

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Computational Analysis of Large-Scale Somatic Alterations in Cancer

Permalink

<https://escholarship.org/uc/item/0tn1g6z1>

ISBN

9798293841622

Author

Khandekar, Azhar

Publication Date

2025-09-17

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Computational Analysis of Large-Scale Somatic Alterations in Cancer

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

in

Bioinformatics and Systems Biology

by

Azhar Khandekar

Committee in charge:

Professor Ludmil Alexandrov, Chair
Professor Vineet Bafna, Co-Chair
Professor Hannah Carter
Professor Lukas Chavez
Professor Katherine Curtius
Professor Scott Lippman

2025

Copyright
Azhar Khandekar, 2025
All rights reserved.

The dissertation of Azhar Khandekar is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

DEDICATION

I would like to dedicate this dissertation to my closest family, particularly my
mom, sister, grandmother, and uncle.

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Table of Contents	v
List of Figures	vii
Acknowledgements	ix
Vita	x
Abstract of the Dissertation	xii
1 Introduction	1
2 Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator	12
2.1 Abstract	13
2.2 Introduction	14
2.3 Implementation	17
2.3.1 Classification of copy-number variations	17
2.3.2 Input data for classifying copy-number variations	21
2.3.3 Subclonal copy-number alterations	21
2.3.4 Classification of structural variants	21
2.3.5 Piecewise constant fitting to distinguish clustered SVs	27
2.3.6 Handling of VCF input	29
2.3.7 Input data for classifying structural variants	29
2.3.8 Evaluating the channel distributions for CN48 and SV32 schema in large pan-cancer datasets	30
2.4 Discussion	34
2.5 Conclusion	35
2.5.1 Future directions	35
3 Exploring the utility of Latent Dirichlet Allocation in extracting mutational signatures 37	
3.1 Introduction	38
3.1.1 Utilizing a Variant of an Existing Implementation of LDA Based on the Gibbs Sampling Algorithm	39
3.1.2 Evaluation of LDA on a synthetic benchmarking dataset of SBS signatures 40	
3.1.3 Conclusion and future directions	42

4	Examining the Role of Extrachromosomal DNA in 1,216 Lung Cancers	43
4.1	Introduction	44
4.1.1	Detailed description of the Sherlock Lung cohort	45
4.1.2	Constructing a highly sensitive pipeline to detect ecDNA that accounts for tumor purity and ploidy	48
4.1.3	Overall prevalence of ecDNA in lung cancer and epidemiological associations	57
4.1.4	Methodological approach to examine differences between ecDNA+ and ecDNA- samples	59
4.1.5	Analyzing the presence of ecDNA with clinical characteristics such as stage and metastasis	59
4.1.6	Examining the global prevalence of ecDNA	63
4.1.7	Calculation of genome instability index score (wGII)	66
4.1.8	Examining the association of ecDNA with genomic features	68
4.1.9	Examining the association of driver genes with ecDNA	72
4.1.10	Examining the association of ecDNA with hotspot driver mutations	74
4.1.11	Evaluating the effect of ecDNA on gene expression	76
4.1.12	Analysis of genome-wide mutational signatures in ecDNA+ vs ecDNA-lung tumors	78
4.1.13	Analysis of mutational signatures on ecDNA	83
4.1.14	Genomic content of ecDNA	86
4.1.15	Co-occurrence of oncogenes on ecDNA	95
4.1.16	Examining the relationship between chromothripsis and ecDNA	95
4.1.17	Clinical impact of ecDNA on patient survival	100
4.1.18	Discussion	106
4.1.19	Conclusion and future directions	106

LIST OF FIGURES

Figure 2.1:	Description of the copy-number classification schema.	19
Figure 2.2:	Converting copy-number segmentation profiles into copy-number mutational vectors.	20
Figure 2.3:	Description of the structural variant classification schema.	25
Figure 2.4:	Classifying Structural Variants into Mutational Vectors.	26
Figure 2.5:	Distribution of CN48 schema channels for TCGA samples profiled with SNP6 array data.	31
Figure 2.6:	Distribution of CN48 schema channels for PCAWG samples profiled with WGS.	32
Figure 2.7:	Distribution of SV32 schema channels for PCAWG samples profiled with WGS.	33
Figure 3.1:	Evaluating the performance of LDA on 20 noiseless scenarios.	41
Figure 4.1:	Sherlock-Lung cohort information	47
Figure 4.2:	Tumor purity in Sherlock and pan-cancer cohorts.	51
Figure 4.3:	Detailed worked example of the hybrid rescaling seed identification methodology for identifying ecDNA	52
Figure 4.4:	Representative examples of the hybrid rescaling seed identification methodology with 6 samples containing EGFR-ecDNA as exemplars.	53
Figure 4.5:	Default method vs rescaled method for identifying ecDNA	55
Figure 4.6:	Gene expression of additional EGFR-ecDNA detected by the rescaled pipeline	56
Figure 4.7:	Landscape of extrachromosomal DNA in lung cancer across histologies, sex, ancestry, countries, and smoking status	58
Figure 4.8:	Association of ecDNA with stage and metastasis	60
Figure 4.9:	Association of ecDNA with stage and metastasis in LUAD samples only	61
Figure 4.10:	Association of ecDNA with passive smoking	62
Figure 4.11:	Distribution of ecDNA harboring lung cancers in never-smokers and smokers by country.	64
Figure 4.12:	Association of ecDNA by country	65
Figure 4.13:	Relationship between ecDNA and overall genomic instability (wGII).	67
Figure 4.14:	Genomic features associated with the presence of ecDNA.	70
Figure 4.15:	Pan-cancer validation of the association of WGD with ecDNA.	71
Figure 4.16:	Association between driver gene alterations and ecDNA	73
Figure 4.17:	Association between ecDNA and driver mutations	75
Figure 4.18:	Differential gene expression analysis of ecDNA+ vs ecDNA- tumors	77
Figure 4.19:	Global mutational signatures of small mutations associated with ecDNA	80
Figure 4.20:	Global mutational signatures of small mutations input as numeric continuous variables	81

Figure 4.21: Global mutational signatures of large mutations associated with ecDNA	82
Figure 4.22: Mutational signatures present on ecDNA vs. flanking regions	84
Figure 4.23: Mutational signatures present on ecDNA vs. chrAmp regions	85
Figure 4.24: Genomic elements on ecDNA in lung cancer	87
Figure 4.25: Enrichment of genomic elements on ecDNA in never-smokers vs smokers	89
Figure 4.26: Enrichment of genomic elements on ecDNA in never-smokers vs smokers with recurrent oncogenes removed	90
Figure 4.27: Copy number of genomic elements on ecDNA	91
Figure 4.28: Oncogenes recurrently amplified via ecDNA in never-smokers and smokers	93
Figure 4.29: MDM2-ecDNA vs TP53 status	94
Figure 4.30: Exemplars of the types of ecDNA present in the Sherlock cohort	97
Figure 4.31: Co-occurrence of chromothripsis and ecDNA at the sample level and genomic sequence level	98
Figure 4.32: Size of chromothripsis regions, ecDNA regions, and overlapping regions	99
Figure 4.33: Impact of ecDNA on overall patient survival	101
Figure 4.34: Detailed Cox regression survival analysis results for never-smokers	102
Figure 4.35: Detailed Cox regression survival analysis results for smokers	103
Figure 4.36: Impact of ecDNA containing MDM2 or EGFR on overall patient survival	104
Figure 4.37: Impact of the type of ecDNA on overall patient survival	105

ACKNOWLEDGEMENTS

This dissertation would not have been possible without my co-authors, friends, and family. I would like to particularly thank my supervisors, Ludmil Alexandrov and Maria Teresa Landi, and my close collaborators and colleagues Tongwu Zhang and Jens Luebeck, as well as many more who helped with various aspects of the work.

Chapter 1, in full, is a reprint of the material as it appears in Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator. Khandekar, Azhar; Vangara, Raviteja; Barnes, Mark; Díaz-Gay, Marcos; Abbasi, Ammal; Bergstrom, Erik N.; Steele, Christopher D.; Pillay, Nischalan; Alexandrov, Ludmil B. BMC Genomics 24, 469 (2023). The dissertation author was the primary investigator and author of this paper.

Chapter 2 contains unpublished material co-authored with Vangara, Raviteja; Barnes, Mark; and Díaz-Gay, Marcos. The dissertation author was the primary author of chapter.

Chapter 3, in part, has been submitted for publication of the material as it may appear in Examining the Role of Extrachromosomal DNA in 1,216 Lung Cancers, Khandekar, Azhar; Hoang, Phuc H.; Luebeck, Jens; Díaz-Gay, Marcos; Zhao, Wei; McElderry, John P.; Hartman, Caleb; Mirafteb, Mona; Lee, Olivia W.; Barnao, Kara M.; Bergstrom, Erik N.; Yang, Yang; Nowak, Martin A.; Rothman, Nathaniel; Homer, Robert; Yang, Soo-Ryum; Lan, Qing; Wedge, David C.; Yang, Lixing; Chanock, Stephen J.; Zhang, Tongwu; Alexandrov, Ludmil B.; Landi, Maria Teresa, Nature Genetics, 2025. The dissertation author was the primary researcher and author of this paper.

VITA

2016	B. S. in Biochemistry <i>cum laude</i> , University of Washington, Seattle
2025	Ph. D. in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

First-author Publications

1. **Khandekar, A.**, Hoang, P.H., Luebeck, J., Díaz-Gay, M., Zhao, W., McElderry, J.P., Hartman, C., Mirafteb, M., Lee, O.W., Barnao, K.M., Bergstrom, E.N., Yang, Y., Nowak, M.A., Rothman, N., Homer, R., Yang, S.R., Lan, Q., Wedge, D.C., Yang, L., Chanock, S.J., Zhang, T., Alexandrov, L.B., Landi, M.T. *Examining the Role of Extrachromosomal DNA in 1,216 Lung Cancers*. *bioRxiv* [Preprint], 2025. doi: 10.1101/2025.06.03.657117.
2. **Khandekar, A.**, Vangara, R., Barnes, M., Díaz-Gay, M., Abbasi, A., Bergstrom, E.N., Steele, C.D., Pillay, N., Alexandrov, L.B. *Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator*. *BMC Genomics*, 2023, 24:469. doi: 10.1186/s12864-023-09584-y.

Co-author Publications

3. Díaz-Gay, M., Zhang, T., Hoang, P.H., Leduc, C., Baine, M.K., Travis, W.D., **Khandekar, A.**, et al. *The mutagenic forces shaping the genomes of lung cancer in never smokers*. *Nature*, 2025.
4. Abbasi, A., Steele, C.D., Bergstrom, E.N., **Khandekar, A.**, et al. *HRProfiler Detects Homologous Recombination Deficiency in Breast and Ovarian Cancers Using Whole-Genome and Whole-Exome Sequencing Data*. *Cancer Research*, 2025; 85(13):2504–2513.
5. Díaz-Gay, M., Dos Santos, W., Moody, S., Kazachkova, M., Abbasi, A., **Khandekar, A.**, et al. *Geographic and age variations in mutational processes in colorectal cancer*. *Nature*, 2025.
6. Senkin, S., Moody, S., Díaz-Gay, M., Abedi-Ardekani, B., Cattiaux, T., **Khandekar, A.**, et al. *Geographic variation of mutagenic exposures in kidney cancer genomes*. *Nature*, 2024; 629(8013):910–918.
7. Mangiante, L., Alcalá, N., Sexton-Oates, A., Di Genova, A., Gonzalez-Perez, A., **Khandekar, A.**, et al. *Multiomic analysis of malignant pleural mesothelioma identifies molecular axes and specialized tumor profiles driving intertumor heterogeneity*. *Nature Genetics*, 2023; 55(4):607–618.

8. Islam, S.M.A., Díaz-Gay, M., Wu, Y., Barnes, M., Vangara, R., Bergstrom, E.N., **Khandekar, A.**, et al. *Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor*. *Cell Genomics*, 2022; 2(11).
9. Steele, C.D., Abbasi, A., Islam, S.M.A., Bowes, A.L., **Khandekar, A.**, et al. *Signatures of copy-number alterations in human cancer*. *Nature*, 2022; 606(7916):984–991.
10. Bergstrom, E.N., Luebeck, J., Petljak, M., **Khandekar, A.**, et al. *Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA*. *Nature*, 2022; 602(7897):510–517.
11. Moody, S., Senkin, S., Islam, S.M.A., Wang, J., Nasrollahzadeh, D., **Khandekar, A.**, et al. *Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence*. *Nature Genetics*, 2021; 53(11):1553–1563.
12. Zhang, T., Joubert, P., Ansari-Pour, N., Zhao, W., Hoang, P.H., **Khandekar, A.**, et al. *Genomic and evolutionary classification of lung cancer in never smokers*. *Nature Genetics*, 2021; 53(9):1348–1359.

ABSTRACT OF THE DISSERTATION

Computational Analysis of Large-Scale Somatic Alterations in Cancer

by

Azhar Khandekar

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2025

Professor Ludmil Alexandrov, Chair
Professor Vineet Bafna, Co-Chair

Genomic alterations affecting 50 or more base pairs—including amplifications, deletions, and rearrangements—are pervasive drivers of cancer evolution. Despite their importance, these large-scale somatic alterations remain poorly understood due to challenges in detecting and interpreting them from short-read sequencing data. They are broadly classified into two major categories: copy number alterations (CNAs), which involve gains or losses of DNA segments and change the number of copies of specific genomic regions, and structural variants (SVs), which rearrange the genome through events such as deletions, duplications, inversions,

and translocations, often disrupting gene structure or regulation. While mutational signature frameworks for single-base substitutions are well established, comparable frameworks for CNAs and SVs are still limited, hindering insights into the mutational processes shaping cancer genomes. In Chapter 1, I present a robust software implementation for systematic mutational signature analysis of large-scale alterations. The tool integrates two complementary classification frameworks—an established schema for SVs and a newly developed framework for CNAs—and supports multiple data types, input formats, and extensive visualization. It standardizes analysis and provides the first broadly accessible platform for large-scale mutational profiling. In Chapter 2, I evaluate Latent Dirichlet Allocation (LDA) as an alternative to nonnegative matrix factorization (NMF) for mutational signature extraction. LDA effectively captures mutational patterns and offers a complementary approach for discovering previously unrecognized processes driving large-scale genomic changes. In Chapter 3, I present the most comprehensive analysis to date of extrachromosomal DNA (ecDNA)—a highly amplified circular DNA structure—in lung cancers from never-smokers and smokers. Using a large, diverse cohort spanning multiple ancestries, I adapted an ecDNA detection algorithm to account for low tumor purity and variable ploidy, common in lung tumors. Genomic and epidemiological analyses revealed key ecDNA-associated features, including recurrent oncogenes, regulatory elements, chromothripsis scars, and unique mutational signatures. Notably, ecDNA was strongly linked to whole-genome doubling, specific driver mutations, and worse survival in never-smokers, emphasizing its clinical relevance. Together, this dissertation delivers essential computational tools and new biological insights, providing the most detailed characterization of ecDNA in lung cancer to date and advancing our understanding of large-scale cancer genome evolution.

1 Introduction

Genomic instability is a defining hallmark of cancer, driving tumor evolution, heterogeneity, and resistance to therapy (Hanahan and Weinberg, 2011). The primary footprints of genomic instability that are left behind in tumor genomes are large-scale alterations, operationally defined as mutations ≥ 50 base pairs (Cosenza et al., 2022). These large-scale alterations are divided into two classes: structural variations (SVs) and copy-number alterations (CNAs), based on their structure and whether they alter the copy-number of a genomic locus or not. Although these two classes of events are interrelated, not all structural variants alter copy-number, and conversely, not all copy-number alterations require prior SVs. For example, copy neutral events such as inversions and reciprocal translocations do not alter the copy-number but are classified as SVs. Similarly, whole-genome doubling (WGD), the doubling of the entire complement of chromosomes, alters the copy-number but is not considered an SV (Eichler et al., 2011). In general, copy-number alterations refer to any amplification or deletion of segments of the genome, whereas the four canonical SV types encompass deletions, inversions, tandem-duplications and translocations (Eichler et al., 2011). Another important distinction between these two classes of large-scale mutations is that SVs require whole-genome sequencing data for detection, whereas copy-number alterations can be

detected from whole-exome sequencing and SNP arrays(Retterer et al., 2015).

Cancer genomic studies to date have been biased towards the study of small mutations for several reasons. First, large mutations are notoriously difficult to detect from sequencing data, much more so than small mutations. This is evident in the considerable variability among not only algorithms(Masood et al., 2024), but also data modalities(Gabrielaite et al., 2021). In fact, most large-scale pan cancer studies that have examined copy-number and structural alterations rely on ensemble approaches that utilize multiple callers(Li et al., 2020; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Another major consideration is that of tumor purity, defined as the fraction of cancerous cells within a sample, which can substantially affect the accuracy of copy-number and structural variant calling by diluting the tumor signal with normal DNA. In low-purity tumors, the observed amplitude of copy-number changes is attenuated due to admixture with diploid normal cells, often leading to underestimation of true copy-number gains or losses. Structural variant detection is similarly impacted, as the supporting read depth for rearrangement breakpoints is reduced in low-purity samples, decreasing sensitivity and increasing false negatives(Aran et al., 2015). Due to the fact that large-scale alterations are often subclonal, many SV and CN events might actually be missed using current approaches(Raine et al., 2016). Second, the interpretation of structural variants (SVs) and the mutational processes that generate them is considerably more complex than that of small mutations, in part because SVs span multiple scales of magnitude—ranging from kilobase-scale deletions to chromosome-arm level rearrangements—and can involve diverse mechanisms such as replication stress, chromothripsis, or double-strand break repair(Cosenza et al., 2022). Additionally, unlike small mutations which follow the infinite sites assumption, where each base is likely only

mutated only once during tumor evolution, multiple SV events can often affect the same genomic locus. Hence, it is difficult to arrive at the sequence of events leading to the end-state CN profile or SV calls(Steele et al., 2022).

Despite these challenges with detection and interpretation, algorithms to detect large-scale mutations have been rapidly advancing and recent work has shown that most large-scale alterations can in fact be detected from short-read sequencing of paired tumor-normal genomes(Choo et al., 2023), and a host of advanced tools have been developed to characterize large-scale mutations(Raine et al., 2016; Deshpande et al., 2019; Chen et al., 2016b; Talevich et al., 2016; Rausch et al., 2012). The signals derived from sequencing data to detect SVs and CNAs differ. SVs can be identified using aberrantly mapping and/or split reads in paired-end sequencing data. copy-number alterations are primarily identified by the read depth (i.e log ratio of tumor to normal reads), and more sophisticated algorithms also incorporate the B-allele frequency (the fraction of sequencing reads (or array probe intensities) that support the non-reference allele at a given germline heterozygous SNP locus) to determine the copy-number of the maternal and paternal alleles. In a diploid genome with no copy-number alteration, the expected BAF is approximately 0.5. Deviations from this value reflect allelic imbalance due to copy-number changes or loss of heterozygosity (LOH). For example, a region with one allele deleted (copy-number = 1) will have BAF values near 0 or 1, depending on which allele is retained. BAF, when analyzed alongside total copy-number (e.g., logR), enables inference of allele-specific copy-number states across the genome. To estimate tumor purity, sophisticated callers such as Battenberg simulate a large number of possible purity–ploidy combinations and computes, for each, the expected BAF and logR values under different copy-number states (e.g., balanced diploid, single-copy

loss, copy-neutral LOH, allele-specific gains). These expected values are then compared to the actual observed values across all heterozygous loci in the tumor. A likelihood score is computed for each combination, and the best-fitting purity–ploidy pair is selected as the maximum likelihood estimate. The derived purity and ploidy estimates can then be used to call subclonal copy-number alterations, another feature of more sophisticated copy-number callers. For example, in Battenberg, subclonality is inferred when the estimated CCF is significantly below 1.0, suggesting the presence of an ancestral population with one copy-number state and a descendant subpopulation with a different one.

A plethora of prior studies have shown the importance of large-scale mutations in cancer. They are ubiquitous across tumor types and have been found to arise at all stages of tumor development, from pre-cancer(Luebeck et al., 2023) to late-stage cancer(Li et al., 2020). 55% of driver mutations uncovered by the Pan-Cancer Analysis of Whole Genomes project represent SVs (Cosenza et al., 2022). Large-scale alterations in specific genes, such as ALK rearrangements, HER2 amplifications, and BCR-ABL1 translocations, can be used to guide therapeutic decisions(Cosenza et al., 2022). In most tumors, more base pairs are altered as a result of large-scale changes than point mutations(Cosenza et al., 2022), yet little is known about the causes and consequences of these changes compared to that of small mutations.

An approach that has proven to be promising in providing insight into the processes generating somatic mutations in cancer is mutational signature analysis, which involves the examination of mutational patterns within a set of cancer genomes and the extraction of mutational signatures that have generated these patterns. Historically, mutational patterns have been predominately examined in the context of small mutational events, which include single base substitutions (SBS), doublet base substitutions (DBS), and small insertions and

deletions (IDs). The success of the mutational signature approaches for small mutations relies on a biologically relevant schema to classify those mutations (Alexandrov et al., 2013). While these schema for small mutations are well developed and characterized in the literature, with an extensive array of tools and studies published (Islam et al., 2022a; Díaz-Gay et al., 2023; Davies et al., 2017), analogous schema for large-scale alterations are lacking.

The concept of copy-number (CN) signatures has emerged as a method to systematically characterize patterns of CNAs across tumors, akin to the well-established framework of mutational signatures for small base substitutions and indels. Two main methodological paradigms have been developed for CN signature extraction. The first is a hypothesis-driven approach, in which CN features are selected to reflect known biological processes. For example, Macintyre et al. designed features such as total copy-number, breakpoint density, oscillating CN segment length, and segment size to identify patterns associated with chromothripsis or replication-associated genome instability in ovarian cancer. Wang et al. later expanded this framework to include metrics that quantify the extent of CN aberrations both globally and locally within individual chromosomes, applying manual feature binning for consistency across datasets. The main limitation of these approaches is that they are i) optimized for specific rearrangement processes in specific cancer types and ii) no publicly available software implementation exists to facilitate the application of these frameworks across studies and cohorts. The second approach is mechanism-agnostic and relies on extracting features that summarize fundamental CN properties—such as loss of heterozygosity (LOH), total CN state, and segment size—without assuming prior biological associations. Steele et al. first applied this strategy in a study of undifferentiated sarcomas, using non-negative matrix factorization (NMF) to decompose CN profiles into signatures and attributions. This approach allows

downstream interpretation of signatures through their correlation with known genomic events, such as genome doubling or ecDNA formation, and has been extended to pan-cancer analyses identifying over 20 distinct CN signatures. The advantage of this mechanism-agnostic approach is that it allows for the discovery of novel biological processes causing CN changes, and is a unified framework that can consistently be applied across different study cohorts and data modalities (Steele et al., 2022).

SV signatures have also been underexplored in the literature, with only two frameworks having been developed to characterize them. The first comprehensive approach was introduced by Nik-Zainal et al., who analyzed 560 breast cancer whole-genome sequences and categorized rearrangements based on SV type (e.g., deletions, tandem duplications, inversions, translocations), size, and clustering status (Nik-Zainal et al., 2016). This study reinforced that SVs can also arise during periods of catastrophic genomic instability, resulting in a burst of rearrangements that manifests as breakpoints that are tightly clustered in genomic space. Such clustered SVs are characteristic of phenomena like chromothripsis, and it is crucial to distinguish them from the canonical non-clustered SV events. Using this classification, they applied non-negative matrix factorization (NMF) to derive a set of six distinct rearrangement signatures, each reflecting a different underlying mutational process. These included signatures associated with homologous recombination deficiency, tandem duplicator phenotype, and complex rearrangement clusters potentially related to replication-based mechanisms. This study was foundational in demonstrating that rearrangement patterns carry clinically relevant information, as two specific SV signatures (one predominantly characterized by tandem-duplications and the other characterized by deletions) were later linked to homologous recombination deficiency (HRD), a process which can nominate cancer patients for PARP

therapy (Davies et al., 2017). More recently, Li et al. extended this concept to a pan-cancer setting by analyzing over 2,700 whole-genome sequenced tumors from the PCAWG dataset (Li et al., 2020). They classified SVs using a complex feature space that included SV type, size, clustering, orientation, and replication timing, and identified 16 rearrangement signatures across diverse cancer types, using both a hierarchical Dirichlet process and non-negative matrix factorization. Due to the complexity of the derived features in this study, no software tool exists for grouping these structural variants in this manner.

In this dissertation, I first provide a robust software implementation of two frameworks for categorizing large-scale mutations in order to facilitate downstream mutational signature analysis. In Chapter 1, I implement a recently developed schema for classifying copy-number alterations that takes into account segment sizes, total copy-number, and the copy-number of each allele. This tool can consume input from a vast array of tools and algorithms to call copy-number, ensuring widespread utility. I provide extensive visualization capabilities for both the input counts matrix and the extracted signatures, and ensure the tool works in both the R environment and Python environment. Next, I implement a widely used structural variant framework that considers size, SV type, and whether or not an SV is clustered. This tool can ingest both VCF and BEDPE format, again ensuring maximum utility. Additionally, both of these implementations operate downstream of the CN and SV calls and seamlessly integrate with downstream methods to deconvolve mutational signatures such as non-negative matrix factorization.

In the vast majority of studies that I described thus far, focused on deciphering mutational signatures from catalogs of somatic mutations, a process known as de novo signature extraction, primarily utilized nonnegative matrix factorization (NMF). However,

alternative approaches such as Latent Dirichlet Allocation (LDA), which is a probabilistic model for discrete count data commonly used in the topic modeling space, have been relatively unexplored. In Chapter 2, I explore the utility of applying LDA to decipher mutational signatures from single-base substitutions, in order to assess its feasibility in deciphering mutational signatures from large-scale alterations.

One specific type of large-scale somatic alteration receiving a lot of renewed interest, that involves both structural alteration and copy-number alterations, is extrachromosomal DNA. Small fragments of circular extrachromosomal DNA (ecDNA), also referred to as double minutes, were first reported in human tumor cells in 1962 (Spriggs et al., 1962). This subsequently spurred decades of research into these particles (Benner et al., 1991; Balaban-Malenbaum and Gilbert, 1977; Nielsen et al., 1993), establishing that ecDNA can integrate back into the chromosome as an homogenously staining region (HSR) (Inazawa et al., 1989), tumor cells could develop methotrexate resistance through amplification of the DHFR gene on ecDNA (Kaufman et al., 1979), and fluorescence in-situ probes (FISH) probes can be used to map ecDNA in tumor cells (Inazawa et al., 1989). However, these early analyses were restricted to cell lines from specific cancer types or to very small cohorts of patients. Genomically characterizing ecDNA across the full landscape of human cancers was hampered by the lack of technological approaches (Benner et al., 1991; Spriggs et al., 1962; Kaufman et al., 1979). In 2017, Mischel and colleagues used computer vision techniques on DAPI-stained metaphases of cells from cell lines, patient-derived xenografts, and tumor tissue to perform unbiased detection and quantification of ecDNA in glioblastoma. These and subsequent results (Kim et al., 2020b; Wu et al., 2019) allowed the development of bioinformatics tools, such as AmpliconArchitect (Deshpande et al., 2019), that can detect and reconstruct the fine structure

of ecDNA from short-read sequencing data. Subsequent application of AmpliconArchitect to pan-cancer genomics datasets confirmed the presence of ecDNA in numerous tumor types (Kim et al., 2020b). Further analyses have shown that ecDNA exhibits distinct properties from linear DNA, including an altered epigenetic landscape (Wu et al., 2019), non-Mendelian inheritance during mitosis (Lange et al., 2022), and interactions with other mutational processes such as APOBEC3 deaminases (Bergstrom et al., 2022). Importantly, circular amplicons are thought to be a potent form of oncogene amplification, harboring known cancer driver genes, and are associated with worse survival outcomes across multiple cancer types (Kim et al., 2020b; Bailey et al., 2024; Kim et al., 2024). Prior work characterizing the epigenetic landscape of ecDNA, predominantly using ATAC-seq and ChIP-seq data (Wu et al., 2019; Hung et al., 2021), demonstrated that extrachromosomal DNA is more open and accessible to transcriptional machinery than linear DNA. Consequently, the expression of genes contained on ecDNA is elevated compared to genes found on linear DNA, even after normalizing for copy-number (Kim et al., 2020b), although this feature of ecDNA is a point of contention. Work done by (Purshouse et al., 2022) found that this is not always the case when she demonstrated that the transcriptional activity per gene copy is comparable between ecDNA and the native chromosomal locus for EGFR-ecDNA. However, this phenomenon was later confirmed experimentally when a study used a pair of isogenic colorectal cancer cell lines derived from the same patient: COLO320DM (MYC amplification on ecDNA) and COLO320HSR (chromosomal MYC amplification on homogeneously staining region (HSR)) and showed that COLO320DM had a nearly 4-fold increase in accumulated RNA read density and nascent RNA from ecDNA, even after accounting for amplicon copy-number, compared to COLO320HSR (Tang et al., 2024). Gene fusions on ecDNA have also been observed in

multiple cancer types, most notably amplifications of the BCR-ABL1 fusion gene in chronic myeloid leukemia (Graux et al., 2004) and PVT-MYC in non-small cell lung cancer (Hung et al., 2021). Enhancer hijacking events, where a novel interaction is created between a co-amplified non-coding regulatory enhancer and an oncogene, has been found to be a prominent feature on ecDNA in multiple studies (Wu et al., 2019; Zhu et al., 2021). Taken together, the genetic and epigenetic landscapes of ecDNA have shown distinct features from the ones observed on linear genomic DNA. Due to its distinct properties from linear DNA and its absence from normal cells, ecDNA is a promising therapeutic target. For example, oral inhibitors of CHK1, which is thought to be essential for ecDNA+ tumors that undergo replication stress due to transcription-replication conflicts, are currently going clinical trials (Tang et al., 2024).

Lung cancer represents the second most common cancer type globally and remains the leading cause of cancer-related mortality worldwide. While tobacco smoking has historically been recognized as the primary risk factor, a substantial proportion of lung cancer cases—approximately 15–25%—occur in individuals who have never smoked. This distinct clinical entity, termed lung cancer in never-smokers (LCINS), constitutes the fifth leading cause of cancer mortality worldwide when considered independently, underscoring its significant public health impact. The clinical and genomic characteristics of LCINS differ notably from those observed in smoking-associated lung cancers. Adenocarcinoma represents the predominant histological subtype in never-smokers, in contrast to the more diverse histological spectrum observed in smokers. Despite the substantial disease burden, the etiology of LCINS remains poorly understood. Only a limited number of risk factors have been identified, including environmental tobacco smoke exposure, radon, air pollution, and occupational carcinogens, yet these account for only a small fraction of cases (Landi et al., 2021). Our

previous study on 232 LCINS (Zhang et al., 2021), found that copy-number alterations could define 3 distinct subtypes. However, this study did not distinguish focal amplifications in the form of extrachromosomal DNA or examine them in detail. Although pan-cancer studies have examined ecDNA in primary lung cancer cohorts, these studies either have been i) limited in sample size ii) failed to distinguish between never-smokers and smokers or iii) primarily analyzed smokers. Hence, in Chapter 3, I deliver the most comprehensive analysis of ecDNA in lung cancer to date, by examining 871 never-smokers and 345 smokers. Utilizing a highly diverse cohort of patients of varying ancestries and geographic locations, I construct a map of the landscape of extrachromosomal DNA by adapting a state-of-the-art algorithm for detecting ecDNA to account for tumors with low purity and a wide range of ploidy, a pervasive issue in lung tumors. I then conduct a series of epidemiological and genomic analyses to determine the key differences between ecDNA+ and ecDNA- tumors, as well as identify sequence-level features of ecDNA such as recurrent oncogenes and other genomic elements contained therein. These analyses reveal both previously known and unknown associations of genomic features with ecDNA, including whole-genome doubling, chromothripsis, and specific driver mutations. Finally, I show that tumors harboring ecDNA in never-smokers are associated with significantly poorer survival compared to those with chromosomal amplifications. Overall, this dissertation provides tools to analyze large-scale alterations in cancer and presents the most comprehensive examination of extrachromosomal DNA in lung cancer to date.

2 Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator

Note: This chapter has been previously published as:

Azhar Khandekar et al. (2023). Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator. *BMC Genomics*. DOI: 10.1186/s12864-023-09584-y

2.1 Abstract

All cancers harbor somatic mutations in their genomes. In principle, mutations affecting between one and fifty base pairs are generally classified as small mutational events. Conversely, large mutational events affect more than fifty base pairs, and, in most cases, they encompass copy-number and structural variants affecting many thousands of base pairs. Prior studies have demonstrated that examining patterns of somatic mutations can be leveraged to provide both biological and clinical insights, thus, resulting in an extensive repertoire of tools for evaluating small mutational events. Recently, classification schemas for examining large-scale mutational events have emerged and shown their utility across the spectrum of human cancers. However, there has been no computationally efficient bioinformatics tool that allows visualizing and exploring these large-scale mutational events. Here, we present a new version of SigProfilerMatrixGenerator that now delivers integrated capabilities for examining large mutational events. The tool provides support for examining copy-number variants and structural variants under two previously developed classification schemas and it supports data from numerous algorithms and data modalities. SigProfilerMatrixGenerator is written in Python with an R wrapper package provided for users that prefer working in an R environment. The new version of SigProfilerMatrixGenerator provides the first standardized bioinformatics tool for optimized exploration and visualization of two previously developed classification schemas for copy-number and structural variants. The tool is freely available at <https://github.com/AlexandrovLab/SigProfilerMatrixGenerator> with an extensive documentation at <https://osf.io/s93d5/wiki/home/>.

2.2 Introduction

Large-scale cancer genomics projects, including, The Cancer Genome Atlas (TCGA) and the Pan-cancer Analysis of Whole Genomes (PCAWG) initiatives, have comprehensively surveyed the molecular landscapes of most types of human cancer (Consortium, 2020; Cancer Genome Atlas Research et al., 2013). These studies have provided a compendium of somatic mutations for each examined cancer genome and revealed both the mutations driving cancer development and the processes generating most somatic mutations within each cancer (Consortium, 2020; Cancer Genome Atlas Research et al., 2013; Alexandrov et al., 2020). One commonly performed type of genomics analysis is the examination of mutational patterns within a set of cancer genomes and the extraction of mutational signatures that have generated these patterns (Alexandrov et al., 2020; Islam et al., 2022a). Historically, mutational patterns have been predominately examined in the context of small mutational events, which include single base substitutions (SBS), doublet base substitutions (DBS), and small insertions and deletions (IDs) (Alexandrov et al., 2020, 2013). Recent studies have also started exploring the patterns of large mutational events, including ones due to copy-number alterations and/or structural variations (Li et al., 2020; Steele et al., 2022). Previously, we developed a computational tool, termed, SigProfilerMatrixGenerator, designed exclusively for examining the mutational patterns of all types of small mutational events (Bergstrom et al., 2019). Here, we present a new version of SigProfilerMatrixGenerator that now provides the capabilities for optimized exploration and visualization of large mutational events.

Large mutational events, generally defined as genomic alterations greater than 50 base pairs, are an important class of somatic aberrations in human cancer (Alexandrov et al.,

2013). In principle, there are two commonly examined and closely interrelated types of large mutational events: *(i)* a structural variation (SV, also known as a genomic rearrangement), where a large-scale genomic segment gets altered; and *(ii)* a copy-number variation (CNV), where the number of DNA copies of a genomic segment gets modified. Not all structural variations are related to CNVs, as SVs do not necessarily alter the copy-number of a genomic segment; examples include copy neutral events such as inversions and reciprocal translocations. Similarly, not all changes in copy-number require prior SVs, as is the case of chromosomal duplications and whole-genome doubling. Importantly, SVs and CNVs also differ in the types of genomics approaches that can detect them. In most cases, comprehensive detection of SVs requires whole-genome sequencing (WGS) data as it relies on either read alignment (Cameron et al., 2019) or genome assembly methods (Cosenza et al., 2022). In contrast, in addition to WGS data, CNVs can be detected from whole-exome sequencing, RNA-sequencing, single-cell sequencings approaches, and genotyping microarrays (Talevich et al., 2016; Chen et al., 2016a; Serin Harmanci et al., 2020).

Deciphering mutational signatures from catalogues of somatic mutations, a process known as *de novo* signature extraction, relies on a biologically meaningful classification of mutational events (Alexandrov et al., 2013). We previously created the mathematical concept of mutational signatures and provided a set of tools for deciphering signatures of small mutational (Islam et al., 2022a; Bergstrom et al., 2019). Mutational patterns of SBSs, DBSs, IDs, have been extensively explored with more than 100 distinct mutational signatures published in the literature (Alexandrov et al., 2020; Degasperi et al., 2022). These signatures reflect the activities of endogenous and/or exogenous mutational processes with approximately half of all signatures being, at least putatively, linked with a proposed etiology

(Alexandrov et al., 2020; Petljak et al., 2019; Riva et al., 2020; Alexandrov et al., 2015). Recently, mutational signature analyses of larger copy-number alterations and structural alterations have emerged (Li et al., 2020; Steele et al., 2022; Degasperi et al., 2020; Drews et al., 2022). A crucial first step in extracting mutational signatures is the derivation of features according to a predefined schema for mutational classification. This step involves transforming the mutational catalogues of a set of cancer genomes into a matrix, which is then amenable to subsequent matrix decomposition techniques (Bergstrom et al., 2019). Here, we present a computational package for classification of large-scale alterations and the generation of mutational matrices for signature decomposition. Two separate classification schemas are implemented: one for copy-number variations and one for structural variations. Both schemas were previously developed and applied to large cohorts of cancer samples (Steele et al., 2022; Degasperi et al., 2020). To the best of our knowledge, SigProfilerMatrixGenerator is the first tool that allows matrix generation and visualization of the CNV scheme used for generating the global reference set of Catalogue of Somatic Mutations in Cancer (COSMIC) copy-number signatures (Steele et al., 2022). SigProfilerMatrixGenerator’s capabilities for analyzing SVs and CNVs are implemented in Python and R, and the tool allows using multiple input formats, including segmentation and browser extensible data paired-end (BEDPE) files generated by commonly used algorithms for detecting copy-number variations and structural variations, respectively. Additionally, SigProfilerMatrixGenerator provides a comprehensive visualization of mutational patterns of large mutational events and an R wrapper package for users that prefer working within the R environment.

