ABSTRACT

Title of Dissertation: STUDYING TUMOR EVOLUTION WITH EMERGING SEQUENCING TECHNOLOGIES

Yuelin Liu
Doctor of Philosophy, 2025

Dissertation Directed by: Professor Mihai Pop
Department of Computer Science

Doctor S. Cenk Sahinalp
National Institutes of Health

The path along which heterogeneous subpopulations within a tumor arise can be modeled as an evolutionary process, and a better understanding of such process may improve cancer diagnostics, prognostics, and treatment outcome. How to construct more accurate tumor phylogenies and how to leverage the inferred phylogenetic relations among tumor subclones in understanding the disease are both key in the study of tumor evolution. The advances in novel sequencing technologies present opportunities to examine crucial aspects of the evolution of a tumor, beyond the types of information short-read genomic and transcriptomic sequencing technologies offer. As such, this work leverages three such technologies - single-cell bisulfite sequencing, long-read sequencing, and spatial transcriptomics - in answering key questions in tumor evolution.

In the first part of the thesis we present the first distance-based method to jointly construct a

tumor phylogeny and identify lineage-informative CpG sites from single-cell methylation sequencing data. The tool provides a fast and accurate way to examine tumor evolution through the methylome at a single-cell resolution and allows the incorporation of matched RNA sequencing information whenever applicable.

Next, we present a novel framework to study both the genomic and epigenomic changes during tumor evolution using Nanopore long-read sequencing data of single-cell derived tumor sublines. Single nucleotide variants, structural variants, copy number alterations, as well as CpG methylation were profiled. The clonal resolution this data set uniquely affords allowed the inference of active mutational processes, confident identification of parallel CNAs, and detection of differentially methylated regions, all at a subclonal level.

Finally, we present a framework to examine tumor subclones at spatial resolution. Melanoma tumors from mice in both control and anti-CTLA-4 treatment group were subjected to both Smart-seq2 single-cell full length transcriptomics sequencing and 10x Visium spatial transcriptomics sequencing. Six tumor subclones were identified using mutations profiled from the Smart-seq2 data, and their proportion were estimated from the spatial transcriptomics slides via cell type deconvolution. The associations between tumor subclones with various stromal and immune cell types were accessed with established spatial statistics. By contrasting subclone-immune activities between responding and non-responding tumors, we show that there exist subclonal differences in their response to the immune microenvironment.

STUDYING TUMOR EVOLUTION WITH EMERGING SEQUENCING TECHNOLOGIES

by

Yuelin Liu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:
Professor Mihai Pop, Chair/Advisor
Dr. S. Cenk Sahinalp, Co-Advisor
Professor Rob Patro
Professor David M. Mount
Professor Stephen M. Mount

© Copyright by
Yuelin Liu
2025

Acknowledgments

The PhD journey has been full of challenges in ways I could not even have previously understood, and there's a long list of people I would like to acknowledge who have helped me get to where I am today.

First, I want to acknowledge my collaborators and members of the Cancer Data Science Laboratory at the National Cancer Institute. In particular, I would like to thank members of the Sahinalp lab, with whom I have had countless productive discussions even when we are not working on the same project. Especially Chih Hao, John, and Jake, who have consistently driven me to and from the Bethesda Campus of the National Institutes of Health. I for sure owe them dinner, and probably a lot of gas money.

I would like to thank past and present students and faculty of the Center for Bioinformatics and Computational Biology at the University of Maryland, who made me immediately at home when I first set foot in IRB-3112. Specifically, I would like to thank Zinat and Noor, who were in the same year as me and have supported each other through losing our first advisors, and Hirak, who imparted with me so much wisdom in my first year when he left for a postdoc, and is doing that once again now as he settles in a faculty position at Vanderbilt.

I would like to thank the past and present members of CS GradCo at the University of Maryland, and all the friends I have met while organizing for it. It is amazing to be among such community-minded people. From GradCo Day, to fall picnic, and to the very brief time we tried to start a jazz band, I cannot believe what we were able to pull off.

I would like to thank my past and present housemates at 48th Avenue who have witnessed ourselves at our best as much as at our worst, all in between Costco runs and Biryani City orders.

I would like to thank the DC swing dance community for bringing me so much joy through our collective love for the art form.

I would like to thank the very dear friends in my life for inspiring me to be the best version of myself, Yang, Tori, Yujia, Daniel, Ann, and especially Jwi, Shramay, and Gibson for making sure I did not forget to eat in the weeks leading up to my defense.

There are three mentors who have played instrumental roles in this process. I would like to thank Dr. Mihai Pop, who has been tremendously supportive during Max's departure from UMD and over the years. He has been a kind mentor and have set an example for all of us students in CBCB. I would like to thank Dr. Lenore Cowen, my undergrad advisor, who changed my life in the literal sense by introducing me to computational biology and being the first to say "I think you can be good at this". And, of course, I would like to thank Dr. S. Cenk Sahinalp, who supervised all the work in this thesis, who has not only taught me a lot about science, but also a whole lot about being a scientist. His guidance, both spoken and unspoken, has been truly invaluable in this journey.

Last but not least, I would like to thank my parents, Fei Dai and Zhuangcheng Liu, who have sacrificed the unimaginable in making me who I am today. Without their unconditional love and support, none of this would ever have been possible.

Table of Contents

Acknowledgements ii

Table of Contents iv

List of Tables viii

List of Figures ix

List of Abbreviations xii

Chapter 1: Introduction 1
- 1.1 Cancer as a heterogeneous disease . . . 1
- 1.2 Modeling how tumor heterogeneity arises as an evolutionary process . . . 1
- 1.3 Advancements in sequencing technologies further understanding of tumor evolution . . . 3
- 1.4 Phylogeny inference using multi-sampled bulk data . . . 5
- 1.5 Phylogeny inference using single-cell data . . . 6
- 1.6 DNA methylation as phylogenetic marker . . . 7
- 1.7 Long-read sequencing technology . . . 9
- 1.8 Sequencing-based spatial transcriptomics . . . 10
- 1.9 Outline of Thesis . . . 11

Chapter 2: Reconstructing tumor phylogenies with single-cell CpG methylation data using `Sgootr` 14
- 2.1 Motivation . . . 14
- 2.2 Methods . . . 15
  - 2.2.1 Biclustering of cells and sites for reducing sparsity-induced noise . . . 17
  - 2.2.2 Likelihood-based sequencing error correction accounting for copy number estimates . . . 19
  - 2.2.3 Expected distance calculation between cell pairs for tree construction . . . 21
  - 2.2.4 Pruning of CpG sites according to a tree-based methylation status persistence measure . . . 24
  - 2.2.5 Inference of migration history from lesion-labeled tumor lineage tree leveraging prior knowledge on migration patterns . . . 28
- 2.3 Results . . . 31
  - 2.3.1 `Sgootr` obtains a simple tumor migration history with scBS-seq and scRNA-seq data from patient CRC01. . . . 32
  - 2.3.2 `Sgootr` infers migration histories simpler than alternative methods for the Bian CRC cohort. . . . 34
  - 2.3.3 `Sgootr`-identified lineage-informative CpG sites construct lineages suggesting simpler migration histories than those identified via FG. . . . 38
  - 2.3.4 `Sgootr` is orders-of-magnitude faster than IQ-TREE. . . . 38

