bailly, A.P., Alexandrov, L.B., Marshall, J., Raine, K., Maddison, M., Anderson, E., Stratton, M.R., et al. (2014). C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. Genome Research *24*, 1624–1636.

Meric-Bernstam, F., Farhangfar, C., Mendelsohn, J., and Mills, G.B. (2013). Building a personalized medicine infrastructure at a major cancer center. Journal of Clinical Oncology *31*, 1849–1857.

Mestan, K.K., Ilkhanoff, L., Mouli, S., and Lin, S. (2011). Genomic sequencing in clinical trials. Journal of Translational Medicine *9*, 222.

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

Miller, C.A., White, B.S., Dees, N.D., Griffith, M., Welch, J.S., Griffith, O.L., Vij, R., Tomasson, M.H., Graubert, T.A., Walter, M.J., et al. (2014). SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. PLoS Computational Biology *10*, e1003665.

Morris, V., Rao, X., Pickering, C., Foo, W.C., Rashid, A., Eterovic, K., Kim, T., Chen, K., Wang, J., Shaw, K., et al. (2017). Comprehensive Genomic Profiling of Metastatic Squamous Cell Carcinoma of the Anal Canal. Molecular Cancer Research *15*, 1542–1550.

Mulligan, J.M., Hill, L.A., Deharo, S., Irwin, G., Boyle, D., Keating, K.E., Raji, O.Y., McDyer, F.A., O'Brien, E., Bylesjo, M., et al. (2014). Identification and Val-

idation of an Anthracycline/Cyclophosphamide–Based Chemotherapy Response Assay in Breast Cancer. JNCI: Journal of the National Cancer Institute *106*, djt335.

Nagourney, R.A., Link, J.S., Blitzer, J.B., Forsthoff, C., and Evans, S.S. (2000). Gemcitabine Plus Cisplatin Repeating Doublet Therapy in Previously Treated, Relapsed Breast Cancer Patients. Journal of Clinical Oncology *18*, 2245–2249.

Narod, S.A., Feunteun, J., Lynch, H.T., Watson, P., Conway, T., Lynch, J., and Lenoir, G.M. (1991). Familial breast-ovarian cancer locus on chromosome 17q12-q23. Lancet (London, England) *338*, 82–83.

National Cancer Institute (2014). BRCA1 and BRCA2: Cancer Risk and Genetic Testing.

Nik-Zainal, S., and Morganella, S. (2017). Mutational Signatures in Breast Cancer: The Problem at the DNA Level. Clin Cancer Res *23*, 2617–2629.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. Cell *149*.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., and Wedge, D.C. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*.

Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M.D., Wendl, M.C., and Ding, L. (2014). MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics *30*, 1015–1016.

Norquist, B., Wurz, K.A., Pennil, C.C., Garcia, R., Gross, J., Sakai, W., Karlan, B.Y., Taniguchi, T., and Swisher, E.M. (2011). Secondary Somatic Mutations Restoring BRCA1/2 Predict Chemotherapy Resistance in Hereditary Ovarian Carcinomas. Journal of Clinical Oncology *29*, 3008–3015.

Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–28.

Nowell P., H.D. (1960). A minute chromosome in human chronic granulocytic leukemia. Science *132*.

Patch, A.-M., Christie, E.L., Etemadmoghadam, D., Garsed, D.W., George, J., Fereday, S., Nones, K., Cowin, P., Alsop, K., and Bailey, P.J. (2015). Whole-genome characterization of chemoresistant ovarian cancer. Nature *521*, 489–494.

Pilati, C., Shinde, J., Alexandrov, L., Assié, G., André, T., Hélias-Rodzewicz, Z., Doucoudray, R., Le, C., Zucman-Rossi, J., Emile, J., et al. (2017). Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. Journal of Pathology *242*.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.-L., Beare, D., Lau, K.W., Greenman, C., et al. (2010a). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184–190.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.-L., Ordóñez, G.R., Bignell, G.R., et al. (2010b). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Polak, P., Kim, J., Braunstein, L.Z., Karlic, R., Haradhavala, N.J., Tiao, G., Rosebrock, D., Livitz, D., Kübler, K., Mouw, K.W., et al. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. Nature Genetics *49*, 1476–1486.

Poon, S.L., Huang, M.N., Choo, Y., McPherson, J.R., Yu, W., Heng, H.L., Gan, A., Myint, S.S., Siew, E.Y., Ler, L.D., et al. (2015). Mutation signatures implicate aristolochic acid in bladder cancer development. Genome Medicine *7*, 38.

Popova, T., Manié, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M., Longy, M., et al. (2012). Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2inactivation. Cancer Res *72*.

Purdom, E., Ho, C., Grasso, C.S., Quist, M.J., Cho, R.J., and Spellman, P. (2013). Methods and challenges in timing chromosomal abnormalities within cancer samples. Bioinformatics *29*, 3113–3120.

Ranjha, L., Howard, S.M., and Cejka, P. (2018). Main steps in DNA double-strand break repair: an introduction to homologous recombination and related processes. Chromosoma *127*, 187–214.

Raphael, B.J., Hruban, R.H., Aguirre, A.J., Moffitt, R.A., Yeh, J.J., Stewart, C., Robertson, A.G., Cherniack, A.D., Gupta, M., Getz, G., et al. (2017). Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell *32*, 185–203.e13.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine *17*, 405–423.

Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., and Ho, T.S. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science *348*, 124–128.

Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G.V., Carter, S.L., Saksena, G., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nature Genetics *45*, 970–976.

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S.D., Mungall, K., Lee, S., Okada, H.M., Qian, J.Q., et al. (2010). De novo assembly and analysis of RNA-seq data. Nature Methods *7*, 909–912.

Robinson, D.R., Wu, Y.-M., Lonigro, R.J., Vats, P., Cobain, E., Everett, J., Cao, X., Rabban, E., Kumar-Sinha, C., Raymond, V., et al. (2017). Integrative clinical genomics of metastatic cancer. Nature *548*.

Robson, M., Im, S.-A., Senkus, E., Xu, B., Domchek, S.M., Masuda, N., Delaloge, S., Li, W., Tung, N., Armstrong, A., et al. (2017). Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. New England Journal of Medicine *377*, 523–533.

Rosales, R.A., Drummond, R.D., Valieris, R., Dias-Neto, E., and Silva, I.T. da (2017). signeR: an empirical Bayesian approach to mutational signature discovery. Bioinformatics *33*, 8–16.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology *17*, 31.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., Sohrab, &., et al. (2014). Pyclone: statistical inference of clonal population structure in cancer. Nature Methods *11*.

Roy, R., Chun, J., and Powell, S.N. (2012). BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nature Reviews Cancer *12*, 68–78.

Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antolín, A.A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2015). In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell *27*, 382–396.

Saunders, C.T., Wong, W.S.W., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817.

Schulze, K., Imbeaud, S., Letouze, E., Alexandrov, L.B., Calderaro, J., Rebouissou, S., Couchy, G., Meiller, C., Shinde, J., Soysouvanh, F., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet *47*, 505–511.

Scully, R., and Livingston, D.M. (2000). In search of the tumour-suppressor functions of BRCA1 and BRCA2. Nature *408*, 429–432.