2.3 Implementation

2.3.1 Classification of copy-number variations

The schema for classifying copy-number variations is based on Steele et al. (Steele et al., 2022) and it utilizes allele-specific copy-number, which quantifies the number of segments for each allele at each variant loci rather than the total number of chromosome copies. In this schema, the copy-number profile of a sample can be represented by a mutational vector with 48 dimensions. Specifically, copy-number segments are categorized into three heterozygosity states: heterozygous segments with total copy-number (TCN) of $A > 0, B > 0$ (numbers reflect the counts for major allele A and minor allele B; Fig. 2.1a), segments with loss of heterozygosity (LOH) with total copy-number of $A > 0, B = 0$ (Fig. 2.1b), and segments with homozygous deletions and TCN of $A = 0, B = 0$ (Fig. 2.1c). Segments are further subclassified into 5 categories based on total copy-number, which reflects the sum of the copies on the major allele A and the copies on the minor allele B: $TCN = 0$, $TCN = 1$, $TCN = 2$, $TCN = 3$ or 4 , $TCN = 5$ to 8 , and $TCN \geq 9$. Each of these total copy-number states accounts for the phenomenon of whole-genome duplication—for example, a diploid ($TCN = 2$) state transitioning to a doubled state ($TCN = 4$), and a subsequent doubling of this state to $TCN = 8$ is accounted for by the $TCN = 5-8$ category (Fig. 2.1a). The categories for total copy-number have been chosen for biological relevance (Fig. 2.1): $TCN = 0$ reflects homozygous deletions, $TCN = 1$ represents a genomic deletion resulting in an LOH, $TCN = 2$ is equivalent to a diploid state including copy neutral LOH (a phenomenon whereby one of two homologous chromosomal regions is lost, but two identical copies of this region still remain; Fig. 2.1b), $TCN = 3$ or 4 reflect a gained state of tri- to tetra-ploidy, $TCN = 5$ to 8

represent a penta- to octo-ploidy state, and $\text{TCN} \geq 9$ represents high-level amplifications such as ones found in samples containing extrachromosomal DNA (ecDNA)(Kim et al., 2020b). Each of the heterozygous and LOH total copy-number categories are additionally subclassified into five additional categories based on the size of their segments: 0–100 kb, 100 kb–1 Mb, 1 Mb–10 Mb, 10 Mb–40 Mb, and >40 Mb. Three size bins are used for the additional subcategorization of homozygous deletions: 0–100 kb, 100 kb–1 Mb, and >1 Mb. The partitioning by segment sizes was chosen to ensure that a sufficient proportion of segments are classified within each category (Steele et al., 2022). This classification allows summarizing copy-number profiles using 48 distinct channels and can be represented using a vector with 48 components. For example, a sample harboring multiple focal amplifications, either contained on linear or extrachromosomal DNA, will have many events in the 9+ total copy-number category and the first 3 size bins (0–100 kb, 100 kb–1 Mb, 1 Mb–10 Mb; Fig. 2.1a, b). Conversely, a sample containing a large number of focal deletions or losses of entire chromosomes or chromosome arms will have numerous events in the LOH category, spanning all size bins (Fig. 2.1c, d). Another example will be a sample with a whole-genome doubling where copy-number changes will primarily encompass segments with large genomic sizes (10 Mb–40 Mb; 40 Mb) and total copy-number between 3 and 4 (Fig. 2.1e, f). Overall, this 48-channel classification schema can effectively summarize a diverse array of copy-number states seen across tumor types (Steele et al., 2022), whether they contain broad or focal events that result in amplifications or deletions.

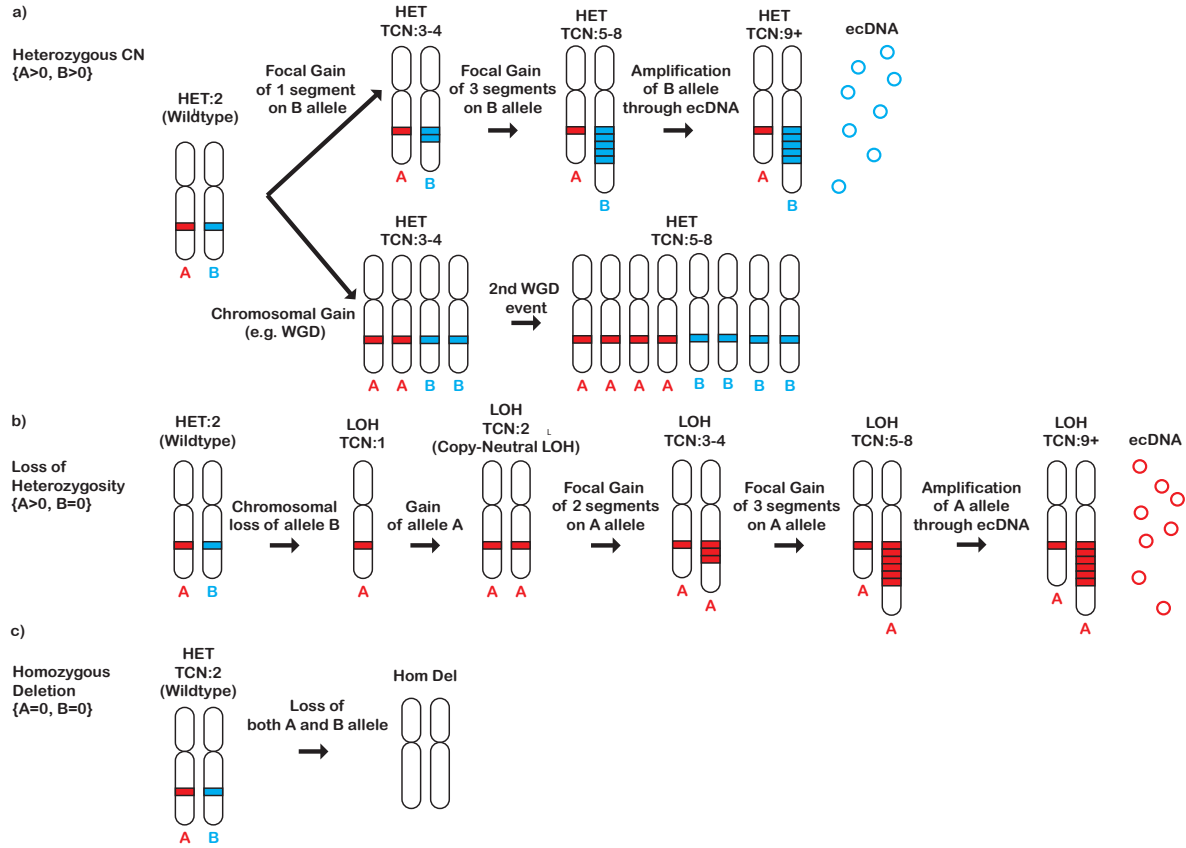


Figure 2.1: Description of the copy-number classification schema. The copy-number classification schema consists of 48 mutually exclusive channels, divided by heterozygosity status, segment size, and total copy-number (TCN). **(a)** In the heterozygous state, both alleles are retained and either one or both alleles can be amplified. This amplification can be focal (top panel) or it can encompass a chromosome or even the whole genome (bottom panel). The heterozygous category is further subdivided based on TCN (TCN=1, TCN=2, TCN=3 or 4, TCN=5 to 8, and TCN≥9). **(b)** In a state of loss of heterozygosity (LOH), one of the alleles is lost. The remaining allele can then be duplicated (i.e., copy-neutral LOH), and undergo additional amplification resulting in higher total copy-number states. The LOH category is further subdivided based on TCN (TCN=1, TCN=2, TCN=3 or 4, TCN=5 to 8, and TCN≥9). The heterozygous and LOH categories are further divided based on the size of the segment: 0–100 kb, 100 kb–1 Mb, 1 Mb–10 Mb, 10 Mb–40 Mb, and >40 Mb. High-level LOH or heterozygous amplifications (e.g., TCN=5 to 8 or TCN≥9) can be carried on extrachromosomal DNA (depicted as red circles) as well as on linear chromosomes. **(c)** Homozygous deletions result in the loss of both alleles and are divided based on the size of the deleted segment: 0–100 kb, 100 kb–1 Mb, and >1 Mb.

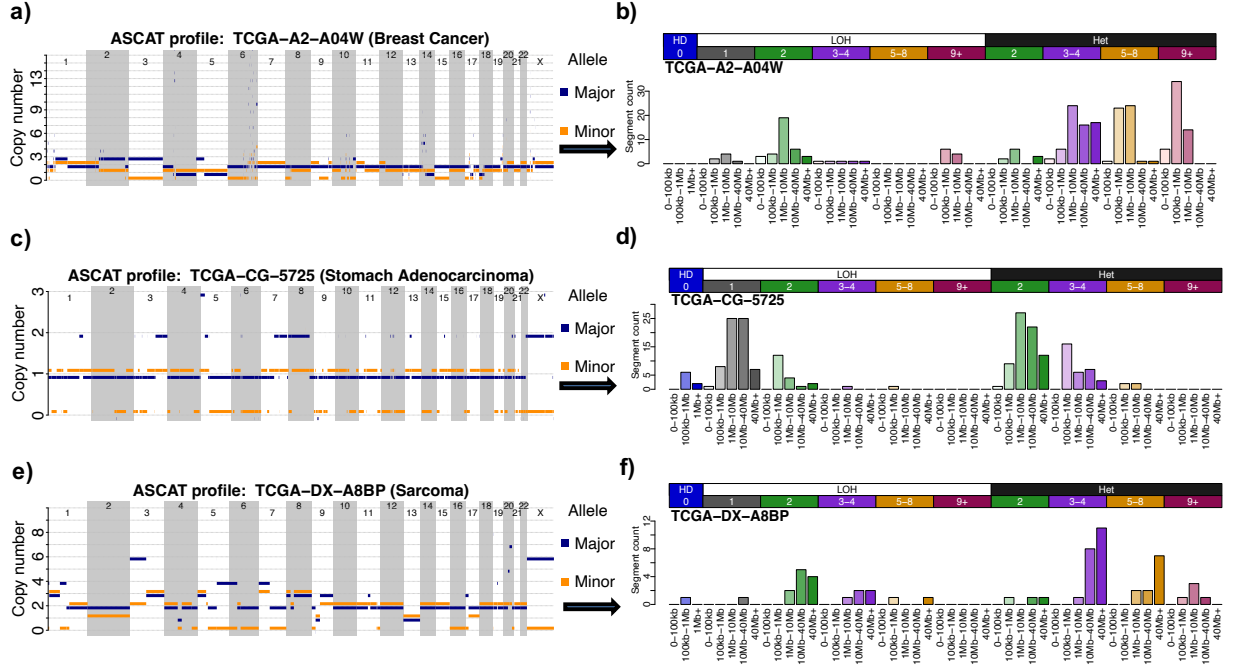


Figure 2.2: Converting copy-number segmentation profiles into copy-number mutational vectors. The CNV classification schema converts a sample's segmentation profile (a, c, e) into a count vector of 48 mutually exclusive components (b, d, f). These components are based on segment size, heterozygosity status, and total copy-number. A breast cancer sample with many highly amplified segments, possibly due to the presence of extrachromosomal DNA, is shown in (a, b). This sample's count vector is characterized by peaks in the 5–8 and 9+ total copy-number categories. A gastric cancer sample with extensive loss of heterozygosity is shown in (c, d). This sample's count vector is characterized by peaks in the LOH category, specifically with a total copy-number of 1 indicating a loss of an allele. A sarcoma sample with a whole-genome duplication event, characterized by peaks in the 3–4 total copy-number category and the 40+ Mb size bin, is shown in (e, f).

2.3.2 Input data for classifying copy-number variations

SigProfilerMatrixGenerator allows examining allele specific CNV data that, at a minimum, include the following information for each CNV segment: chromosome, start coordinate, end coordinate, and copy-number of both the minor and major alleles. Output files from the following tools for detecting CNVs are automatically supported: ASCAT(Raine et al., 2016), ABSOLUTE(Carter et al., 2012), Sequenza(Favero et al., 2015), FACETS(Shen and Seshan, 2016), Battenberg(Nik-Zainal et al., 2016), and PURPLE(Shale et al., 2022). Additionally, custom segmentation files from other CNV detection tools can be used if these files contain the aforementioned information.

2.3.3 Subclonal copy-number alterations

Recently, more advanced copy-number callers such as Battenberg(Nik-Zainal et al., 2016) determine not only the allele-specific copy-numbers for the dominant clone, but also for subclones, defined as a population of tumour cells that carry a unique subset of copy-number alterations. The tool seamlessly accounts for this type of data in a manner analogous to that of single-base substitutions, in that it adds these subclonal events on top of the clonal events, so the final count matrices represent the total counts of all clonal and subclonal alterations.

2.3.4 Classification of structural variants

A classification schema consisting of 32 features, based on (Nik-Zainal et al., 2016), is used to construct a mutational vector with 32 dimensions for each sample. In principle, each structural variant consists of two breakpoints which are at single-base resolution, where a

breakpoint is defined as a junction that indicates a structurally variable genomic segment greater than 50 base pairs (Cosenza et al., 2022). Breakpoints are typically detected using three signals from aligned sequencing reads: depth of sequence coverage, discordant read-pairs, and split read-pairs (Wala et al., 2018; Chen et al., 2016a; Rausch et al., 2012). Breakpoints can also be detected via genome assembly, where reads are assembled into contigs, the contigs are aligned to the reference genome, and these alignments are analyzed for structural variants (Cosenza et al., 2022). The previously developed classification of structural variants considers the following canonical SVs: tandem duplications, deletions, inversions, and translocations (Fig. 2.3). A tandem duplication refers to a segment of genomic material that has been duplicated and inserted on the same chromosome adjacent to the original segment (Fig. 2.3a). It should be noted that a tandem duplication is not necessarily the same as a copy-number amplification. For example, ecDNA copy-number amplifications are not tandem duplications as they are not inserted adjacent to the original chromosome segment. A somatic deletion is an event that has removed a set of existing base-pairs from a given location of a chromosome (Fig. 2.3b). An inversion is when a segment of the chromosome breaks off and reattaches at the same locus but in a reverse orientation (Fig. 2.3c). A translocation event occurs when a piece of one chromosome breaks off and some (or all) fragments from that piece re-attach to either another chromosome or to a different locus of the same chromosome (Fig. 2.3d). The classification schema bins all SVs, apart from translocations, according to the size of the event in base pairs: 0–10 kb, 10 kb–100 kb, 100 kb–1 Mb, 1 Mb–10 Mb, and >10 Mb (Nik-Zainal et al., 2016). Translocations, which may involve more than one chromosome, are not binned by size because they can be either balanced (where there is no net loss of genetic material on the chromosomes involved and thus the size can be described by one number) or unbalanced

(where there is a net loss or gain of genetic material on the chromosomes involved and thus the sizes of the segments cannot be described by just one number). Note that whether a translocation is balanced or unbalanced is not considered in this classification schema. The different types of SVs are then further divided into clustered and non-clustered events to account for the non-random distribution of these events along the genome. Clustered events are defined as events that occur closer to each other on a chromosome than purely expected by chance. These clusters often arise as a result of complex events, such as chromothripsis [30] or chromoplexy [31], generating many breakpoints in a single instantaneous event as opposed to the gradual accumulation of events over many cell cycles which results in more dispersed non-clustered events. Clusters of breakpoints can also form as a result of other mechanisms, including, for example, rearrangement hotspots in the genome (Glodzik et al., 2017). Clustering of SVs is determined based on a previously developed algorithm that utilizes the Potts' filter method (Winkler and Liebscher, 2002). This method segments a chromosome based on inter-mutational distance of SV breakpoints, and if the average distance in a particular segment is less than 10 times the average inter-mutational distance in the sample, all breakpoints in the segment are considered clustered. A minimum of 10 breakpoints must be present for a given segment to be considered clustered, otherwise all breakpoints in that segment are considered non-clustered. An example of a whole-genome sequenced bone cancer with a highly rearranged genome that contains chromosomes with clustered events as well as chromosomes with only non-clustered events is shown in Fig. 2.4a. For instance, in this sample, chromosome 12 contains a high number of SV breakpoints in close proximity to one another (Fig. 2.4b) and the SV pattern of this chromosome can be summarized in a vector with 32 components containing a high number of clustered SVs (Fig. 2.4d). In contrast,

chromosome 8 has SV breakpoints randomly scattered throughout the chromosome (Fig. 2.4c) and the SV pattern of chromosome 8 is exclusively one of non-clustered SVs (Fig. 2.4e).

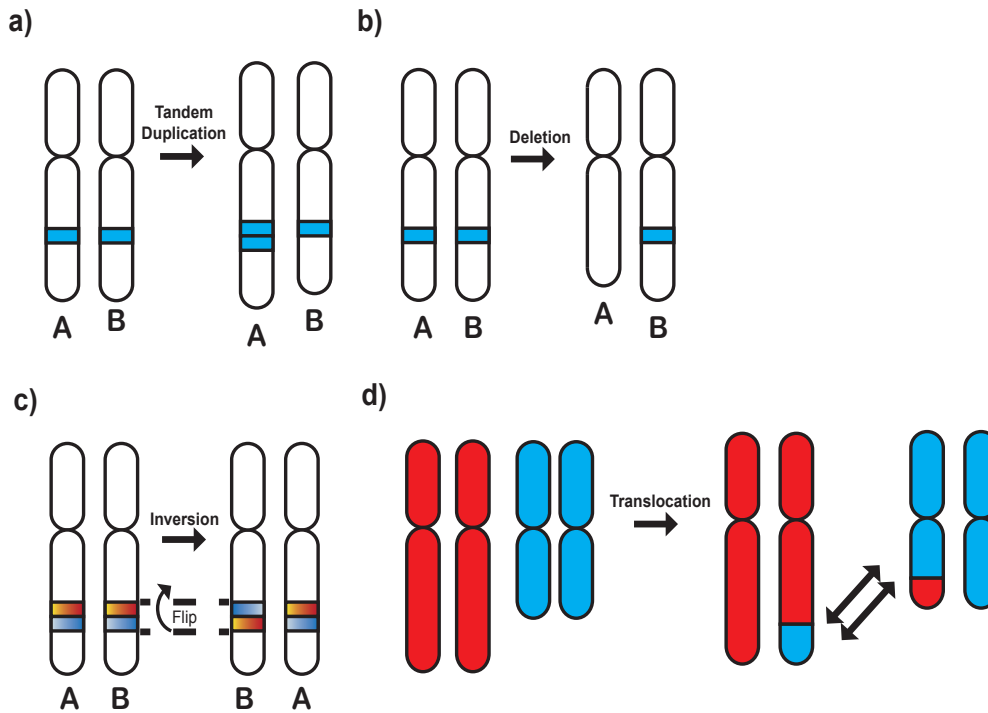


Figure 2.3: Description of the structural variant classification schema. (a) *Tandem duplication of a segment containing the A allele. A tandem duplication occurs when a segment is duplicated and inserted adjacent to the original chromosomal segment.* (b) *Deletion of the segment containing the A allele. A deletion occurs when there is a loss of genetic material from a chromosome.* (c) *An inversion of the segment containing the B allele. An inversion occurs when a segment breaks off and reattaches in a reverse orientation within the same chromosome.* (d) *A translocation of a chromosomal segment. A translocation event occurs when a piece of one chromosome breaks off and some (or all) fragments from that piece re-attach to either another chromosome or to a different locus of the same chromosome.*

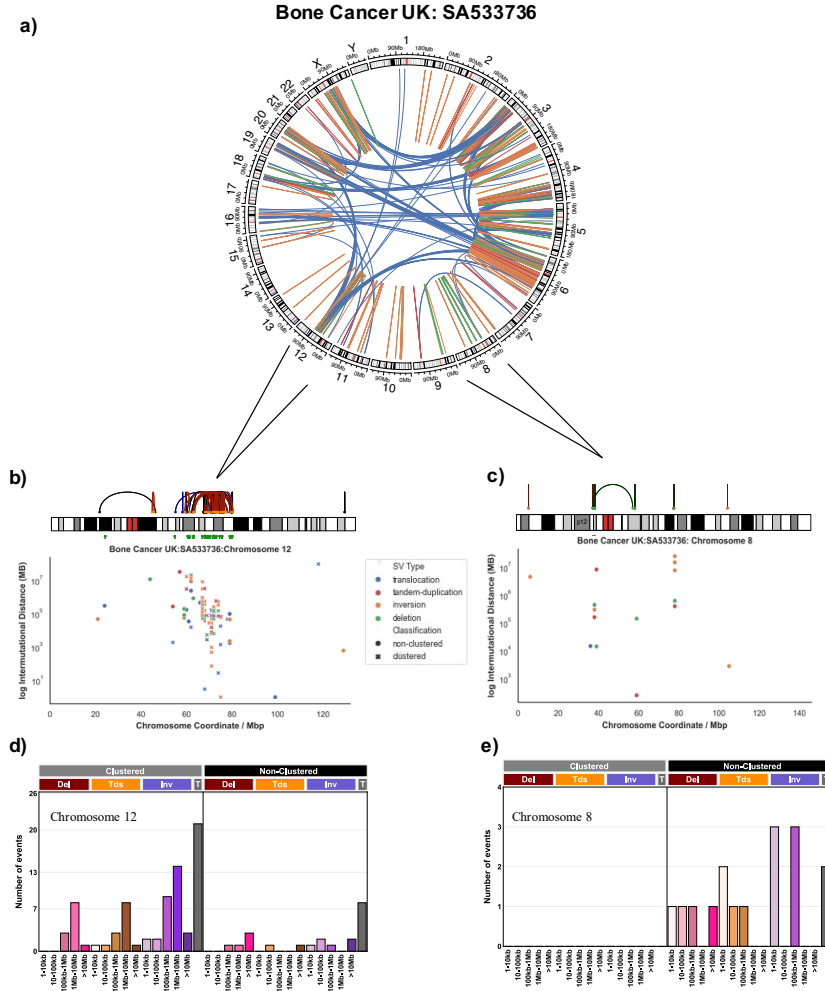


Figure 2.4: Classifying Structural Variants into Mutational Vectors. (a) An example of a bone cancer sample from PCAWG with a highly rearranged genome consisting of both clustered and non-clustered structural variants (SVs) is shown as a Circos plot representation. (b) Zooming into SVs specifically found on chromosome 12 in the bone cancer sample. SVs are shown as a linear representation (top) and as a rainfall plot (bottom). The rainfall plot depicts all breakpoints on chromosome 12 according to their genomic coordinate (x-axis) and the $\log_{10}$ inter-mutational distance (y-axis), which is the distance to the breakpoint immediately preceding it. The tendency of breakpoints to cluster in a specific genomic region on chromosome 12 due to a chromothripsis event is evident in all representations. (c) Zooming into SVs specifically found on chromosome 8 in the bone cancer sample. SVs are shown as a linear representation (top) and as a rainfall plot (bottom). The rainfall plot depicts all breakpoints on chromosome 8 according to their genomic coordinate (x-axis) and the $\log_{10}$ inter-mutational distance (y-axis), which is the distance to the breakpoint immediately preceding it. There are no clustered SVs on chromosome 8 because, per the SV classification schema, clustering requires a minimum of ten breakpoints in a segment of a chromosome. (d) The SV classification schema is applied to the SVs found on chromosome 12 in the bone cancer sample. SVs are classified by event type (denoted by color) and are binned according to size (0–10 kb, 10 kb–100 kb, 100 kb–1 Mb, 1 Mb–10 Mb, and > 10 Mb). (e) The SV classification schema is applied to the SVs found on chromosome 8 in the bone cancer sample. SVs are classified by event type (denoted by color) and are binned according to size (0–10 kb, 10 kb–100 kb, 100 kb–1 Mb, 1 Mb–10 Mb, and > 10 Mb).

2.3.5 Piecewise constant fitting to distinguish clustered SVs

Here I will provide exact details of the Python implementation to discriminate between clustered SV events (those arising from a single catastrophic event in a cell cycle such as chromothripsis) from non-clustered SV events (canonical SV events arising from double-strand breaks and error-prone DNA repair), and show why this particular implementation can be applied to any cancer genome from any cancer type.

In order to distinguish clustered SVs from unclustered SVs, a piecewise constant fitting algorithm (PCF) was used (Winkler and Liebscher, 2002). This methodology was first developed for a cohort of 560 breast cancers (Nik-Zainal et al., 2016), and here I provide a fast, robust and flexible implementation in Python that can be applied to cohort of samples from any cancer type. First, all pairs of SV breakpoints were considered as individual breakpoints and sorted according to reference genomic coordinate in each sample. The intermutational distance (IMD), defined as the distance from one breakpoint to the breakpoint immediately preceding it, was calculated for each breakpoint. These inter-mutational distances (y) are then segmented based on PCF, where the parameter λ is a penalty for controlling the number of segments, the parameter $kmin$ is the minimal number of breakpoints on a chromosome to perform segmentation (set to 10 by default):

$$\min_{\mathbf{z}} H_{\lambda}(\mathbf{z} \mid \mathbf{y}) = \min_{z_1, \dots, z_n} \left\{ \sum_{i=1}^n (y_i - z_i)^2 + \lambda \left| \{ i \mid z_i \neq z_{i+1} \} \right| \right\}. \quad (2.1)$$

λ is determined by multiplying gamma sdev(set to 25 by default) by the mean absolute deviation (MAD), which serves as a robust estimate of the variability (inherent noise) of the data. For a dataset $X = \{x_1, x_2, \dots, x_n\}$, the MAD is calculated as:

$$MAD = \text{median}(|x_i - \text{median}(X)|) \quad (2.2)$$

Thus, the noisier the data (higher MAD), the greater the penalty for introducing a new segment and vice versa (lower MAD corresponds to lower penalty for creating a new segment). The primary advantage of the MAD is the robustness to outliers, as the presence of extreme values will not affect the MAD as much as the standard deviation. The idea is to adapt the segmentation penalty to the inherent noise in the data. If the data is noisy and jumps around a lot, a stronger penalty is required for creating a new segment in order to prevent over-segmentation. Only large changes in the IMD should warrant a new segment. Conversely, if the distances are consistent with little noise, a lower penalty is appropriate. This flexible adaptation to the data ensures this methodology is applicable to all cancer genomes. Once the data is segmented, an overall sample-dependent mean intermutational distance is calculated for each sample by summing all the IMDs and dividing by the total. If the mean inter-mutational distance is 10X less in a particular segment than the expected distance, all breakpoints (and hence SV events corresponding to those breakpoints) are considered clustered, otherwise, they are considered unclustered. Note that if there are less than 10 breakpoints on the chromosome, all the SV events corresponding to those breakpoints are considered unclustered. An output file with the clustered/unclustered status for each SV event is provided by the tool.

2.3.6 Handling of VCF input

The majority of somatic structural variant calls are output in BEDPE format, but recently, VCF format is also growing in popularity. In fact, in the case of SV calls such as those generated by the Gnomad project, the primary output is VCF. In an effort to maintain the tool and make it as widely applicable as possible, I introduced an update to handle SV calls in VCF format as well (specifically VCF versions 4.1, 4.2, and 4.3). Internally, this is done by first converting the VCF to BEDPE format. This functionality has been thoroughly tested with VCF calls generated from Manta, Delly, and Lumpy.

2.3.7 Input data for classifying structural variants

SigProfilerMatrixGenerator allows examining SV data that contains genomics information for each of the two breakpoints of a structural variant. In principle, the tool can process files in browser extensible data paired-end (BEDPE) format that, at a minimum, contain the following six columns: chrom1, start1, end1, chrom2, start2, and end2. Here, the genomics coordinates of the first breakpoint are annotated as chrom1, start1, and end1, while the genomics coordinates of the second breakpoint are provided as chrom2, start2, and end2. If the type of SV has been predetermined, then its annotation can be provided using a column named svclass. Otherwise, the columns strand1 and strand2, which indicate the strands of the read mate-pairs, are required. If the mates are on the same chromosome, the convention followed is inversion (+/- or -/+), deletion (+/+), and tandem-duplication (-/-). If mates are on different chromosomes, the SV is automatically classified as a translocation. SigProfilerMatrixGenerator supports SV in BEDPE format, which is utilized by most bioinformatics

tools for detecting SVs, as well as being the native output files from BRASS(Nik-Zainal et al., 2016).

2.3.8 Evaluating the channel distributions for CN48 and SV32 schema in large pan-cancer datasets

In order to determine the applicability and utility of the SV32 and CN48 schemas, and to test the speed and accuracy of the generation of matrices for large mutations by SigProfilerMatrixGenerator, I applied the tool to large, pan-cancer datasets. First, I applied the tool to 9873 samples that were profiled with SNP arrays in TCGA, and showed that every channel was represented across this pan-cancer dataset (Fig. 2.5). I also applied it to 2600 samples in PCAWG that were whole-genome sequenced (Fig. 2.6), and here too every channel was represented, indicating the applicability of the CN48 schema across data modalities. Lastly, I applied the SV matrix generation to the same 2600 whole-genome sequenced samples in PCAWG (Fig. 2.7), and here too every channel was represented and the tool generated the matrix in a matter of seconds.

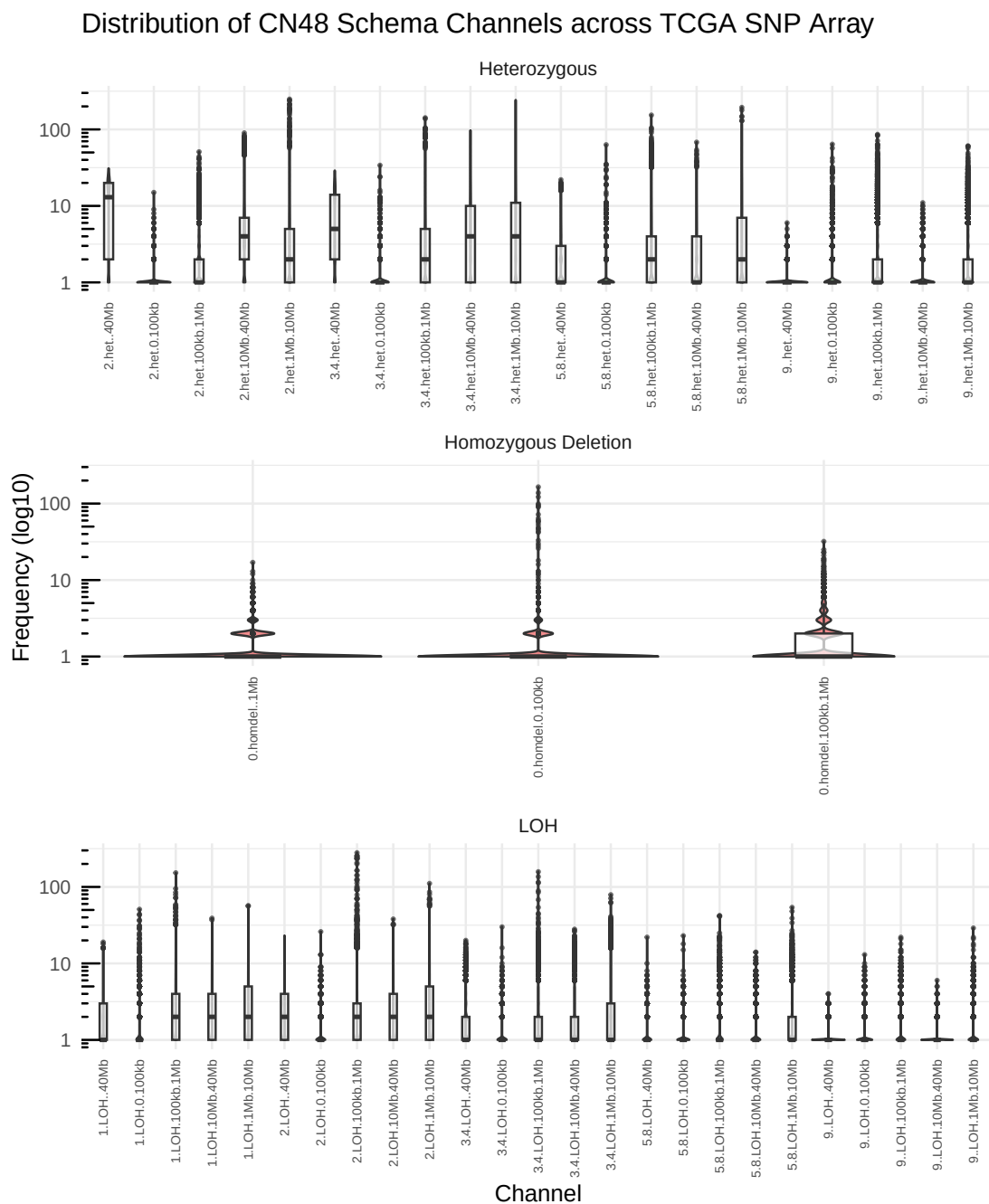


Figure 2.5: Distribution of CN48 schema channels for TCGA samples profiled with SNP6 array data. Application of CNVMatrixGenerator to TCGA samples profiled with SNP6 array data. Each of the 48 channels are split by the heterozygosity status, and the distribution of the log10 counts is shown on the y-axis.

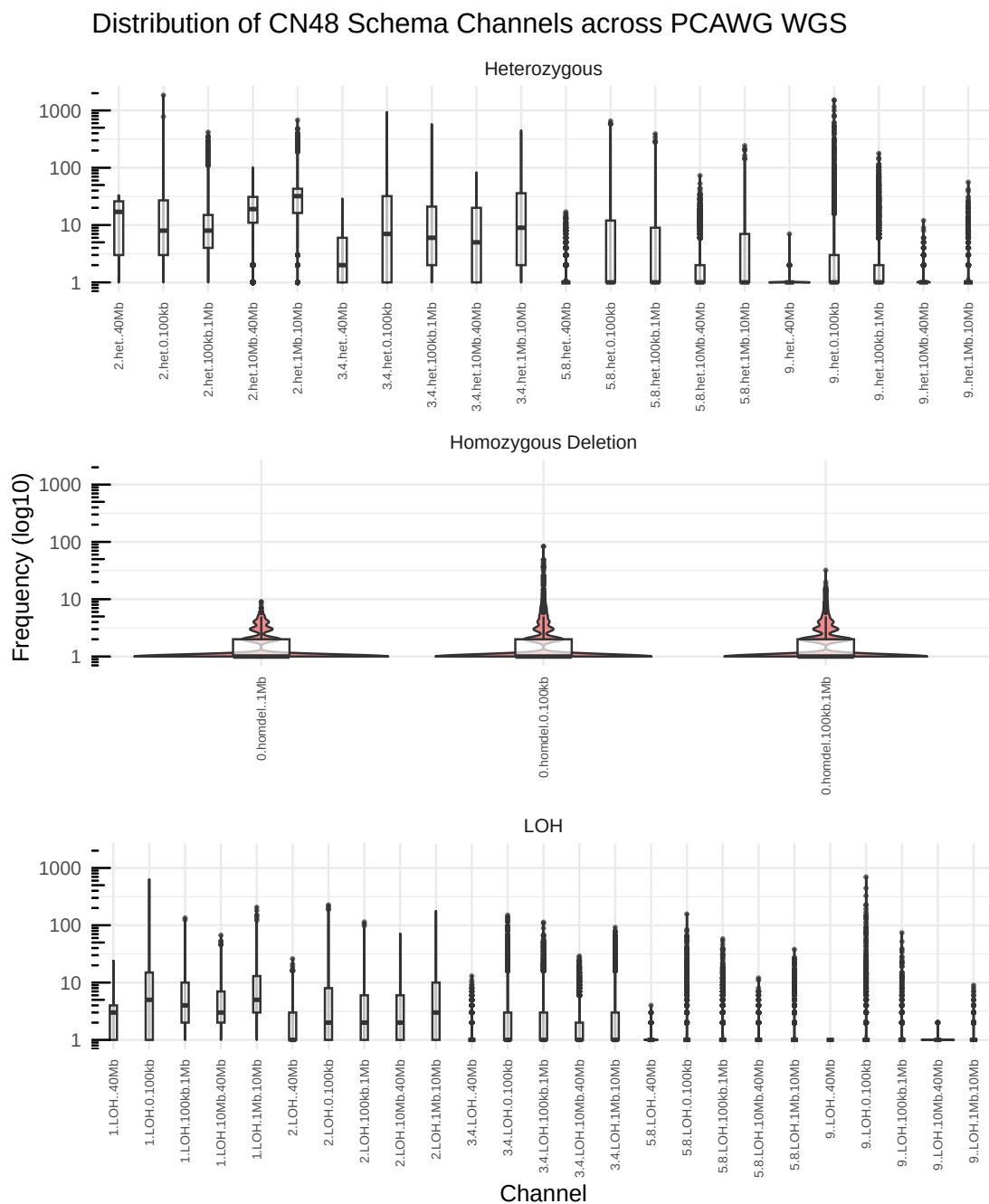


Figure 2.6: Distribution of CN48 schema channels for PCAWG samples profiled with WGS. Application of *CNVMatrixGenerator* to PCAWG WGS samples. The CN48 schema is split by the heterozygosity status, and the distribution of the log10 counts is shown on the y-axis.