2.3.5 `Sgootr`-identified lineage-informative CpG sites are enriched in inter-CGI regions. . . . 39
2.3.6 Genes associated with `Sgootr`-identified lineage-informative CpG sites are enriched for pan-cancer and CRC-related terms. . . . 39
2.4 Discussion . . . 40

Chapter 3: Studying melanoma evolution with Nanopore long-read sequencing data 42
3.1 Motivation . . . 42
3.2 Results . . . 43
3.2.1 Long-read sequencing of single-cell derived subclones of B2905 mouse melanoma model . . . 43
3.2.2 Constructed phylogeny reveals phenotypically-distinct major clades . . . 46
3.2.3 Phylogeny-guided harmonization with `TreeHarmonizer` provides evolutionary timing of genomic variants . . . 48
3.2.4 Evolutionary dynamics of somatic SNVs and mutational processes . . . 49
3.2.5 Somatic SVs dynamics suggest early drivers and late genomic instability . . . 51
3.2.6 Recurrent CNAs in independent lineages suggest parallel adaptation process . . . 53
3.2.7 Methylation changes contribute to differentiation of an aggressive clonal lineage 56
3.2.8 Systematic analysis of multiple variation types provides insights into B2905 lineages differentiation . . . 59
3.2.9 Curated set of subclonal variant calls enables benchmarking of long-read tools . 62
3.3 Discussion . . . 64

Chapter 4: Studying melanoma subclones at spatial resolution 66
4.1 Spatial transcriptomics profiling of the control and anti-CTLA-4 treated tumors. . . . 66
4.2 Identifying tumor subclones from Smart-seq2 full-length scRNA-seq data of the control and anti-CTLA-4 treated tumors. . . . 66
4.3 Estimating tumor subclone proportions on the ST slides via deconvolution using SpaCET 68
4.4 Validation of SpaCET-deconvolved tumor subclone proportions . . . 69
4.5 Analysis of SpaCET-deconvolved tumor subclone proportions . . . 73
4.5.1 SpaCET-deconvolved proportions reveal differences in subclone growth dynamics between treatment responders and non-responders . . . 73

Chapter 5: Concluding remarks 76

Appendix A: Supplemental Material for Chapter 2 80
A.1 Biclustering: benchmarking on simulated data and a case study on CRC01 . . . 80
A.1.1 Running time and memory benchmark on simulated data . . . 81
A.1.2 Biclustering case study on CRC01 from the Bian cohort . . . 85
A.2 Benchmarking `Sgootr` on simulated data . . . 89
A.2.1 Expected distance calculation on simulated data . . . 89
A.2.2 Iterative pruning on simulated data . . . 94
A.3 Application of `Sgootr` on single cells from U87MG *ex vivo* tree from the RETrace study . . . 99

A.4 Application of Sgootr on the complete set of 9 metastatic CRC patients made available by Bian and colleagues . . . 102
A.5 Application of Sgootr on the cohort, replacing NJ with FastME 2.0 . . . 111
A.6 Patient CRC01 tumor lineage tree constructed using mutation . . . 115
A.7 Application of Sgootr to multi-regional GBM patient MGH105 MscRRBS data by Chaligne and collegues. . . . 116
A.8 How to choose p, P(1^c), P(0^c), and P(mixed) for Sgootr . . . 118
A.9 How to choose κ for Sgootr . . . 119
A.10 Experiment details . . . 120
A.10.1 Sgootr . . . 120
A.10.2 FastME 2.0 . . . 120
A.10.3 IQ-TREE . . . 121
A.10.4 FG . . . 121

Appendix B: Supplemental Material for Chapter 3 124
B.1 Methods . . . 124
B.1.1 Constructing tumor phylogeny using long read sequencing data from 23 sublines 124
B.1.2 Determining the phenotypes of the 4 major clades . . . 125
B.1.3 Assessing the effectiveness of TreeHarmonizer in determining the placement of variants impacted by filtering during tree construction . . . 128
B.1.4 Mutational signature analysis along the tumor phylogeny . . . 129
B.1.5 Computing dN/dS values along the tumor phylogeny . . . 130
B.1.6 Pathway enrichment analysis . . . 131
B.1.7 Determining genes impacted by SVs . . . 131
B.1.8 Determining the set of cancer-relevant mouse genes-of-interest according to the COSMIC Gene Census . . . 133
B.1.9 Timing analysis of chr4 copy-neutral LOH . . . 134
B.1.10 Subclone-specific differential methylation analysis . . . 137
B.1.11 Permutation test on the number of DMRs obtained . . . 137
B.1.12 Correlation between the number of DMRs and cumulative SNVs at a branch in the phylogeny . . . 138
B.1.13 Differential gene expression analysis of subline scRNA-seq data . . . 138
B.1.14 Discovering key regions where changes in methylation is monotonic along the evolutionary trajectory . . . 139
B.1.15 Permutation test on the number of genes associated with monotonic methylation changes along the evolutionary trajectory . . . 140
B.1.16 Calling copy number from bWES and Smart-Seq2 full-length scRNA-seq data on the 23 sublines . . . 140
B.2 Technology comparison: long-read, bWES, and scRNA-seq . . . 141
B.3 Additional Supplementary Figures and Tables . . . 145

Appendix C: Supplemental Material for Chapter 4 161
C.1 SpaCET deconvolution results on 17 ST slides . . . 161
C.1.1 Control STs . . . 161
C.1.2 Treated STs (Non-responders) . . . 165

C.1.3 Treated STs (Responders) 168
C.2 Validation of SpaCET-deconvolved tumor subclone proportions on 17 ST slides 171
C.2.1 Control STs 171
C.2.2 Treated STs (Non-responders) 174
C.2.3 Treated STs (Responders) 177

Bibliography 180

List of Tables

A.1 Feature sizes of the biclustering formulation model as a function of the number of cells, v, and the number of CpG sites, u in the input. 81
A.2 Densities of the original binary matrix and the subsampled matrices. 87
A.3 Data summary of the Bian metastatic CRC cohort. 102
A.4 Overlap coefficients of GO terms associated with different genomic locations (CpG island and inter-CGI) discovered by GOrilla and DAVID. 110
A.5 Spearman correlation of CpG site distance from closest CpG island and homogeneity. . 111

B.1 Phenotypes of the 4 major clades. 126
B.2 DMRs associated with the red clade that overlap putative promoter or putative proximal enhancers of known genes. 157

List of Figures

1.1 Illustration of the Nowell's theory of tumor evolution. . . . 3
1.2 Illustration of tumor phylogeny reconstruction from multi-sampled bulk data. . . . 5
1.3 Illustration of tumor phylogeny reconstruction from single-cell data. . . . 7

2.1 Overview of `Sgootr` . . . 16
2.2 Application of `Sgootr` to patient CRC01 scTrio-seq data by Bian and colleagues. . . . 33
2.3 Overview of results from the Bian metastatic CRC cohort. . . . 35