Scully, R., Chen, J., Plug, A., Xiao, Y., Weaver, D., Feunteun, J., Ashley, T., and Livingston, D.M. (1997a). Association of BRCA1 with Rad51 in mitotic and meiotic cells. Cell *88*, 265–275.

Scully, R., Chen, J., Ochs, R.L., Keegan, K., Hoekstra, M., Feunteun, J., and Livingston, D.M. (1997b). Dynamic changes of BRCA1 subnuclear location and phosphorylation state are initiated by DNA damage. Cell *90*, 425–435.

Segovia, R., Tam, A.S., and Stirling, P.C. (2015). Dissecting genetic and environmental mutation signatures with model organisms. Trends in Genetics *31*, 465–474.

Serrano-Fernández, P., Debniak, T., Górski, B., Bogdanova, N., Dörk, T., Cybulski, C., Huzarski, T., Byrski, T., Gronwald, J., Wokołorczyk, D., et al. (2009). Synergistic interaction of variants in CHEK2 and BRCA2 on breast cancer risk. Breast Cancer Research and Treatment *117*, 161–165.

Seyfried, T.N., and Huysentruyt, L.C. (2013). On the origin of cancer metastasis. Critical Reviews in Oncogenesis *18*, 43–73.

Sheffield, B.S., Tinker, A.V., Shen, Y., Hwang, H., Li-Chang, H.H., Pleasance, E., Ch'ng, C., Lum, A., Lorette, J., and McConnell, Y.J. (2015). Personalized oncogenomics: clinical experience with malignant peritoneal mesothelioma using whole genome sequencing. PloS One *10*, e0119689.

Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. PLOS Genetics *11*, e1005657.

Sirohi, B., Arnedos, M., Popat, S., Ashley, S., Nerurkar, A., Walsh, G., Johnston, S., and Smith, I.E. (2008). Platinum-based chemotherapy in triple-negative breast cancer. Annals of Oncology *19*.

Stan Development Team (2017). shinystan: Interactive Visual and Numerical Diagnostics and Posterior Analysis for Bayesian Models.

Stecklein, S.R., and Sharma, P. (2014). Tumor homologous recombination deficiency assays: another step closer to clinical application? Breast Cancer Research : BCR *16*, 409.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.

Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Zainal, S.N., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. Nature *486*, 400.

Supek, F., and Lehner, B. (2017). Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. Cell *170*, 534–547.e23.

Surgeon General (1964). Report of the Advisory Committee to the Surgeon General of the Public Health Service. US Department of Health, Education and Welfare, Public Health Service Publication.

Swisher, E.M., Sakai, W., Karlan, B.Y., Wurz, K., Urban, N., and Taniguchi, T. (2008). Secondary BRCA1 mutations in BRCA1-mutated ovarian carcinomas with platinum resistance. Cancer Research *68*, 2581–2586.

Szikriszt, B., Póti, Á., Pipek, O., Krzystanek, M., Kanu, N., Molnár, J., Ribli, D., Szeltner, Z., Tusnády, G.E., Csabai, I., et al. (2016). A comprehensive survey of the mutagenic impact of common cancer cytotoxics. Genome Biology *17*, 99.

Tattini, L., D'Aurizio, R., and Magi, A. (2015). Detection of Genomic Structural Variants from Next-Generation Sequencing Data. Frontiers in Bioengineering and Biotechnology *3*, 92.

Telli, M.L., Timms, K.M., Reid, J., Hennessy, B., Mills, G.B., Jensen, K.C., Szallasi, Z., Barry, W.T., Winer, E.P., and Tung, N.M. (2016). Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. Clinical Cancer Research *22*, 3764–3773.

The Cancer Genome Atlas (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*.

Thibodeau, S.N., Bren, G., and Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. Science *260*, 816–819.

Timms, K.M., Abkevich, V., Hughes, E., Neff, C., Reid, J., Morris, B., Kalva, S., Potter, J., Tran, T.V., and Chen, J. (2014). Association of BRCA1/2 defects with

genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. Breast Cancer Research *16*, 1.

Tomasetti, C., Li, L., and Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science *355*.

Tomita-Mitchell, A., Kat, A.G., Marcelino, L.A., Li-Sucholeiki, X.C., Goodluck-Griffith, J., and Thilly, W.G. (2000). Mismatch repair deficient human cells: spontaneous and MNNG-induced mutational spectra in the HPRT gene. Mutation Research *450*, 125–138.

Tsuneizumi, M., Emi, M., Hirano, A., Utada, Y., Tsumagari, K., Takahashi, K., Kasumi, F., Akiyama, F., Sakamoto, G., Kazui, T., et al. (2002). Association of allelic loss at 8p22 with poor prognosis among breast cancer cases treated with high-dose adjuvant chemotherapy. Cancer Letters *180*, 75–82.

Tutt, A., Ellis, P., Kilburn, L., Gilett, C., Pinder, S., Abraham, J., Barrett, S., Barrett-Lee, P., Chan, S., and Cheang, M. (2015). Abstract S3-01: The TNT trial: A randomized phase III trial of carboplatin (C) compared with docetaxel (D) for patients with metastatic or recurrent locally advanced triple negative or BRCA1/2 breast cancer (CRUK/07/012). Cancer Research *75*, S3–01.

Tuxen, I., Yde, C., Mau-Sørensen, M., Santoni-Rugiu, E., Lassen, U., and Nielsen, F. (2016). Copenhagen prospective personalized oncology (CoPPO): Genomic profiling to select patients for phase 1 trials. Annals of Oncology *27*.

Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., Policicchio, E., Damiano Urso, E., Agostini, M., Genuardi, M., et al. (2017). A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. EBioMedicine *20*.

Von Minckwitz, G., Hahnen, E., Fasching, P.A., Hauke, J., Schneeweiss, A., Salat, C., Rezai, M., Blohmer, J.U., Zahm, D.M., and Jackisch, C. (2014). Pathological com-

plete response (pCR) rates after carboplatin-containing neoadjuvant chemotherapy in patients with germline BRCA (gBRCA) mutation and triple-negative breast cancer (TNBC): Results from GeparSixto. In ASCO Annual Meeting Proceedings, p. 1005.

Waddell, N., Pajic, M., Patch, A.M., Chang, D.K., Kassahn, K.S., Bailey, P., Johns, A.L., Miller, D., Nones, K., Quek, K., et al. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. Nature *518*, 495–501.

Wang, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H.N., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nature Genetics *46*, 573–582.

Wang, Y.K., Bashashati, A., Anglesio, M.S., Cochrane, D.R., Grewal, D.S., Ha, G., McPherson, A., Horlings, H.M., Senz, J., and Prentice, L.M. (2017). Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. Nature Genetics.

Watkins, J.A., Irshad, S., Grigoriadis, A., and Tutt, A.N.J. (2014). Genomic scars as biomarkers of homologous recombination deficiency and drug response in breast and ovarian cancers. Breast Cancer Research *16*, 3405.

Weigelt, B., Comino-Méndez, I., Bruijn, I. de, Tian, L., Meisel, J.L., García-Murillas, I., Fribbens, C., Cutts, R., Martelotto, L.G., Ng, C.K.Y., et al. (2017). Diverse BRCA1 and BRCA2 Reversion Mutations in Circulating Cell-Free DNA of Therapy-Resistant Breast or Ovarian Cancer. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research *23*, 6708–6720.