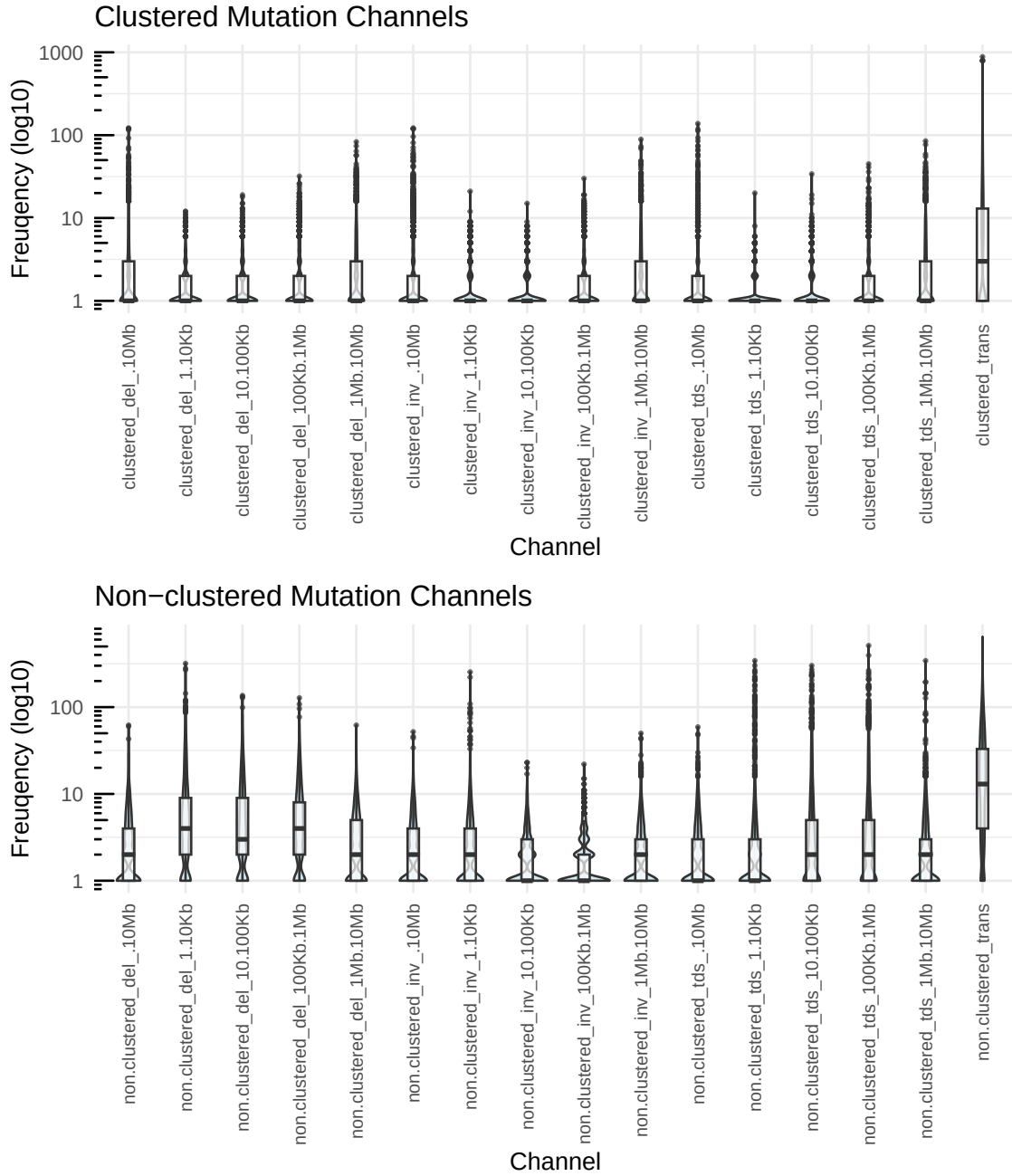


Figure 2.7: Distribution of SV32 schema channels for PCAWG samples profiled with WGS. Application of the SV32 schema to 2500 PCAWG WGS samples. The SV32 schema is split into clustered and non-clustered events, and the distribution of the log10 counts is shown on the y-axis

2.4 Discussion

The newly developed version of SigProfilerMatrixGenerator allows transforming a set of mutational catalogues of copy-number changes and structural rearrangements into matrices amenable to decomposition, including, subsequent mutational signature analysis. The tool provides support for two previously developed (Steele et al., 2022; Nik-Zainal et al., 2016) classification schemas for large mutational events and seamlessly integrates with other components of the SigProfiler software suite, such as downstream signature extraction with SigProfilerExtractor (Islam et al., 2022a) and visualization of both mutational patterns and signatures with SigProfilerPlotting (Bergstrom et al., 2019). Plots for CNV and SV patterns can now be generated for each cancer sample (as shown in Figs. 2 and 4), and plots for CNV and SV signatures are automatically generated following signature extraction from a cohort of samples. This enables a streamlined workflow for end-to-end analysis of mutational signatures of large-scale events. Additionally, SigProfilerMatrixGenerator rapidly scales to large datasets. For example, the tool can generate an SV count matrix for all 2,658 PCAWG samples in 3.6 s and a CNV count matrix for the entire TCGA array data (9,875 samples) in 14.3 s. SigProfilerMatrixGenerator is also the first tool to provide support for the 48 channel CNV schema across a wide variety of popular tools for detecting CNV. Importantly, this schema can be applied across several data modalities, including whole-genome sequencing, whole-exome sequencing, RNA-sequencing, single-cell sequencing approaches, and genotyping microarrays. In addition, SigProfilerMatrixGenerator is the first Python package that provides support for the 32 channel SV schema in a fast and intuitive manner with minimal preprocessing, and the only package to provide support for SV and CNV schemas in both a Python and R

environment.

2.5 Conclusion

A breadth of computational tools exists for exploring the patterns for small mutational events, including our initial implementation of SigProfilerMatrixGenerator(Bergstrom et al., 2019). We recently demonstrated that a classification of CNVs into 48 channels provides the means to better elucidate and understand the mutational processes operative in human cancer(Steele et al., 2022). Similarly, we and others have previously demonstrated that the classification of SVs into 32 channels can be used to understand the mutational processes giving rise to SVs across multiple cancer types(Degasperi et al., 2020). Our newly developed version of SigProfilerMatrixGenerator provides the capability to examine these classification schemas from cancer genomics sequencing data. The tool can scale to large datasets and will serve as foundation for future analysis of both mutational patterns and mutational signatures of large mutational events.

2.5.1 Future directions

The schema for classifying copy-number and structural variants presented in this chapter can be expanded upon in several ways. A framework specifically for whole-genome sequencing data can be developed. WGS data has increased resolution of copy-number alterations compared to SNP array data and WES data. Hence, the features corresponding to the size bins can be increased. The use of whole-genome sequencing data also enables the detection of structural alterations, meaning that both SV and CN events can be integrated

together in the feature set. Additionally, certain mutational processes likely leave footprints of both small mutations and large mutations. An example of this is homologous recombination deficiency, where SBS3 and specific SV signatures of tandem-duplications are imprinted upon the genome by this process. Hence, an integrated signature framework where all classes of mutations, large and small, are combined together in the initial counts matrix prior to decomposition would be promising future direction, with the caveat that specialized normalization schemes would have to be developed in order to overcome the challenge of the vastly different frequency distributions of small and large mutations in tumor genomes.

Acknowledgements

I would especially like to acknowledge Chris Steele and Nischalan Pillay for their help and guidance with copy-number signatures, and Raviteja Vangara for his help with the SVMMatrixGenerator tool.

Chapter 2, in full, is a reprint of the material as it appears in Visualizing and exploring patterns of large mutational events with SigProfilerMatrixGenerator. Khandekar, Azhar; Vangara, Raviteja; Barnes, Mark; Díaz-Gay, Marcos; Abbasi, Ammal; Bergstrom, Erik N.; Steele, Christopher D.; Pillay, Nischalan; Alexandrov, Ludmil B. BMC Genomics 24, 469 (2023). The dissertation author was the primary investigator and author of this paper.

3 Exploring the utility of Latent Dirichlet Allocation in extracting mutational signatures

3.1 Introduction

Prior analyses of mutational signatures for large-scale alterations have been exclusively based on nonnegative matrix factorization (NMF) approaches, which were originally developed for single base substitutions (SBS). NMF has been shown to provide interpretable mutational signatures using the SBS96 mutational context (Alexandrov et al., 2013). In contrast to single base substitutions, large-scale mutations pose a distinct problem of being non-additive events that overlap in genomic space, making interpretability of the signatures difficult. In this chapter, I explore alternative Bayesian methods for signature extraction based on generative statistical modeling. Specifically I examine the utility of Latent Dirichlet Allocation (LDA) algorithm, which is a probabilistic model for discrete count data commonly used in the topic modeling space. In this context, LDA can extract “meaningful” topics given a collection of documents (termed a corpus) using word frequencies, as well as weights associated with each topic. Importantly, LDA has been shown to output more interpretable topics than NMF in this context, as it deals with the issue of polysemy (multiple words that can be assigned to multiple topics) (Blei et al., 2003). LDA was first successfully applied to genomics data in the widely used tool STRUCTURE, which analyzes population structure by discovering populations and calculating admixture proportions using genotyping microarray data (Pritchard et al., 2000).

From a cancer biology perspective, LDA can be used to model cohorts of cancer patients as a collection of cancer genomes, where the somatic mutations in a genome are classified according to a biologically relevant schema (termed channels), and mutation counts falling within each channel are analogous to word frequencies. The weights of the topics derived from LDA can be considered the signature activities, which can serve as an approximation

of the number of mutations in a sample imprinted by a given signature. Importantly, both the distributions of the mutational channels over a signature and of the signatures within a sample are determined by Dirichlet priors, and these priors can be adjusted to reflect biological plausibility. The majority of existing LDA implementations either use the online variational Bayes algorithm (VB-LDA) (Hoffman et al., 2010) or a collapsed Gibbs sampling algorithm (Griffiths and Steyvers, 2004). Importantly, it has been shown that the variational Bayes algorithm is a numerical approximation to NMF, when the NMF objective function utilizes Kullback-Leibler divergence, if a uniform Dirichlet prior is assumed for VB-LDA.

3.1.1 Utilizing a Variant of an Existing Implementation of LDA Based on the Gibbs Sampling Algorithm

In order to evaluate the initial performance of the base LDA model in accurately extracting mutational signatures, I ran a custom version of an existing LDA implementation from the Python package `tomotopy`. Given a set of samples D , where each sample d is represented as a bag-of-mutations across N_d mutation channels and each mutation channel is represented as ω , the channel distribution of each signature and the signature distribution in each sample are modeled according to the following equation:

$$p(D \mid \alpha, \beta) = \prod_{d=1}^M \int p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn} \mid \theta_d) p(\omega_{dn} \mid z_{dn}, \beta) \right) d\theta_d \quad (3.1)$$

Here, the parameter α reflects the prior assumption about the signature distribution (i.e., activities) in each sample, and the parameter β reflects the prior assumption about the channel distribution in each topic (signature). θ_d represents the per-sample signature

proportions, ω_{dn} is an observed channel in a particular genome, z_{dn} is the latent signature assignment for that channel, and N_d is the total number of observed channels in sample d .

A symmetric α prior assumes that each signature is equally likely to occur in a sample. In contrast, a symmetric β prior assumes that each channel is equally likely to be assigned to a signature. Because these priors can be customized to reflect biological plausibility, the LDA model provides greater flexibility than other blind source separation approaches—especially non-negative matrix factorization (NMF). For example, instead of assigning channels to signatures with a uniform distribution, β can be adjusted to reflect the tendency of particular channels to associate more strongly with specific signatures.

Although this implementation converges more slowly than others such as variational Bayes, it can compute each iteration faster. Hence, we increased the number of iterations from 1000 to 10000 because the accuracy of Gibbs sampling methods increases with the number of iterations. The α parameter was set to 0.06 and the β parameter was set to 0.1.

3.1.2 Evaluation of LDA on a synthetic benchmarking dataset of SBS signatures

The dataset encompasses a total of 33 distinct scenarios with 1,900 simulated substitution signatures operative in more than 60,000 unique synthetic genomes. Importantly, this dataset has already been used to evaluate another 13 tools for *de novo* extraction of mutational signatures and applying our LDA approach will allow us to put it in the context of existing methods. The model was forced to extract the known number of ground truth signatures across 20 synthetic scenarios of single base substitutions resulting in identical

values for precision and recall. A true positive signature is defined as a signature with a cosine similarity > 0.90 to a ground truth signature. A false positive signature is defined as a signature with a maximum cosine similarity of less than 0.90 with all ground-truth signatures. The cutoff of a cosine similarity 0.90 is highly unlikely to happen purely by chance ($p\text{-value} = 5.90 \times 10^{-9}$) as two random nonnegative vectors are expected to have an average cosine similarity of 0.75 purely by chance (?). False negative signatures are defined as ground truth signatures that were not detected in the data. These metrics will enable the calculation of the algorithm’s precision, sensitivity, and F1 score.

The average F1 scores were computed for each of the 20 noiseless scenarios, and we compared the performance of LDA with the performance of the top five performing tools for extracting mutational signatures of single base substitutions (Figure 1). Although the LDA algorithm performed well on easy scenarios and was better than around half of the 14 published tools we benchmarked, it had poor performance on medium and hard scenarios, suggesting a need for further optimization of parameters and additional algorithmic refinement.

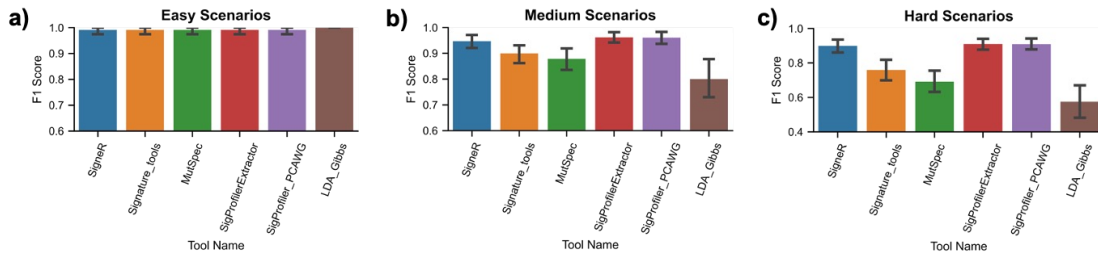


Figure 3.1: Evaluating the performance of LDA on 20 noiseless scenarios. Average F1 scores were calculated across 20 synthetic datasets for LDA with collapsed Gibb’s sampling method and the top five performing tools for extracting single base substitutions. Scenarios were divided into easy, medium, or hard based on the total number of operative signatures in each scenario. Values reflect average F1 scores and error bars correspond to 95% confidence intervals.

3.1.3 Conclusion and future directions

The base LDA algorithm using Gibbs sampling, without any optimizations or modifications, performs well on easy and medium scenarios but does not perform well on hard scenarios, compared to other state-of-the-art tools with optimizations. The α and β parameters are critical in LDA, and in my initial results I used $\alpha = 0.06$ and $\beta = 0.1$. Hence, I used symmetric priors, which assume that every signature has equal prior probability to be present in a sample (controlled by α) and every channel has equal prior probability to belong to a signature (controlled by β). While symmetric priors provide a reasonable starting point by not imposing prior preferences for specific signatures or mutation types, future work could explore asymmetric priors to better capture biological reality where certain signatures (such as the aging signature) are more prevalent across samples. Some criterion for automatic selection of k with LDA, using a metric like BIC, is also a future consideration. Ultimately, employing LDA on large mutations (SV and CN) will likely yield the most benefit as its flexibility and ability to handle sparse data will be advantageous over NMF.

Acknowledgements

I would like to thank Raviteja Vangara and Marcos Diaz-Gay for their help with various aspects of this chapter.

4 Examining the Role of Extrachromosomal DNA in 1,216 Lung Cancers

Note: This chapter has been submitted as a preprint as:

Azhar Khandekar, Phuc H. Hoang, Jens Luebeck, Marcos Díaz-Gay, Wei Zhao, John P. McElderry, Caleb Hartman, Mona Miraftab, Olivia W. Lee, Kara M. Barnao, Erik N. Bergstrom, Yang Yang, Martin A. Nowak, Nathaniel Rothman, Robert Homer, Soo-Ryum Yang, Qing Lan, David C. Wedge, Lixing Yang, Stephen J. Chanock, Tongwu Zhang, Ludmil B. Alexandrov, Maria Teresa Landi. *Examining the Role of Extrachromosomal DNA in 1,216 Lung Cancers*. doi: 10.1101/2025.06.03.657117

4.1 Introduction

Extrachromosomal DNA (ecDNA) is a potent mechanism for oncogene amplification in human cancer (Yan et al., 2024), exhibiting unique characteristics such as a circular structure (Wu et al., 2019), non-Mendelian inheritance (Bafna and Mischel, 2022; Lange et al., 2022), and an altered epigenetic and transcriptional landscape (Hung et al., 2021). Prior studies have linked ecDNA to worse overall survival in pan-cancer analyses, highlighting its clinical significance (Kim et al., 2020b; Bailey et al., 2024; Kim et al., 2024) and potential as a therapeutic target (Tang et al., 2024). Recent advances in computational techniques (Deshpande et al., 2019) have enabled large-scale detection of ecDNA from whole-genome sequencing (WGS) data. However, only three pan-cancer studies have examined ecDNA in primary lung cancer cohorts with over 100 samples, focusing predominantly on patients of European descent (Kim et al., 2020b; Bailey et al., 2024; Kim et al., 2024). Two studies (Kim et al., 2020b, 2024) analyzed lung cancer samples from the United States, reporting ecDNA prevalence rates of about 21% in 36 lung adenocarcinomas (LUAD) and 27% in 47 lung squamous cell carcinomas (LUSC). Additionally, a UK lung cancer study (Bailey et al., 2024) analyzed 14,778 patients across 39 tumor types, including 718 LUAD tumors and 378 LUSC tumors, detecting ecDNA in 8.2% of LUAD and 22.4% of LUSC cases. All three studies (Kim et al., 2020b, 2024; Bailey et al., 2024) reported lung cancer findings as part of broader pan-cancer analyses and did not distinguish between lung cancers in subjects who have never-smoked (LCINS) and subjects who have smoked (LCSS). Notably, based on the high prevalence of the tobacco-associated (Islam et al., 2022b) mutational signature SBS4 in these studies, more than 84% of the analyzed lung cancer cases were from individuals with a history of

smoking. Despite these prior studies, the clinical relevance of ecDNA in lung cancer remains unclear, including its prevalence across histologies, geographical areas, genetic ancestries, and biological sexes. Its role in never-smokers, smokers, and subjects who were exposed to passive smoking also requires further exploration.

4.1.1 Detailed description of the Sherlock Lung cohort

The Sherlock-Lung dataset consisted of whole-genome sequencing (WGS) data from 1,216 treatment-naïve lung cancers from diverse patient populations. This large international cohort enabled a comprehensive evaluation of ecDNA’s association with genomic alterations, as well as its prevalence across smoking status, biological sexes, ancestries, and 18 areas defined primarily by country of residence (Fig. 4.11). The dataset includes 871 lung cancers in never-smokers (LCINS), sourced from the Sherlock-Lung, the Environment And Genetics in Lung Cancer Etiology (EAGLE)(Consonni et al., 2018), The Cancer Genome Atlas (TCGA)(ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), and other publicly available cohorts(Imielinski et al., 2012; Lee et al., 2019; Leong et al., 2019), representing patients from 17 geographical areas(Landi et al., 2021). Additionally, it comprises 345 lung cancers in smokers (LCSS), drawn from EAGLE(Consonni et al., 2018), TCGA(ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020), and other publicly available cohorts(Imielinski et al., 2012; Lee et al., 2019; Leong et al., 2019), spanning five geographical areas(Landi et al., 2021). The majority of tumors were lung adenocarcinomas (LUAD; $n = 1,023$, 84.1%) or lung squamous cell carcinomas (LUSC; $n = 67$, 5.5%), with carcinoids ($n = 61$, 5%, exclusively in LCINS) and other rarer histological

lung cancer subtypes ($n = 65$, 5.3%) also represented. Amongst Sherlock-Lung LCINS, 250 were exposed to secondhand tobacco smoke, while 208 had documented non-exposure (Fig. 4.1). The majority of tumors in never-smokers were early stage (Stage I-II, $n=437$). In addition to the whole-genome sequencing data, transcriptomic data was available for 866 patients.

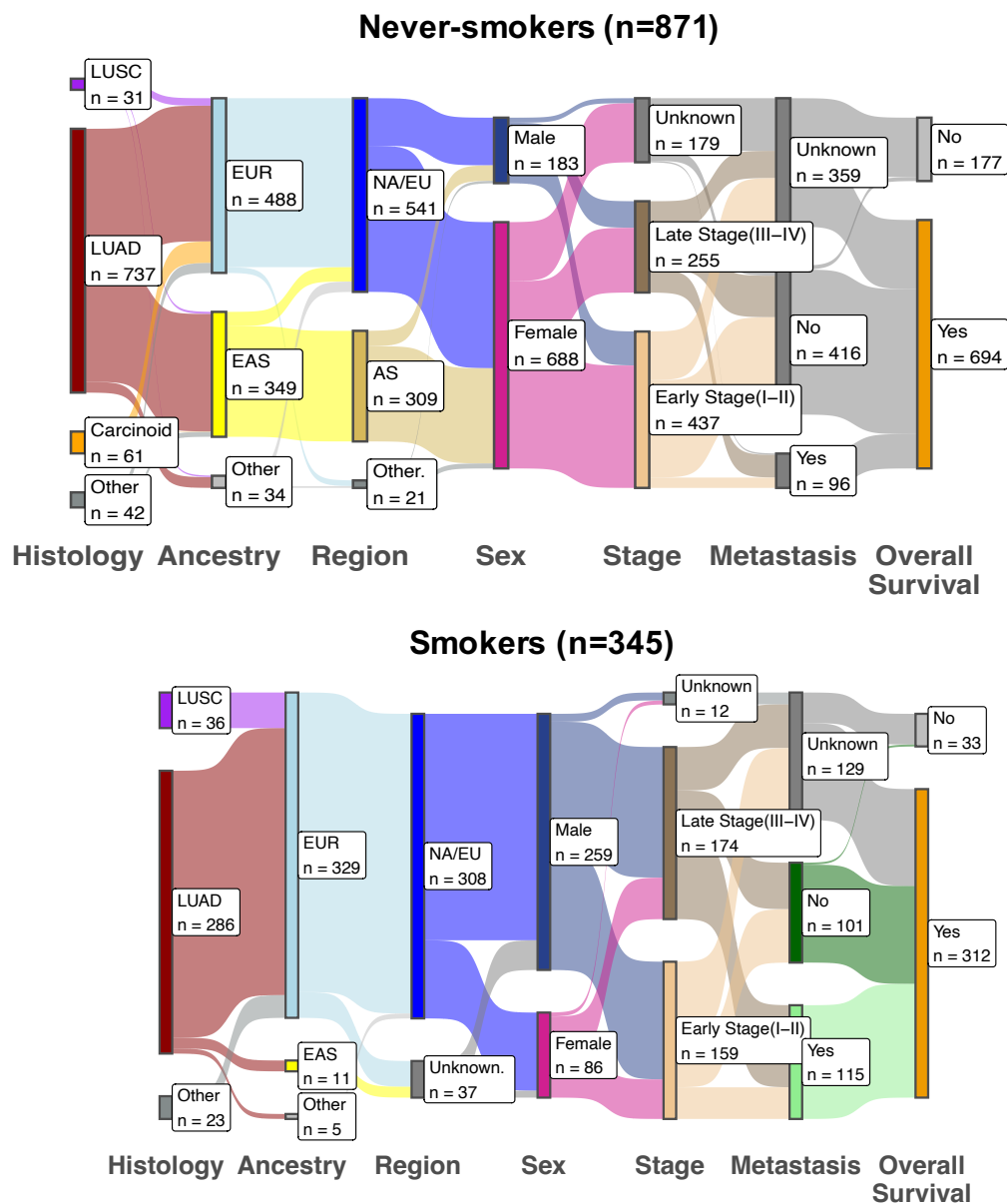


Figure 4.1: Sherlock-Lung cohort information Sankey diagrams showing the numbers of samples with histology, ancestry, geographic region, sex and the clinical variables stage, metastasis and overall survival for never-smokers (top) and smokers (bottom). NA/EU: North America or Europe region, AS: Asia region EAS: East Asian genetic ancestry super-sample, EUR: European genetic ancestry super-sample, LUAD: lung adenocarcinoma, LUSC: lung squamous cell carcinoma.

4.1.2 Constructing a highly sensitive pipeline to detect ecDNA that accounts for tumor purity and ploidy

In order to generate a map of extrachromosomal DNA across the cohort, I applied a modified version of the AmpliconArchitect pipeline (Deshpande et al., 2019). As a first step, this pipeline identifies a set of 'seed regions' to further explore, and these regions, by default, have to have a total copy number ≥ 4.5 . Consequently, the pipeline is highly reliant on the robust and accurate estimation of total copy number at the outset. The pipeline then jointly utilizes internal copy number calls and structural variant calls in these regions to construct a breakpoint graph where the nodes are amplified segments and the edges are SV's connecting the segments. Copy number calling from short-read sequencing is a complex procedure, and considerable variability exists between algorithms (Masood et al., 2024). Differences in the segmentation algorithm (e.g Hidden Markov model, circular binary segmentation) used can lead to differences in the overall segmentation profile (i.e # of segments). Additionally, the determination of a consensus tumor purity and tumor ploidy can also lead to differences in the estimation of total copy number of those segments. Most copy number callers rely on read depth, and some (e.g Battenberg, (Raine et al., 2016)) also incorporate the B-allele frequency, which allows for estimating the copy number each allele as it uses reads covering germline SNPS to determine the relative ratios of one allele to another. In addition, the presence of subclones often needs to be accounted for and can lead to non-integer copy-number estimates for a locus (e.g., if 10% of cells harbor a high-level amplification at *EGFR* with a total copy number (TCN) of 10, while the remaining 90% are diploid, the resulting bulk estimate is $0.10 \times 10 + 0.90 \times 2 = 2.8$ copies). The input copy number calls can be from any copy

number caller, but the default is CNVKit v. 0.9.6 (Talevich et al., 2016). Note that CNVkit, by default, assumes a sample purity of 100% and a ploidy of 2, as it does not correct for either purity or ploidy. Importantly, orthogonal validation of the AmpliconSuite pipeline has mainly been done in the context of cell lines, which are known to have high purity (Deshpande et al., 2019). However, a systematic pan-cancer analysis of tumor purity by (Aran et al., 2015) demonstrated that lung tissue is one of the tissue types with the lowest median purity. This analysis was done using SNP6 array in TCGA, but I confirmed this finding using the consensus copy number calls and purity estimation obtained from PCAWG (Fig. 4.2a). In the Sherlock-Lung cohort, a wide range of tumor purities were observed for both never-smokers and smokers (Fig. 4.2b). Utilizing a set of consensus tumor purity estimates from (Aran et al., 2015), I confirmed that Battenberg provides robust tumor purity estimates (Fig. 4.2d), as these purity estimates were highly concordant ($R^2 = 0.89$, $p = 8.96 \times 10^{-34}$). Hence, I developed an approach that combines the strengths of Battenberg (estimating purity and ploidy by using both BAF and read depth) and CNVkit (sensitive segmentation based on read depth) (Fig. 4.3). Specifically, CNVKit v. 0.9.6 (Talevich et al., 2016) was run in tumor-normal mode to call somatic CNVs against matched normal whole-genome-sequenced samples for each patient, using the log ratio of tumor to normal read depths ($\log R$) and the circular binary segmentation (CBS) algorithm. I then rescaled the total copy-number estimates using the purity and ploidy inferred by Battenberg according to Eq. 4.1:

$$\text{TCN} = \frac{\psi(2^{\log R}) - 2(1 - \rho)}{\rho}, \quad (4.1)$$

where ψ is the tumor ploidy, ρ is the tumor purity, and $\log R$ is the log ratio of tumor

to normal read depth. These copy-number calls were used to identify putative *seed* regions. I then applied AmpliconArchitect v. 1.3 (Deshpande et al., 2019) and AmpliconClassifier v. 1.3.1 to classify amplicons according to their most probable mechanism of formation (ecDNA, breakage–fusion–bridge, complex non-cyclic, or linear) and determine the genomic regions involved. The downsampling parameter was set to 40x as the Sherlock samples had increased coverage of 100x in tumor. Since the total copy number estimates were adjusted prior to input, the default copy number threshold for seeding was changed from 4.5 to 9 to more accurately capture high-level amplifications. Visual inspection and verification were performed for all ecDNA amplicon regions.

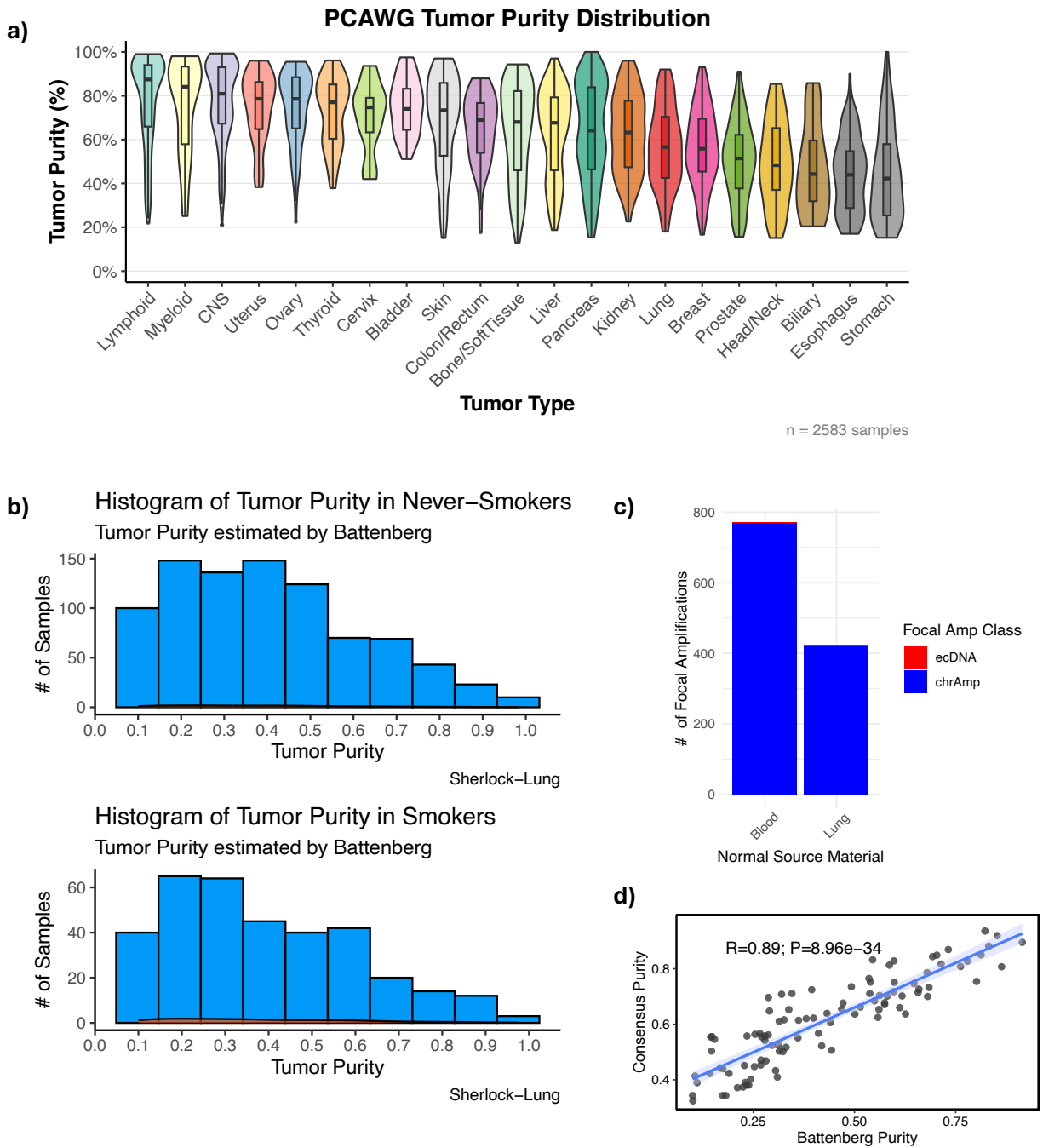


Figure 4.2: Tumor purity in Sherlock and pan-cancer cohorts. (a) Tumor purity distributions derived from 2583 samples in the PCAWG dataset, stratified by tumor type. Copy number calling and purity estimation was done using a consensus approach for these samples. (b) Tumor purity distributions for never-smokers and smokers in the Sherlock cohort. (c) Results from running the ecDNA detection pipeline on normal tissue for 1216 samples. The number of focal amplifications detected is indicated on the y-axis. (d) Scatter plot showing the concordance of tumor purity estimates derived from Battenberg vs. the consensus purity derived from orthogonal methods for the TCGA samples in Aran et al.

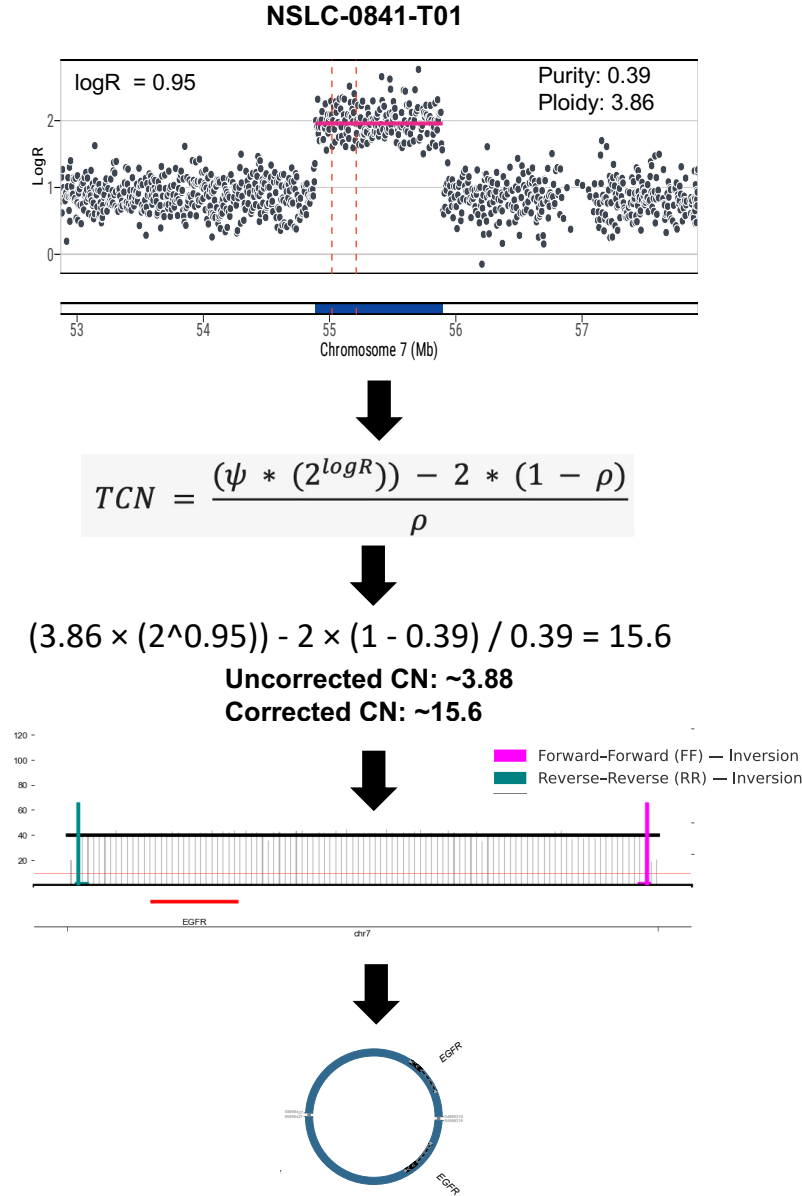


Figure 4.3: Detailed worked example of the hybrid rescaling seed identification methodology for identifying ecDNA Schematic showing a detailed worked example for the rescaled pipeline that considers tumor purity and ploidy. The logR track for the region containing the *EGFR* gene on chr7 and the purity and ploidy computed from Battenberg is indicated in one tumor as example (left). The exact location of the *EGFR* gene is shown by the dashed red lines. The equation that simultaneously accounts for sample purity and ploidy when computing the total copy number is shown, as well as the total copy number estimate when assuming a sample purity of 1.0 and ploidy of 2.0 (Uncorrected CN) vs. considering the purity and ploidy (Corrected CN). The Sashimi plot indicates the structural variants (colored arcs) and smoothed coverage around the *EGFR* region, which was annotated as ecDNA by AmpliconClassifier.

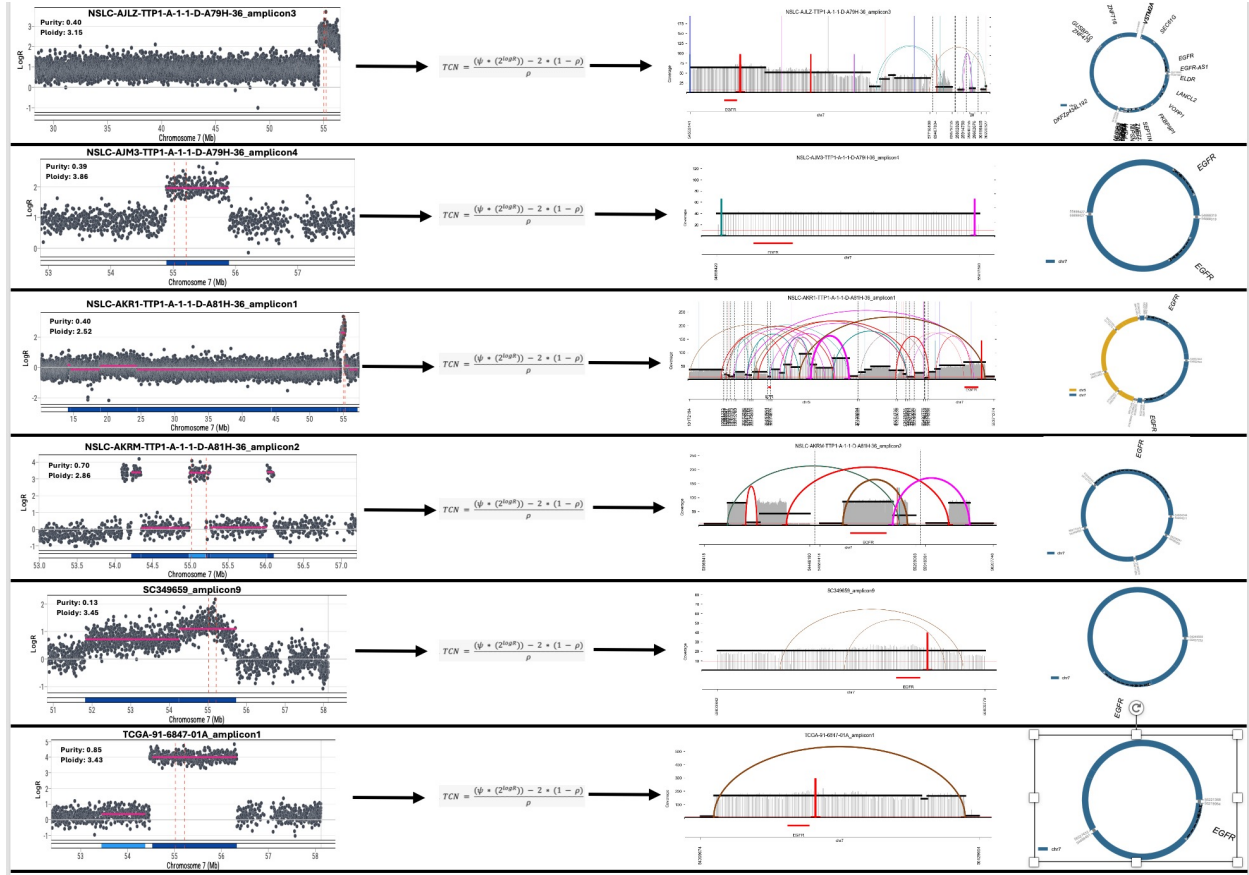


Figure 4.4: Representative example of the hybrid rescaling seed identification methodology with 6 samples containing EGFR ecDNA as exemplars. The logR (the log2 of the ratio of tumor read depth to normal read depth) is displayed (y-axis) for the chromosome region containing the EGFR oncogene (x-axis, red dashed line) as well as the surrounding region (left). The sample purity and ploidy as estimated by Battenberg is also indicated. The total copy number for each segment is calculated using the equation shown (middle), where the sample purity, ploidy and logR serve as inputs. The raw read coverage (y-axis) and structural variant breakpoints detected by AmpliconArchitect in the same region are indicated by the Sashimi plot. AmpliconClassifier reconstructed an ecDNA containing EGFR in all six of these samples after using the rescaling method for seed discovery, whereas the default method missed these seed regions, and hence did not detect these ecDNA.