3.1 An overview of the long-read approach and analysis of 23 subclones of a mouse melanoma tumor. . . . 45
3.2 Phylogenetic tree and mutational landscape of 23 single-cell derived sublines. . . . 47
3.3 Patterns and dynamics of somatic SVs. . . . 52
3.4 Genomic regions recurrently affected by CNAs suggest parallel evolution. . . . 54
3.5 DMRs detected at subclonal resolution. . . . 57
3.6 A model of B2905 melanoma evolution informed by analyses of long-read sequencing data on the 23 sublines. . . . 60
3.7 Subline-derived pseudobulk samples enables benchmarking the performances of long-read variant callers at various clonality. . . . 63

4.1 Growth dynamics of the 17 tumors from which the ST slides were derived. . . . 67
4.2 Mutations profiled from Smart-seq2 full-length scRNA-seq data of control and treated tumors reveal 6 major tumor subclones. . . . 68
4.3 Validation of tumor subclone proportions estimated by SpaCET using subclone-specific mutations detected on ST slides. . . . 70
4.4 Deconvolved subclone proportions inform differences in tumor growth dynamics. . . . 74

A.1 Patient CRC01 cell and site coverate level distribution . . . 80
A.2 Running time and memory behavior on 200 simulated square binary matrices. . . . 84
A.3 Mean and standard deviation of wall clock time-to-convergence in hours, optimality gap at 20 hours, and resulting submatrix density across the 10 subsamples, over a range of (α, β) value pairs. . . . 88
A.4 Probability mass functions of A. $\mathbf{ZTPois}(\lambda)$, where $\lambda = \{1, 2, 3, 4, 5, 6\}$ and B. the empirical per-site read coverage distribution from patient CRC01. . . . 91
A.5 Baseline distance measure and expected distance measure on $13,500$ sets of read count vectors, simulated from cell pairs with known normalized ground truth distance $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$, with level of read coverage depth modeled by ZTPois(λ), where $\lambda = \{1, 2, 3, 4, 5, 6\}$. . . . 92
A.6 Baseline distance measure and expected distance measure on $2,250$ sets of read count vectors, simulated from cell pairs with known normalized ground truth distance $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$, with level of read coverage depth modeled by the empirical read coverage distribution gathered from the CRC01 sample. . . . 94
A.7 Topologies of the simulated clone trees. . . . 95
A.8 Results on 10 simulated clone trees from `Sgootr` and baseline method. . . . 98

A.9 Application of Sgootr to scRRBS data from U87MG *ex vivo* tree by the RETrace study. . . . 101
A.10 Sgootr-constructed ($\kappa = .1$) tumor lineage trees (t^*s) and Sgootr-inferred tumor migration histories for the 9 patients in the Bian cohort . . . 104
A.11 RF distances across iterations for each patient in the metastatic CRC cohort applying Sgootr with different κ values. . . . 105
A.12 GOrilla -identified GO terms with significant (<.05) q-values in enrichment analysis of non-pseudo, protein coding genes whose gene bodies and 1000bp upstream promoter regions contain the Sgootr-identified lineage-informative CpG sites, stratified by whether they locate in CpG island or inter-CGI regions. . . . 107
A.13 DAVID-identified GO terms with significant (<.05) q-values in enrichment analysis of non-pseudo, protein coding genes whose gene bodies and 1000bp upstream promoter regions contain the Sgootr-identified lineage-informative CpG sites, stratified by whether they locate in CpG island or inter-CGI regions. . . . 109
A.14 Median number of bp to the closest CGI from CpG sites used to construct the methylation lineage tree at each iteration for all CRC patients and GBM patient MGH105. . . 110
A.15 Sgootr-constructed ($\kappa = .1$) tumor lineage trees (t^*s) and Sgootr-inferred tumor migration histories for the 9 patients in the Bian cohort, replacing NJ with FastME 2.0. 113
A.16 Number of migration events and BAMDPO of migration histories obtained via applying Sgootr with FastME 2.0 in comparison to NJ. . . . 113
A.17 RF distances across iterations for each patient in the Bian metastatic CRC cohort applying Sgootr - substituting NJ with FastME 2.0 - with different κ values. . . . 114
A.18 Tumor lineage tree based on SNVs/INDELs obtained from scRNA-seq data from patient CRC01. . . . 116
A.19 Application of Sgootr to patient MGH105 MscRRBS data by Chaligne and colleagues. 117
A.20 FG frequencies (averaged across 5 runs) for the Bian cohort. . . . 122
A.21 Pearson product-moment correlation coefficients between FG frequencies across sites between downsampled runs for the Bian cohort. . . . 123

B.1 DGI of sublines in C57BL/6 and nude mice. . . . 127
B.2 Mutational signature activities along the tumor phylogeny. . . . 130
B.3 SNVs along the tumor phylogeny. . . . 132
B.4 SNV evidence for the timing clonal CNA events. . . . 135
B.5 Comparison among long-read, bWES, and scRNA-seq data of the 23 sublines. . . . 142
B.6 SNVs and SVs placed by TreeHarmonizer. . . . 145
B.7 SNV density and SV events present in individual sublines in the green clade. . . . 146
B.8 SNV density and SV events present in individual sublines in the blue clade. . . . 147
B.9 SNV density and SV events present in individual sublines in the red clade. . . . 148
B.10 SV-disrupted genes and gene fusion annotated by Padfoot. . . . 149
B.11 Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (1/7). . . . 150
B.12 Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (3/7). . . . 152

B.13 Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (4/7). 153
B.14 Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (5/7). 154
B.15 Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (6/7). 155
B.16 Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (7/7). 156
B.17 Permutation tests on the number of DMRs obtained for all subtree sizes of the phylogeny in Figure 3.2a. 158
B.18 Associating truncating *Dnah6* mutation with (haplotype-specific) methylation of *Dnah6* promoter in major clades. 159
B.19 CNA profiles of 23 sublines obtained using long-read, bWES, and scRNA-seq data. . . 160

List of Abbreviations

AID	Activation-induced cytidine deaminase
ANOVA	Analysis of variance
ASM	Allele-specific methylation
BAMDPO	Binary adjacency matrix distance from partial order
bp	Base pair(s)
bWES	Bulk whole-exome sequencing
cDNA	complementary DNA
CGC	Cancer Gene Census
CGI	CpG island
chr	Chromosome
CLL	Chronic lymphocitic leukemia
COSMIC	Catalogue Of Somatic Mutations In Cancer
CpG	Cytosine-guanine
CNA	Copy number aberration
CRC	Colorectal cancer
dbSNP	The Single Nucleotide Polymorphism Database
DGI	Delay of growth initiation
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
FG	Four gametes
GATK	Genomic Analysis Toolkit
GBM	Glioblastoma
GENCODE	Encyclopedia of genes and gene variants
GO	Gene Ontology
HSD	Honestly Significant Difference
ICB	Immune checkpoint blockade
INDEL	Insertion-deletion
kb	kilo-base pairs
LN	Lymph node metastasis
LOH	Loss of heterozygosity