Weinberg, R. (2013). The biology of cancer (Garland science).

Weymann, D., Laskin, J., Roscoe, R., Schrader, K.A., Chia, S., Yip, S., Cheung, W.Y., Gelmon, K.A., Karsan, A., Renouf, D.J., et al. (2017). The cost and cost tra-

jectory of whole-genome analysis guiding treatment of patients with advanced cancers. Molecular Genetics & Genomic Medicine *5*, 251–260.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.

Wong, H.-l., Yang, K.C., Shen, Y., Zhao, E.Y., Loree, J.M., Kennecke, H.F., Kalloger, S.E., Karasinska, J.M., Lim, H.J., Mungall, A.J., et al. (2018). Molecular characterization of metastatic pancreatic neuroendocrine tumors (PNETs) using whole-genome and transcriptome sequencing. Molecular Case Studies *4*, a002329.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., et al. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789–792.

Wu, J., Ho, C., Laskin, J., Gavin, D., Mak, P., Duncan, K., French, J., McGahan, C., Reid, S., Chia, S., et al. (2013). The development of a standardized software platform to support provincial population-based cancer outcomes units for multiple tumour sites: OaSIS - Outcomes and Surveillance Integration System. Studies in Health Technology and Informatics *183*, 98–103.

Wynder, E., and Graham, E. (1950). Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. Journal of the American Medical Association *143*, 329–336.

Xu, H., Di Antonio, M., McKinney, S., Mathew, V., Ho, B., O'Neil, N.J., Santos, N.D., Silvester, J., Wei, V., Garcia, J., et al. (2017). CX-5461 is a DNA G-quadruplex stabilizer with selective lethality in BRCA1/2 deficient tumours. Nature Communications *8*, 14432.

Yang, D., Khan, S., Sun, Y., Hess, K., Shmulevich, I., Sood, A.K., and Zhang, W. (2011). Association of BRCA1 and BRCA2 mutations with survival, chemotherapy

sensitivity, and gene mutator phenotype in patients with ovarian cancer. JAMA *306*, 1557–1565.

Yates, L.R., Knappskog, S., Wedge, D., Farmery, J.H.R., Gonzalez, S., Martincorena, I., Alexandrov, L.B., Van Loo, P., Haugland, H.K., Lilleng, P.K., et al. (2017). Genomic Evolution of Breast Cancer Metastasis and Relapse. Cancer Cell *32*, 169–184.e7.

Yip, S., Miao, J., Cahill, D.P., Iafrate, A.J., Aldape, K., Nutt, C.L., and Louis, D.N. (2009). MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance. Clinical Cancer Research : An Official Journal of the American Association for Cancer Research *15*, 4622–4629.

Yoshida, K., and Miki, Y. (2004). Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. Cancer Science *95*, 866–871.

Zare, F., Dow, M., Monteleone, N., Hosny, A., and Nabavi, S. (2017). An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. BMC Bioinformatics *18*, 286.

Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nature Medicine.

Zhao, E.Y., Shen, Y., Pleasance, E., Kasaian, K., Leelakumari, S., Jones, M., Bose, P., Ch'ng, C., Reisle, C., Eirew, P., et al. (2017). Homologous Recombination Deficiency and Platinum-Based Therapy Outcomes in Advanced Breast Cancer. Clinical Cancer Research *23*, 7521–7530.

Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al.

(2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature Biotechnology *34*, 303–311.

Zolkind, P., and Uppaluri, R. (2017). Checkpoint immunotherapy in head and neck cancers. Cancer and Metastasis Reviews *36*, 475–489.

APPENDICES

A1 NOTE ON NUCLEIC ACID NOMENCLATURE

Mutations are denoted by their base change. For example C>T denotes a base change from cytosine to thymine. Additionally, the trinucleotide context of a base change is described in one of two ways. The first is to provide the entire altered trinucleotide, for example a TCT > TTT mutation. The other is too indicate the context separately, for example a C>T mutation in a TpCpT context. Here, the p between bases denotes a phosphodiester bond.

In addition, expanded nomenclature is occasionally used. For example, a CpApY trinucleotide includes both CpApC or CpApT. A complete table of the expanded nomenclature is provided in table A.1.

Table A.1: The expanded nomenclature for nucleic acid naming.

Letter	Meaning	Nucleotide(s)
A	Adenine	A
T	Thymine	T
G	Guanine	G
C	Cytosine	C
R	Purine	G or A
Y	Pyrimidine	T or C
M	Amino	A or C
K	Keto	G or T
S	Strong	G or C
W	Weak	A or T
H	Not Guanine	A or C or T

Letter	Meaning	Nucleotide(s)
B	Not Adenine	G or T or C
V	Not Thymine	G or C or A
D	Not Cytosine	G or T or A
N	Any base	G or T or A or C

A2 APPENDIX TABLES

Table A.2: Significance tests for differences in mutation signatures across molecular subtypes. Multiple Kruskal-Wallis non-parametric tests were performed to identify variation in NMF-derived *de novo* mutation signatures across five breast cancer molecular subtypes (Luminal A, Luminal B, Her2-Amplified, Basal, and Normal-like). P-values were adjusted for false discovery rate, and revealed statistically significant subtype variability in three signatures: V3, V8, and V9.

	Chi-squared	Degrees of freedom	p	Adjusted p
V1	5.3	4	0.26	0.28
V2	10	4	0.035	0.075
V3	18	4	0.0015	0.0045
V4	9.9	4	0.042	0.075
V5	9.2	4	0.056	0.085
V6	7.3	4	0.12	0.16
V7	5	4	0.28	0.28
V8	18	4	0.0015	0.0045
V9	25	4	6.1e-05	0.00055

Table A.3: Sample details for whole genome sequencing of multiply-sequenced tumours.

ID	Occurrence	Stage	Sex	Diagnosis	Prep	Depth	Purity
P01	Primary	Adult	F	Colorectal Cancer	FFPE	41x	60
P01	Metastatic	Adult	F	Colorectal Cancer	OCT	92x	47
P02	Primary	Adult	F	Breast Cancer	FFPE	46x	80
P02	Metastatic	Adult	F	Breast Cancer	OCT	99x	80
P03	Primary	Adult	M	Colorectal Cancer	FFPE	49x	70
P03	Metastatic	Adult	M	Colorectal Cancer	OCT	46x	79
P04	Primary	Adult	F	Appendix Cancer	FFPE	42x	60
P04	Metastatic	Adult	F	Appendix Cancer	OCT	87x	81
P05	Primary	Adult	M	Colorectal Cancer	FFPE	50x	65
P05	Metastatic	Adult	M	Colorectal Cancer	OCT	92x	73
P06	Primary	Adult	F	Ovarian granulosa	FFPE	48x	80
P06	Metastatic	Adult	F	Ovarian granulosa	OCT	112x	90
P07	Primary	Adult	F	Breast Cancer	FFPE	37x	75
P07	Metastatic	Adult	F	Breast Cancer	OCT	105x	47