This modified pipeline enabled low-purity samples to be analyzed, as it removed the effect of tumor purity on ecDNA detection (Fig. 4.5a). Consequently, more ecDNA⁺ samples were detected (Fig. 4.5a) and all 1 216 samples were included in the ecDNA analysis (Fig. 4.5a). Of the additional ecDNA identified, eight were EGFR-ecDNA (seven in LCINS and one in LCSS), and—except for one—each also harbored a driver mutation (Fig. 4.4). Among the eight EGFR-ecDNA⁺ samples, five had RNA-seq data, which showed very high *EGFR* expression (Fig. 4.6, consistent with the suspected ecDNA mode of amplification). Intriguingly, all 7 of these samples also contained an EGFR driver mutation Table ??.

These EGFR-ecDNA were predominantly simple ecDNA with no SVs detected within or around the region, likely generated by the simple deletion-excision method. It should be noted that strikingly similar EGFR-ecDNA have been found in glioblastoma and validated with orthogonal methods(Noorani et al., 2024). These ecDNA also contained mutations (EGFRvIII), were not co-amplified with other oncogenes, were hypothesized to be subclonal as timing analysis indicated they appeared before clonal expansion, and were simple in structure with no scars of chromothripsis evident.

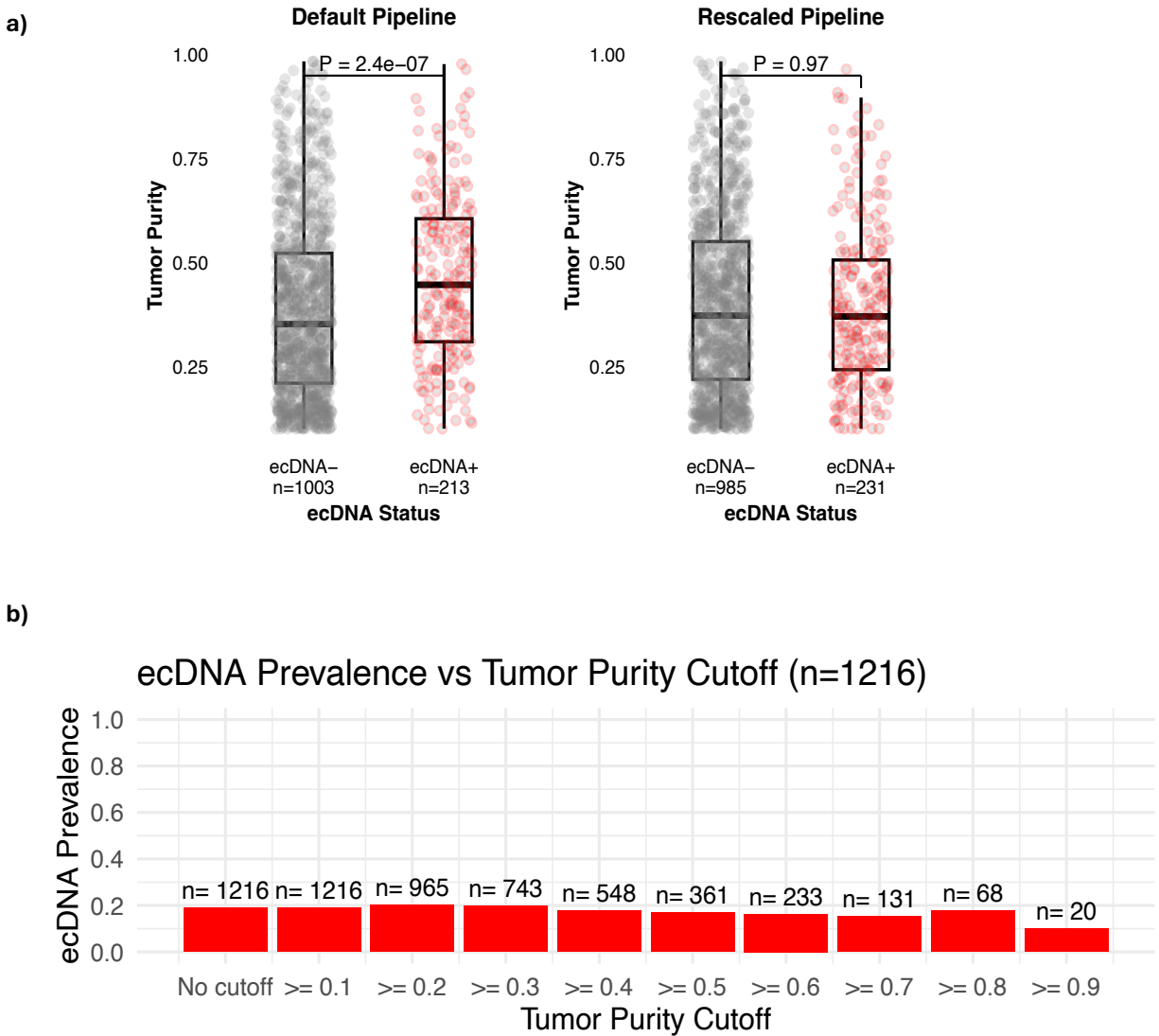


Figure 4.5: Default method vs rescaled method for identifying ecDNA *a)* Distribution of tumor purities for samples with an ecDNA annotation for both the default (left) and rescaled (right) pipelines. P-value for a Mann-Whitney test between ecDNA+ and ecDNA- samples is indicated. The total number of samples in each group is also indicated. *b)* The ecDNA prevalence (y-axis), which is the proportion of ecDNA+ samples over the total number of samples, for all possible tumor purity cutoffs applied to samples (x-axis). The total number of samples available for analysis after each tumor purity cutoff is applied is also indicated.

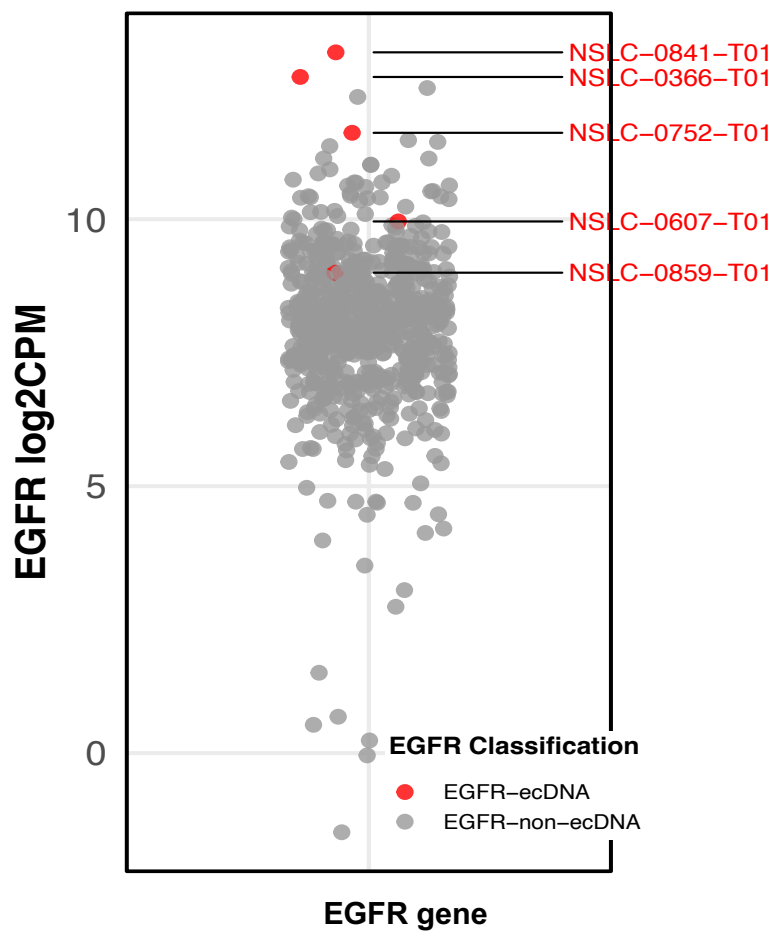


Figure 4.6: Gene expression of additional EGFR-ecDNA detected by the rescaled pipeline. Gene expression ($\log_2$ CPM) of the EGFR gene for all samples with RNA-seq data available. LCINS samples with EGFR-ecDNA are indicated in red.

4.1.3 Overall prevalence of ecDNA in lung cancer and epidemiological associations

A total of 231 samples harbored at least one ecDNA (18.9%), with similar prevalence between never-smokers (17%) and smokers (23%; q-value: 0.21; Fig. 4.7). The overall prevalence was consistent with previous estimates of ecDNA prevalence in lung cancer derived from pan-cancer studies, but these studies did not distinguish between smokers and never-smokers. This overall prevalence of ecDNA in lung cancer was higher than that reported for pan-cancer (14%, (Kim et al., 2020b)).

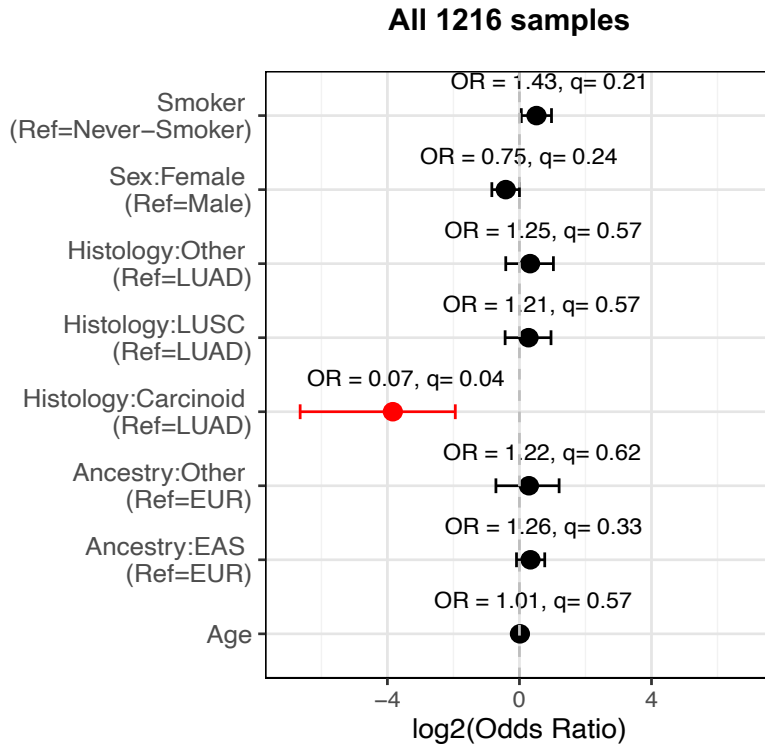
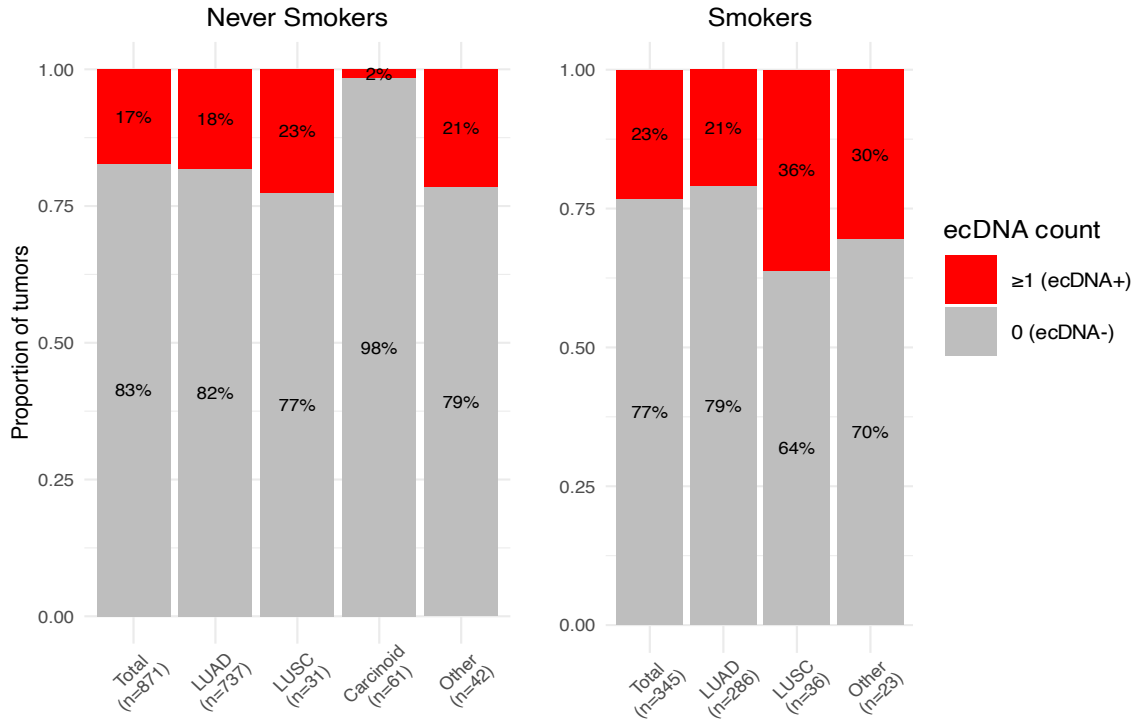


Figure 4.7: Landscape of extrachromosomal DNA in lung cancer across histologies, sex, ancestry, countries, and smoking status The number and proportion of tumors harboring ecDNA among LCINS and LCSS, stratified by histology (top). Samples with 1 or more ecDNA were classified as ecDNA+, and samples with no ecDNA were classified as ecDNA-. Forest plot of the full cohort (LCINS and LCSS) from a multivariate logistic regression analysis of key epidemiological and clinical features and the presence of ecDNA in the tumors (bottom).

4.1.4 Methodological approach to examine differences between ecDNA+ and ecDNA- samples

In order to analyze differences between ecDNA+ samples and ecDNA- samples, I utilized multivariate logistic regression with two classes: samples that had ≥ 1 ecDNA detected, and samples that had no ecDNA detected. This technique allowed (i) the inclusion of covariates such as age, sex, histology, and ancestry, which were important to correct for when working with a diverse cohort, and (ii) greater interpretability of the coefficients (odds and odds ratios), as the goal of this analysis was to understand what features are associated with ecDNA. For the remainder of this chapter, I separate the analyses presented into never-smokers and smokers as these are entirely distinct types of lung cancer at the genomic level.

4.1.5 Analyzing the presence of ecDNA with clinical characteristics such as stage and metastasis

Due to the fact that not all variables were collected for every sample, I ran separate models for passive smoking, tumor stage, and presence of metastases in both LCINS and LCSS so the statistical modeling of the other core variables (age, sex, histology, smoking and ancestry) would not be affected by missing data for these three variables. ecDNA was not associated with tumor stage (Fig. 4.8a-b), subsequent development of metastases (Fig. 4.8c-d) or passive smoking status (Fig. 4.10). These findings remained consistent when analyses were restricted to only LUAD cases in either LCSS or LCINS (Fig. 4.9, Fig. 4.10).

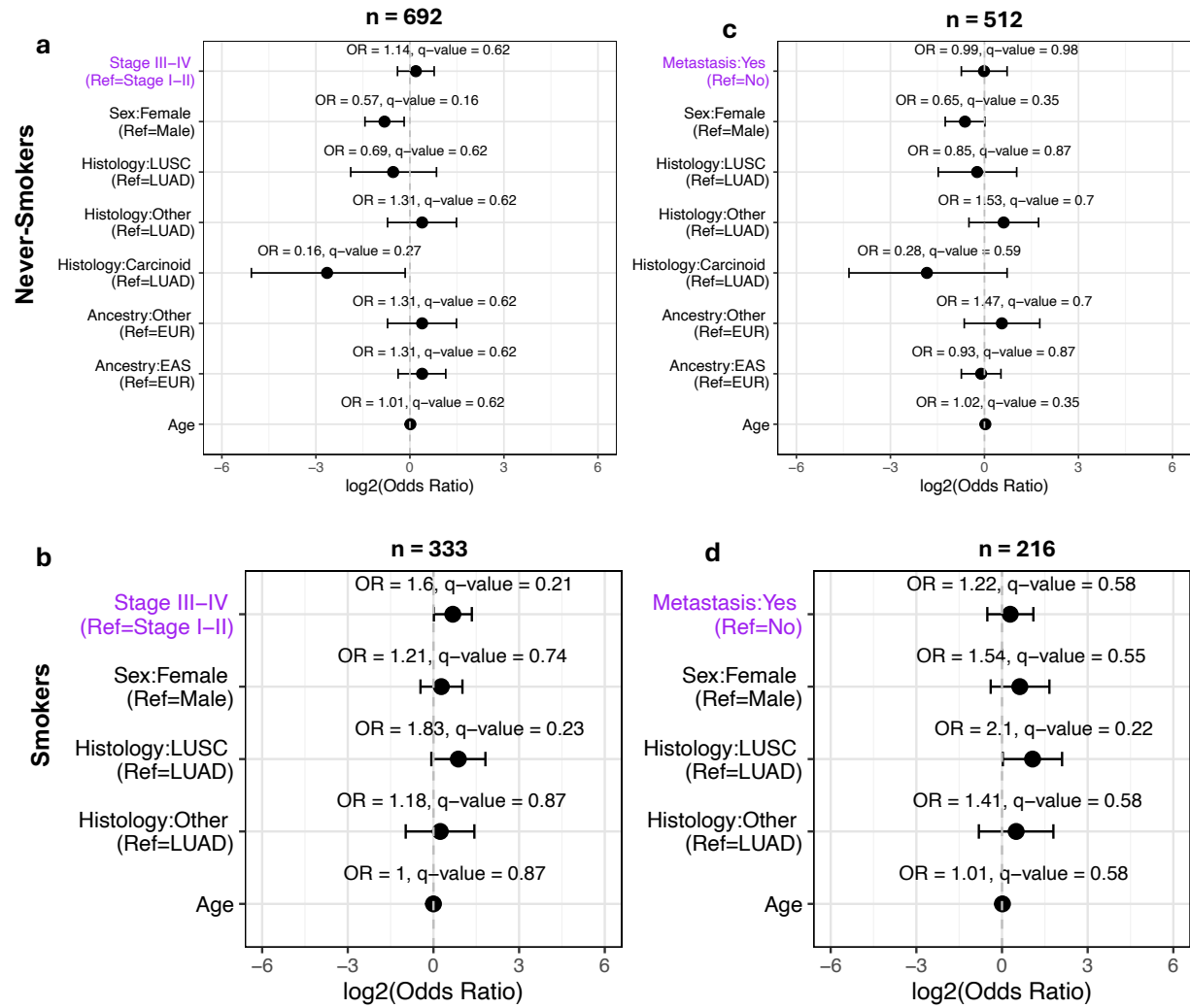


Figure 4.8: Association of ecDNA with stage and metastasis Forest plots for multivariate logistic regression models testing the clinical variables stage and metastasis with the presence of ecDNA in the tumors, for both all never-smokers (top) or LUAD only never-smokers (bottom).

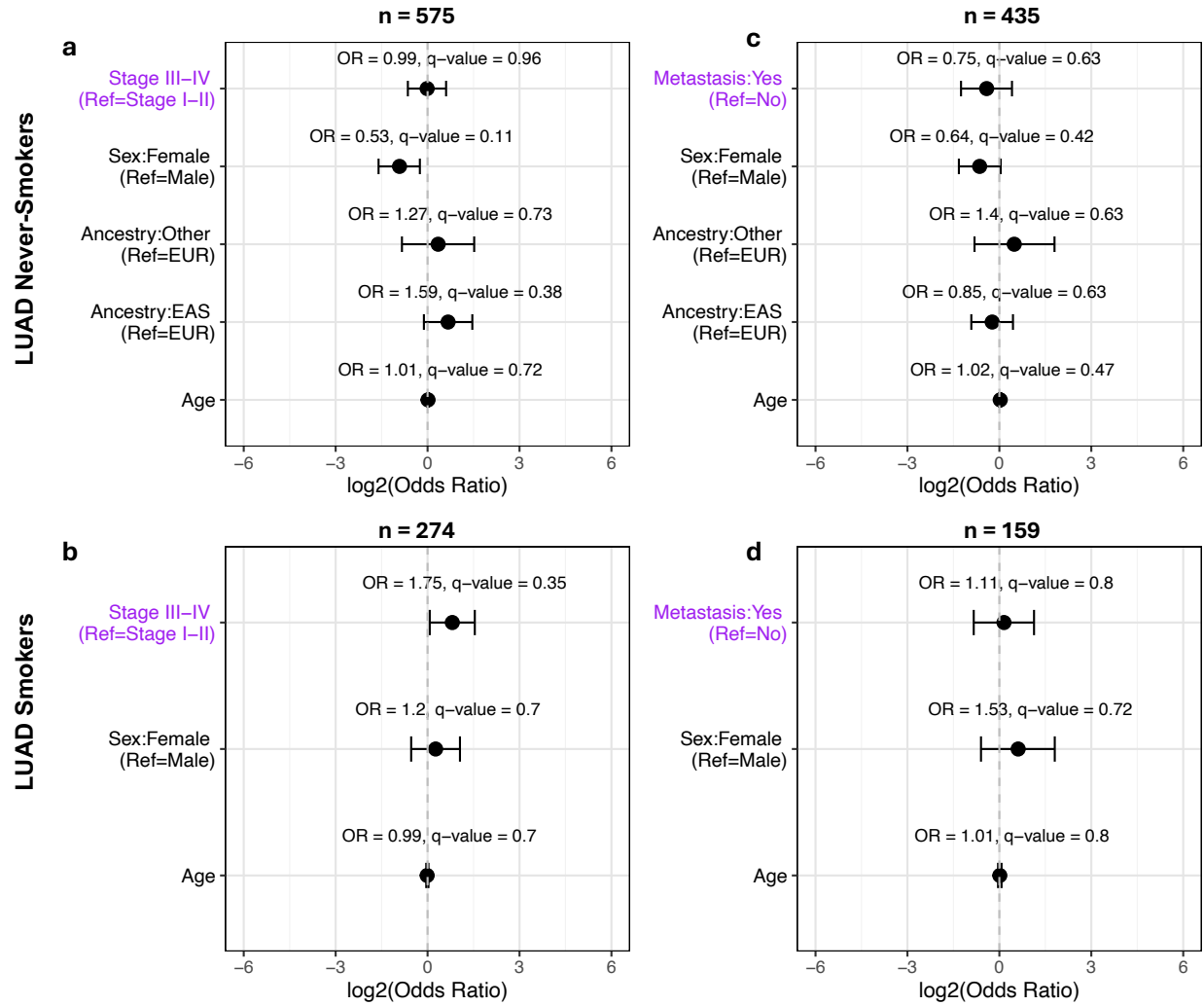


Figure 4.9: Association of ecDNA with stage and metastasis in LUAD samples only
 Forest plots for multivariate logistic regression models testing the clinical variables stage and metastasis with the presence of ecDNA in the tumors, for both all never-smokers (top) and smokers (bottom).

Passive Smoking & ecDNA in never-smokers

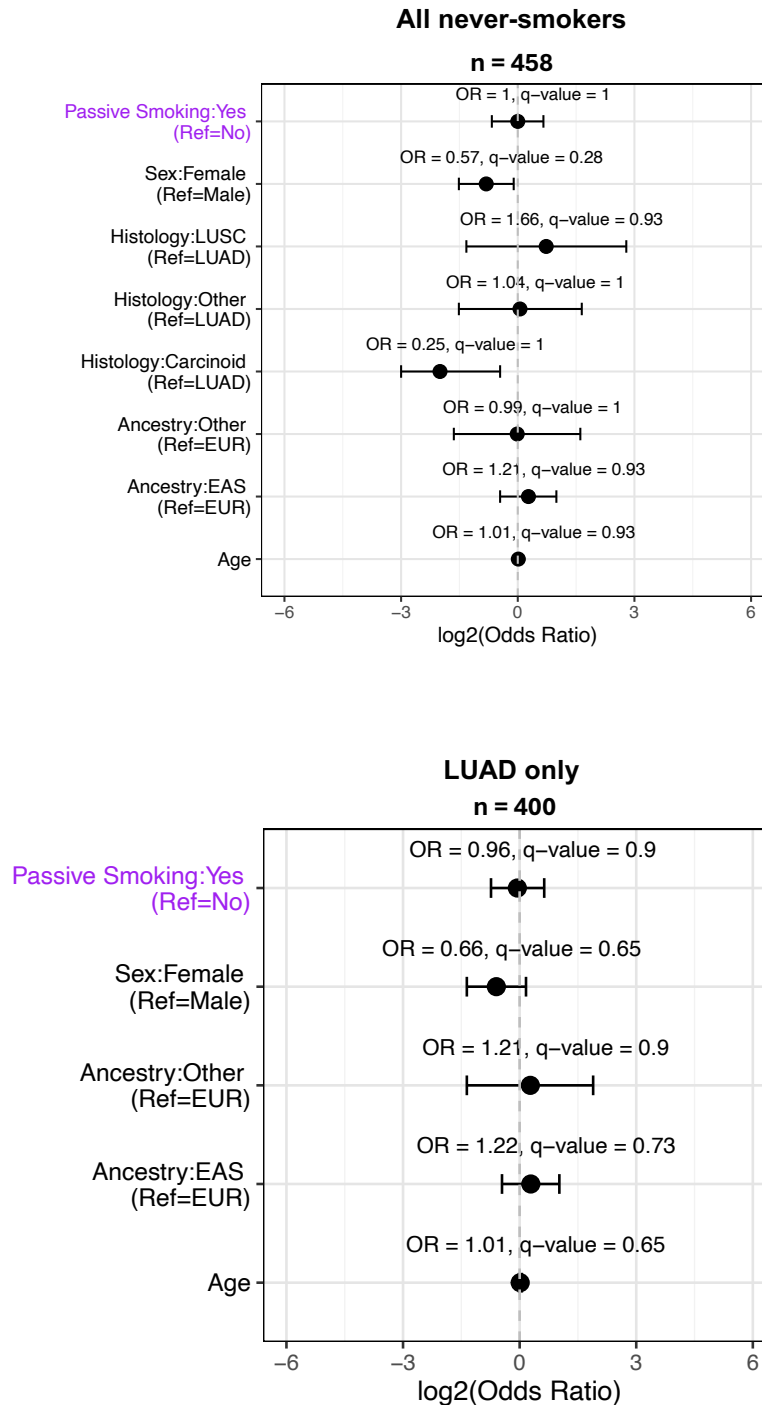


Figure 4.10: Association of ecDNA with passive smoking Forest plots for multivariate logistic regression models testing the passive smoking status with the presence of ecDNA in the tumors, for both all never-smokers (top) or LUAD only never-smokers (bottom).

4.1.6 Examining the global prevalence of ecDNA

The geographical differences in the incidence of certain types of cancer are hypothesized to be caused by differential exposures to certain types of carcinogens. Prior studies have comprehensively examined differences in mutational patterns between patients across countries (Moody et al., 2021; Senkin et al., 2024; Díaz-Gay et al., 2025), including those of small and large mutations, but there has been no explicit examination of the prevalence of ecDNA across countries. I conducted a one vs. all comparison for each country using logistic regression with covariates, where the ecDNA status of a patient is the outcome. ecDNA was present at similar rates across all geographic locations (Fig. 4.11), and as expected, there was no enrichment of ecDNA+ samples present in any particular country (Fig. 4.12). Note that here I applied Firth's penalized logistic regression using the `logistf` package in R to model the association between ecDNA status and geographic regions, in order to reduce small-sample bias and address potential issues of data separation.

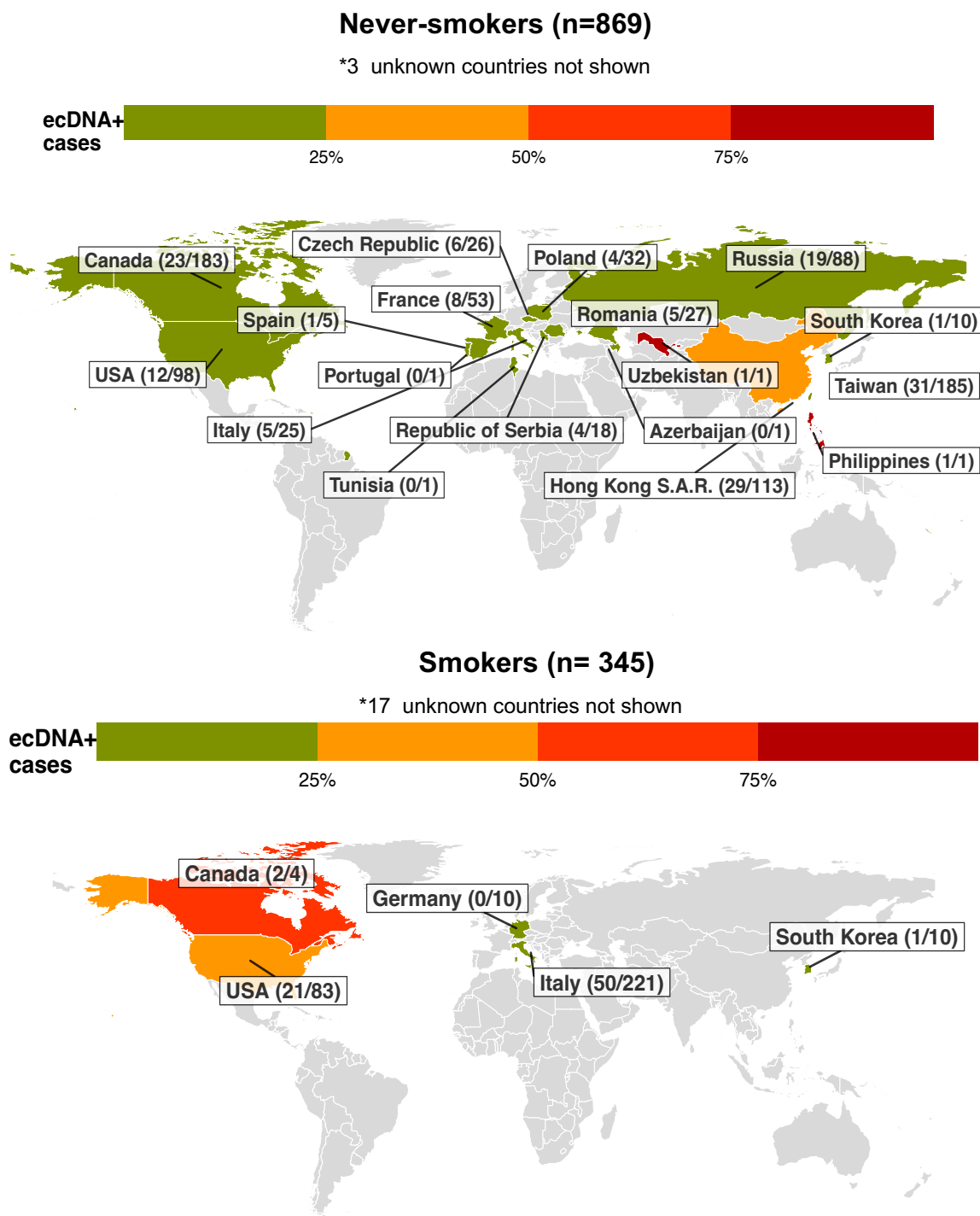


Figure 4.11: Distribution of ecDNA harboring lung cancers in never-smokers and smokers by country. Maps showing the worldwide prevalence of lung cancers harboring ecDNA by country for the never-smokers cohort (top) and smokers cohort (bottom). Fractions indicate the number of ecDNA+ samples in each country over the total number of samples from that country. Volcano plot of a multivariate one-vs-all logistic regression model for each country, with ecDNA status as the outcome, for c) never-smokers and d) smokers. An FDR q-value threshold of 0.05 is indicated with the dashed red line.

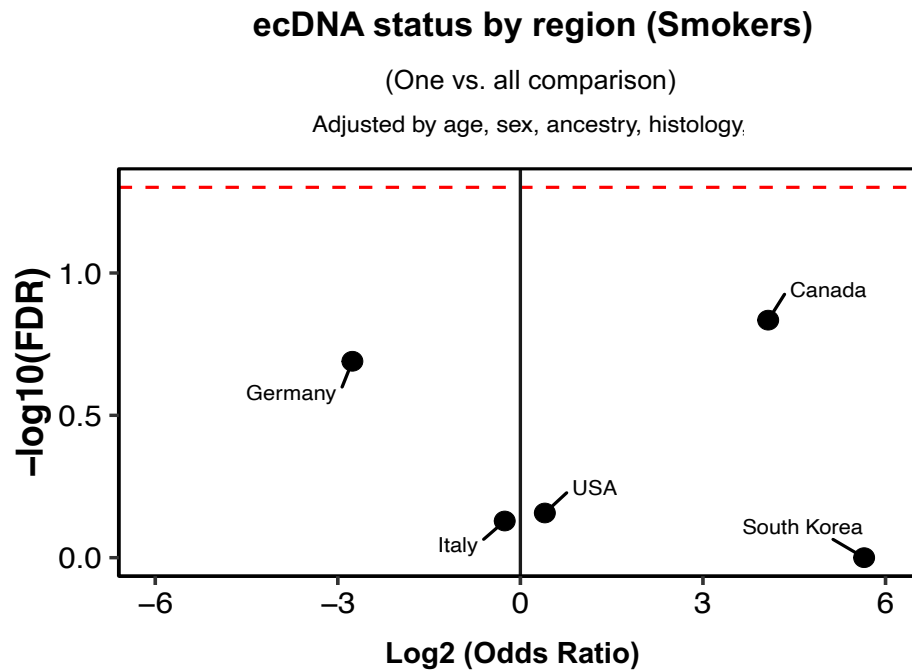
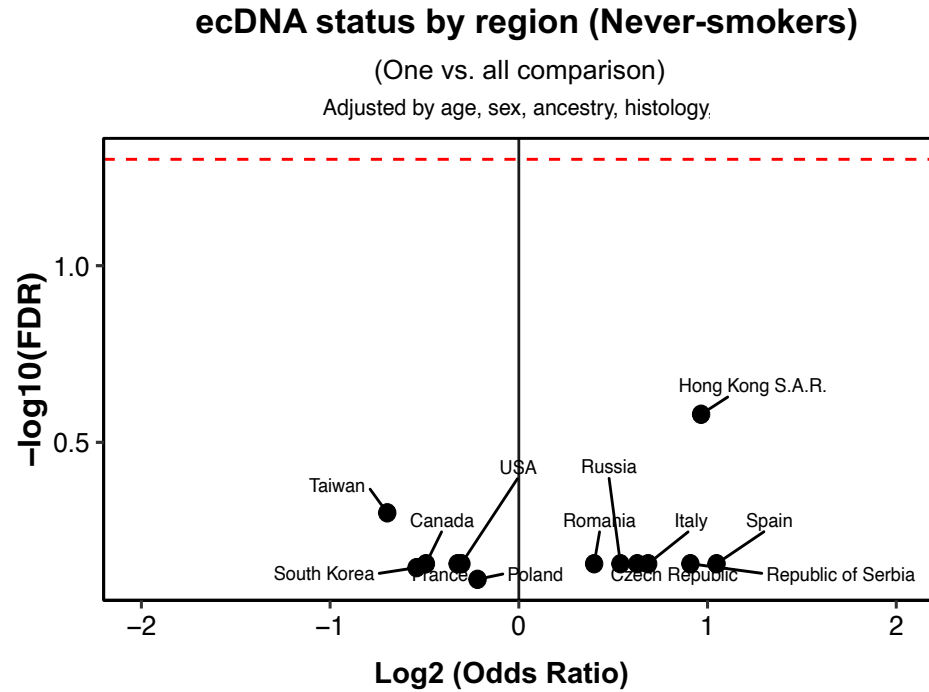


Figure 4.12: Association of ecDNA status with country Volcano plots of a multivariate one-vs-all logistic regression model for each country, with ecDNA status as the outcome, for never-smokers (top) and smokers (bottom).

4.1.7 Calculation of genome instability index score (wGII)

ecDNA is associated with overall genomic instability (Bafna and Mischel, 2022). Before I present the associations of ecDNA with specific processes known to either cause, or form as a consequence of, genome instability, I will first describe a metric used to quantify overall genome instability in tumors, the genome instability index score (wGII). To calculate the genome instability index score (wGII), the method devised in (Bailey et al., 2024) was used. The overall ploidy was calculated as the length-weighted average of all copy number segments. The proportion of segments that were aberrant (where total copy number was greater than the ploidy or less than the ploidy) was computed for each chromosome, these proportions were summed, and the total was divided by 22. This metric was included as a covariate in the following logistic regression models in order to distinguish the effect of ecDNA amplification specifically, without the confounding effect of overall genome instability. The proportion of samples with ecDNA seemed to increase with wGII (Fig. 4.13).