ML	Liver metastasis
MP	Post-treatment liver metastasis
mRNA	Messenger RNA
MRCA	Most-recent common ancestor
MscRRBS	Multiplexed single-cell reduced representation bisulfite sequencing
NC	Normal colon tissue
ND	Not designated
NJ	Neighbor joining
ONC	(Proto-)Oncogene
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PT	Primary tumor
RF	Robinson-Foulds
RNA	Ribonucleic acid
ROS	Reactive oxygen species
SBS	Single base substitution
scBS-seq	Single-cell bisulfite sequencing
scRNA-seq	Single-cell RNA sequencing
scRRBS	Single-cell reduced representation bisulfite sequencing
scTrio-seq	Single-cell triple omics sequencing
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
ST	Spatial transcriptomics
SV	Structural variant
TCGA	The Cancer Genome Atlas Program
TSG	Tumor Suppressor Gene
T2T	Telomere-to-Telomere
UV	Ultraviolet (radiation)
WGS	Whole-genome sequencing
5mC	5-methylcytosine

Chapter 1: Introduction

1.1 Cancer as a heterogeneous disease

Cancer is a group of diseases characterized by the abnormal growth of cells with the potential to spread to other parts of the body. It is highly heterogeneous. Different tumor types are characterized by the primary tissue or cell type of origin. Treatment decisions for different tumor types are frequently specific to the primary site, since tumors in different tissues likely are caused by different endogenous or exogenous processes. Within each tumor type, there could be further molecular subclassifications based on histological or genomic profiling. Distinguishing among different tumor subtypes is important, as they may differ in invasiveness, metastatic potential, proneness to relapse, prognosis, and response to specific treatment. These types of differences we call inter-tumor heterogeneity.

Looking within a single tumor, there is extensive heterogeneity as well. There may be cellular subpopulations that differ by genetic or non-genetic characteristics. We call each group of these cells subclones. From a more clinically relevant perspective, how well a cancer patient responds to a type of treatment may be explained by the tumor's subclone composition. For example, treatment resistance and relapse may be caused by the presence of a stubborn subclone unaffected by treatment. If we can better characterize such intra-tumor heterogeneity, we may achieve more accurate cancer diagnostics, prognostics, and more effective and targeted treatment.

The latter, intra-tumor heterogeneity, will be the focus of this thesis.

1.2 Modeling how tumor heterogeneity arises as an evolutionary process

The model with which we study how such intra-tumor heterogeneity arises has its root in species evolution. In his revolutionary work in 1859, The Origin of Species, Charles Darwin provided a

framework which enabled the understanding of somatic selection, diversification and extinction [1]. His famous 1837 drawing of an evolutionary tree diagram was the first to describe the phylogenetic relations among multiple species.

Incidentally, exactly a hundred years after Darwin's publication of The Origin of Species, in 1959, a predoctoral fellow by the name of David Hungerford was working in a genetics lab in Philadelphia, when he observed under the microscope that the white blood cells from patients with chronic myelogenous leukemia had an abnormally short chromosome 22. Together with a pathologist at Univeristy of Pennsylvania named Peter Nowell, they published this observation in 1960 [2], and named the abnormality the Philadelphia chromosome after where the discovery was made.

It may be an indisputable fact at the time of the writing of this thesis, but back then, cancer was not believed to be a genetic disease, and no consistent theory of how cancer arises was recognized. The discovery of Philadelphia chromosome provided the first piece of evidence that, alterations in one's DNA may lead to cancer.

More than a decade later, using a novel staining technology, Dr. Janet Rowley at the University of Chicago observed that the part missing from the end of chromosome 22 seems to be relocated to the end of one copy of chormosome 9, causing a translocation event that resulted in a gene fusion [3]. This discovery eventually led to the development of the targeted therapy Imatinib in the early two thousands, which is a landmark drug shown to be very effective in maintaining remission in CML patients.

With the increasingly well established idea that cancer is caused by aberrations in one's DNA, Nowell, who co-discovered the Philadelphia chromosome, went on to formalize the foundation of the theory of tumor evolution in 1976 [4]. Much like species evolution where heritable traits were passed down from ancestor to offsprings, somatic alterations that are heritable from parent to daughter cells

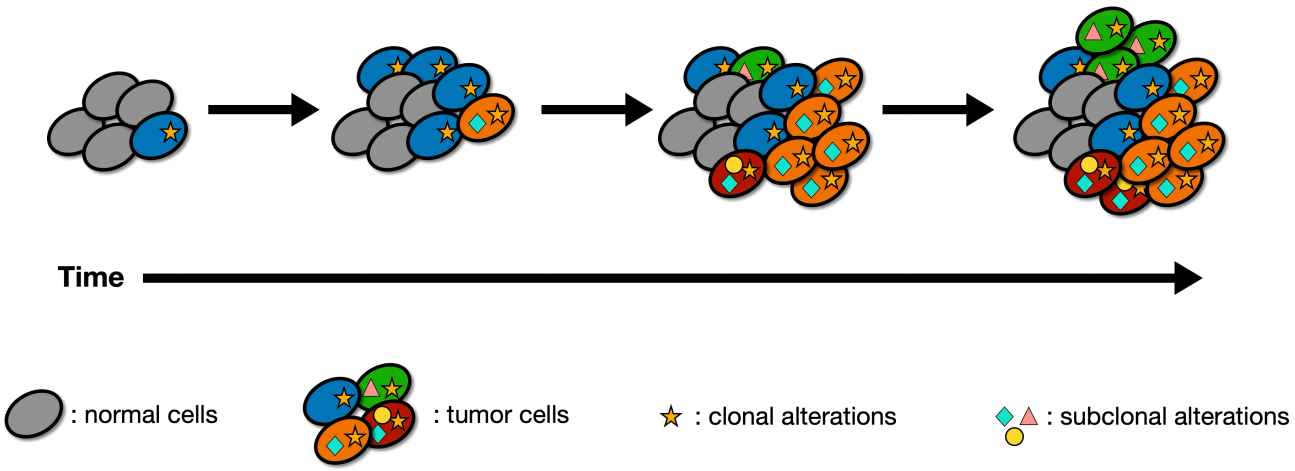


Figure 1.1: **Illustration of the Nowell's theory of tumor evolution.** Clonal alterations occur in the cancer founder cell, which initiates the abnormal cell growth. As the clonal population expands, additional subclonal alterations also occur and are inherited by daughter cells during cell replication. The accumulation of such somatic alterations over time results in a heterogeneous tumor.

during cell replication accumulate over the course of tumor development, resulting in a heterogeneous cell population (Figure 1.1). Under the "infinite sites assumption" [5], that states that, as the length of the genome is large, the probability of the same nucleotide change at a particular locus in distinct samples happening independently is negligibly small, the observation of shared somatic mutations in distinct samples at different frequencies serve as a piece of evidence for the relative timing of when the mutation occurred along the course of the development of the tumor.