ID	Occurrence	Stage	Sex	Diagnosis	Prep	Depth	Purity
P08	Primary	Adult	F	Breast Cancer	FFPE	53x	80
P08	Metastatic	Adult	F	Breast Cancer	OCT	95x	35
P09	Primary	Adult	F	Endometrial Cancer	FFPE	44x	55
P09	Metastatic	Adult	F	Endometrial Cancer	OCT	96x	80
P10	Primary	Adult	F	Breast Cancer	FFPE	40x	70
P10	Metastatic	Adult	F	Breast Cancer	OCT	122x	63
P11	Primary	Adult	F	Breast Cancer	FFPE	43x	70
P11	Metastatic	Adult	F	Breast Cancer	OCT	91x	86
P12	Primary	Adult	F	Breast Cancer	FFPE	41x	65
P12	Metastatic	Adult	F	Breast Cancer	OCT	105x	70
P13	Primary	Adult	F	Breast Cancer	FFPE	54x	70
P13	Metastatic	Adult	F	Breast Cancer	OCT	97x	76
P14	Primary	Adult	F	Lung Cancer	FFPE	46x	60
P14	Metastatic	Adult	F	Lung Cancer	OCT	100x	64
P15	Primary	Adult	F	Breast Cancer	FFPE	39x	65
P15	Metastatic	Adult	F	Breast Cancer	OCT	97x	69
P16	Primary	Adult	F	Breast Cancer	FFPE	39x	60
P16	Metastatic	Adult	F	Breast Cancer	OCT	93x	63
P17	Primary	Pediatric	M	Neuroblastoma	FFPE	41x	90
P17	Metastatic	Pediatric	M	Neuroblastoma	FF	89x	69
P18	Primary	Adult	F	Breast Cancer	FFPE	31x	50
P18	Metastatic	Adult	F	Breast Cancer	OCT	96x	70
P19	Primary	Pediatric	M	Sarcoma	FFPE	54x	95
P19	Metastatic	Pediatric	M	Sarcoma	FF	102x	65
P20	Primary	Adult	F	Breast Cancer	FFPE	44x	70
P20	Metastatic	Adult	F	Breast Cancer	OCT	110x	90
P21	Primary	Adult	F	Adenocarcinoma of the lung	OCT	91x	20
P21	Metastatic	Adult	F	Adenocarcinoma of the lung	OCT	67x	51
P22	Primary	Adult	F	Cholangiocarcinoma	FF	86x	48
P22	Metastatic	Adult	F	Cholangiocarcinoma	OCT	79x	58
P23	Primary	Adult	M	Pancreatic Adenocarcinoma	FA	86x	25
P23	Metastatic	Adult	M	Pancreatic Adenocarcinoma	OCT	84x	49
P24	Primary	Adult	M	Metastatic lung cancer	OCT	100x	48
P24	Metastatic	Adult	M	Metastatic lung cancer	OCT	82x	35

A3 APPENDIX FIGURES

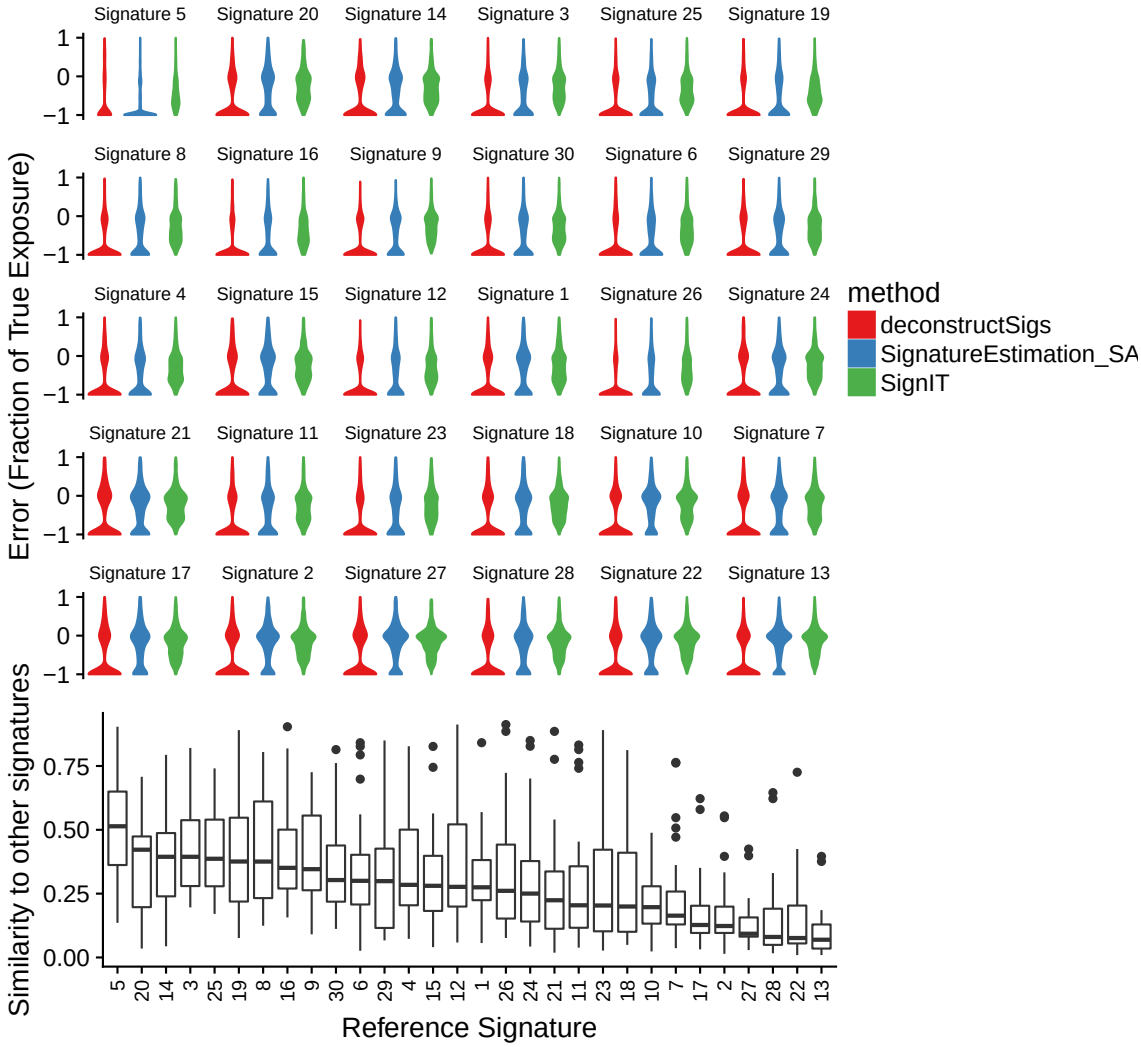

Figure A.1: Underestimated mutation signature exposures from simulated data. Simulated mutation catalogs with known exposure vectors were generated under various conditions and their exposures were re-estimated using deconstructSigs, SignatureEstimation, and SignIT. For every signature with non-zero simulated exposure, the error ratio was computed as $\frac{e-\epsilon}{\epsilon}$, where e is the estimated exposure and ϵ is the true exposure. Mutation signatures were ordered by their median similarity to other reference signatures. We observed frequent underestimation of mutation signatures by all methods, but errors were greatest in deconstructSigs and smallest in SignIT exposures. Signatures which are highly similar to other signatures were most likely to be underestimated.