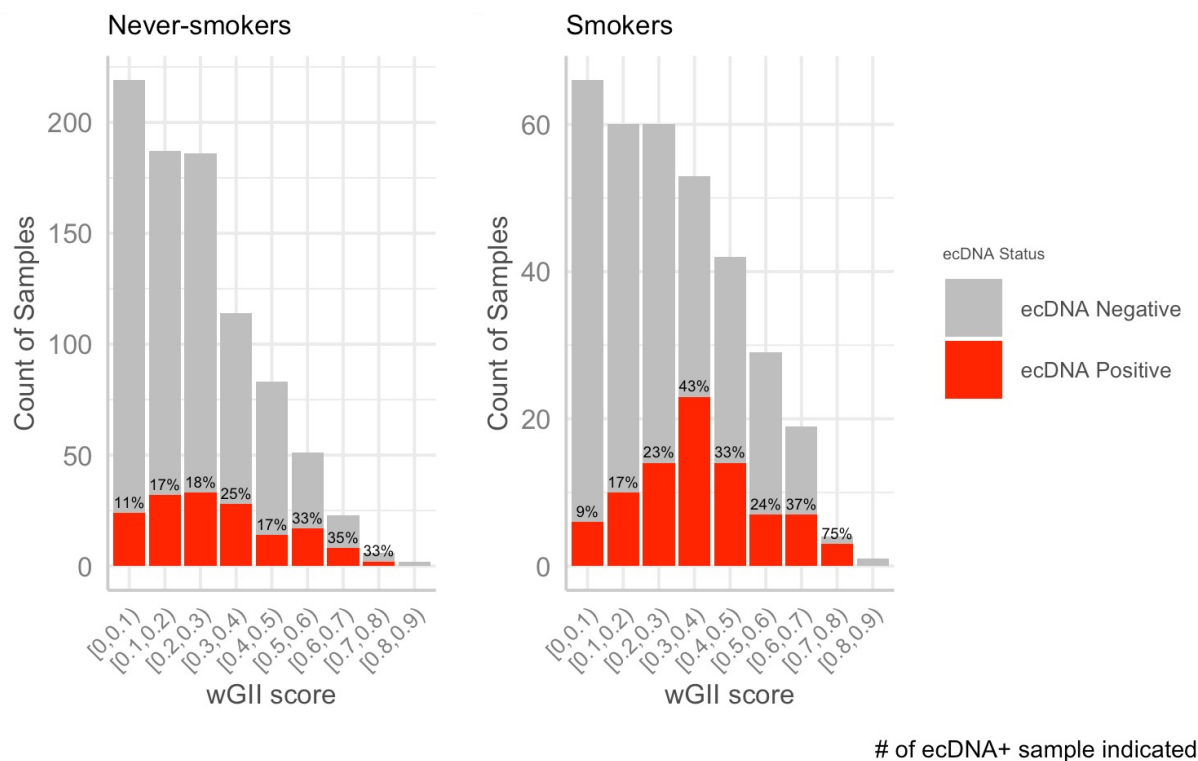


Figure 4.13: Relationship between ecDNA and overall genomic instability (wGII). Stacked bar plot showing the percentage of ecDNA+ samples stratified by bins of genomic instability for never-smokers (top) and smokers (bottom). To calculate the genome instability index score (wGII), the overall ploidy was calculated as the length-weighted average of all copy number segments. The proportion of segments that were aberrant (where total copy number was greater than the ploidy or less than the ploidy) was computed for each chromosome, these proportions were summed, and the total was divided by 22.

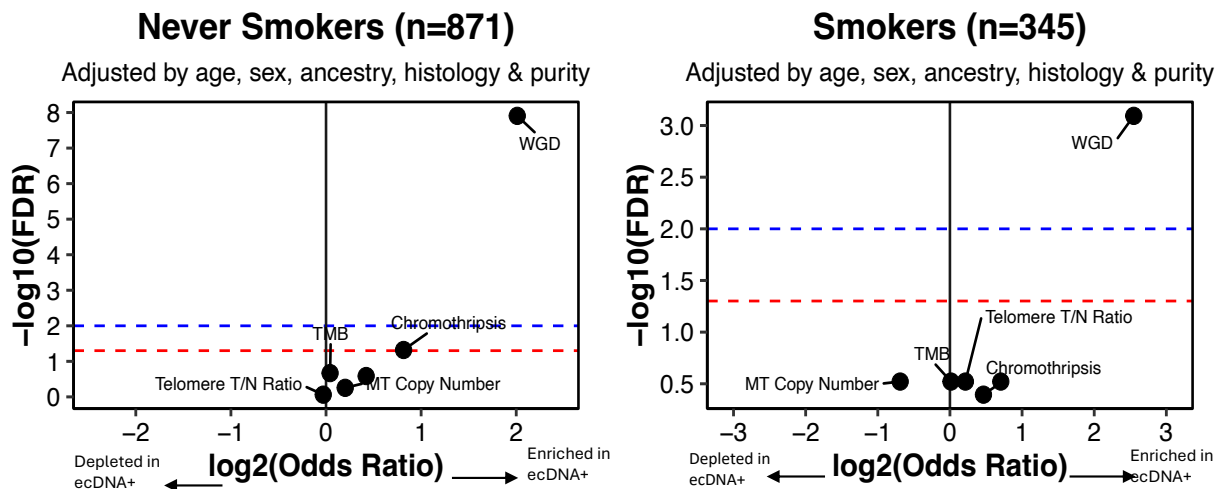
4.1.8 Examining the association of ecDNA with genomic features

To examine the association between ecDNA and genomic features previously linked to genome instability (Bielski et al., 2018; Cortés-Ciriano et al., 2020; Umbreit et al., 2020; Tang et al., 2021), we evaluated the relationship of ecDNA with whole-genome doubling (WGD), chromothripsis, tumor mutational burden, mitochondrial copy number, overall genome instability index (wGII), and telomere length tumor/normal ratio. For each genomic feature, I ran an independent logistic regression model with covariates in order to deal with the issue of multicollinearity between features (e.g WGD is known to be associated with chromothripsis). Amongst these features, only WGD showed a significant association with ecDNA in both LCSS and LCINS (Fig. 4.14), as determined by logistic regression adjusted for age, sex, ancestry, histology, and tumor purity. Specifically, ecDNA was present in 7.1% of LCINS and 8.9% of LCSS without WGD, but its prevalence increased markedly in WGD-positive tumors from LCINS (24.9%, OR=4.0; $q=1.24 \times 10^{-8}$) and LCSS (30.3%, OR=5.85; $q=8.06 \times 10^{-4}$). The same association was observed when only LUAD was examined (Fig. 4.14). Moreover, I further validated in a pan-cancer dataset of 974 samples with high-confidence WGD annotation (Bielski et al., 2018), using logistic regression adjusted for age, purity, sex, and cancer type (OR=2.81; $q=1.63 \times 10^{-6}$; (Fig. 4.15).

Mitochondrial copy number was estimated per sample using the ratio of mitochondrial read depth to nuclear genome depth, adjusted for tumor ploidy and purity, following the procedure in Reznik et al. (Reznik et al., 2016). Telomere length was estimated in kilobases using TelSeq (v.0.0.2) (Ding et al., 2014). As was previously done (Diaz-Gay et al., 2024), we used seven as the threshold for the number of TTAGGG/CCCTAA repeats in a read for

the read to be considered telomeric. The TelSeq calculation was done individually for each read group within a sample, and the total number of reads in each read group was used as weight to calculate the average TL for each sample. The telomere length ratio (log2 scale) between tumor and normal tissue was then calculated for each patient and served as the independent variable in multivariable logistic regressions. Samples were considered to have a whole genome doubling event if the major copy number was ≥ 3 for more than 50% of copy number segments.

Genomic features associated with ecDNA



Genomic features associated with ecDNA

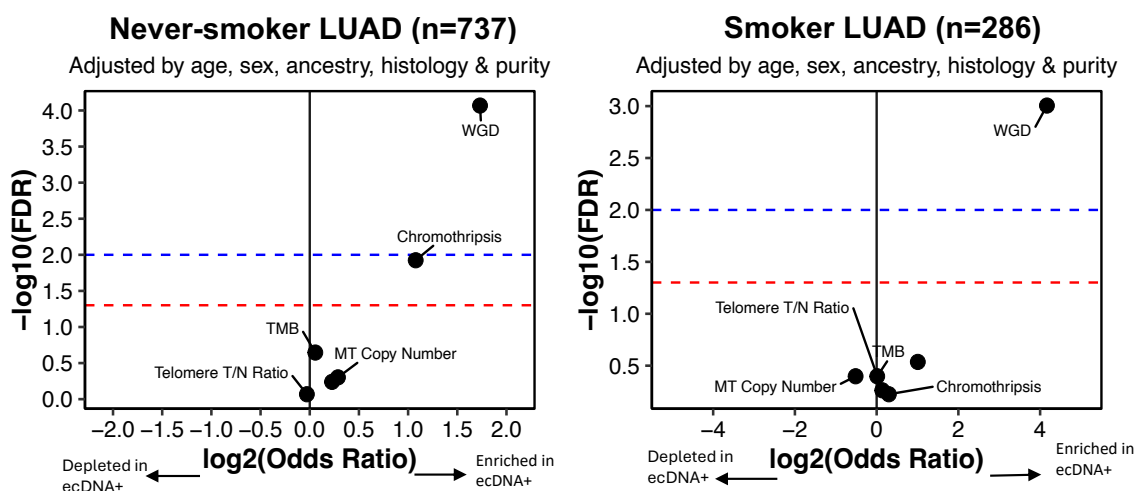


Figure 4.14: Genomic features associated with the presence of ecDNA. Volcano plot of logistic regression models of genomic features in association with ecDNA status for LCINS (left) and lung adenocarcinoma in LCSS (right). The same analysis repeated for LUAD samples only is shown below. In all volcano plots, the x-axes reflect the log₂ odds ratio, and the y-axes correspond to the log₁₀ FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line.

Pan-cancer analysis of ecDNA status from PCAWG/TCGA (default pipeline) n=974

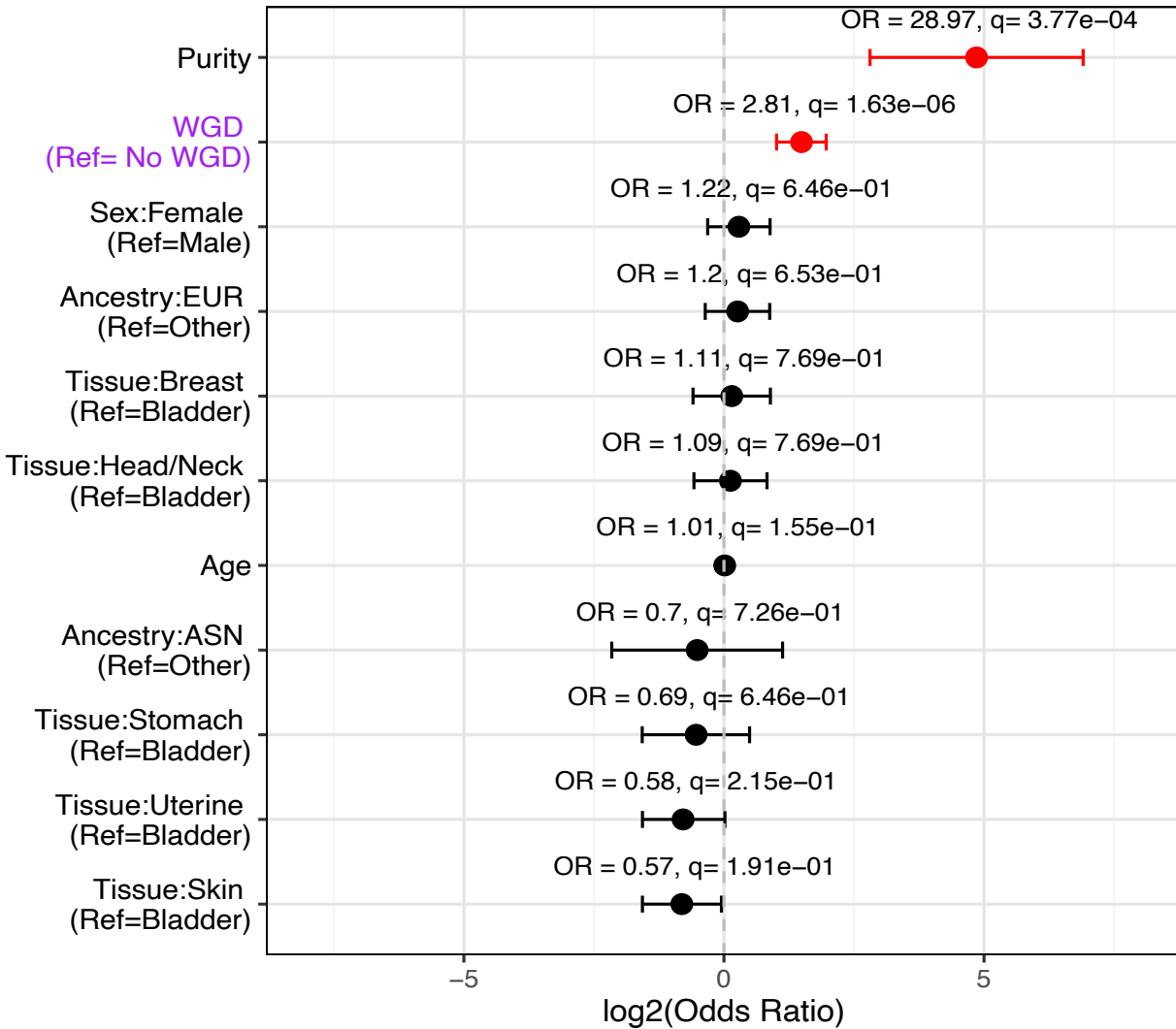


Figure 4.15: Pan-cancer validation of the association of WGD with ecDNA. Forest plot of multivariate logistic regression modeling the presence or absence of ecDNA in relation to whole-genome doubling in a pan-cancer dataset comprising four different tumor types from the PCAWG project, adjusted for the covariates of age, purity, sex, and tissue type. All significant associations are indicated in red. This public callset was generated using the default AmpliconArchitect pipeline.

4.1.9 Examining the association of driver genes with ecDNA

To evaluate the relationship between ecDNA presence and lung cancer driver genes, I analyzed driver genes altered in at least 10 samples. Using multivariate logistic regression, adjusted for age, sex, tumor purity, ancestry, and histology, no significant associations were found between ecDNA and individual driver gene alterations (Fig. 4.16). Notably, TP53 is not associated with ecDNA presence. Although this association has been reported in other tumor types(Luebeck et al., 2023), the sample sizes in these studies were limited and the analysis may have been confounded by the fact that TP53 alteration is also associated with whole-genome doubling and chromothripsis. Taken together, this suggests that in both LCINS and LCSS, ecDNA is primarily formed via TP53 independent mechanisms. The IntOGen pipeline (v.2020.02.0123)⁴⁰, which combines seven state-of-the-art computational methods, was employed with default parameters to detect signals of positive selection in the mutational patterns of driver genes across the cohort. The status of each driver gene (altered or unaltered) was then input into a logistic regression model with covariates age, sex, and tumor purity, and ecDNA status was the output. These sets of genes were considered altered if there was a focal deletion overlapping the gene, an inactivating (stop or missense) mutation, or an activating gain of function mutation.

Driver gene alterations associated with ecDNA

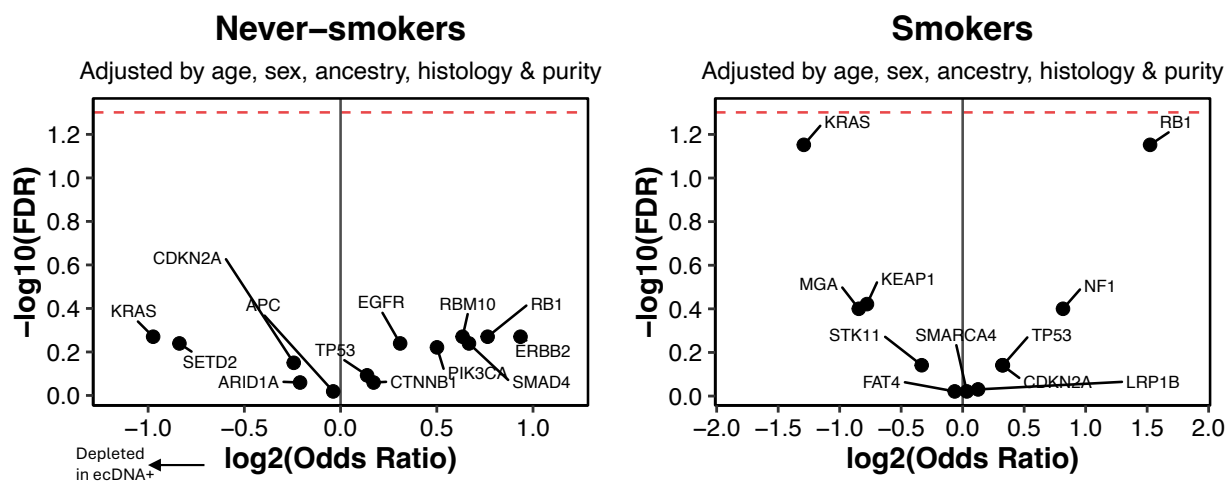
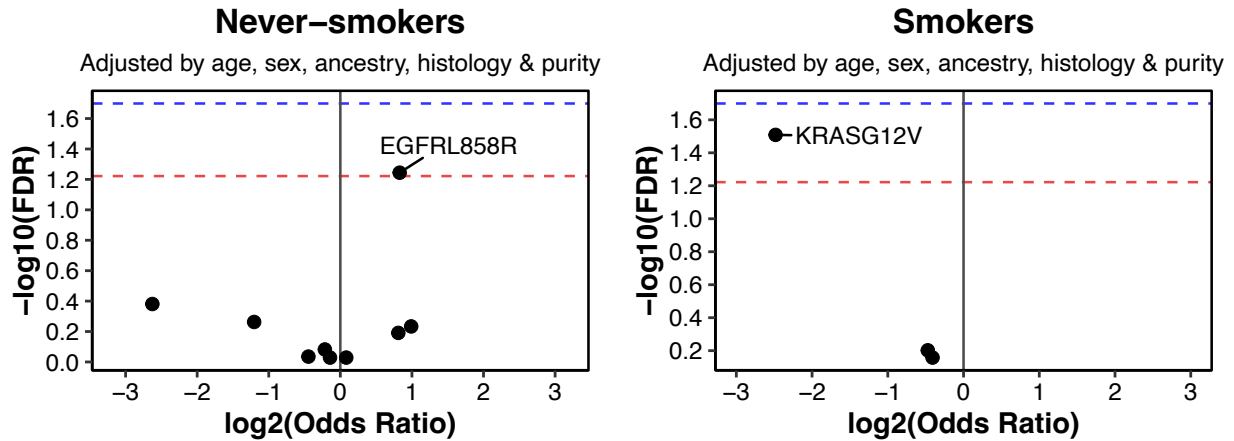


Figure 4.16: Association between driver gene alterations and ecDNA Volcano plot of a logistic regression model of driver gene alterations in a sample in association with ecDNA status for LCINS (left) and LCSS (right). The x-axes reflect the $\log_2$ odds ratio, and the y-axes correspond to the $\log_{10}$ FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line.

4.1.10 Examining the association of ecDNA with hotspot driver mutations

Using multivariate logistic regression, adjusted for age, sex, tumor purity, ancestry, and histology, I showed an enrichment of EGFR L858R mutations (OR=1.78; $q=0.05$; (Fig. 4.17)) in never-smokers with ecDNA and a depletion of KRAS G12V mutations in smokers with ecDNA (OR=0.18; $q=0.03$; (Fig. 4.17)). Overall, only 15.5% of never-smokers without a EGFR L858R mutation contained ecDNA, but 23.8% of never-smokers with a EGFR L858R mutation contained ecDNA. To identify driver mutations within the set of identified driver genes, we implemented a rigorous and multifaceted strategy, considering multiple criteria: (i) the presence of truncating mutations specifically in genes annotated as tumor suppressors, (ii) the recurrence of missense mutations in a minimum of 3 samples, (iii) mutations designated as “Likely drivers” with a boostDM score(Muinos et al., 2021) exceeding 0.5, (iv) mutations categorized as either “Oncogenic” or “Likely Oncogenic” based on the criteria established by OncoKB(Chakravarty et al., 2017), an expert-guided precision oncology knowledge base, (v) mutations previously recognized as drivers in the TCGA MC3 drivers paper(Bailey et al., 2018), and (vi) missense mutations characterized as “likely pathogenic” in genes that are annotated as tumor suppressors, as described in Cheng et al.(Cheng et al., 2021) Any mutation meeting one or more of these criteria was recognized as a potential driver mutation.

Hotspot driver mutations associated with ecDNA



EGFR L858R mutation in never-smokers

OR=1.78, q=0.05, Adjusted by age, sex, ancestry, histology, purity

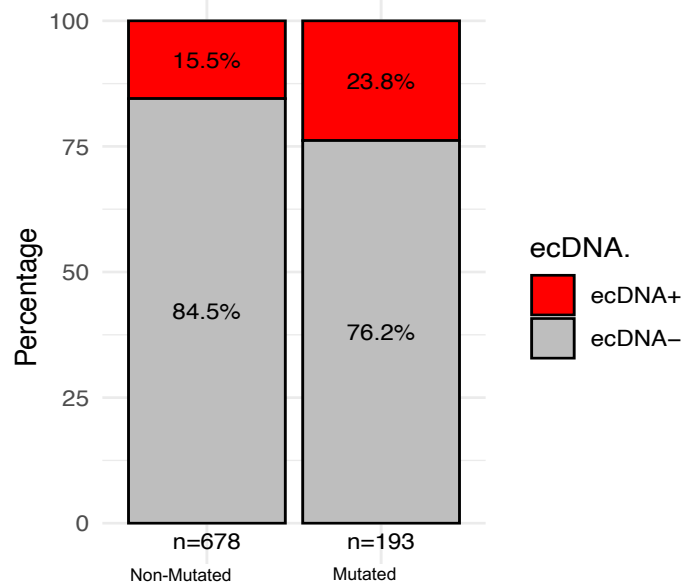


Figure 4.17: Association between ecDNA and driver mutations Volcano plot of a logistic regression model of hotspot driver mutations in a sample in association with ecDNA status (top) for LCINS (left) and LCSS (right). The x-axes reflect the log2 odds ratio, and the y-axes correspond to the log10 FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line. Bar plot showing the proportion of samples with an EGFR L858R mutation, stratified by ecDNA status (bottom).

4.1.11 Evaluating the effect of ecDNA on gene expression

In order to determine if the presence of ecDNA in a sample results in different gene expression patterns of individual genes, I set up this analysis as a standard differential expression analysis conditioned on whether a sample has ecDNA or not, with the gene expression of log CPM as input. Specifically, differential expression analysis was performed using the limma-voom pipeline. A design matrix was constructed using a cell-means parameterization ($0 + \text{ecDNA status}$) with additional covariates for gender, age, histology, and tumor purity to account for potential confounding effects. Gender was coded with female as the reference level, histology with adenocarcinoma as reference, and ancestry with European (EUR) as reference. The voom transformation was applied to model the mean-variance relationship in the count data and generate precision weights for linear modeling. Contrasts were defined to compare ecDNA-positive versus ecDNA-negative tumors. Linear models were fit using empirical Bayes moderation of the standard errors. Two statistical approaches were employed: (1) treat method with a log-fold change threshold of 1, and (2) standard eBayes moderation. Differential expression was assessed using a false discovery rate (FDR) adjusted p-value threshold of 0.05.

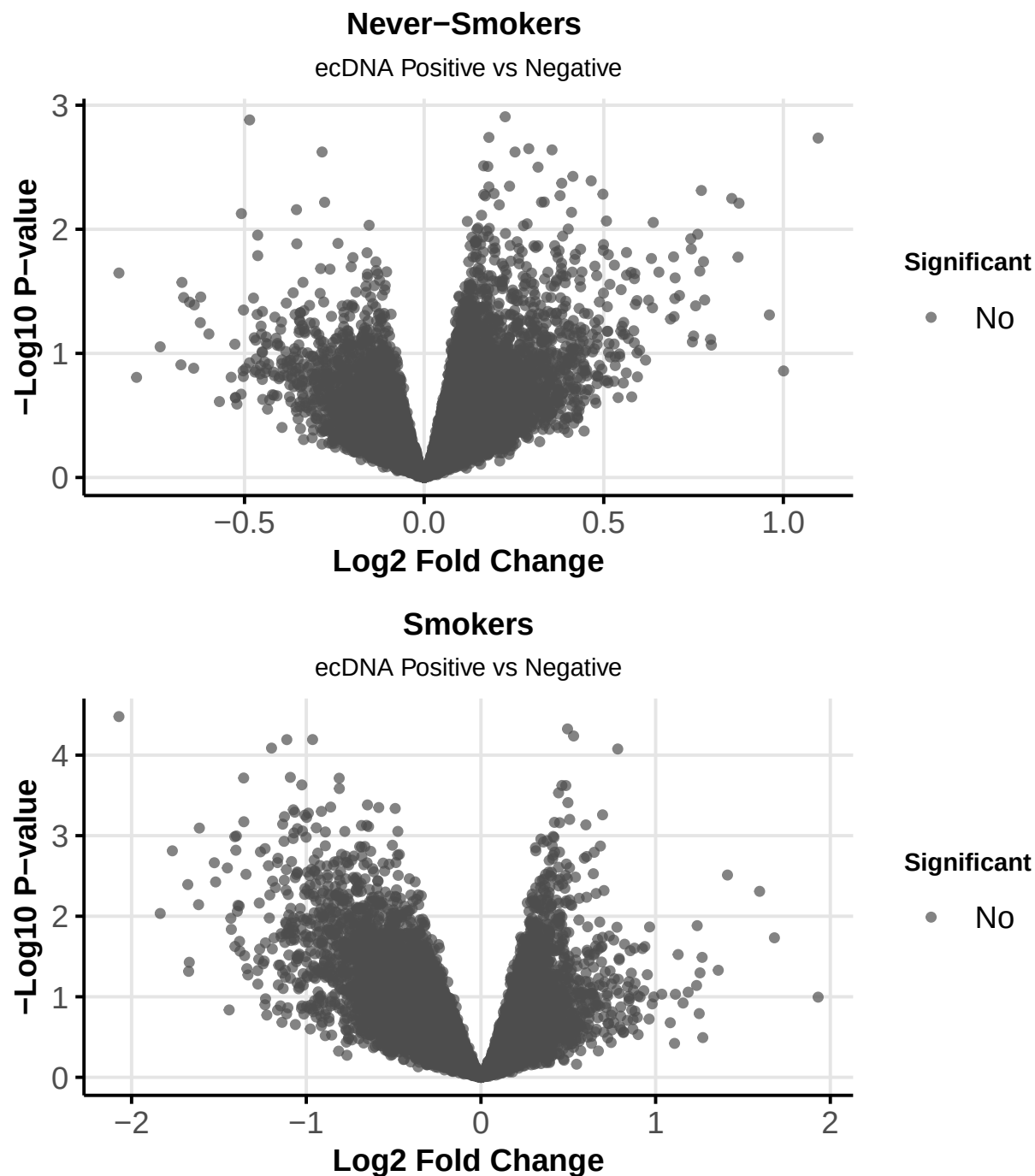


Figure 4.18: Differential gene expression analysis of ecDNA+ vs ecDNA- tumors
Volcano plot showing the results of a differential gene expression analysis. Each point represents a gene, the log2 fold change is shown on the x-axis, and the log10 p-value is shown on the y-axis.

4.1.12 Analysis of genome-wide mutational signatures in ecDNA+ vs ecDNA- lung tumors

Mutational signature analysis provides insights into both the endogenous and exogenous mutational processes that have been active throughout the lineage of a cancer genome (Islam et al., 2022b). I investigated the association between ecDNA presence and genome-wide mutational signatures in order to gain insight on the mutational process associated with ecDNA+ tumors. As expected (Kim et al., 2024), multiple copy-number and structural variant signatures were associated with ecDNA presence similarly in LCSS and LCINS (Fig. 4.21). Additionally, in LCINS, ecDNA-positive samples were enriched for clock-like signatures (SBS1, SBS5, ID1, ID2) and, as previously reported (Bergstrom et al., 2022; Hung et al., 2024), APOBEC-associated signatures (SBS2, SBS13; (Fig. 4.19). The association with clock-like signatures suggests ecDNA+ tumors undergo accelerated cell division, although whether its the case that rapidly dividing cells provide a more permissive background for ecDNA formation or the ecDNA itself drives accelerated cell division is unclear. In LCSS, enrichment was observed only for SBS3 (Fig. 4.19), a signature previously associated with homologous recombination deficiency (Davies et al., 2017). The association with clock-like signatures suggests ecDNA+ tumors undergo accelerated cell division, although whether its the case that rapidly dividing cells provide a more permissive background for ecDNA formation or the ecDNA itself drives accelerated cell division is unclear. The mutational signature attributions were determined as was previously done, based on the methodology in Díaz-Gay et al. (Diaz-Gay et al., 2024). SigProfilerMatrixGenerator (Khandekar et al., 2023; Bergstrom et al., 2019) was used to create the mutational matrix for all types of somatic

mutations. The deconvolution of mutational signatures was performed by SigProfilerExtractor (v.1.1.3)(Islam et al., 2022b) using the whole-genome sequencing setting and the COSMIC mutational signatures (v3.2) as reference. Signature extractions were performed separately for LCINS and LCSS. Signature attributions were input into a multivariate logistic regression model as either present or absent if the signature was present in less than 50% of the samples. If the signature was present in $\geq 50\%$ samples (e.g., SBS1 and SBS5), it was input as either above median or below median. For SBS1 and SBS5, the signature attributions were also input as continuous numeric variables to confirm the significant associations found with inputting them as above median or below median.

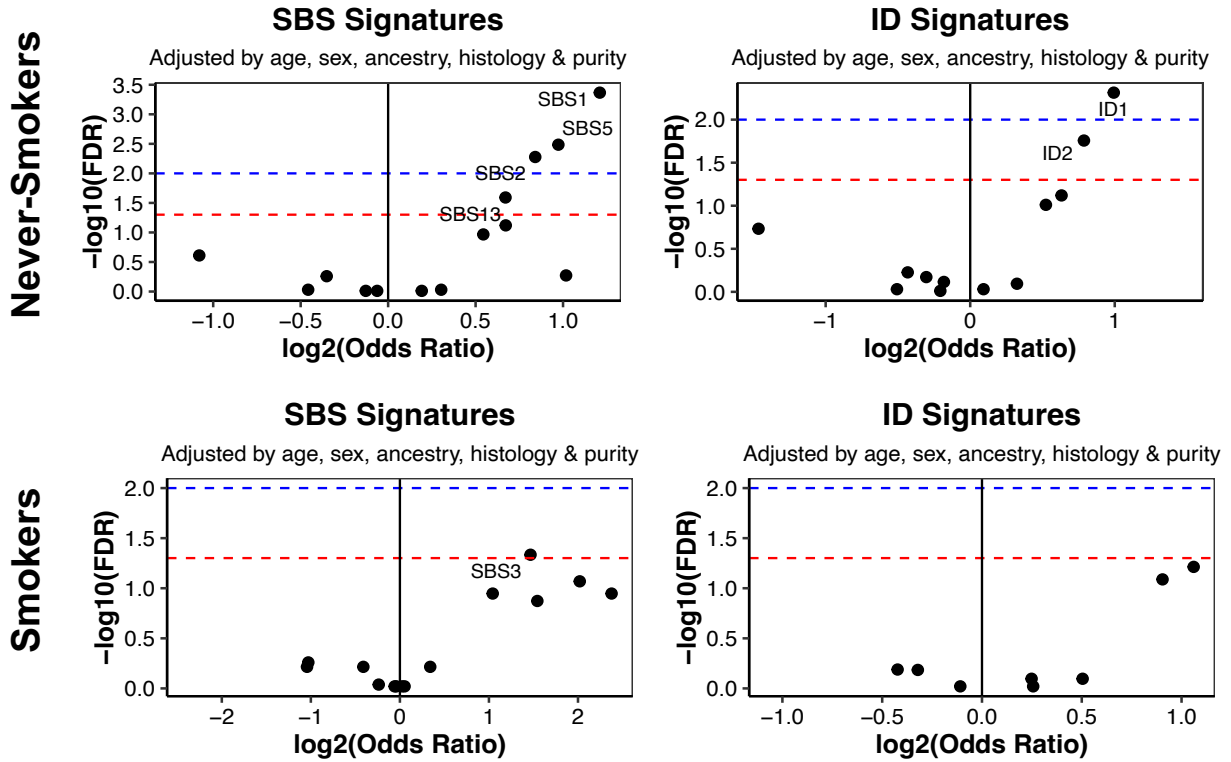


Figure 4.19: Global mutational signatures of small mutations associated with ecDNA Volcano plots of logistic regression models of the activities of Single base substitution (SBS) and Indel (ID) in association with ecDNA in lung cancers of never-smokers (top) and smokers (bottom). The logistic regression models were adjusted for age, sex, ancestry, histology, and purity. In all volcano plots, the x-axes reflect the $\log_2$ odds ratio, and the y-axes correspond to the $\log_{10}$ FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line.

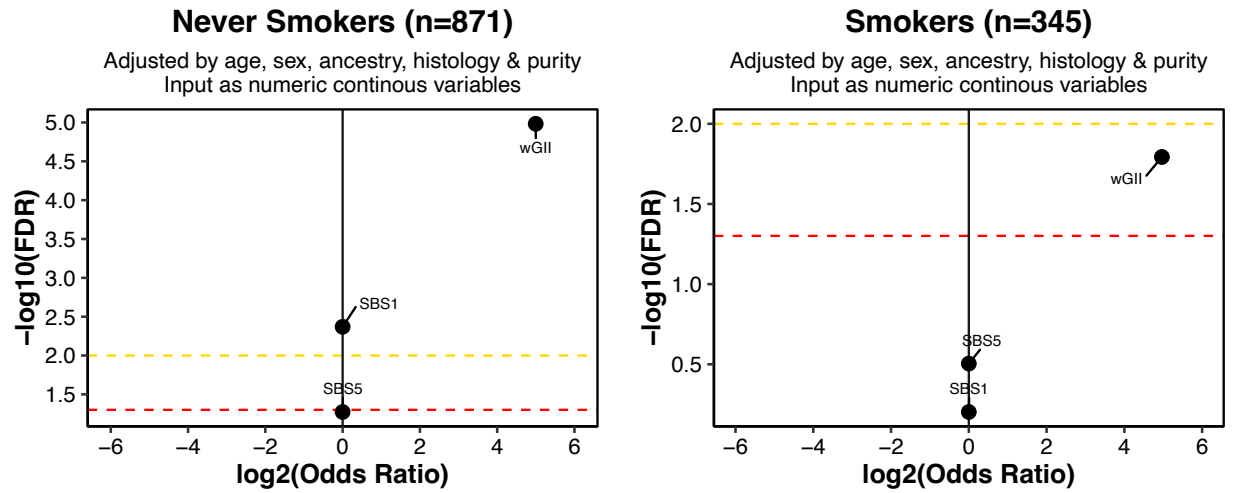


Figure 4.20: Global mutational signatures of small mutations input as numeric continuous variables Volcano plots of logistic regression models of the activities of Single base substitution (SBS) signatures 1 and 5 input as continuous numeric variables rather than as above/below median or present/absent.

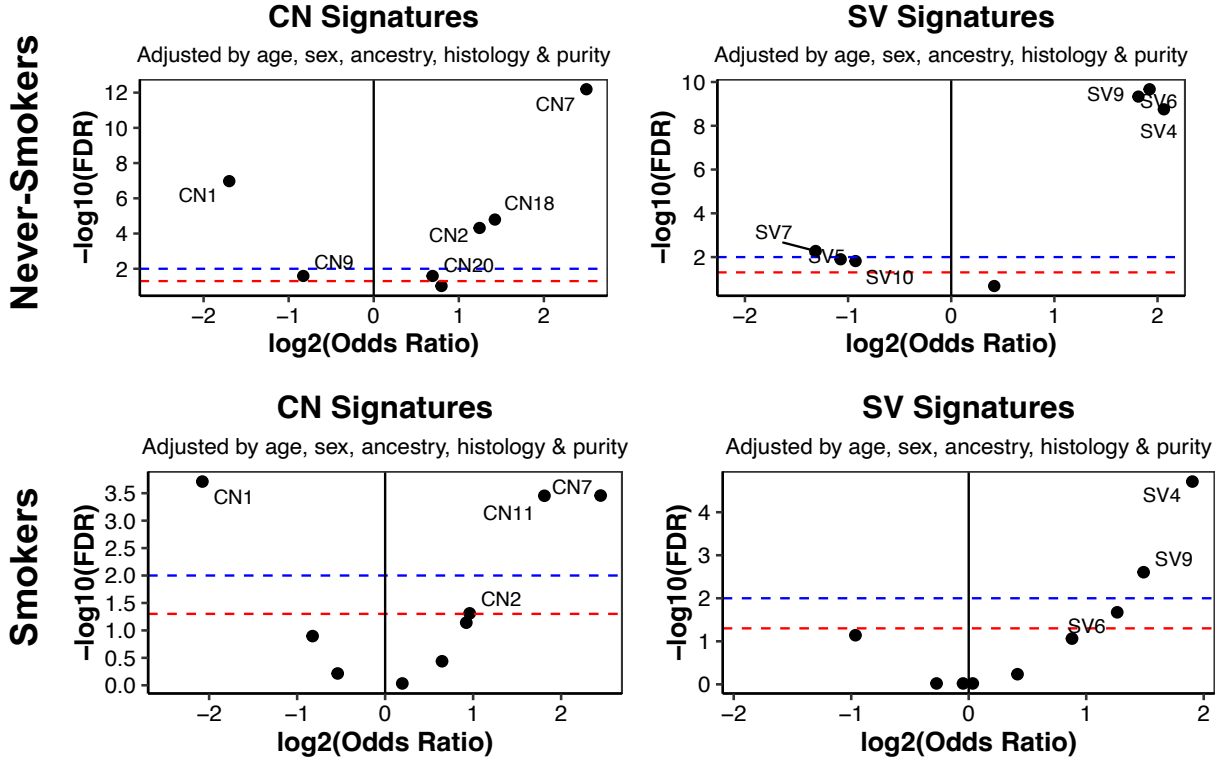


Figure 4.21: Global mutational signatures of large mutations associated with ecDNA. Volcano plots of logistic regression models of the activities of copy number (CN) and structural rearrangement (SV) signatures in association with ecDNA in lung cancers of never-smokers (top) and smokers (bottom). The logistic regression models were adjusted for age, sex, ancestry, histology, and purity. In all volcano plots, the x-axes reflect the log2 odds ratio, and the y-axes correspond to the log10 FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line.