1.3 Advancements in sequencing technologies further understanding of tumor evolution

While traditionally, tumor evolution has been conceptualized as the gradual accumulation of point mutations under selective pressures [4], more recently, distinct mutational processes that lead to the rapid accumulation of structural variants during catastrophic events — such as chromothripsis and

chromoplexy — have been increasingly recognized as critical drivers of tumor initiation and adaptation [6, 7]. Furthermore, tumor evolution can also be shaped by non-genetic, or even non-heritable, determinants—especially epigenetic alterations such as DNA methylation or histone acetylation [8]. The advancement in our understanding of tumor evolution largely owes to the advances in sequencing technologies, which have allowed us to study somatic alterations at the base pair level. Novel sequencing technologies afford us with more, and new types of information about tumor samples, all the while with decreasing sequencing cost.

Specifically, the three main pieces of work discussed in this thesis will respectively leverage tumor data generated with three novel sequencing technologies introduced within the past decade. First, single-cell bisulfite sequencing [9], introduced in 2014, generates CpG methylation information within individual cells in a bisulfite-treated tumor sample. This technology is further augmented by the scTrio-seq protocol in 2016 [10], that additionally profiled RNA-seq data from the same single cells. Next, Oxford Nanopore Technology long read sequencing, which was made commercially available in 2015, profiles ultra long reads at the kilo basepair to mega basepair scale while allowing simultaneous detection of genetic and epigenetic base modifications. Lastly, spatial transcriptomics, first reported in 2016 [11], provides gene expression read outs from spot with known spatial coordinates on a tumor slide.

The discovery of the Philadelphia chromosome, discussed in the previous section, not only was an important event in the development of the theory of tumor evolution, the story also highlights the importance of constantly adopting and leveraging novel technologies, such as the staining method used to discover the translocation event. This thesis will discuss how we leverage new opportunities and address the challenges brought by the three novel sequencing technologies in our study of tumor evolution.

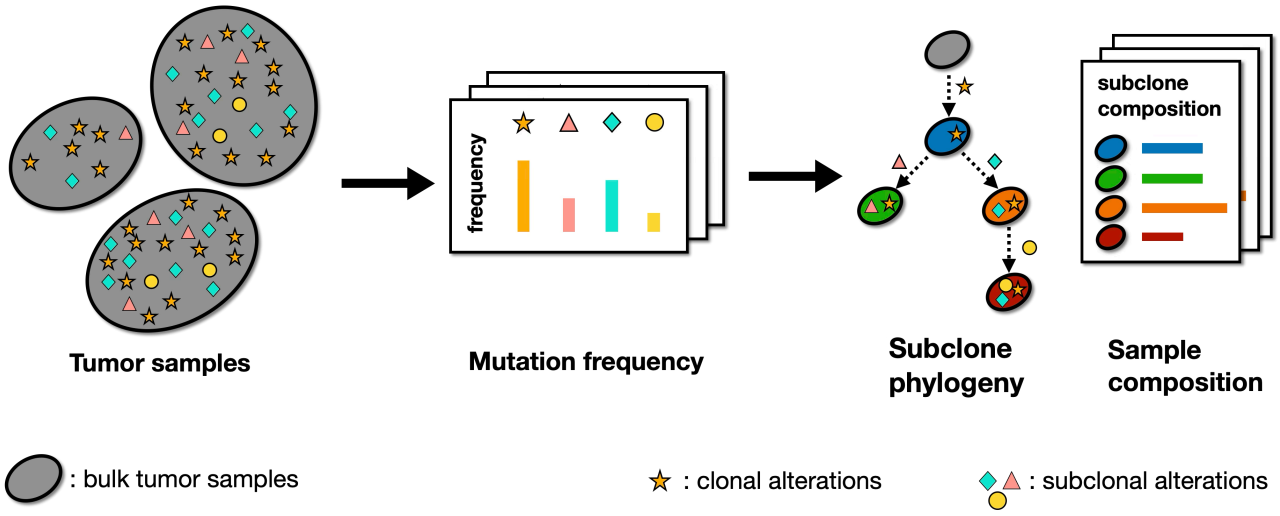


Figure 1.2: **Illustration of tumor phylogeny reconstruction from multi-sampled bulk data.**

1.4 Phylogeny inference using multi-sampled bulk data

One way to study the evolutionary process that is tumor evolution is through phylogeny reconstruction. The problem of algorithmically inferring a phylogenetic tree that may be interpreted as a tumor progression history from multiple bulk samples per patient has been widely studied in the past decade. In such experiments, the tumor samples obtained from the patient are subjected to bulk sequencing, and in each sample, for each unique point mutation, the number of sequencing reads with the reference allele, and that with the variant allele are obtained. Under some reasonable assumptions that the mutations are heterozygous from a diploid genome, the observed frequency of each mutation per sample can be computed. Then, the subclone phylogeny could be obtained by either - first explicitly clustering the mutations by frequencies then determining their order along the phylogeny - or - incorporating some error model to find the most likely tree that resulted in the observed variant allele frequencies (Figure 1.2).

Methods developed to address this problem include: TuMult [12], PhyloSub [13], PyClone [14],

SciClone [15], MEDICC [16,17], PhyloWGS [18], AncesTree [19], CITUP [20], Canopy [21], Treeomics [22] and others. Tumor phylogenies and the implied intra-tumor heterogeneity are occasionally used to make clinical decisions [23]. For understanding basic cancer biology, a key question is whether tumor progression is more branched (leading to a wider and shallower tree) or more linear (leading to a narrower and deeper tree) [24]. For metastases, results on the question of branched vs. linear evolution are mixed (e.g., [25–27]). Hong and colleagues showed that results are sensitive to the method used for phylogeny inference [28]. Complicating matters further is that most of the methods listed above use bulk sequencing data as input, and El-Kebir and colleagues showed that under reasonable and models, the optimal phylogeny from bulk sequencing data may not be unique [29]. Therefore, recent efforts have looked beyond bulk sequencing data to elucidate the progression history of a tumor. In a major step towards this goal, Quinn and colleagues introduced technologies for cell lineage tracing and applied those technologies to study single-cell data in xenograft tumor models. To our knowledge, lineage tracing is not yet possible in human cancer samples, but analysis of single-cell data of other types is possible [30].