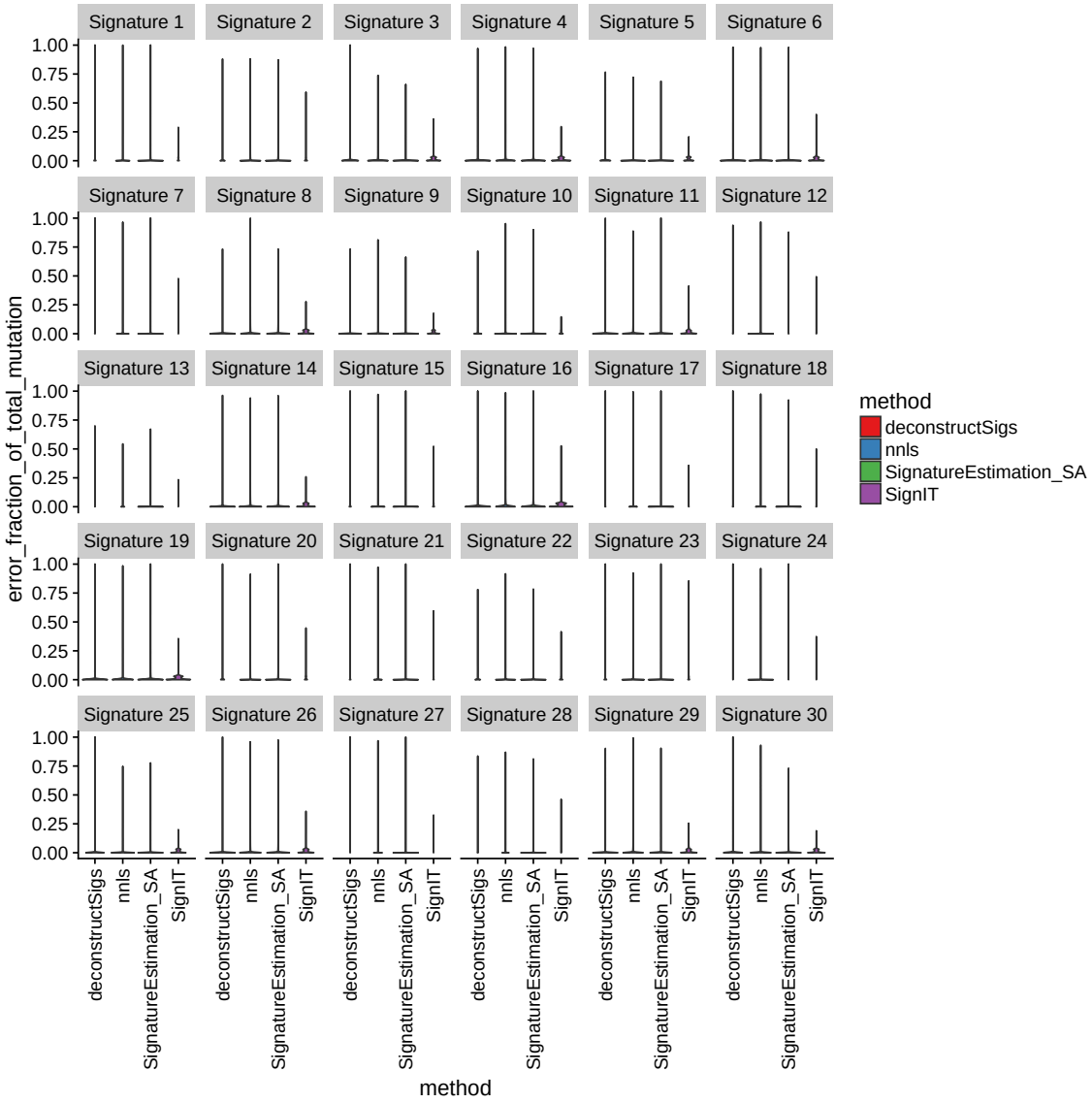

Figure A.2: Overestimated mutation signature exposures from simulated data. Simulated mutation catalogs with known exposure vectors were generated under various conditions and their exposures were re-estimated using deconstructSigs, SignatureEstimation, and SignIT. For every signature with zero simulated exposure, the overestimation error was computed as the estimated exposure divided by the total mutation burden. While all methods exhibited exposure errors, SignIT overestimated exposures with lower magnitude.