4.1.13 Analysis of mutational signatures on ecDNA

In the previous section, I focused on global (genome-wide) mutational signatures. Here, I focus on mutational signatures specifically on ecDNA. Our previous paper, (Bergstrom et al., 2022) revealed recurrent APOBEC mutagenesis of ecDNA, where ecDNA appeared to be attacked by APOBEC after formation, as was shown by analyzing the distances from mutations to the nearest SV breakpoint. I build upon this previous finding by applying a different approach and showing this phenomenon holds true in lung cancer. First, each mutation was probabilistically assigned to a mutational signature based on the methodology developed in (Díaz-Gay et al., 2023). Next, I extracted all mutations that overlapped with ecDNA regions, and assigned each mutation to a signature by taking the highest probability signature as determined by SigProfilerAssignment. I then generated two different sets of background mutations to compare against, those from region immediately flanking the left-most ecDNA breakpoint and the right-most ecDNA breakpoint (which I refer to as 'flanking'), and those overlapping other chromosomal amplified regions (which I refer to as chrAmp). In the case of the flanking region, I take a 10 KB window as prior research has shown that kataegis frequently occurs within 10 kb of detected SV breakpoints. Proportions of mutations belonging to a specific signature are computed for all comparisons in order to account for the differences in size between ecDNA regions, flanking regions, and chromosomal amplification regions. A paired Wilcoxon signed-rank test was used to account for sample to sample variability in mutational signatures. In order to ensure that there are enough mutations for comparison, I only take samples where ecDNA amplicons have > 10 mutations. Notably, this analysis showed that SBS signatures 2 and 13, the APOBEC signatures, were

significantly elevated p ($p < 0.001$) on ecDNA compared to both flanking regions (Fig. 4.22) and other chromosomally amplified regions (Fig. 4.23), suggesting evidence of recurrent APOBEC hypermutation after ecDNA formation and ongoing ecDNA evolution.

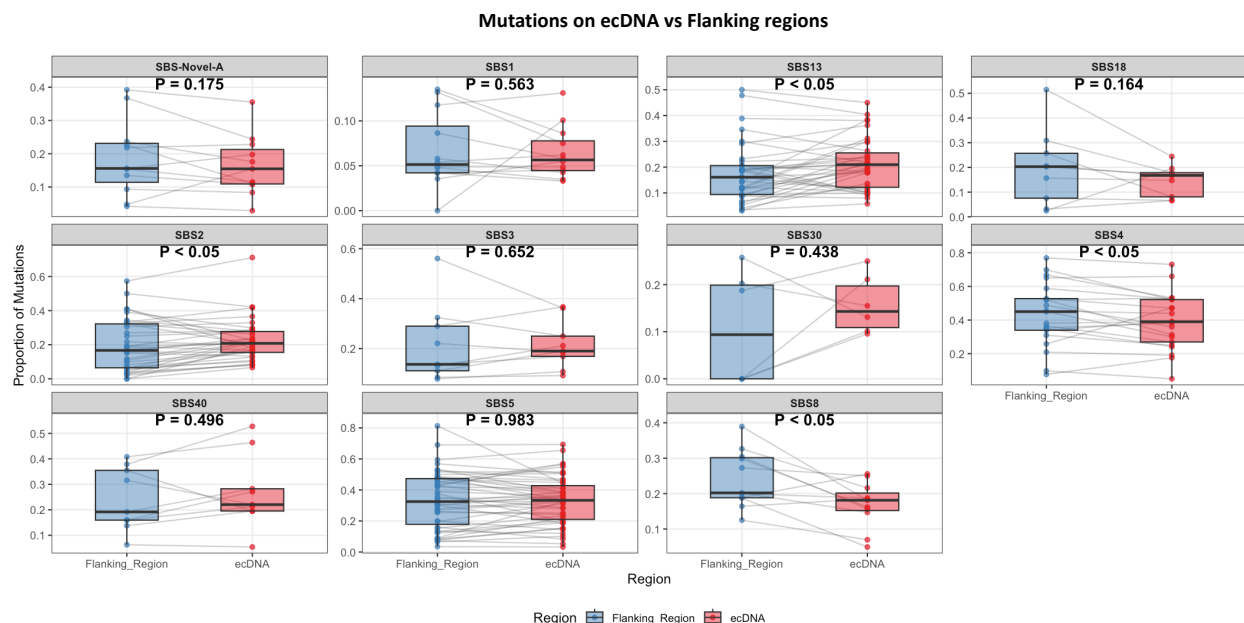


Figure 4.22: Global mutational signatures of large mutations associated with ecDNA Volcano plots of logistic regression models of the activities of copy number (CN) and structural rearrangement (SV) signatures in association with ecDNA in lung cancers of never-smokers (top) and smokers (bottom). The logistic regression models were adjusted for age, sex, ancestry, histology, and purity. In all volcano plots, the x-axes reflect the $\log_2$ odds ratio, and the y-axes correspond to the $\log_{10}$ FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line.

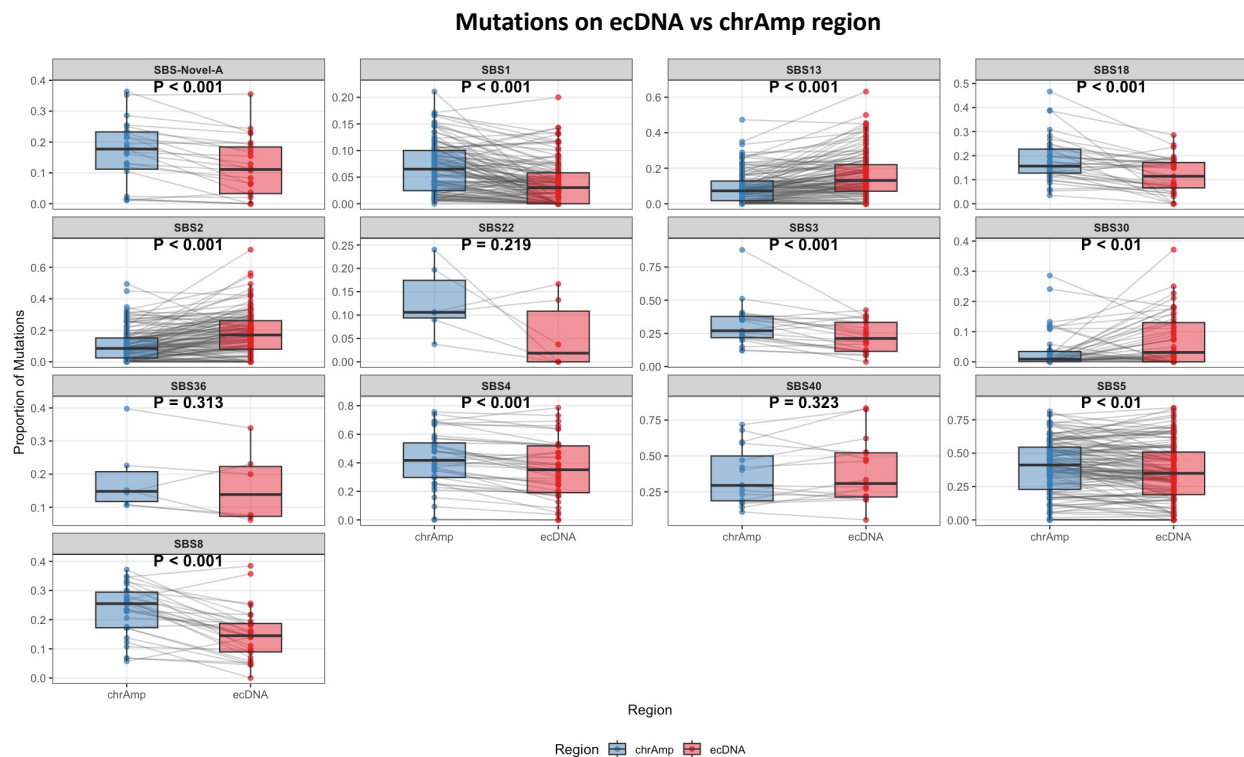


Figure 4.23: Mutational signatures present on ecDNA vs. chrAmp regions Volcano plots of logistic regression models of the activities of copy number (CN) and structural rearrangement (SV) signatures in association with ecDNA in lung cancers of never-smokers (top) and smokers (bottom). The logistic regression models were adjusted for age, sex, ancestry, histology, and purity. In all volcano plots, the x-axes reflect the $\log_2$ odds ratio, and the y-axes correspond to the $\log_{10}$ FDR q-value. An FDR q-value threshold of 0.05 is indicated with the dashed red line, an FDR q-value threshold of 0.01 is indicated with the dashed blue line.

4.1.14 Genomic content of ecDNA

Previous studies have shown that ecDNA can harbor oncogenes, regulatory elements, and immunomodulatory genes in varying configurations (Bafna and Mischel, 2022; Luebeck et al., 2023; Bailey et al., 2024; Wu et al., 2019; Zhu et al., 2021). In our cohort, 151 (17%) *never-smokers* and 80 (23%) *smokers* harbored ecDNA, with a total of 236 ecDNAs in *LCINS* and 123 in *LCSS*, reflecting multiple ecDNAs in some tumors. The composition of ecDNA in lung cancer, particularly never-smokers, is unknown. In order to examine ecDNA at the sequence level, I first quantified the varying configurations of genomic elements on all ecDNA in both never-smokers and smokers (FIG). Here, I defined a immunomodulatory gene as a gene that had an immunomodulatory function (GO terms: 0006968, 0002228, 0042267, 0001906, 0001909, 0002698, 0001910, 0031341, 0002367, 0002695, 0050866, 0051250, 0050777), following the procedure in Bailey et. al. In order to define regulatory regions, lung enhancer coordinates were downloaded from EnhancerAtlas. Each ecDNA region was then overlapped with these enhancers using a 100% containment criterion (i.e. the entire enhancer had to lie within the ecDNA). Separately, a set of lung-cancer promoter regions was generated via an R script that extracted the 1,000 bp immediately upstream of each gene’s transcription start site.

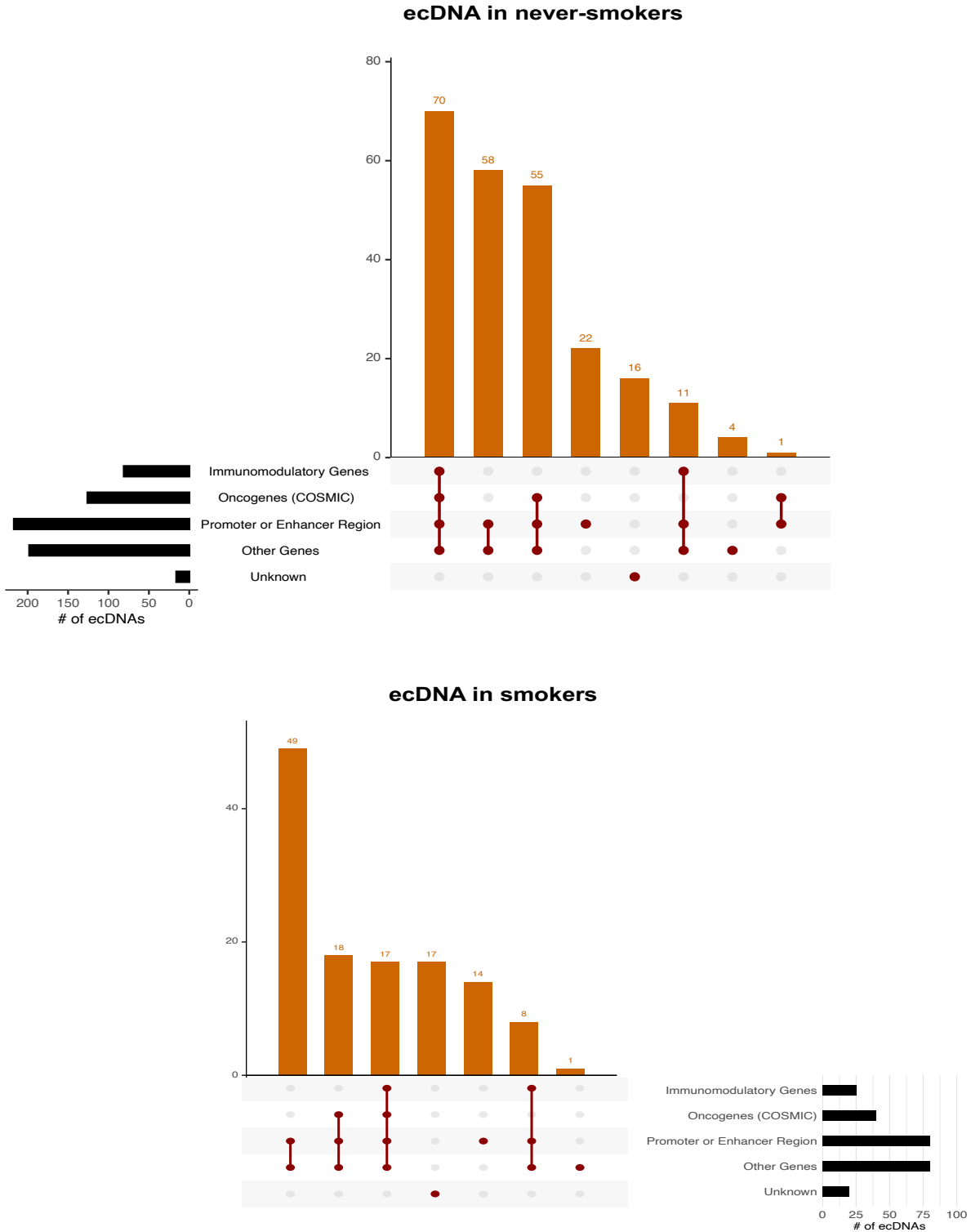


Figure 4.24: Genomic elements on ecDNA in lung cancer Upset plot indicating the number of ecDNA with oncogenes, immunomodulatory genes, promoter or enhancer regions, other genes that are not oncogenes, or unidentified genomic elements (Unknown) or any combination thereof for LCINS (top) and LCSS (bottom).

Although oncogenes are the most frequent genomic element found on ecDNA, Bailey et. al conducted a overrepresentation analysis and found that ecDNA lacking oncogenes are enriched for immunomodulatory genes compared to focal amplifications, and samples containing immunomodulatory ecDNAs have a significant depletion of T cells, indicating a weaker immune response. Although there were no ecDNA containing in our cohort containing only immunomodulatory genes, over half of ecDNA (81/236) contained an immunomodulatory gene as cargo. Recently, the presence of regulatory regions on ecDNA has also garnered a significant amount of interest. (Hung et al., 2021) showed that ecDNAs containing promoters or enhancers can interact in trans with other ecDNAs, forming hubs. This spatial reorganization caused by ecDNA can drive increased gene expression, if, for example, an enhancer on one ecDNA interacts with the promoter of another ecDNA. In our cohort, 22 of 236 (9.3%) ecDNAs in never-smokers and 14 of 80 (17.5%) ecDNAs in smokers contained only regulatory or promoter elements, which is higher than the 22 out 579 (3%) reported at the pan-cancer level in Bailey et. al. Notably, ecDNAs in never-smokers were enriched for oncogenes ($p < 0.001$) and immunomodulatory genes ($p=0.008$) compared to smokers, despite the higher prevalence of ecDNA in smokers. This result remained consistent even after removing all the ecDNAs with recurrently amplified oncogenes such as MDM2 ($p=0.002$, 4.26), suggesting this was not driven purely by a subset of recurrent ecDNAs. The median copy number estimate of an ecDNA can serve as a proxy for the degree of selective pressure on it and its clonal prevalence in the tumor. ecDNAs containing both oncogenes and immunomodulatory genes or oncogenes only had significantly higher median copy number than other ecDNAs (Wilcoxon test, $p=0.0015$, 4.27).

ecDNA content by smoking status

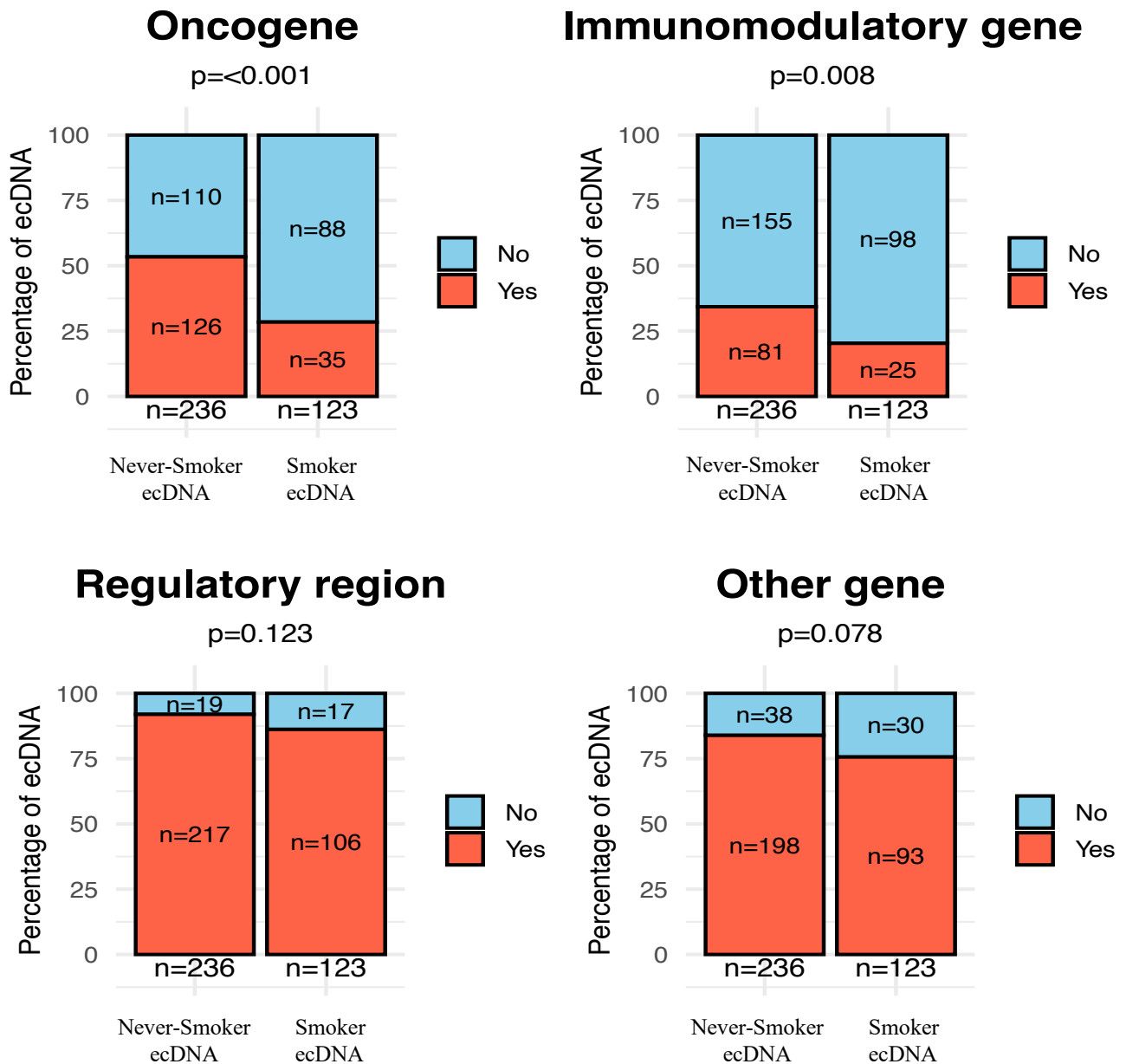


Figure 4.25: Enrichment of genomic elements on ecDNA in never-smokers vs smokers
Bar plot indicating the proportion of ecDNA containing a particular genomic element in never-smokers vs smokers. Each plot represents a one vs all. comparison (e.g oncogene-containing ecDNA vs. all other ecDNAs)

ecDNA content w/ all recurrent oncogenes removed

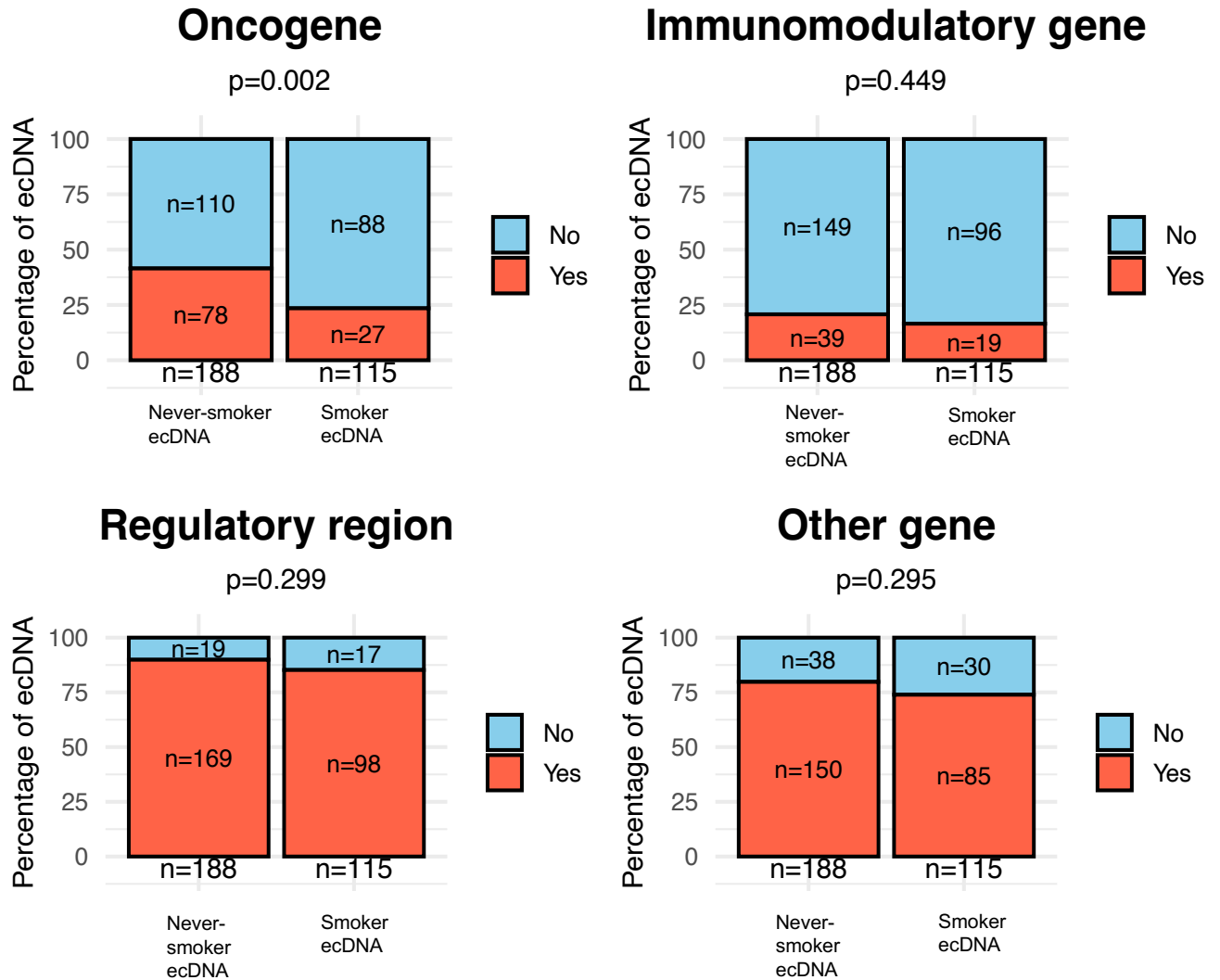


Figure 4.26: Enrichment of genomic elements on ecDNA in never-smokers vs smokers with recurrent oncogenes removed Bar plot indicating the proportion of ecDNA containing a particular genomic element in never-smokers vs smokers, with all recurrent oncogenes removed from the analysis. Each plot represents a one vs all. comparison (e.g oncogene-containing ecDNA vs. all other ecDNAs)

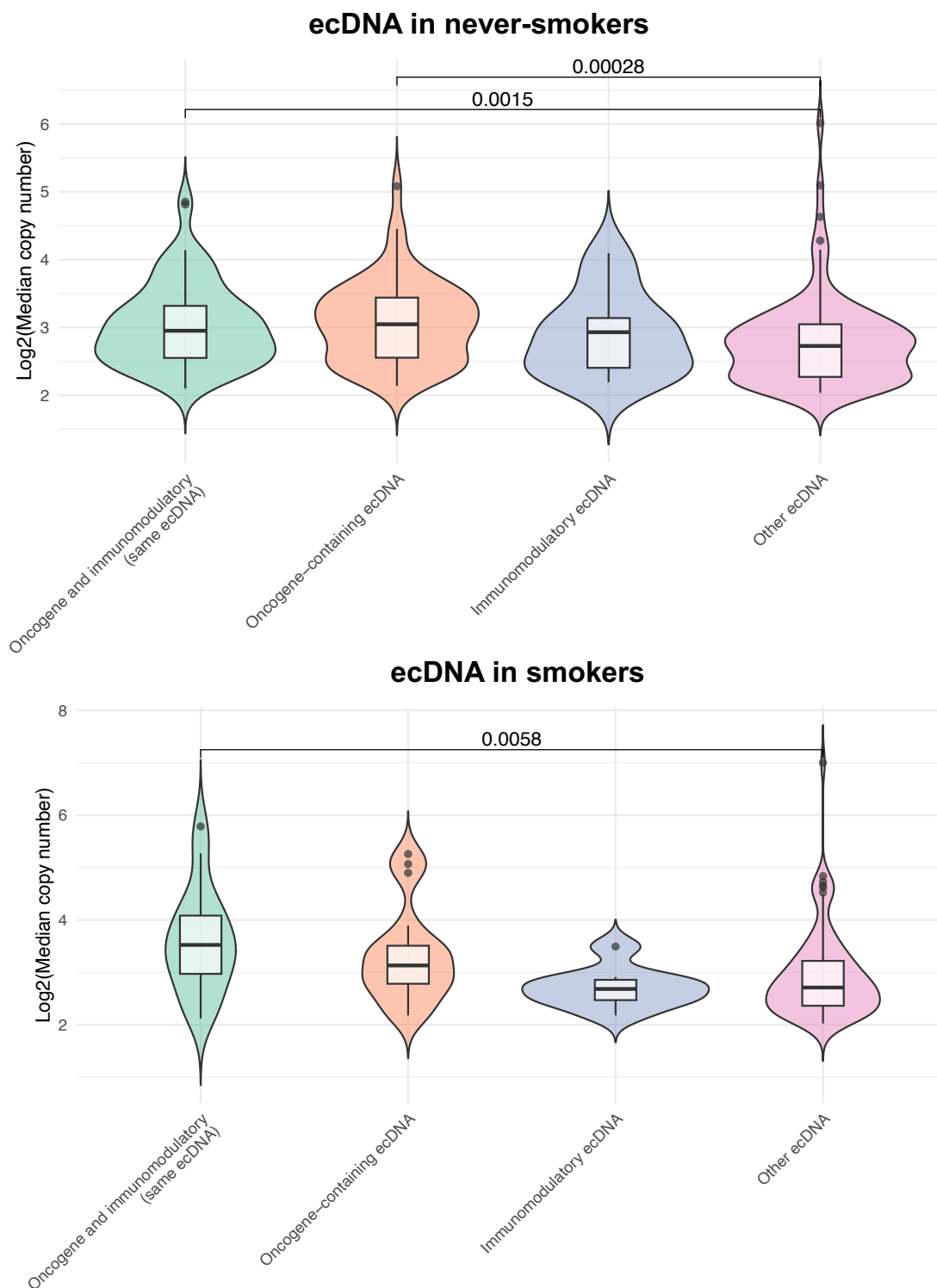


Figure 4.27: Copy number of genomic elements on ecDNA *Log2 median copy number* (*y-axis*) for each class of ecDNA (*x-axis*) in LCINS and LCSS.

Due to the clear enrichment of oncogenes on ecDNA, I next examined patterns of recurrent ecDNA amplification. The MDM2 gene stood out as recurrent hotspot for ecDNA amplification, as it was the most frequently amplified oncogene on ecDNA in both LCINS (n=36) and LCSS (n=5); (Fig. 4.28a). Given its prominence, I examined the relationship between ecDNA-driven MDM2 amplification and TP53 alterations, as MDM2 negatively regulates TP53 through ubiquitination(Chene, 2003). I found that MDM2 amplification on ecDNA was mutually exclusive with TP53 alterations in never-smokers (OR = 7.51, p = 0.01, Fig. 4.28), using logistic regression where the presence of MDM2-ecDNA is the binary outcome and a samples TP53 status is the independent variable, with age, sex, tumor purity, histology, and ancestry as covariates. The determination of TP53 status included both bi-allelic and mono-allelic alterations, and point mutations, indels and copy number deletions were taken into account. In LCINS, the next four most prevalent oncogenes that did not co-occur with MDM2 were TERT (n=10), CCND1 (n=8), and EGFR (n=7; Fig. 4.28a). In LCSS, no oncogene besides MDM2 was amplified on ecDNA in more than two samples (Fig. 4.28a). Indeed, oncogenes amplified via ecDNA appeared to reach higher copy number levels than those amplified via other means (Fig. 4.28b).

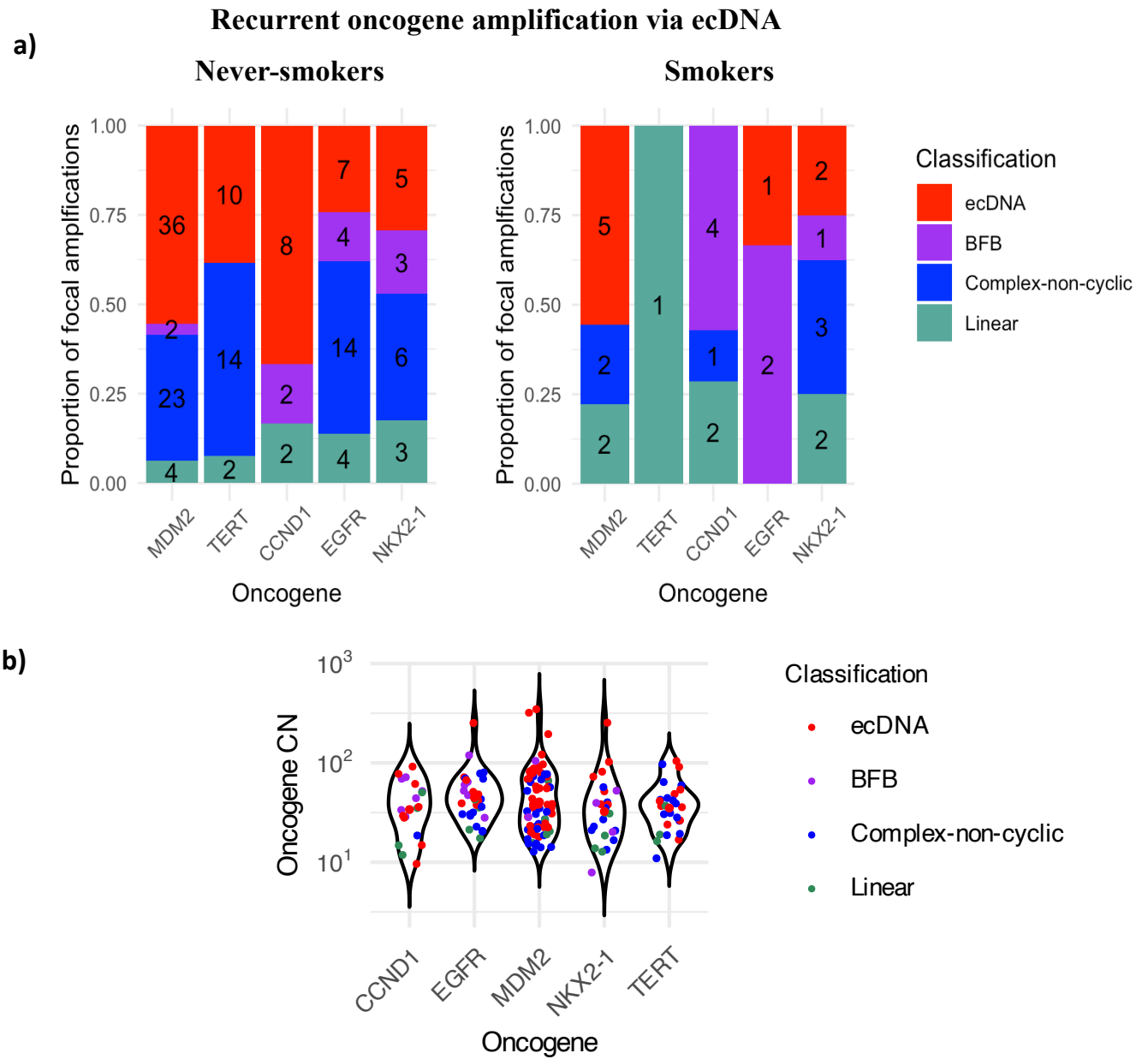
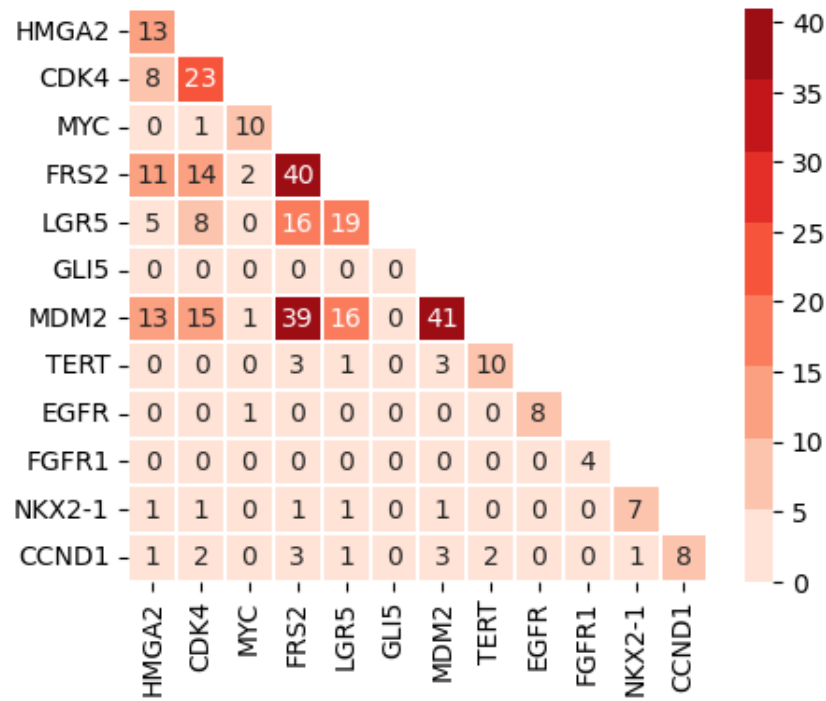


Figure 4.28: Oncogenes recurrently amplified via ecDNA in never-smokers and smokers
a) Bar plot showing the counts and proportions of the most frequently ecDNA-amplified oncogenes
b) Median copy number of each gene stratified by the mode of amplification

Co-occurrence of oncogenes on ecDNA



MDM2-ecDNA vs TP53 Status

OR=7.51, q=0.01, Adjusted by age, sex, ancestry, histology, purity

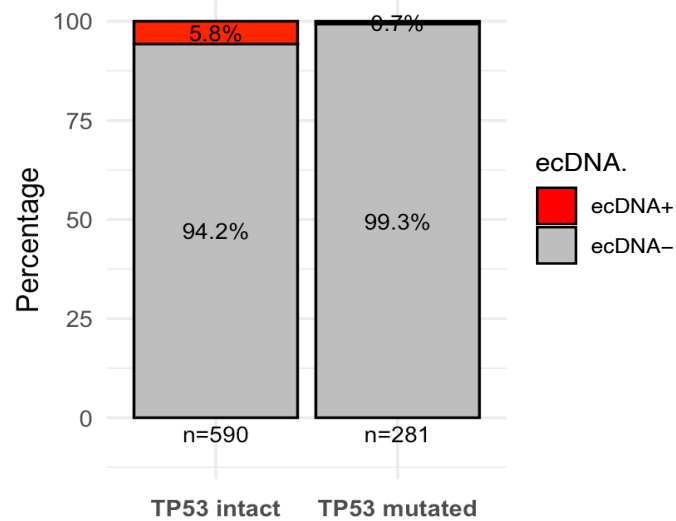


Figure 4.29: Co-occurrence of oncogenes on ecDNA and MDM2-ecDNA vs TP53 Status
 Bar plot showing the counts and proportions of the most frequently ecDNA-amplified oncogenes

4.1.15 Co-occurrence of oncogenes on ecDNA

Given that prior work suggested that passenger gene co-amplifications with oncogene amplifications on ecDNA can create collateral therapeutic vulnerabilities in cancer (Bei et al., 2024), I analyzed co-occurrence patterns of oncogenes detected on ecDNA across the cohort (Fig. 4.29). Here, each cell represents the number of samples in which a given pair of oncogenes was found on the same ecDNA element. Notably, the most common co-occurrences involved well-known oncogenic drivers such as MDM2 and FRS2 (39 samples), MDM2 and CDK4 (15 samples), and MDM2 and HMGA2 (13 samples).

4.1.16 Examining the relationship between chromothripsis and ecDNA

In order to detect chromothripsis regions in genomes, I applied the computational algorithm ShatterSeek (Cortés-Ciriano et al., 2020), which employs a set of statistical criteria to identify chromothripsis given input copy-number and structural-variant calls. I used the most stringent criteria corresponding to high-confidence chromothripsis calls. Specifically, these criteria were:

- (i) a cluster of structural variants (> 6 of DUP, DEL, h2hINV, or t2tINV),
- (ii) oscillating CNV between two states (> 7 CNV events),
- (iii) chromosomal enrichment and distribution of DNA breakpoints ($p < 0.05$),
- (iv) randomness of fragment joins ($p > 0.05$) and/or ≥ 4 inter-chromosomal rearrangements between multiple chromosomes.