1.5 Phylogeny inference using single-cell data

The recent rise of single-cell sequencing technology empowers more accurate tumor lineage inference by allowing the examination of intratumor heterogeneity at a cellular resolution [31–36] (Figure 1.3). However, since single-cell sequencing data are derived from a limited amount of genetic material, the signals obtained are more scarce than those from bulk sequencing [37]. Genomic events commonly used in bulk tumor lineage inference, such as SNVs and CNAs, are observed with less frequency or measured with less confidence in single cell datasets. Single-cell tumor lineage inferences

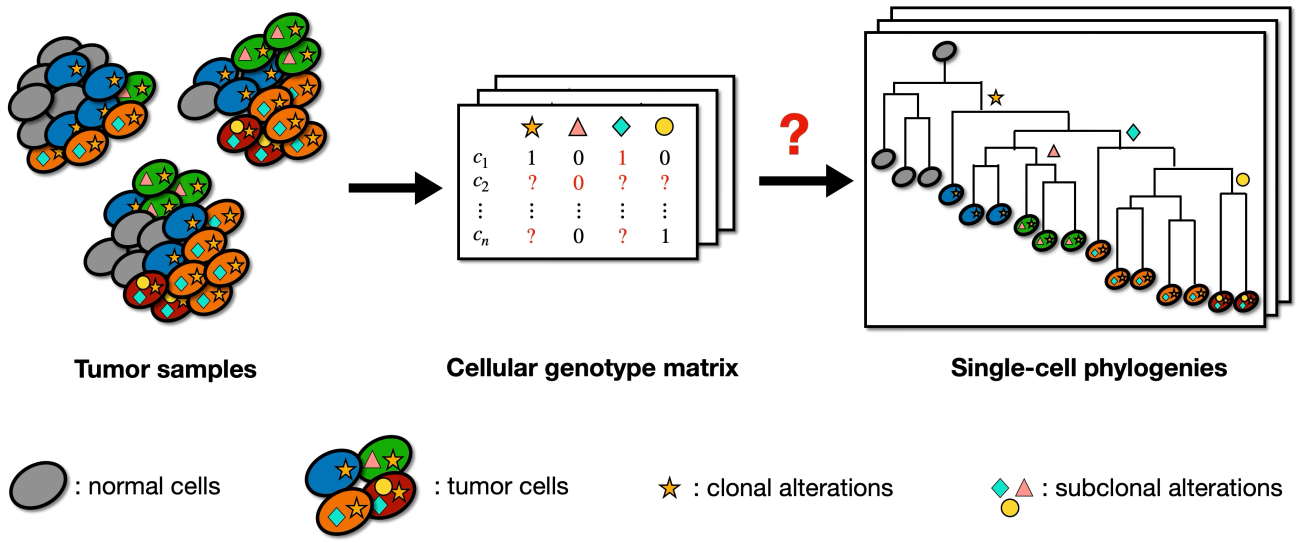


Figure 1.3: **Illustration of tumor phylogeny reconstruction from single-cell data.**

using SNVs often suffer from false negatives in the datasets, which calls for data imputation or aggregation; on the other hand, tumor lineage inference using CNAs called from single-cell sequencing data often involves ambiguities in the determination of chromosomal breakpoints that may mislead the conclusion.

1.6 DNA methylation as phylogenetic marker

More recently, there have been growing interests in examining heritable and non-genetic determinants in tumor evolution [38], as while genomic alterations such as SNVs and CNAs have been conventionally used to infer tumor phylogenies, other non-genomic alterations can also influence clonal phenotype and fitness. DNA methylation, the addition of a methyl group to cytosine, which results in the formation of 5mC especially in the context of CpG sites, is an epigenetic marker that has been extensively studied for its role in regulating gene expression and maintaining cellular memory [39]. As Gaiti and colleagues importantly noted, the methylation maintenance machinery has been shown to have an error rate four orders of magnitude higher than that for DNA replication [40, 41] and there-

fore may provide a greater amount of observable evidence for tumor evolution than alternative data modalities, which is a critical asset especially in the single-cell context [42]. Recent studies have leveraged CpG methylation in single cells sampled from CRC [43], CLL [42], GBM and IDH-mutant glioma [44] tumors to construct lineages, showing CpG methylation to be a valuable signal for lineage inference.

That said, using CpG methylation to construct tumor lineage tree has its challenges. First, single cell methylation data has shallow read coverage. In each cell, most of the covered sites have only a couple of read support. The shallow read coverage also means it is very likely that the reads do not cover all alleles present. Furthermore, methylation is known to be a generally more transient signal than mutations, and in a recent work by Meir and colleagues [45], they show that, while the methylation statuses at some CpG sites are persistently inherited during cell replication, others seem to be independently resampled at each generation. For the goal of constructing tumor phylogenies using methylation as features, it is intuitive that only sites of the persistency model could harbor changes in methylation status that would be lineage-informative, while those of the mixture model add noise at best.

As such, in Chapter 2 of this thesis, we aim to answer the following questions. Can we leverage single-cell methylation data to infer tumor evolutionary trajectories? Can we in the meantime identify CpG sites that are lineage-informative? And finally, can we uncover biological insights from these results?

1.7 Long-read sequencing technology

Long read sequencing technologies, such as those from PacBio HiFi and ONT, have pushed the frontier of genomic analysis: with read lengths ranging from tens to hundreds of kilobases, with ultra-long reads reaching a maximum read length > 4 million bases, long read sequencing data has enabled more accurate characterization of genomes, especially of the highly repetitive regions and large SVs, which have been difficult to assess with short reads with lengths of the order of hundreds of bases. Indeed, long-read technology has empowered the near-complete assembly of the human genome by the T2T Consortium, elucidating centromeric satellite arrays, recent tandem duplications, and short arms of acrocentric chromosomes [46]. Similarly, it has further empowered accurate read-based haplotype phasing, which is the disambiguation of maternally- and paternally-inherited alleles using heterozygous SNVs. This is intuitive because long reads are likely to each span multiple polymorphic sites, leading to the construction of phase blocks of more substantial lengths, resolving allele-specific characteristics across a longer range of genomic coordinates. New tools have been developed specifically for phasing long reads [47–51].

The results discussed in Chapter 3 were based on data generated by ONT PromethION R10 flow cells. As a piece of genetic molecule passes through the protein nanopore, the change in electrical current caused by the sequence of nucleic acid bases is monitored and recorded. Neural network-based basecallers, such as `guppy`, and hidden Markov model-based caller, such as `Nanopolish`, have been developed to translate the recorded raw signals back into the genetic sequences that generated them. Furthermore, many basecallers have been trained to call base modifications like the methylation of the 5mC directly from the Nanopore raw signals. The advantage presented by the ability to directly call the methylation status of CpG sites from raw sequencing signals is two folds: first, it allows the

simultaneous assessment of many characteristics of the genetic material; secondly, the data will not have bias and artifact introduced to the genetic material by the additional round of bisulfite-treatment, which has so far been the gold-standard of methylation sequencing.

The capability of ONT long-read sequencing to jointly profile different types of genomic and epigenomic variants motivates the work discussed in Chapter 3. Expanding upon what we established in Chapter 2, that epigenomic variants also play a part in tumor evolution, the main goal of this proof-of-concept study is to showcase how cohering different types of somatic variants along a tumor phylogeny may reveal a holistic model of tumor evolution that describe the interplay among them.

1.8 Sequencing-based spatial transcriptomics

Spatial transcriptomics techonologies have enabled studying the spatial organizations of heterogeneous cell populations within a tissue. In particular, the 10x Genomics Visium platform - an extension of a previous generation of barcoding-based spatial transcriptomics technology [11] - is one that offers transcriptome-scale information at the resolutions of 1 to 10 cells per spot. One of the most cited success stories of the use of the Visium technology was produced by Maynard and colleagues [52]. In their work, Maynard and colleagues used the spatial information afforded by the Visium technology to study the six-layered human dorsolateral prefrontal cortex tissue and identified clinically relevant layer-enriched gene expression signatures, overcoming the difficulties in associating whole-cell expression with anatomical locations with single-cell, single-nucleus RNA-sequencing, or microscopy-based technologies. The spatial information adds another dimension to genomic analysis.