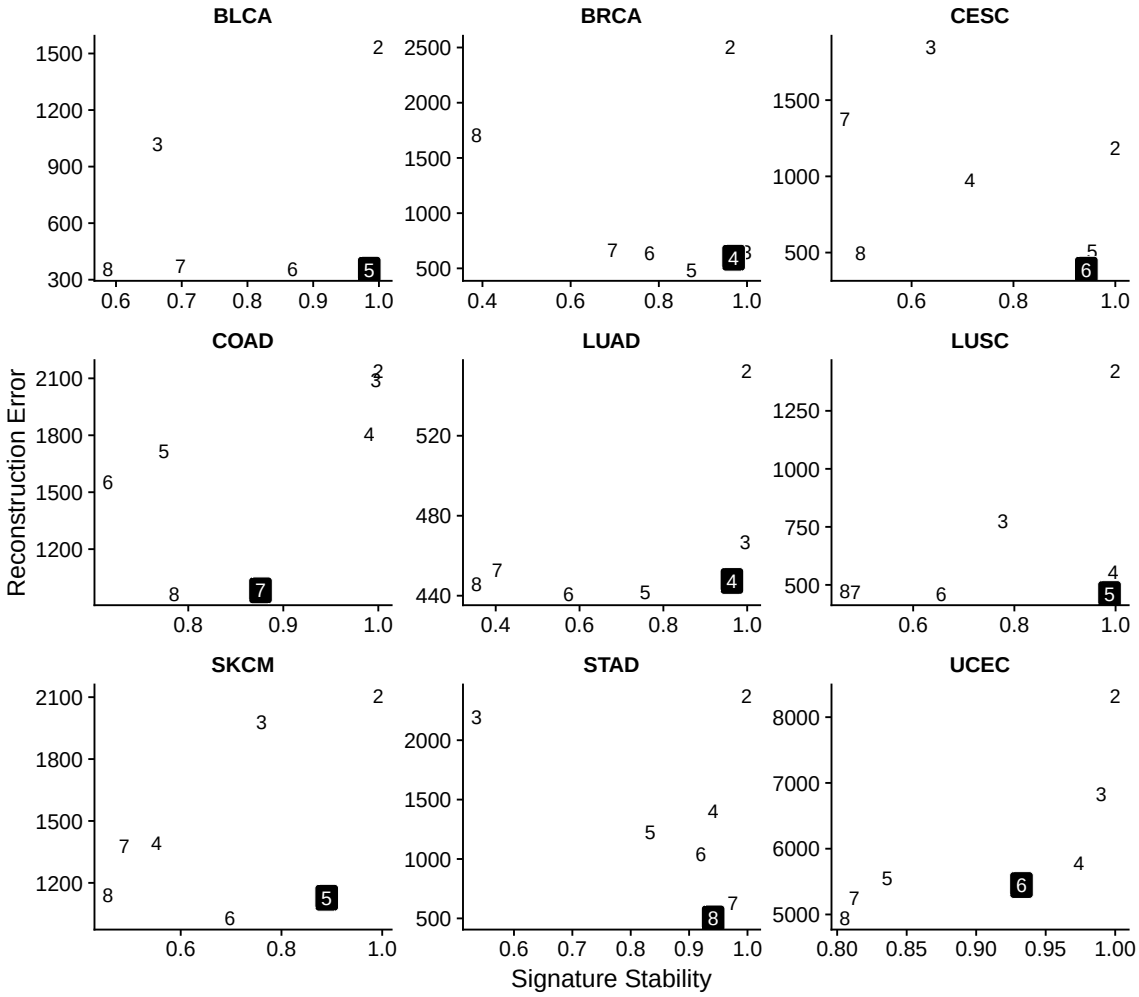

Figure A.3: Model selection for mutation signature analysis of nine cohorts of The Cancer Genome Atlas. To select the number of signatures, NMF was performed for models varying from 2 to 8 mutation signatures. To estimate signature stability, each NMF algorithm was run with 1000 Monte Carlo resimulated mutation catalog matrices. In each cohort, the model containing a number of signatures maximizing signature stability while minimizing Frobenius reconstruction error was chosen. Chosen models are indicated with a black box.

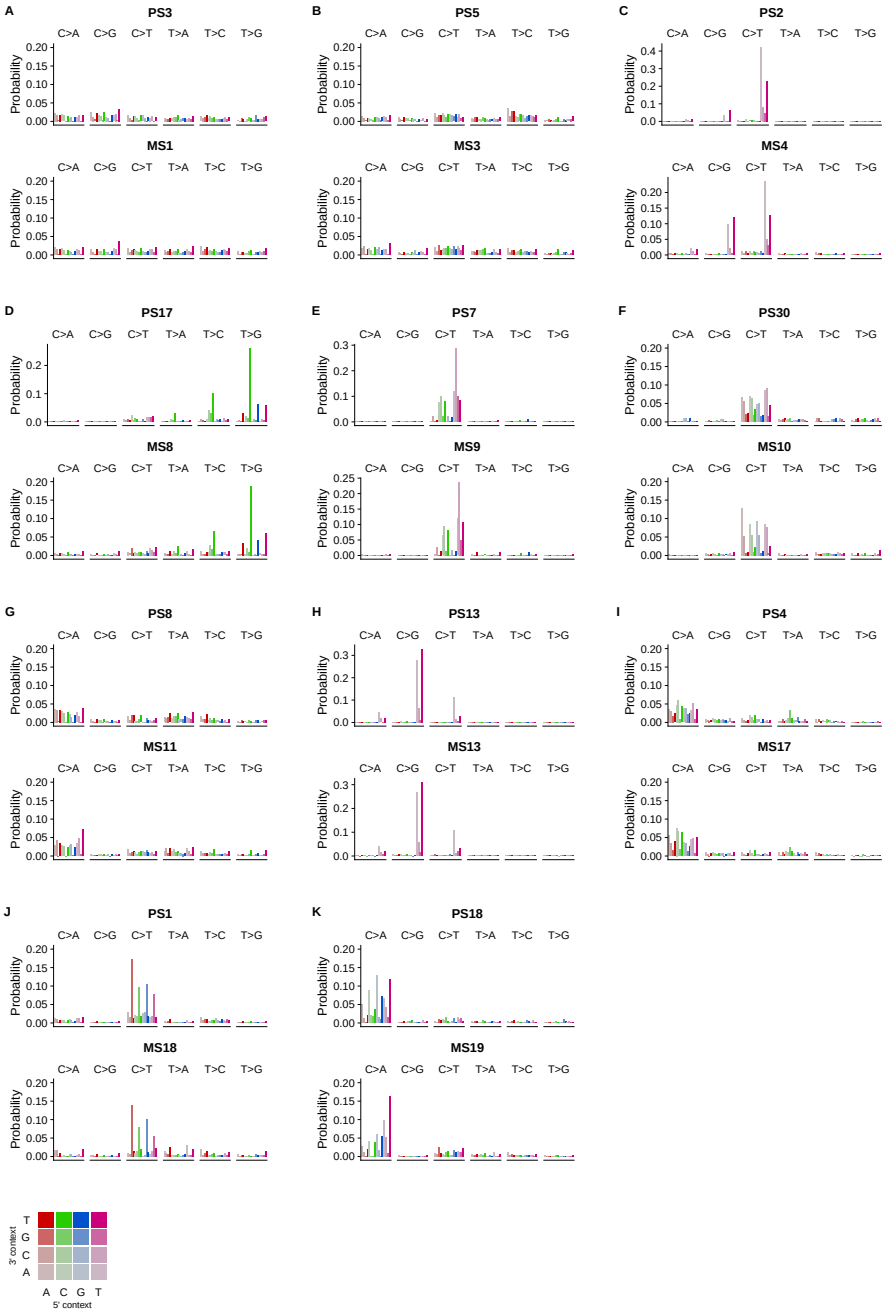

Figure A.4: Matching *de novo* mutation signatures to previously identified known signatures. Mutation signatures were deciphered *de novo* from cancertype-specific cohorts of metastatic cancer whole genomes. Signatures were clustered across cohorts to yield a set of independent signatures. Those closely resembling primary signatures were mapped accordingly. Here, PS stands for primary signature, and the numbers correspond to the 30 COSMIC mutation signatures.