The input CNV calls were generated using CNVkit in tumor-normal mode with default parameters, and the input structural-variant calls were generated using a union of Meerkat (v. 0.189) and Manta (v. 1.6.0) calls with the recommended filtering. Overall, I detected at least one high confidence chromothripsis events in 557 never-smokers (64%) and 236 smokers (68%) (Fig. 4.31). These rates of chromothripsis in lung cancer are much higher than the overall rate of chromothripsis reported in pan-cancer studies, which is 29%(Cosenza et al., 2022). 115 never-smokers and 60 smokers respectively had both a chromothripsis event and an ecDNA event (Fig. 4.31). Prior studies have indicated that chromothripsis may directly contribute to the formation of certain ecDNA(Shoshani et al., 2021). In samples where both ecDNA and chromothripsis were detected, overlap between ecDNA segments and chromothripsis regions was rare, occurring more frequently in LCINS—observed in 31 of 115 LCINS cases compared to 4 of 60 LCSS (OR=5.01, p=0.02 comparing LCINS vs. LCSS; (Fig. 4.31). In order to determine whether or not the absence of chromothripsis is due to the loss of all or parts of the chromosome involved, I examined the regions immediately surrounding the leftmost and rightmost ecDNA breakpoint for evidence of loss of heterozygosity. The vast majority of segments (82.7%) exhibited no signs of LOH and were confirmed to be diploid with a copy number of 2. Taken together, these results suggest the majority of ecDNA in lung cancer are generated through chromothripsis independent mechanisms.

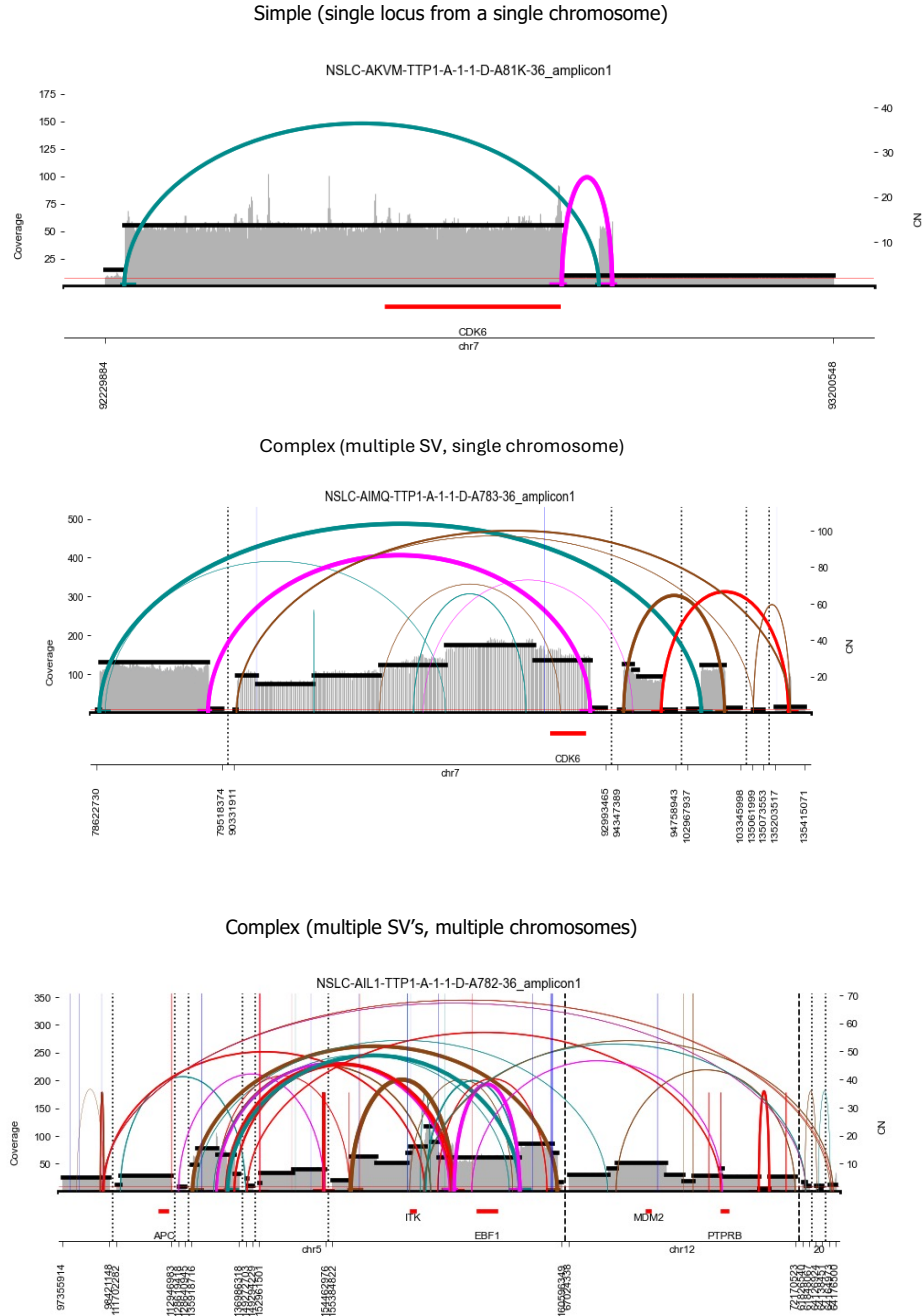


Figure 4.30: Exemplars of the types of ecDNA present in the Sherlock cohort A simple ecDNA amplicon composed of a single focal amplification and structural variant originating from a single locus on chromosome 7, encompassing the CDK6 oncogene (Top). A more complex ecDNA amplicon involving multiple structural variants within a single chromosome (chr7), also targeting CDK6 (Middle). A highly complex ecDNA structure involving multiple structural variants across multiple chromosomes (chr5 and chr12), including rearranged oncogenes such as EBF1, MDM2, PTPRB, and RYK (Bottom). Arcs denote rearrangement junctions; coverage depth and estimated copy number (CN) are plotted along the y-axes.

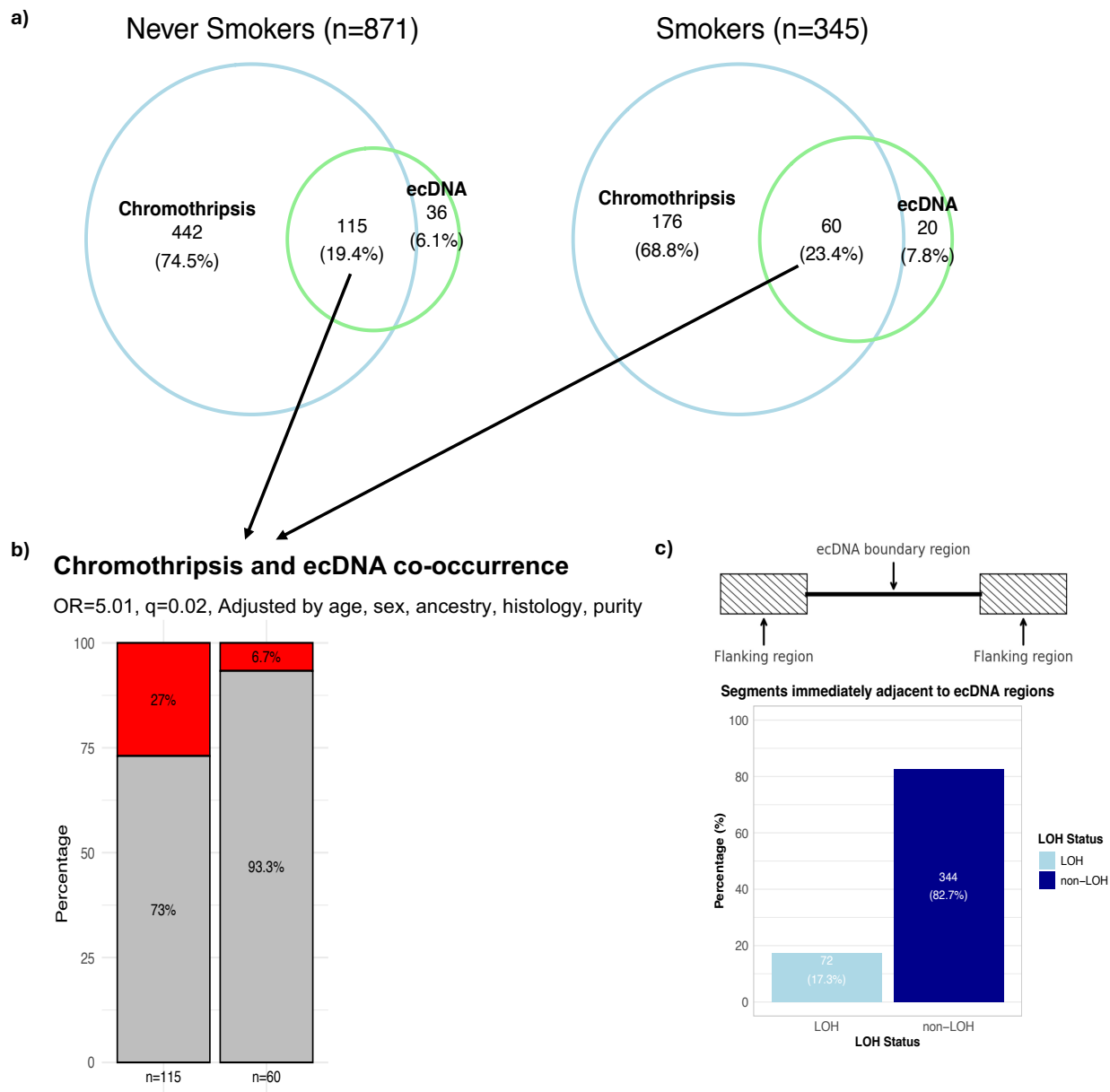


Figure 4.31: Co-occurrence of chromothripsis and ecDNA at the sample level and genomic sequence level **a)** Venn diagrams indicating the number and percentage of samples with chromothripsis events, ecDNA events, and both ecDNA and chromothripsis events. **b)** Bar plot indicating the proportion of samples that contained overlapping chromothripsis and ecDNA regions (red) and the proportion of all samples that did not contain any overlap (gray) for all samples that had both an ecDNA and chromothripsis event detected. **c)** Bar plot showing the number and percentage of segments immediately adjacent to ecDNA breakpoints that either exhibited LOH or did not exhibit LOH.

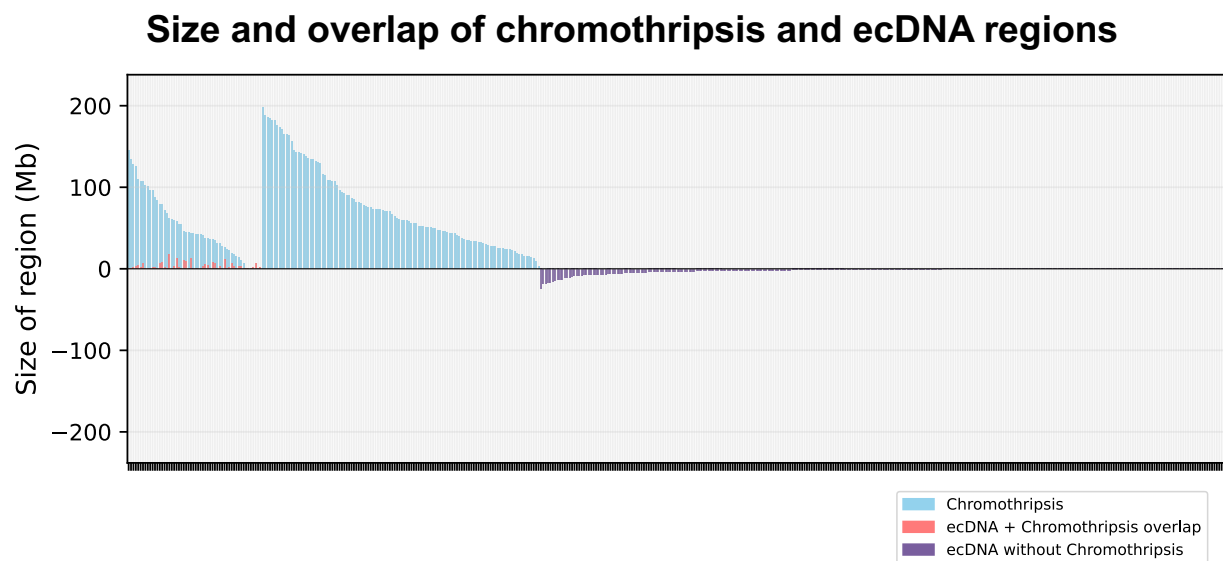


Figure 4.32: Size of chromothripsis regions, ecDNA regions, and overlapping regions
Bar plot showing the size of all ecDNA and chromothripsis regions detected. Blue bars indicate a chromothripsis region with no overlapping ecDNA regions. Red bars indicate a chromothripsis region with an overlapping ecDNA region contained within it. Purple bars indicate ecDNA regions with no overlap with any chromothripsis regions.

4.1.17 Clinical impact of ecDNA on patient survival

ecDNA has been associated with worse overall patient survival in prior studies, both as part of pan-cancer analyses (Bailey et al., 2024; Kim et al., 2020a, 2024) and in individual tumor subtypes (Chapman et al., 2023). To determine whether tumors harboring ecDNA had worse outcomes than those harboring only chromosomal amplifications, I performed survival analysis for all patients where clinical metadata (tumor stage and overall survival) were available using a Cox multivariate regression model, controlling for tumor stage, sex, age, histology, ancestry, and the overall genome instability index (wGII). TP53 status was not included in the model because it did not have a significant association with ecDNA in our cohort, as reported in a prior section. The analysis was conducted separately for never-smokers (n=693) and smokers (333), where each patient was stratified as having ecDNA, other focal amplifications, or no focal amplifications. Indeed, this analysis revealed that ecDNA-positive lung cancers had worse overall survival than those with only chromosomal amplifications in never-smokers (hazard ratio=2.17; p=0.025). In smokers, the effect was not significant (hazard ratio=1.65; p=0.01), however, the majority of smokers were not treated with any therapy, in contrast to the never-smokers.

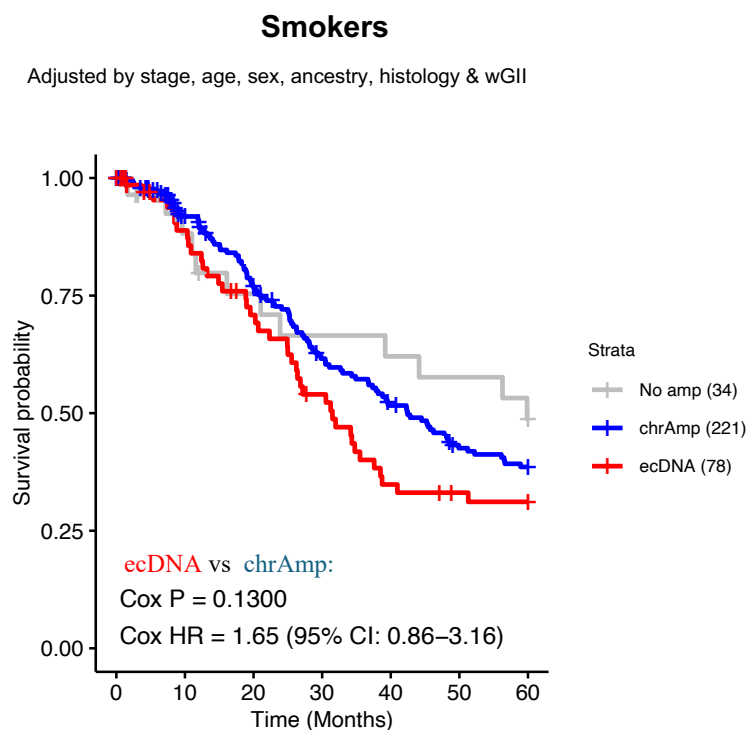
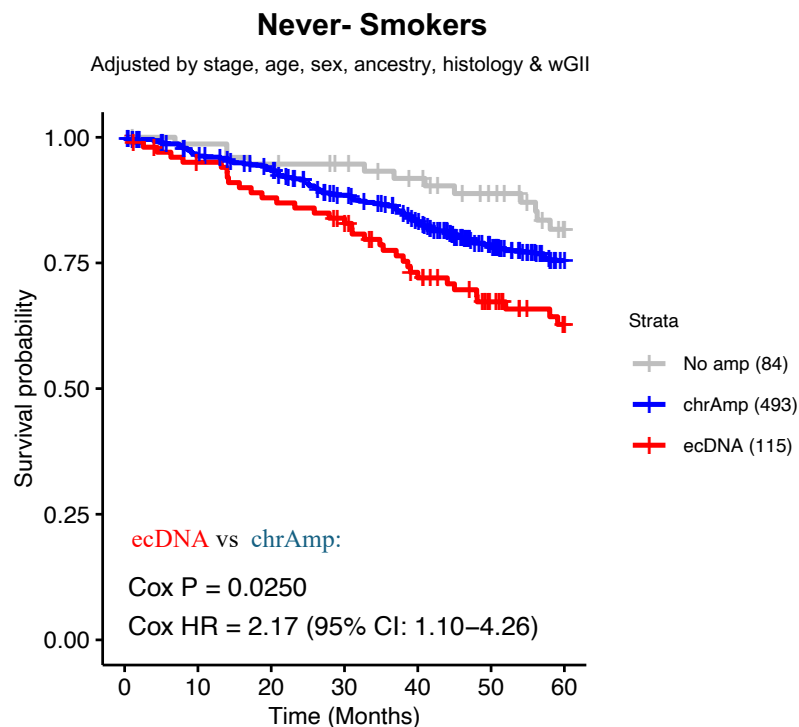


Figure 4.33: Impact of ecDNA on overall patient survival Kaplan–Meier survival curves for 5-year overall survival stratified by the mode of amplification status for lung cancers in never-smokers (left) and smokers (right). *P*-values for significance and HRs of the difference were calculated using two-sided Cox proportional-hazards regression with adjustment for age, sex, tumor stage, ancestry, wGII and histology and are indicated within each plot.

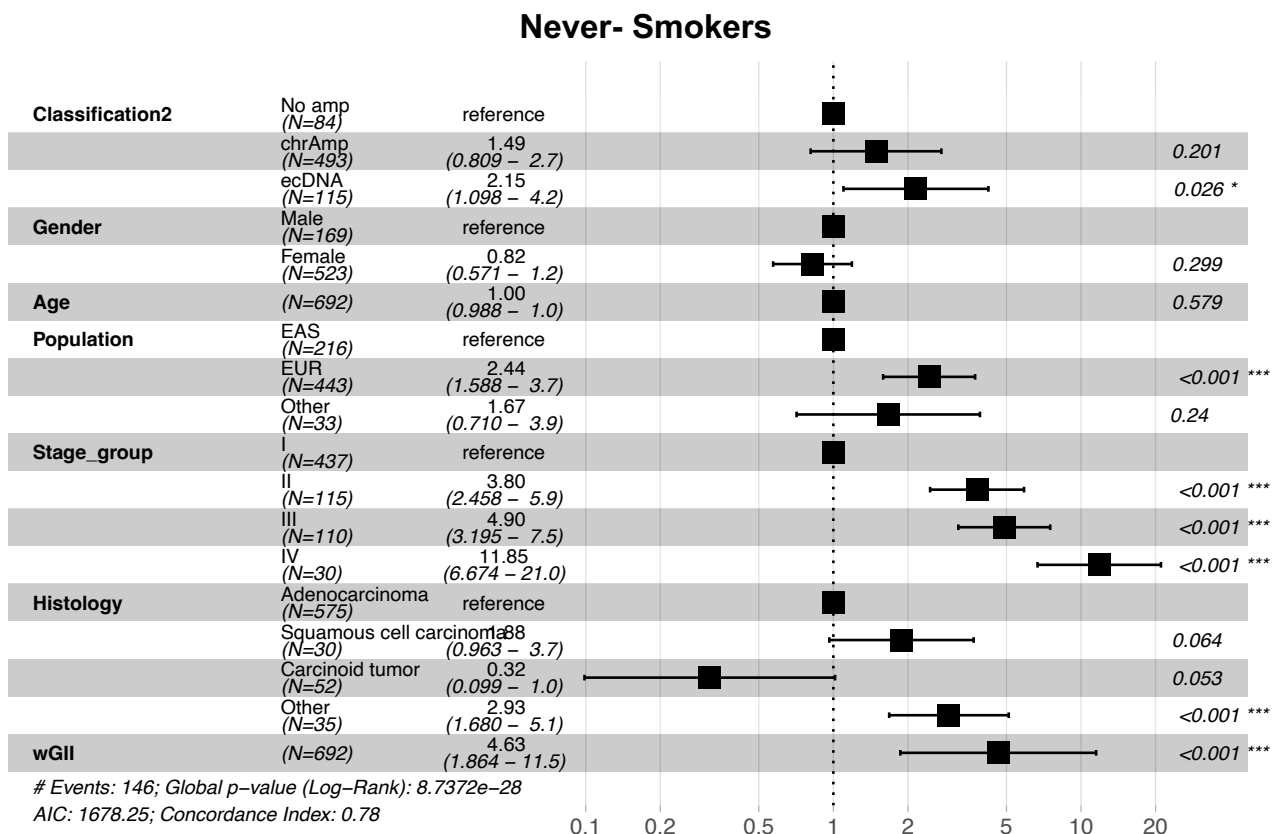


Figure 4.34: Detailed Cox regression survival analysis results for never-smokers Detailed results of a Cox regression model used to assess the survival differences of samples with ecDNA vs. samples with chrAmp. P-values, hazard ratios, and confidence intervals are shown for all covariates included in the model.

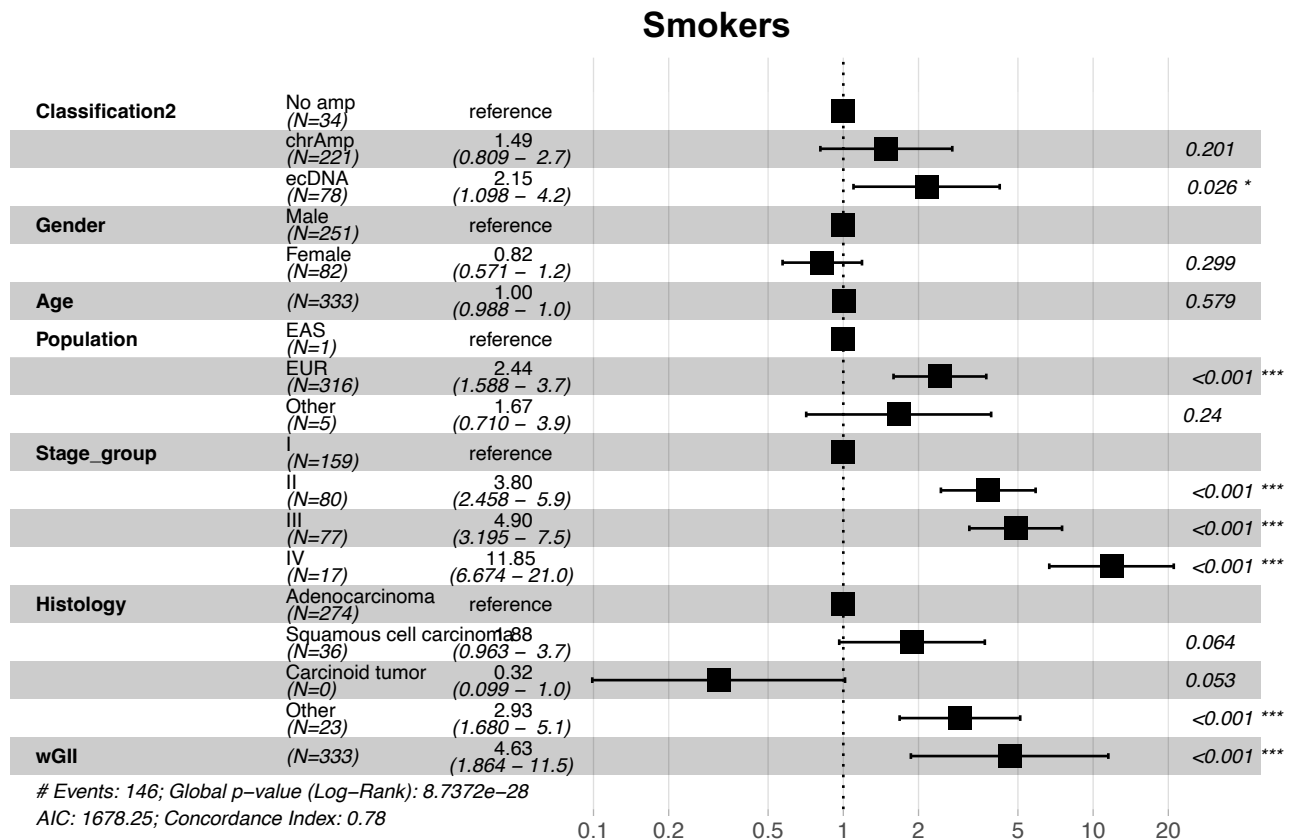


Figure 4.35: Detailed Cox regression survival analysis results for smokers Kaplan–Meier survival curves for 5-year overall survival stratified by the mode of amplification status for lung cancers in never-smokers (left) and smokers (right). P-values for significance and HRs of the difference were calculated using two-sided Cox proportional-hazards regression with adjustment for age, sex, tumor stage, ancestry, wGII and histology and are indicated within each plot.

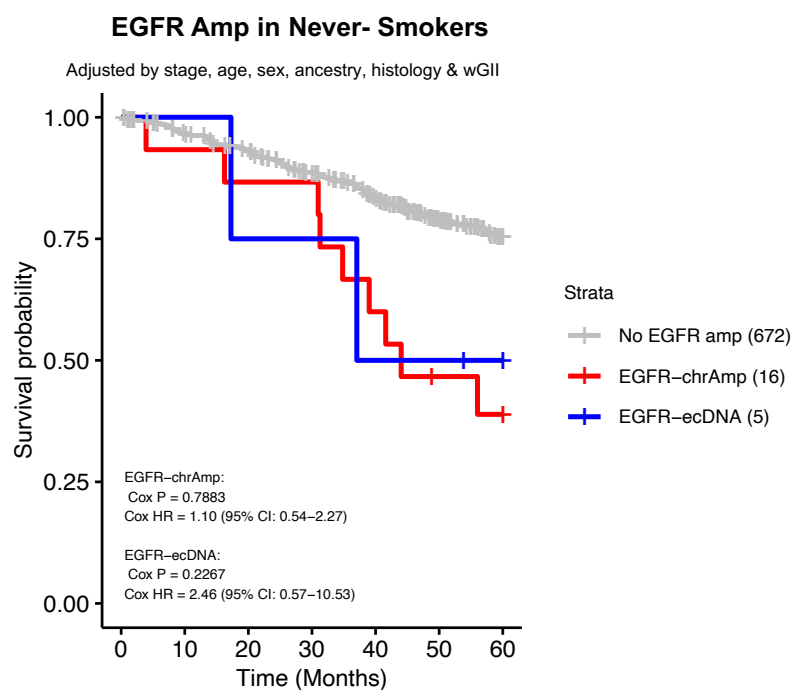
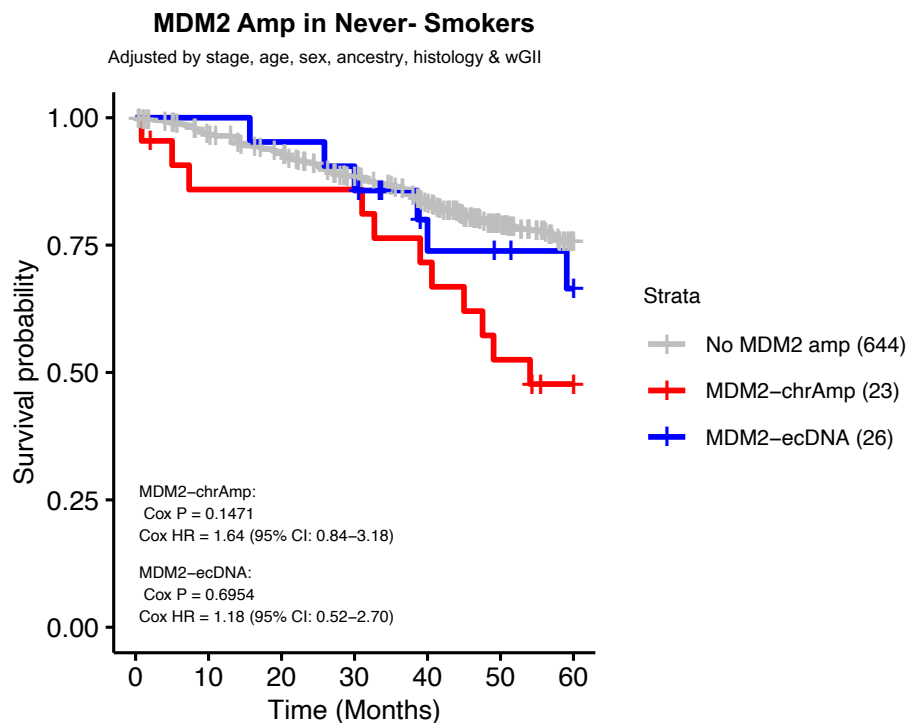


Figure 4.36: Impact of ecDNA containing MDM2 or EGFR on overall patient survival Kaplan–Meier survival curves for 5-year overall survival stratified by the mode of amplification status for the MDM2 oncogene specifically in lung cancers in never-smokers (top) and for the EGFR oncogene (bottom). Samples without any amplifications in these genes were considered the reference category. P-values for significance and HRs were calculated using two-sided Cox proportional-hazards regression with adjustment for age, sex, tumor stage, ancestry, wGII and histology and are indicated within each plot.

Overall Survival by ecDNA type (simple vs complex)

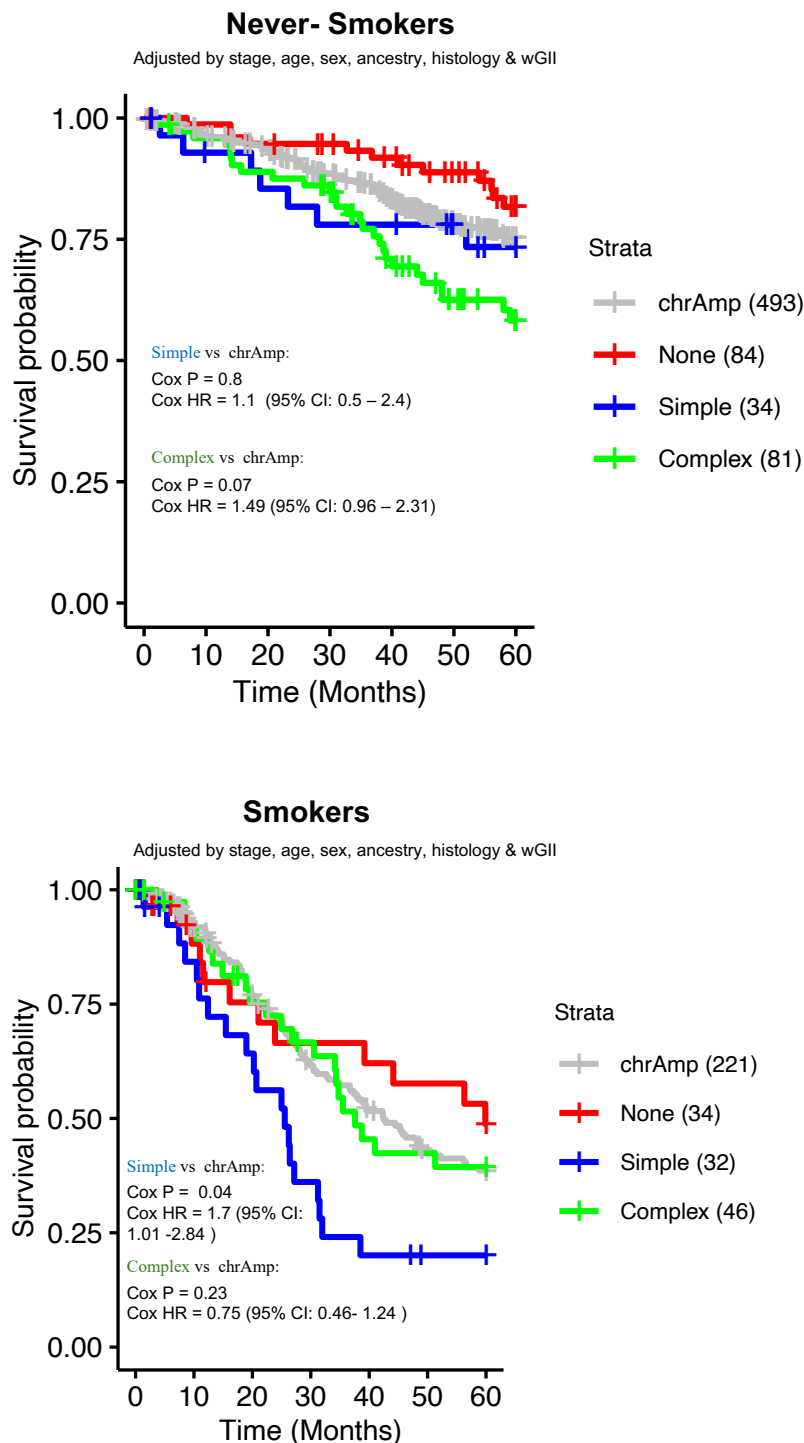


Figure 4.37: Impact of the type of ecDNA on overall patient survival Kaplan–Meier survival curves for 5-year overall survival stratified by the type of ecDNA (simple vs complex) for never-smokers (top) and smokers (bottom). Samples with chromosomal amplifications were considered the reference category for comparison. P-values for significance and HRs were calculated using two-sided Cox proportional-hazards regression with adjustment for age, sex, tumor stage, ancestry, wGII and histology and are indicated within each plot.

4.1.18 Discussion

ecDNA is a complex and poorly understood phenomenon, both in terms of causes and consequences. The size and diversity of this cohort allowed for robust statistical associations of a wide array of features associated with ecDNA, albeit without establishing causality. The lack of associations (e.g with TP53) and negative associations (e.g with KRAS in smokers) are also both unexpected and informative. Although most aspects of ecDNA biology (e.g prevalence rates, survival effect, enrichment of oncogenes, enrichment of APOBEC mutations on ecDNA) found in prior studies were also confirmed in the context of lung cancer, some findings, such as the transcriptional advantage of genes on ecDNA, were not confirmed in this study.

4.1.19 Conclusion and future directions

In conclusion, in this chapter, I conducted the most comprehensive examination of ecDNA in lung cancer. I mapped the full spectrum of ecDNA in the largest dataset of whole-genome sequenced lung cancers to date, from many geographical regions and ancestry, and with unique information on tobacco smoking status. Using a highly sensitive ecDNA detection pipeline that identifies both clonal and subclonal ecDNA by correcting for tumor purity and ploidy, I established the overall prevalence of ecDNA in lung cancer was 19 percent, with a increased prevalence in smokers (24 percent) compared to never-smokers (17 percent). The pipeline I developed also uncovered EGFR-ecDNA in 8 samples, as well as other putative ecDNA missed by the default pipeline. I also determined that among the numerous features associated with genome instability, its whole-genome doubling (in both never-smokers and

smokers) and chromothripsis (in smokers) that are most strongly coupled with ecDNA presence in never-smokers. The WGD association was validated in an independent pan-cancer cohort, suggesting that most ecDNA in treatment-naïve lung cancer are likely to be a byproduct of genomic instability. Intriguingly, ecDNA in never-smokers were enriched with oncogenes and immunomodulatory genes in comparison to smokers, indicating that a subset of LCINS may be driven by ecDNA amplifications. The APOBEC mutational signature is strongly enriched on ecDNA regions compared to background regions, and the clock-like signature SBS1 is markedly increased in ecDNA+ samples, indicating a higher rate of cell division. A significant proportion of ecDNA detected in this study were generated from chromothripsis-independent and TP53-independent mechanisms. Lastly, tumors carrying ecDNA showed worse overall survival in comparison to those with focal chromosomal amplification, particularly in LCINS. Overall, these results indicate that ecDNA carrying oncogenes could contribute to driving LCINS and affecting prognosis.

Although this study and several prior studies have examined features associated with ecDNA, the timing of ecDNA formation with respect to other events such whole-genome doubling is unknown. Since ecDNA does not follow the assumptions that current tools for timing mutational events are based on, such as the assumption of Mendelian inheritance, specialized methods will need to be developed in this space. Although this study did not find any clear differences in the immune landscape of ecDNA+ vs ecDNA- samples, the interaction of ecDNA with the immune system will also need to be examined in a more detailed manner, such as the immune landscape specifically of samples containing immunomodulatory genes on ecDNA. The detection of ecDNA from short-read sequencing data is not only challenging, but also impractical for clinical use as it requires whole-genome sequencing data. Thus, the

detection of ecDNA from different data modalities such as cell-free DNA is a promising avenue of investigation as well. This study focused on a samples sequenced at a single-time point, but the evolution of ecDNA using samples from multiple time points in a patient, specifically in a large cohort of lung cancers, has not been assessed to date. The genomic landscape of ecDNA after therapy, for example, would be important to examine.

Acknowledgments

I would like to acknowledge all the co-authors listed in the preprint. Specifically, I would like to acknowledge Tongwu Zhang for assistance with data analysis and visualization, and Jens Luebeck for helpful feedback regarding ecDNA detection and the AmpliconSuite pipeline. I would like to particularly thank my supervisors, Ludmil Alexandrov and Maria Teresa Landi, as well for funding and guidance.

Chapter 4, in part, has been submitted for publication of the material as it may appear in Examining the Role of Extrachromosomal DNA in 1,216 Lung Cancers Khandekar, Azhar; Hoang, Phuc H.; Luebeck, Jens; Díaz-Gay, Marcos; Zhao, Wei; McElderry, John P.; Hartman, Caleb; Mirafteb, Mona; Lee, Olivia W.; Barnao, Kara M.; Bergstrom, Erik N.; Yang, Yang; Nowak, Martin A.; Rothman, Nathaniel; Homer, Robert; Yang, Soo-Ryum; Lan, Qing; Wedge, David C.; Yang, Lixing; Chanock, Stephen J.; Zhang, Tongwu; Alexandrov, Ludmil B.; Landi, Maria Teresa, Nature Genetics, 2025. The dissertation author was the primary researcher and author of this paper.