During sequencing, the prepared tumor tissue sample is overlaid over the 6.5mm-by-6.5mm capture area, which is composed of a field of around 5,000 spots 55μm in diameter and 100μm apart.

Spatially barcoded oligonucleotides present on each spot then binds to the mRNA from the tissue, and the reverse transcribed barcoded cDNA are collected to be processed for downstream Illumina short-read sequencing. It is important to note that, the Visium technology uses the 3' instead of whole transcript method; as a result, while gene expression can be reliably quantified, one cannot attempt to call expressed variants from Visium data due to the lack of full transcript information.

In Chapter 4 of this thesis, we examine how the addition of spatial information further our understanding of intra-tumor heterogeneity. Specifically, recent works have established that spatial technologies can further elucidate cell-cell interactions, which previously can only be inferred from co-expression of gene markers [53–58], by confirming their physical contact evidenced by their spatial co-location [59–67]. We ask the question, with additional information from full-length scRNA-seq data, can we extend our current knowledge on tumor-immune interaction to the subclonal level? This work explores how heterogeneous subclones, characterized by various types of somatic alterations described in Chapter 3, may interact with the tumor microenvironment differently, and therefore may have differential response to immunity.

1.9 Outline of Thesis

First, in Chapter 2, we establish the possibility to reconstruct tumor phylogenies using data from single-cell bisulfite-treated sequencing technology that profiles CpG methylation information on a single-cell level, highlighting the valuable phylogenetic signals in epigenetic data that have been underexplored. The newly developed method, Sgootr, jointly infers a tumor phylogeny and the amount of phylogenetic information in each CpG site from single-cell CpG methylation data. The effectiveness of the tool in inferring tumor phylogenies that suggest parsimonious migration histories was demon-

strated in simulated, U87MG glioma cell line, and multi-regionally sampled colorectal cancer data sets, respectively. The migration hisotry Sgootr inferred from the metastatic colorectal cancer patient, specifically, was more parsimonious than the one inferred from copy number changes in the original work.

Next, in Chapter 3, we present a novel framework to study both the genomic and epigenomic changes during tumor evolution using Nanopore long-read sequencing data of single-cell derived tumor sublines. From the B2905 cell line derived from the M4 mouse melanoma model, which is known to represent RAS-mutant transitory TCGA human melanoma with varied responses to anti-CTLA-4 treatment, 23 single cells were isolated and derived into sublines, which were then subjected to Nanopore long-read sequencing. SNVs, SVs, CNAs, as well as CpG methylation were profiled. A tumor phylogeny was constructed using a filtered set of mutations, revealing 4 major clades with distinct phenotypes validated via orthogonal experiments. The evolutionary timing of the various types of somatic variants was resolved against the tree, allowing the characterization of key alterations in terms of their clonality, and of the major clades in terms of the key alterations that may have contributed to their phenotypes. Critically, the clonal resolution this data set uniquely affords allowed the inference of active mutational processes, confident identification of parallel CNAs, and detection of differentially methylated regions, all at a subclonal level. Besides the biological insights, this data set also serves as a valuable resource for the community to generate pseudobulk of various clonality for benchmarking computational methods developed for heterogeneous cancer data.

In Chapter 4, we present a framework to examine tumor subclones at spatial resolution. Using the B2905 mouse malanoma cell line discussed in the previous section, heterogenous mixtures of cells from the cell line culture are injected into mice that are divided into control and treatment group. Anti-CTLA-4 immune checkpoint blockade therapy was administered to the mice in the treated group,

and tumors are resected for sequencing after various amounts of time. For each treatment condition, some tumors were subject to scRNA-seq via the Smart-seq2 full length transcriptomics sequencing, and the rest were subject to 10x Genomics Visium spatial transcriptomics sequencing. Six tumor subclones were identified using mutations profiled from the Smart-seq2 data, and subclone-specific expression profiles were used in cell type deconvolution of the spatial data by SpaCET. We validated the estimation of subclonal proportions at each spatial spot by examining their association with the few clone-specific mutations detectable directly on the spatial slides. The associations between tumor subclones with various stromal and immune cell types were accessed with established spatial statistics. By contrasting subclone-immune activities between treatment responders and non-responders, we show that differences in tumor response to anti-CTLA-4 treatment may relate to subclone-specific activities in its microenvironment.

Chapter 5 provides the conclusion to the thesis.

Chapter 2: Reconstructing tumor phylogenies with single-cell CpG methylation data using `Sgootr`

2.1 Motivation

There are two main challenges in constructing single-cell CpG methylation-based tumor lineage trees. First, detailed examination reveals that data obtained via single-cell bisulfite sequencing (scBS-seq) [43], single-cell reduced representation bisulfite sequencing (scRRBS) [68, 69], or multiplexed scRRBS (MscRRBS) [42] protocols exhibit high level of sparsity. Cells with subpar bisulfite conversion rates have information at few CpG sites, and of the millions of unique CpG sites across all cells in a dataset, less than one out of a hundred are sequenced in a sufficient number of cells to be useful in lineage reconstruction. Furthermore, even when a CpG site is sequenced in a cell, it is oftentimes covered by less than a few sequencing reads, which is not likely to capture heterozygosity in aneuploid cancers.

Besides sparsity, another key challenge is that not all CpG sites have their methylation statuses stably retained during tumor evolution. In particular, Meir and colleagues posited two models of methylation dynamics: *mixture*, where the methylation status of a CpG site is resampled (from the parental status) in each cell replication, and *persistency*, where that is an exact copy of the parent cell [45]. It is the CpG sites whose status follows the persistency model would be informative in tumor lineage reconstruction because the maintenance of information from the parental generation is the necessary condition for the infinite sites assumption [5, 70] and Dollo parsimony [71], which form the basis for tumor lineage inference tools based on mutation profiles. However, as of now, there is no known set of CpG sites that follow either inheritance model, highlighting the necessity to perform

intentional CpG site selection while reconstructing tumor lineages by CpG methylation.

To address these two challenges, in this work, we introduce `Sgootr` (Single-cell Genomic methylatiOn tumOr Tree Reconstruction, Figure 2.1), the first distance-based computational method to jointly select informative CpG sites and reconstruct tumor lineages from single-cell methylation data.

2.2 Methods

`Sgootr` (Figure 2.1) consists of five components: (i) optional biclustering of cells and sites for reducing sparsity-induced noise; (ii) likelihood-based sequencing error correction accounting for copy number estimates; (iii) expected distance calculation between cell pairs for tree construction; (iv) pruning of CpG sites according to a tree-based methylation status persistence measure; and (v) inference of migration history from the lesion-labeled tree leveraging prior knowledge on migration patterns. Components (iii) and (iv) are iteratively applied. Code for this work is available on GitHub as a `Snakemake` [72] workflow. Results from this work are reproducible, and intermediate experiment outputs of `Sgootr` are available for further analysis.