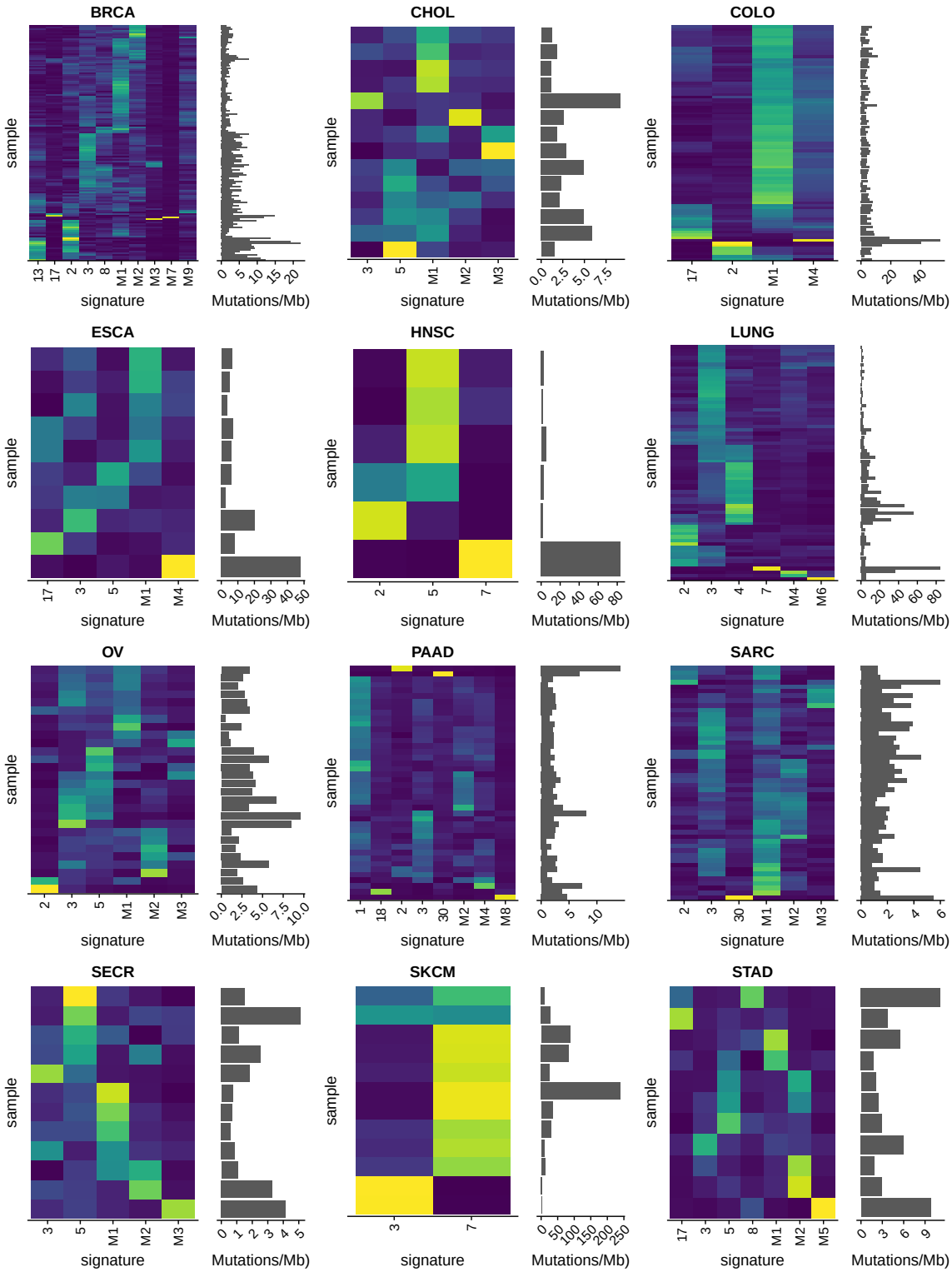

Figure A.5: Clustering of mutation signatures across multiple cancer cohorts into a common consensus signature set. Exposures for present signatures are shown per sample for all cohorts. Exposures were normalized by dividing by total mutation burden per sample such that the exposures of each sample (each row) sum to 1. Total mutation burden for the corresponding sample is also shown shown. Rows are ordered based on hierarchical clustering of the fractional exposures.

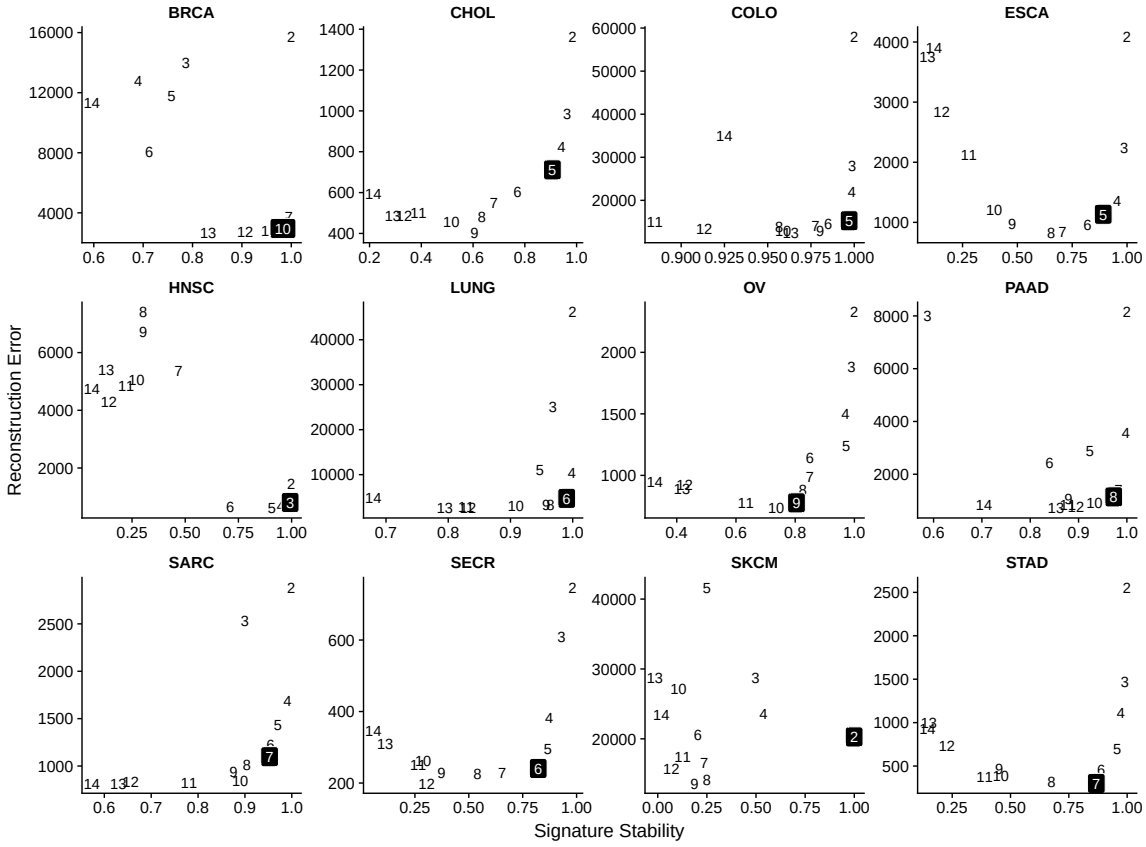

Figure A.6: Mutation signatures were successfully deciphered across 12 cancer cohorts. For each cohort, the number of signatures to be inferred was selected by jointly minimizing reconstruction error and maximizing signature stability. Chosen models are indicated with a black box.