Bibliography

- L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*, 3(1):246–59, 2013. doi: 10.1016/j.celrep.2012.12.008. URL <https://www.ncbi.nlm.nih.gov/pubmed/23318258>.
- L. B. Alexandrov, P. H. Jones, D. C. Wedge, J. E. Sale, P. J. Campbell, S. Nik-Zainal, and M. R. Stratton. Clock-like mutational processes in human somatic cells. *Nat Genet*, 47(12):1402–7, 2015. doi: 10.1038/ng.3441. URL <https://www.ncbi.nlm.nih.gov/pubmed/26551669>.
- L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, PcaWG Mutational Signatures Working Group, G. Getz, S. G. Rozen, M. R. Stratton, and PcaWG Consortium. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, 2020. doi: 10.1038/s41586-020-1943-3. URL <https://www.ncbi.nlm.nih.gov/pubmed/32025018>.
- D. Aran, M. Sirota, and A. J. Butte. Systematic pan-cancer analysis of tumour purity. *Nat Commun*, 6:8971, 2015. doi: 10.1038/ncomms9971.

- V. Bafna and P. S. Mischel. Extrachromosomal dna in cancer. *Annu Rev Genomics Hum Genet*, 23:29–52, 2022. doi: 10.1146/annurev-genom-120821-100535.
- C. Bailey, O. Pich, K. Thol, T. B. K. Watkins, J. Luebeck, A. Rowan, G. Stavrou, N. E. Weiser, B. Dameracharla, R. Bentham, et al. Origins and impact of extrachromosomal dna. *Nature*, 635(8037):193–200, 2024. doi: 10.1038/s41586-024-08107-3.
- M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 174(4):1034–1035, 2018. doi: 10.1016/j.cell.2018.07.034.
- G. Balaban-Malenbaum and F. Gilbert. Double minute chromosomes and the homogeneously staining regions in chromosomes of a human neuroblastoma cell line. *Science*, 198(4318):739–741, 1977. doi: 10.1126/science.71759.
- Yi Bei, Luca Bramé, Marieluise Kirchner, Raphaela Fritsche-Guenther, Severine Kunz, Animesh Bhattacharya, Mara-Camelia Rusu, Dennis Gürgen, Frank P. B. Dubios, Julia K. C. Köppke, Jutta Proba, Nadine Wittstruck, Olga Alexandra Sidorova, Rocío Chamorro González, Heathcliff Dorado Garcia, Lotte Brückner, Robin Xu, Mădălina Giurgiu, Elias Rodriguez-Fos, Qinghao Yu, Bastiaan Spanjaard, Richard P. Koche, Clemens A. Schmitt, Johannes H. Schulte, Angelika Eggert, Kerstin Haase, Jennifer Kirwan, Anja I. H. Hagemann, Philipp Mertins, Jan R. Dörr, and Anton G. Henssen. Passenger gene coamplifications create collateral therapeutic vulnerabilities in cancer. *Cancer Discovery*, 14(3):492–507, March 2024. doi: 10.1158/2159-8290.CD-23-1189.

- S. E. Benner, G. M. Wahl, and D. D. Von Hoff. Double minute chromosomes and homogeneously staining regions in tumors taken directly from patients versus in human tumor cell lines. *Anticancer Drugs*, 2(1):11–25, 1991. doi: 10.1097/00001813-199102000-00002.
- E. N. Bergstrom, M. N. Huang, U. Mahto, M. Barnes, M. R. Stratton, S. G. Rozen, and L. B. Alexandrov. Sigproflermatrixgenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics*, 20(1):685, 2019. doi: 10.1186/s12864-019-6041-2. URL <https://www.ncbi.nlm.nih.gov/pubmed/31470794>.
- E. N. Bergstrom, J. Luebeck, M. Petljak, A. Khandekar, M. Barnes, T. Zhang, C. D. Steele, N. Pillay, M. T. Landi, V. Bafna, et al. Mapping clustered mutations in cancer reveals apobec3 mutagenesis of ecDNA. *Nature*, 602(7897):510–517, 2022. doi: 10.1038/s41586-022-04398-6.
- C. M. Bielski, A. Zehir, A. V. Penson, M. T. A. Donoghue, W. Chatila, J. Armenia, M. T. Chang, A. M. Schram, P. Jonsson, C. Bandlamudi, et al. Genome doubling shapes the evolution and prognosis of advanced cancers. *Nat Genet*, 50(8):1189–1195, 2018. doi: 10.1038/s41588-018-0165-1.
- David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- D. L. Cameron, L. Di Stefano, and A. T. Papenfuss. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun*, 10(1):3240, 2019. doi: 10.1038/s41467-019-11146-4. URL <https://www.ncbi.nlm.nih.gov/pubmed/31324872>.

- Network Cancer Genome Atlas Research, J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10):1113–20, 2013. doi: 10.1038/ng.2764. URL <https://www.ncbi.nlm.nih.gov/pubmed/24071849>.
- Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, and Barbara A Weir. Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5): 413–421, 2012.
- D. Chakravarty, J. Gao, S. M. Phillips, R. Kundra, H. Zhang, J. Wang, J. E. Rudolph, R. Yaeger, T. Soumerai, M. H. Nissan, et al. Oncokb: A precision oncology knowledge base. *JCO Precis Oncol*, 2017:PO.17.00011, 2017. doi: 10.1200/PO.17.00011.
- O. S. Chapman, J. Luebeck, S. Sridhar, I. T. Wong, D. Dixit, S. Wang, G. Prasad, U. Rajkumar, M. S. Pagadala, J. D. Larson, et al. Circular extrachromosomal dna promotes tumor heterogeneity in high-risk medulloblastoma. *Nat Genet*, 55(12):2189–2199, 2023. doi: 10.1038/s41588-023-01551-3.
- X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Kallberg, A. J. Cox, S. Kruglyak, and C. T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–2, 2016a. doi: 10.1093/bioinformatics/btv710. URL <https://www.ncbi.nlm.nih.gov/pubmed/26647377>.
- X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Kallberg, A. J. Cox, S. Kruglyak, and C. T. Saunders. Manta: rapid detection of structural variants and indels

- for germline and cancer sequencing applications. *Bioinformatics*, 32(8):1220–1222, 2016b. doi: 10.1093/bioinformatics/btv710.
- P. Chene. Inhibiting the p53-mdm2 interaction: an important target for cancer therapy. *Nat Rev Cancer*, 3(2):102–109, 2003. doi: 10.1038/nrc991.
- F. Cheng, J. Zhao, Y. Wang, W. Lu, Z. Liu, Y. Zhou, W. R. Martin, R. Wang, J. Huang, T. Hao, et al. Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat Genet*, 53(3):342–353, 2021. doi: 10.1038/s41588-020-00774-y.
- Z. N. Choo, J. M. Behr, A. Deshpande, K. Hadi, X. Yao, H. Tian, K. Takai, G. Zakusilo, J. Rosiene, A. Da Cruz Paula, et al. Most large structural variants in cancer genomes can be detected without long reads. *Nat Genet*, 55(12):2139–2148, 2023. doi: 10.1038/s41588-023-01540-6.
- D. Consonni, M. Carugno, S. De Matteis, F. Nordio, G. Randi, M. Bazzano, N. E. Caporaso, M. A. Tucker, P. A. Bertazzi, A. C. Pesatori, et al. Outdoor particulate matter (pm10) exposure and lung cancer risk in the eagle study. *PLoS One*, 13(9):e0203539, 2018. doi: 10.1371/journal.pone.0203539.
- Icgc Tcga Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020. doi: 10.1038/s41586-020-1969-6. URL <https://www.ncbi.nlm.nih.gov/pubmed/32025007>.
- Isidro Cortés-Ciriano, Jake June-Koo Lee, Ruibin Xi, Dhawal Jain, Youngsook L. Jung, Lixing Yang, Dmitry Gordenin, Leszek J. Klimczak, Cheng-Zhong Zhang, David S. Pellman, PCAWG Structural Variation Working Group, PCAWG Consortium, et al. Comprehensive

- analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, 52(3):331–341, 2020. doi: 10.1038/s41588-019-0576-7.
- M. R. Cosenza, B. Rodriguez-Martin, and J. O. Korb. Structural variation in cancer: Role, prevalence, and mechanisms. *Annu Rev Genomics Hum Genet*, 23:123–152, 2022. doi: 10.1146/annurev-genom-120121-101149. URL <https://www.ncbi.nlm.nih.gov/pubmed/35655332>.
- H. Davies, D. Glodzik, S. Morganella, L. R. Yates, J. Staaf, X. Zou, M. Ramakrishna, S. Martin, S. Boyault, A. M. Sieuwerts, et al. Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nat Med*, 23(4):517–525, 2017. doi: 10.1038/nm.4292.
- A. Degasperi, X. Zou, T. D. Amarante, A. Martinez-Martinez, G. C. C. Koh, J. M. L. Dias, L. Heskin, L. Chmelova, G. Rinaldi, V. Y. W. Wang, A. S. Nanda, A. Bernstein, S. E. Momen, J. Young, D. Perez-Gil, Y. Memari, C. Badja, S. Shooter, J. Czarnecki, M. A. Brown, H. R. Davies, Consortium Genomics England Research, and S. Nik-Zainal. Substitution mutational signatures in whole-genome-sequenced cancers in the uk population. *Science*, 376(6591), 2022. doi: 10.1126/science.abl9283. URL <https://www.ncbi.nlm.nih.gov/pubmed/35949260>.
- Andrea Degasperi, Tauanne Dias Amarante, Jan Czarnecki, Scott Shooter, Xueqing Zou, Dominik Glodzik, Sandro Morganella, Arjun S Nanda, Cherif Badja, and Gene Koh. A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nature cancer*, 1(2):249–263, 2020.

- V. Deshpande, J. Luebeck, N. D. Nguyen, M. Bakhtiari, K. M. Turner, R. Schwab, H. Carter, P. S. Mischel, and V. Bafna. Exploring the landscape of focal amplifications in cancer using ampliconarchitect. *Nat Commun*, 10(1):392, 2019. doi: 10.1038/s41467-018-08200-y.
- M. Diaz-Gay, T. Zhang, P. H. Hoang, A. Khandekar, W. Zhao, C. D. Steele, B. Otlu, S. P. Nandi, R. Vangara, E. N. Bergstrom, et al. The mutagenic forces shaping the genomic landscape of lung cancer in never smokers. *medRxiv*, page 2024.05.15.24307318, 2024. doi: 10.1101/2024.05.15.24307318.
- Marcos Díaz-Gay, Wellington dos Santos, Sarah Moody, Mariya Kazachkova, Ammal Abbasi, Christopher D. Steele, Raviteja Vangara, Sergey Senkin, Jingwei Wang, Simon Fitzgerald, Erik N. Bergstrom, Azhar Khandekar, Burçak Otlu, Behnoush Abedi-Ardekani, Ana Carolina de Carvalho, Thomas Cattiaux, Ricardo Cortez Cardoso Penha, Valérie Gaborieau, Priscilia Chopard, Beata Świątkowska, Jolanta Lissowska, Katarzyna Roszkowska-Purska, Antonio Huertas-Salgado, Tatsuhiro Shibata, Satoshi Shiba, Surasak Sangkhathat, Taned Chitapanarux, Gholamreza Roshandel, Patricia Ashton-Prolla, Daniel C. Damin, Francine H. de Oliveira, Laura Humphreys, Trevor D. Lawley, Sandra Perdomo, Michael R. Stratton, Paul Brennan, and Ludmil B. Alexandrov. Geographic and age variations in mutational processes in colorectal cancer. *Nature*, 643:230–240, 2025. doi: 10.1038/s41586-025-09025-8. Published 23 April 2025.
- Z. Ding, M. Mangino, A. Aviv, T. Spector, R. Durbin, and UK10K Consortium. Estimating telomere length from whole genome sequence data. *Nucleic Acids Res*, 42(9):e75, 2014. doi: 10.1093/nar/gku181.

R. M. Drews, B. Hernando, M. Tarabichi, K. Haase, T. Lesluyes, P. S. Smith, L. Morrill Gavarro, D. L. Couturier, L. Liu, M. Schneider, J. D. Brenton, P. Van Loo, G. Macintyre, and F. Markowitz. A pan-cancer compendium of chromosomal instability. *Nature*, 606(7916):976–983, 2022. doi: 10.1038/s41586-022-04789-9. URL <https://www.ncbi.nlm.nih.gov/pubmed/35705807>.

Marcos Díaz-Gay, Raviteja Vangara, Mark Barnes, Xi Wang, S M Ashiqul Islam, Ian Vermes, Stephen Duke, Nithish Bharadhwaj Narasimman, Ting Yang, Zichen Jiang, Sarah Moody, Sergey Senkin, Paul Brennan, Michael R Stratton, and Ludmil B Alexandrov. Assigning mutational signatures to individual samples and individual somatic mutations with sigproflerassignment. *Bioinformatics*, 39(12):btad756, 12 2023.

Evan E. Eichler et al. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 2011.

F. Favero, T. Joshi, A. M. Marquard, N. J. Birkbak, M. Krzystanek, Q. Li, Z. Szallasi, and A. C. Eklund. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol*, 26(1):64–70, 2015. doi: 10.1093/annonc/mdu479. URL <https://www.ncbi.nlm.nih.gov/pubmed/25319062>.

Migle Gabrielaite, Mathias Husted Torp, Malthe Sebro Rasmussen, Sergio Andreu-Sánchez, Filipe Garrett Vieira, Christina Bligaard Pedersen, Savvas Kinalis, Majbritt Busk Madsen, Miyako Kodama, Gül Sude Demircan, Arman Simonyan, Christina Westmose Yde, Lars Rønn Olsen, Rasmus L. Marvig, Olga Østrup, Maria Rossing, Finn Cilius Nielsen, Ole Winther, and Frederik Otzen Bagger. A comparison of tools for copy-number variation

- detection in germline whole exome and whole genome sequencing data. *Frontiers in Genetics*, 12:689788, 2021. doi: 10.3389/fgene.2021.689788. Editor: Dimitrios H Roukos.
- D. Glodzik, S. Morganello, H. Davies, P. T. Simpson, Y. Li, X. Zou, J. Diez-Perez, J. Staaf, L. B. Alexandrov, M. Smid, A. B. Brinkman, I. H. Rye, H. Russnes, K. Raine, C. A. Purdie, S. R. Lakhani, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. Martens, A. L. Borresen-Dale, A. L. Richardson, G. Kong, A. Viari, D. Easton, G. Evan, P. J. Campbell, M. R. Stratton, and S. Nik-Zainal. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat Genet*, 49(3):341–348, 2017. doi: 10.1038/ng.3771. URL <https://www.ncbi.nlm.nih.gov/pubmed/28112740>.
- C. Graux, J. Cools, C. Melotte, H. Quentmeier, A. Ferrando, R. Levine, J. R. Vermeesch, M. Stul, B. Dutta, N. Boeckx, A. Bosly, P. Heimann, A. Uyttebroeck, N. Mentens, R. Somers, R. A. MacLeod, H. G. Drexler, A. T. Look, D. G. Gilliland, L. Michaux, P. Vandenberghe, I. Wlodarska, P. Marynen, and A. Hagemeijer. Fusion of nup214 to abl1 on amplified episomes in t-cell acute lymphoblastic leukemia. *Nat Genet*, 36(10):1084–1089, 2004. doi: 10.1038/ng1425.
- Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the national academy of sciences*, 101(suppl_1):5228–5235, 2004.
- Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, March 4 2011. doi: 10.1016/j.cell.2011.02.013.

- Matthew Hoffman, Francis R Bach, and David M Blei. Online learning for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 856–864, 2010.
- K. L. Hung, K. E. Yost, L. Xie, Q. Shi, K. Helmsauer, J. Luebeck, R. Schopflin, J. T. Lange, R. Chamorro Gonzalez, N. E. Weiser, et al. ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature*, 600(7890):731–736, 2021. doi: 10.1038/s41586-021-04116-8.
- K. L. Hung, M. G. Jones, I. T. Wong, E. J. Curtis, J. T. Lange, B. J. He, J. Luebeck, R. Schmargon, E. Scanu, L. Bruckner, et al. Coordinated inheritance of extrachromosomal dnAs in cancer cells. *Nature*, 635(8037):201–209, 2024. doi: 10.1038/s41586-024-07861-8.
- ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, 2020. doi: 10.1038/s41586-020-1969-6.
- M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–1120, 2012. doi: 10.1016/j.cell.2012.08.029.
- J. Inazawa, T. Abe, K. Inoue, H. Nishigaki, S. Horiike, M. Taniwaki, S. Misawa, and T. Takino. Simultaneous existence of double minute chromosomes and a homogeneously staining region in a retinoblastoma cell line (y79) and amplification of n-myc at hsr. *Cancer Genet Cytogenet*, 37(1):133–137, 1989. doi: 10.1016/0165-4608(89)90082-4.
- S. M. A. Islam, M. Diaz-Gay, Y. Wu, M. Barnes, R. Vangara, E. N. Bergstrom, Y. He, M. Vella, J. Wang, J. W. Teague, P. Clapham, S. Moody, S. Senkin, Y. R. Li, L. Riva, T. Zhang, A. J. Gruber, C. D. Steele, B. Otlu, A. Khandekar, A. Abbasi, L. Humphreys, N. Syulyukina,

- S. W. Brady, B. S. Alexandrov, N. Pillay, J. Zhang, D. J. Adams, I. Martincorena, D. C. Wedge, M. T. Landi, P. Brennan, M. R. Stratton, S. G. Rozen, and L. B. Alexandrov. Uncovering novel mutational signatures by de novo extraction with sigproflerextractor. *Cell Genom*, 2(11):None, 2022a. doi: 10.1016/j.xgen.2022.100179. URL <https://www.ncbi.nlm.nih.gov/pubmed/36388765>.
- S. M. A. Islam, M. Diaz-Gay, Y. Wu, M. Barnes, R. Vangara, E. N. Bergstrom, Y. He, M. Vella, J. Wang, J. W. Teague, et al. Uncovering novel mutational signatures by de novo extraction with sigproflerextractor. *Cell Genom*, 2(11):100179, 2022b. doi: 10.1016/j.xgen.2022.100179.
- R. J. Kaufman, P. C. Brown, and R. T. Schimke. Amplified dihydrofolate reductase genes in unstably methotrexate-resistant cells are associated with double minute chromosomes. *Proc Natl Acad Sci U S A*, 76(11):5669–5673, 1979. doi: 10.1073/pnas.76.11.5669.
- Azhar Khandekar, Raviteja Vangara, Mark Barnes, Marcos Díaz-Gay, Ammal Abbasi, Erik N. Bergstrom, Christopher D. Steele, Nischalan Pillay, and Ludmil B. Alexandrov. Visualizing and exploring patterns of large mutational events with sigproflermatrixgenerator. *BMC Genomics*, 24(1):469, 2023. doi: 10.1186/s12864-023-09584-y.
- H. Kim, N. P. Nguyen, K. Turner, S. Wu, A. D. Gujar, J. Luebeck, J. Liu, V. Deshpande, U. Rajkumar, S. Namburi, S. B. Amin, E. Yi, F. Menghi, J. H. Schulte, A. G. Henssen, H. Y. Chang, C. R. Beck, P. S. Mischel, V. Bafna, and R. G. W. Verhaak. Extrachromosomal dna is associated with oncogene amplification and poor outcome across multiple cancers.

- Nat Genet*, 52(9):891–897, 2020a. doi: 10.1038/s41588-020-0678-2. URL <https://www.ncbi.nlm.nih.gov/pubmed/32807987>.
- H. Kim, N. P. Nguyen, K. Turner, S. Wu, A. D. Gujar, J. Luebeck, J. Liu, V. Deshpande, U. Rajkumar, S. Namburi, et al. Extrachromosomal dna is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet*, 52(9):891–897, 2020b. doi: 10.1038/s41588-020-0678-2.
- H. Kim, S. Kim, T. Wade, E. Yeo, A. Lipsa, A. Golebiewska, K. C. Johnson, S. An, J. Ko, Y. Nam, et al. Mapping extrachromosomal dna amplifications during cancer progression. *Nat Genet*, 56(11):2447–2454, 2024. doi: 10.1038/s41588-024-01949-7.
- M. T. Landi, N. C. Synnott, J. Rosenbaum, T. Zhang, B. Zhu, J. Shi, W. Zhao, M. Kebede, J. Sang, J. Choi, et al. Tracing lung cancer risk factors through mutational signatures in never-smokers. *Am J Epidemiol*, 190(6):962–976, 2021. doi: 10.1093/aje/kwaa234.
- J. T. Lange, J. C. Rose, C. Y. Chen, Y. Pichugin, L. Xie, J. Tang, K. L. Hung, K. E. Yost, Q. Shi, M. L. Erb, et al. The evolutionary dynamics of extrachromosomal dna in human cancers. *Nat Genet*, 54(10):1527–1533, 2022. doi: 10.1038/s41588-022-01177-x.
- J. J. Lee, S. Park, H. Park, S. Kim, J. Lee, J. Lee, J. Youk, K. Yi, Y. An, I. K. Park, et al. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell*, 177(7):1842–1857.e21, 2019. doi: 10.1016/j.cell.2019.05.013.
- T. L. Leong, V. Gayevskiy, D. P. Steinfort, M. R. De Massy, A. Gonzalez-Rajal, K. D. Marini, E. Stone, V. Chin, A. Havryk, M. Plit, et al. Deep multi-region whole-genome sequencing

- reveals heterogeneity and gene-by-environment interactions in treatment-naive, metastatic lung cancer. *Oncogene*, 38(10):1661–1675, 2019. doi: 10.1038/s41388-018-0536-1.
- Y. Li, N. D. Roberts, J. A. Wala, O. Shapira, S. E. Schumacher, K. Kumar, E. Khurana, S. Waszak, J. O. Korb, J. E. Haber, M. Imielinski, Pcawg Structural Variation Working Group, J. Weischenfeldt, R. Beroukhi, P. J. Campbell, and Pcawg Consortium. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, 2020. doi: 10.1038/s41586-019-1913-9. URL <https://www.ncbi.nlm.nih.gov/pubmed/32025012>.
- J. Luebeck, A. W. T. Ng, P. C. Galipeau, X. Li, C. A. Sanchez, A. C. Katz-Sumercorn, H. Kim, S. Jammula, Y. He, S. M. Lippman, et al. Extrachromosomal dna in the cancerous transformation of barrett’s oesophagus. *Nature*, 616(7958):798–805, 2023. doi: 10.1038/s41586-023-05937-5.
- Daniial Masood, Luyao Ren, Cu Nguyen, Francesco G. Brundu, Lily Zheng, Yongmei Zhao, Erich Jaeger, Yong Li, Seong Won Cha, Aaron Halpern, Sean Truong, Michael Virata, Chunhua Yan, Qingrong Chen, Andy Pang, Reyes Alberto, Chunlin Xiao, Zhaowei Yang, Wanqiu Chen, Charles Wang, Frank Cross, Severine Catreux, Leming Shi, Julia A. Beaver, Wenming Xiao, and Daoud M. Meerzaman. Evaluation of somatic copy number variation detection by ngs technologies and bioinformatics tools on a hyper-diploid cancer genome. *Genome Biology*, 25(1):163, 2024. doi: 10.1186/s13059-024-03294-8.
- Sarah Moody, Sergey Senkin, S. M. Ashiqul Islam, Jaeil Wang, Dariush Nasrollahzadeh, Rita Cortez Cardoso Penha, Simon Fitzgerald, Erik N. Bergstrom, James R. Atkins, Yuan He, Azhar Khandekar, Kate Smith-Byrne, Carla Carreira, Valerie Gaborieau, Claire La-

- timer, Elizabeth Thomas, Irina Abnizova, Paolo E. Bucciarelli, Duncan Jones, Jennifer W. Teague, Bahman Abedi-Ardekani, Sandrine Serra, Jean-Yves Scoazec, Hamid Saffar, Farhad Azmoudeh-Ardalan, Masoud Sotoudeh, Abdol Nikmanesh, Hossein Poustchi, Amir Nivavarani, Shayesteh Gharavi, Mark Eden, Paul Richman, Luis S. Campos, R. C. Fitzgerald, Larissa F. Ribeiro, Sandra C. Soares-Lima, Christopher Dзамalala, Benjamin T. Mmbaga, Tatsuhiro Shibata, Derek Menya, Amanda M. Goldstein, Nan Hu, Reza Malekzadeh, Azita Fazel, Valerie McCormack, Judith McKay, Sandra Perdomo, Gilles Scelo, Emma Chanudet, Laura Humphreys, Ludmil B. Alexandrov, Paul Brennan, and Michael R. Stratton. Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. *Nature Genetics*, 53(11):1553–1563, 2021. doi: 10.1038/s41588-021-00928-6. Epub 2021 Oct 18.
- F. Muinos, F. Martinez-Jimenez, O. Pich, A. Gonzalez-Perez, and N. Lopez-Bigas. In silico saturation mutagenesis of cancer genes. *Nature*, 596(7872):428–432, 2021. doi: 10.1038/s41586-021-03771-1.
- J. Nielsen, J. T. Walsh, D. R. Degen, S. M. Drabek, and J. R. McGill. Evidence of gene amplification in the form of double minute chromosomes is frequently observed in lung cancer. *Cancer Genetics*, 65(2):120–124, 1993. doi: 10.1016/0165-4608(93)90219-C.
- S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge, P. Van Loo, Y. S. Ju, M. Smid, A. B. Brinkman, S. Morganella, M. R. Aure, O. C. Lingjaerde, A. Langerod, M. Ringner, S. M. Ahn, S. Boyault, J. E. Brock, A. Broeks, A. Butler, C. Desmedt, L. Dirix, S. Dronov, A. Fatima, J. A. Foekens, M. Gerstung, G. K. Hooijer, S. J. Jang, D. R. Jones, H. Y. Kim, T. A.

King, S. Krishnamurthy, H. J. Lee, J. Y. Lee, Y. Li, S. McLaren, A. Menzies, V. Mustonen, S. O’Meara, I. Pauporte, X. Pivot, C. A. Purdie, K. Raine, K. Ramakrishnan, F. G. Rodriguez-Gonzalez, G. Romieu, A. M. Sieuwerts, P. T. Simpson, R. Shepherd, L. Stebbings, O. A. Stefansson, J. Teague, S. Tommasi, I. Treilleux, G. G. Van den Eynden, P. Vermeulen, A. Vincent-Salomon, L. Yates, C. Caldas, L. van’t Veer, A. Tutt, S. Knappskog, B. K. Tan, J. Jonkers, A. Borg, N. T. Ueno, C. Sotiriou, A. Viari, P. A. Futreal, P. J. Campbell, P. N. Span, S. Van Laere, S. R. Lakhani, J. E. Eyfjord, A. M. Thompson, E. Birney, H. G. Stunnenberg, M. J. van de Vijver, J. W. Martens, A. L. Borresen-Dale, A. L. Richardson, G. Kong, G. Thomas, and M. R. Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016. doi: 10.1038/nature17676. URL <https://www.ncbi.nlm.nih.gov/pubmed/27135926>.

Imran Noorani, Magnus Haughey, Jens Luebeck, Andrew Rowan, Eva Grönroos, Francesco Terenzi, Ivy Tsz-Lo Wong, and ... Extrachromosomal dna driven oncogene spatial heterogeneity and evolution in glioblastoma. *bioRxiv*, 2024.10.22.619657, 2024. doi: 10.1101/2024.10.22.619657. URL <https://www.biorxiv.org/content/10.1101/2024.10.22.619657v1.full.pdf>. Preprint – not peer reviewed.

M. Petljak, L. B. Alexandrov, J. S. Brammeld, S. Price, D. C. Wedge, S. Grossmann, K. J. Dawson, Y. S. Ju, F. Iorio, J. M. C. Tubio, C. C. Koh, I. Georgakopoulos-Soares, B. Rodriguez-Martin, B. Otlu, S. O’Meara, A. P. Butler, A. Menzies, S. G. Bhosle, K. Raine, D. R. Jones, J. W. Teague, K. Beal, C. Latimer, L. O’Neill, J. Zamora, E. Anderson, N. Patel, M. Maddison, B. L. Ng, J. Graham, M. J. Garnett, U. McDermott, S. Nik-Zainal, P. J. Campbell, and M. R. Stratton. Characterizing mutational signatures in human cancer

- cell lines reveals episodic apobec mutagenesis. *Cell*, 176(6):1282–1294 e20, 2019. doi: 10.1016/j.cell.2019.02.012. URL <https://www.ncbi.nlm.nih.gov/pubmed/30849372>.
- Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.
- Karin Purshouse, Elias T. Friman, Shelagh Boyle, Pooran Singh Dewari, Vivien Grant, Alhafidz Hamdan, Gillian M. Morrison, Paul M. Brennan, Sjoerd V. Beentjes, Steven M. Pollard, and Wendy A. Bickmore. Oncogene expression from extrachromosomal dna is driven by copy number amplification and does not require spatial clustering in glioblastoma stem cells. *eLife*, 11:e80207, 2022. doi: 10.7554/eLife.80207.
- Keiran M Raine, Peter Van Loo, David C Wedge, David Jones, Andrew Menzies, Adam P Butler, Jon W Teague, Patrick Tarpey, Serena Nik-Zainal, and Peter J Campbell. ascatngs: Identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Current protocols in bioinformatics*, 56(1):15.9. 1–15.9. 17, 2016.
- T. Rausch, T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18): i333–i339, 2012. doi: 10.1093/bioinformatics/bts378. URL <https://www.ncbi.nlm.nih.gov/pubmed/22962449>.
- Kyle Retterer, Julie Scuffins, Daniel Schmidt, Rachel Lewis, Daniel Pineda-Alvarez, Amanda Stafford, Lindsay Schmidt, Stephanie Warren, Federica Gibellini, Anastasia Kondakova, Amanda Blair, Sherri Bale, Ludmila Matyakhina, Jeanne Meck, Swaroop Aradhya, and Eden Haverfield. Assessing copy number from exome sequencing and exome array cgh

- based on cnv spectrum in a large clinical cohort. *Genetics in Medicine*, 17(8):623–629, 2015. doi: 10.1038/gim.2014.160. URL <https://doi.org/10.1038/gim.2014.160>.
- E. Reznik, M. L. Miller, Y. Senbabaoglu, N. Riaz, J. Sarungbam, S. K. Tickoo, H. A. Al-Ahmadie, W. Lee, V. E. Seshan, A. A. Hakimi, and C. Sander. Mitochondrial dna copy number variation across human cancers. *Elife*, 5:e10769, 2016. doi: 10.7554/eLife.10769.
- L. Riva, A. R. Pandiri, Y. R. Li, A. Droop, J. Hewinson, M. A. Quail, V. Iyer, R. Shepherd, R. A. Herbert, P. J. Campbell, R. C. Sills, L. B. Alexandrov, A. Balmain, and D. J. Adams. The mutational signature profile of known and suspected human carcinogens in mice. *Nat Genet*, 52(11):1189–1197, 2020. doi: 10.1038/s41588-020-0692-4. URL <https://www.ncbi.nlm.nih.gov/pubmed/32989322>.
- Sergey Senkin, Sarah Moody, Marcos Díaz-Gay, Behnoush Abedi-Ardekani, Thomas Cattiaux, Alberto Ferreiro-Iglesias, Jingwei Wang, Simon Fitzgerald, Mariya Kazachkova, Raviteja Vangara, Anh P. Le, Erik N. Bergstrom, Azhar Khandekar, Burçak Otlı, Simran Cheema, Claire Latimer, Elizabeth Thomas, James R. Atkins, Kate Smith-Byrne, Ricardo Cortez Cardoso Penha, Carla Carreira, Priscilia Chopard, Valérie Gaborieau, Pekka Keski-Rahkonen, Duncan Jones, Jennifer W. Teague, Stéphane Ferlicot, Farnaz Asgari, Surasak Sangkhathat, Wipawee Attawettayanon, Beata Świątkowska, Jolanta Jarmalaite, Rasa Sabaliauskaite, Tatsuhiro Shibata, Akimasa Fukagawa, Daniel Mates, Voichița Jinga, Radu Rascu, Milena Mijuskovic, Suzana Savic, Svetlana Milosavljevic, Jennifer M. S. Bartlett, Mario Albert, Latsamy Phouthavongsy, Patricia Ashton-Prolla, Marcia R. Botton, Bruno Silva Neto, Sônia M. Bezerra, Marie P. Curado, Sergio C. Zequi, Rômulo M. Reis, Eduardo F. Faria, Neusa S. de Menezes, Rosane S. Ferrari, Robert E.

- Banks, Nelhatikshani S. Vasudev, Dmitry Zaridze, Alex Mukeriya, Olga Shangina, Victor Matveev, Lenka Foretova, Michaela Navratilova, Irena Holcatova, Anna Hornakova, Václav Janout, Mark P. Purdue, Nathan Rothman, Stephen J. Chanock, Per Magne Ueland, Mattias Johansson, Judith McKay, Gilles Scelo, Emma Chanudet, Laura Humphreys, Ana Carolina de Carvalho, Sandra Perdomo, Ludmil B. Alexandrov, Michael R. Stratton, and Paul Brennan. Geographic variation of mutagenic exposures in kidney cancer genomes. *Nature*, 629(8013):910–918, 2024. doi: 10.1038/s41586-024-07368-2. Epub 2024 May 1.
- A. Serin Harmanci, A. O. Harmanci, and X. Zhou. Casper identifies and visualizes cnv events by integrative analysis of single-cell or bulk rna-sequencing data. *Nat Commun*, 11(1):89, 2020. doi: 10.1038/s41467-019-13779-x. URL <https://www.ncbi.nlm.nih.gov/pubmed/31900397>.
- Charles Shale, Daniel L Cameron, Jonathan Baber, Marie Wong, Mark J Cowley, Anthony T Papenfuss, Edwin Cuppen, and Peter Priestley. Unscrambling cancer genomes via integrated analysis of structural variation and copy number. *Cell Genomics*, 2(4):100112, 2022.
- R. Shen and V. E. Seshan. Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput dna sequencing. *Nucleic Acids Res*, 44(16):e131, 2016. doi: 10.1093/nar/gkw520. URL <https://www.ncbi.nlm.nih.gov/pubmed/27270079>.
- O. Shoshani, S. F. Brunner, R. Yaeger, P. Ly, Y. Nechemia-Arbely, D. H. Kim, R. Fang, G. A. Castillon, M. Yu, J. S. Z. Li, et al. Chromothripsis drives the evolution of gene amplification in cancer. *Nature*, 591(7848):137–141, 2021. doi: 10.1038/s41586-020-03064-z.

- A. I. Spriggs, M. M. Boddington, and C. M. Clarke. Chromosomes of human cancer cells. *Br Med J*, 2(5317):1431–1435, 1962. doi: 10.1136/bmj.2.5317.1431.
- C. D. Steele, A. Abbasi, S. M. A. Islam, A. L. Bowes, A. Khandekar, K. Haase, S. Hames-Fathi, D. Ajayi, A. Verfaillie, P. Dhami, A. McLatchie, M. Lechner, N. Light, A. Shlien, D. Malkin, A. Feber, P. Proszek, T. Lesluyes, F. Mertens, A. M. Flanagan, M. Tarabichi, P. Van Loo, L. B. Alexandrov, and N. Pillay. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, 2022. doi: 10.1038/s41586-022-04738-6. URL <https://www.ncbi.nlm.nih.gov/pubmed/35705804>.
- E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian. Cnvkit: Genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS Comput Biol*, 12(4): e1004873, 2016. doi: 10.1371/journal.pcbi.1004873. URL <https://www.ncbi.nlm.nih.gov/pubmed/27100738>.
- J. Tang, N. E. Weiser, G. Wang, S. Chowdhry, E. J. Curtis, Y. Zhao, I. T. Wong, G. K. Marinov, R. Li, P. Hanoian, et al. Enhancing transcription-replication conflict targets ecdna-positive cancers. *Nature*, 635(8037):210–218, 2024. doi: 10.1038/s41586-024-07802-5.
- M. Tang, E. Bolderson, K. J. O’Byrne, and D. J. Richard. Tumor hypoxia drives genomic instability. *Front Cell Dev Biol*, 9:626229, 2021. doi: 10.3389/fcell.2021.626229.
- N. T. Umbreit, C. Z. Zhang, L. D. Lynch, L. J. Blaine, A. M. Cheng, R. Tourdot, L. Sun, H. F. Almubarak, K. Judge, T. J. Mitchell, et al. Mechanisms generating cancer genome complexity from a single cell division error. *Science*, 368(6488):eaba0712, 2020. doi: 10.1126/science.aba0712.

- J. A. Wala, P. Bandopadhyay, N. F. Greenwald, R. O'Rourke, T. Sharpe, C. Stewart, S. Schumacher, Y. Li, J. Weischenfeldt, X. Yao, C. Nusbaum, P. Campbell, G. Getz, M. Meyerson, C. Z. Zhang, M. Imielinski, and R. Beroukhim. Svaba: genome-wide detection of structural variants and indels by local assembly. *Genome Res*, 28(4):581–591, 2018. doi: 10.1101/gr.221028.117. URL <https://www.ncbi.nlm.nih.gov/pubmed/29535149>.
- Gerhard Winkler and Volkmar Liebscher. Smoothers for discontinuous signals. *Journal of Nonparametric Statistics*, 14(1-2):203–222, 2002.
- S. Wu, K. M. Turner, N. Nguyen, R. Raviram, M. Erb, J. Santini, J. Luebeck, U. Rajkumar, Y. Diao, B. Li, et al. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature*, 575(7784):699–703, 2019. doi: 10.1038/s41586-019-1763-5.
- X. Yan, P. Mischel, and H. Chang. Extrachromosomal dna in cancer. *Nat Rev Cancer*, 24(4): 261–273, 2024. doi: 10.1038/s41568-024-00669-8.
- T. Zhang, P. Joubert, N. Ansari-Pour, W. Zhao, P. H. Hoang, R. Lokanga, A. L. Moye, J. Rosenbaum, A. Gonzalez-Perez, F. Martinez-Jimenez, et al. Genomic and evolutionary classification of lung cancer in never smokers. *Nat Genet*, 53(9):1348–1359, 2021. doi: 10.1038/s41588-021-00920-0.
- Y. Zhu, A. D. Gujar, C. H. Wong, H. Tjong, C. Y. Ngan, L. Gong, Y. A. Chen, H. Kim, J. Liu, M. Li, et al. Oncogenic extrachromosomal dna functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell*, 39(5):694–707.e7, 2021. doi: 10.1016/j.ccell.2021.03.006.