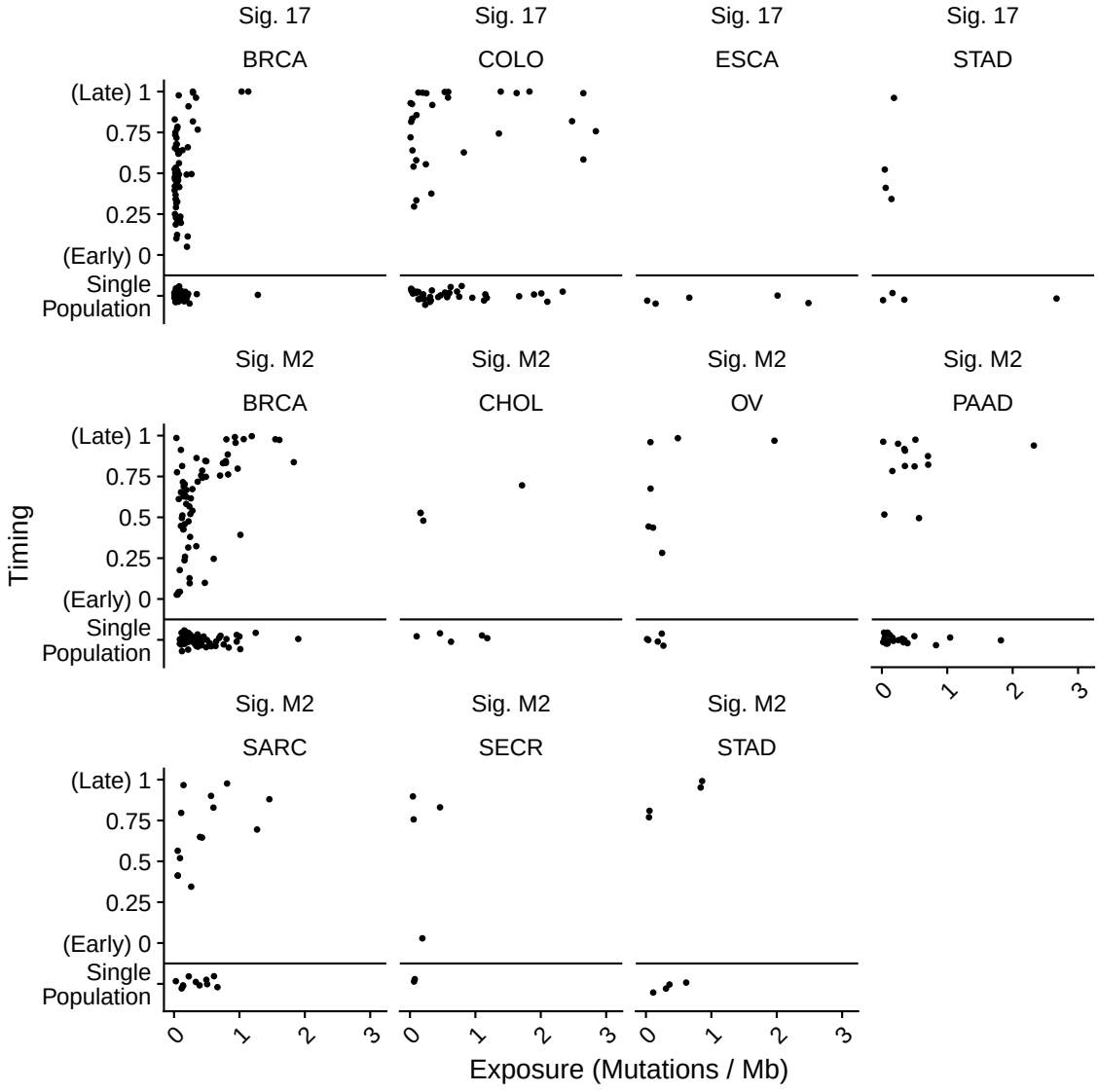

Figure A.7: Late-arising mutation signatures across cancer types. Temporal dissection of mutation signatures deciphered *de novo* from metastatic cancers revealed two late-arising mutation signatures. Signatures 17 and M2 were observed in late-arising mutational subpopulations across various cancer types.

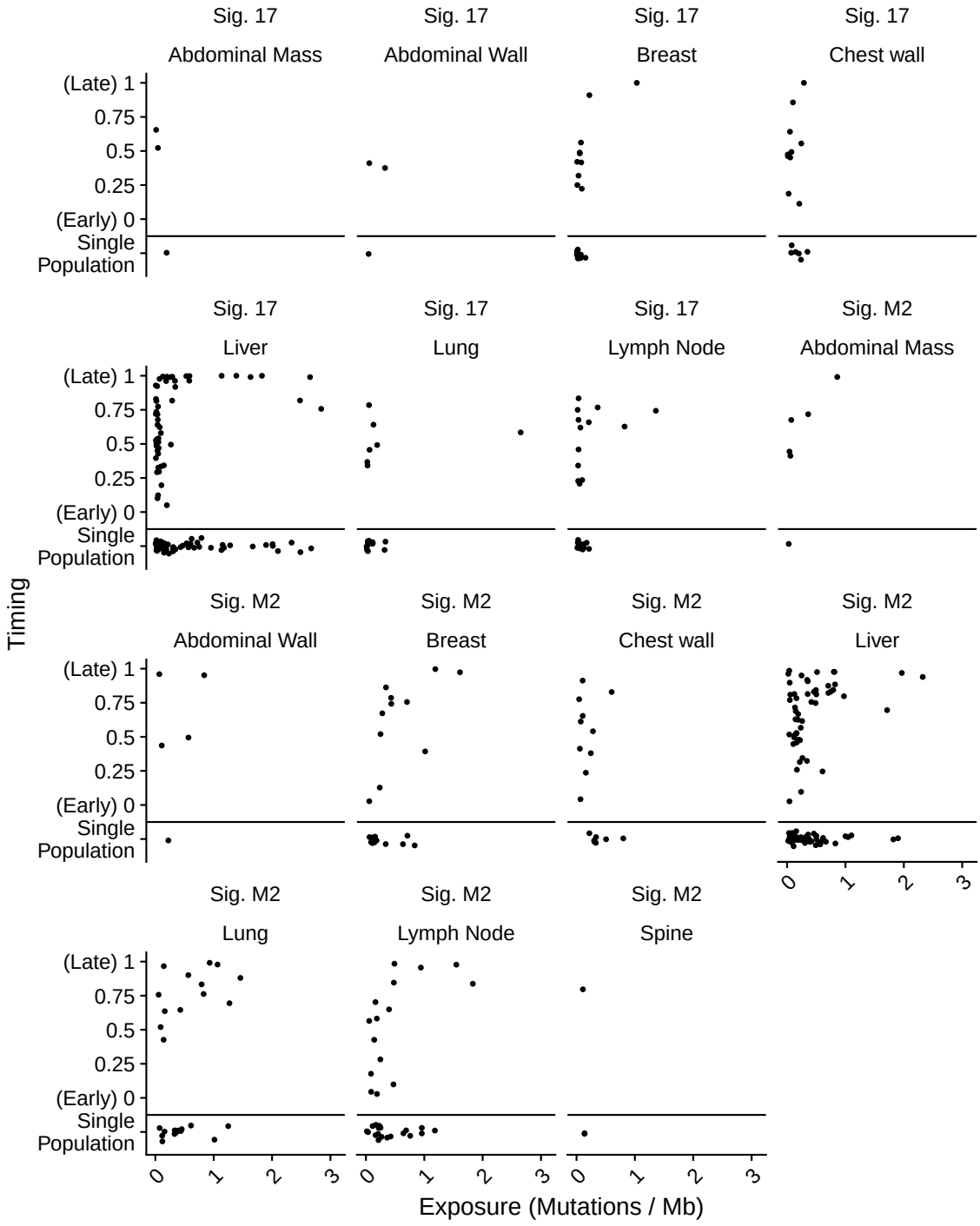

Figure A.8: Late-arising mutation signatures across biopsy sites. Temporal dissection of mutation signatures deciphered *de novo* from metastatic cancers revealed two late-arising mutation signatures. Signatures 17 and M2 were observed in late-arising mutational subpopulations across various biopsy sites.