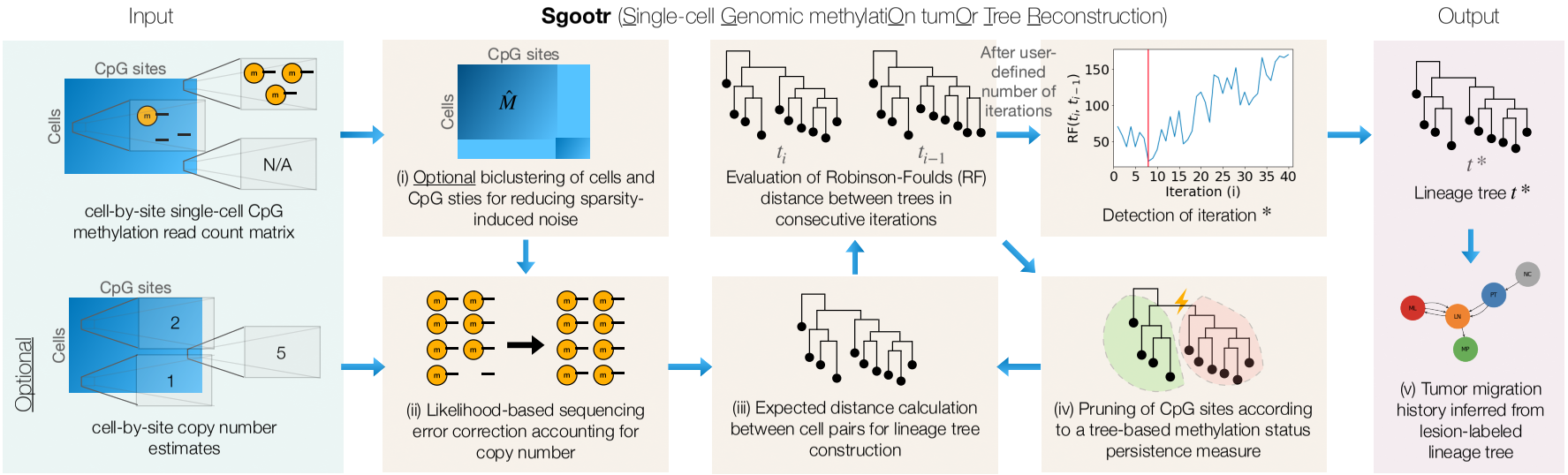


Figure 2.1: **Overview of `Sgootr`.** `Sgootr` leverages single-cell methylation sequencing data from tumor samples, incorporating copy number information when available, to jointly infer a single-cell tumor lineage tree and identify CpG sites that may harbor lineage-informative methylation changes.

2.2.1 Biclustering of cells and sites for reducing sparsity-induced noise

`Sgootr` requires a fairly accurate estimate of the differences between the methylation status of distinct CpG sites across all cell pairs. Prior to submitting the input dataset to `Sgootr`, it is advisable that one filters out cells for which the number of CpG sites with non-zero read coverage is low and CpG sites with non-zero coverage in only a few cells. For example, for the Bian metastatic CRC dataset [43] we will analyze further in depth in the Results section, we apply the following heuristic filters to the cells and sites: (i) among all input cells, remove low-quality ones with coverage in fewer than 4 million CpG sites; (ii) remove CpG sites covered in less than $\frac{2}{3}$ of the remaining cells. For that particular dataset, we additionally removed CpG sites on sex chromosomes, which may confound the findings, and Chromosome 21 peri-centromeric regions (hg19, Chr21:9825000-9828000), which we discovered to be an overlooked alignment artifact. After filtering the input for reducing sparsity, we perform sequencing error correction as described in the following subsection. Since this may further introduce sparsity, in an additional post error correction step, we may remove sites covering $< \frac{2}{3}$ of cells, and then remove cells with coverage in $< \frac{2}{3}$ of the remaining CpG sites. Overall, our filtering approach aims to make sure that the cells remaining are well-covered, and that each cell pair has at least $\frac{1}{3}$ of the selected CpG sites as shared dimensions along which cell-to-cell distance can be measured.

Our main results on the Bian [43] dataset demonstrate the heuristic filtering scheme described above is effective in practice; however, in principle, the filtering of cells and CpG sites should ideally be coordinated in order to minimize information loss. For that purpose, we present an integer linear program (ILP)-based biclustering formulation as follows. Let the site coverage data be represented in a cell-by-site matrix $M_{v \times u}$, where v is the number of cells and u is the number of CpG sites; $m_{ij} = 1$ if site j is covered (by at least one read) in cell i, and $m_{ij} = 0$ otherwise. Let α and β respectively be the

specified fraction of cells and sites to be kept. Given M, α, and β, we wish to compute a biclustering - namely, a selection of rows of columns - of M, $\hat{M}$, so that we maximize the number of CpG sites covered in the resulting $\lfloor \alpha v \rfloor \times \lfloor \beta u \rfloor$ submatrix in which rows and columns respectively correspond to the selected cells and sites.

To solve this biclustering problem, let $C \in \{0,1\}^v, S \in \{0,1\}^u$ be (unknown) binary vectors respectively indicating whether a cell or site is kept. In addition, let $A \in \{0,1\}^{v \times u}$ denote (unknown) binary matrix such that $a_{ij} = 1$ if and only if $c_i = 1$ and $s_j = 1$ (i.e., if cell i and site j are both kept). The ILP formulation is:

$$
\begin{aligned}
\text{maximize:} \quad & \textstyle\sum_{i=1}^{v} \sum_{j=1}^{u} a_{ij} m_{ij} \\
\text{subject to:} \quad & a_{ij} \in \{0,1\} \\
& c_i \in \{0,1\} \\
& s_j \in \{0,1\} \\
& a_{ij} \leq c_i \\
& a_{ij} \leq s_j \\
& c_i + s_j - 1 \leq a_{ij} \\
& \textstyle\sum_{i=1}^{v} c_i = \lfloor \alpha v \rfloor \\
& \textstyle\sum_{j=1}^{u} s_j = \lfloor \beta u \rfloor
\end{aligned}
$$

The output submatrix $\hat{M}$ is obtained by taking cells $\{i | c_i = 1\}$ and sites $\{j | s_j = 1\}$ from M.

We implement the above formulation as an optional step in our software using the Gurobi ILP solver, a licensed software which is free to use for academic purposes [73]. The performance of our implementation in terms of run time, memory, and accuracy (i.e. optimality gap) with respect to the input size and other parameters, as well as a strategy on setting α and β are explored in detail in

Supplemental Section A.1 (Supplemental Tables A.1,A.2, Supplemental Figures A.2,A.1,A.3). For a typical computational environment, the current implementation of the biclustering formulation may not be practical for the scale of the single-cell CpG methylation datasets analyzed in this study (see Supplemental Section A.1). However, since for given α and β values the ILP formulation produces a submatrix no sparser than that produced by the heuristic filters, it should be the preferred choice whenever it is feasible. In Supplemental Section A.1, we show that memory consumption is likely to be the bottleneck, and provide a way to estimate memory consumption of our biclustering implementation given the size of the input so that the users may determine whether performing biclustering may be appropriate for their own datasets of interest. It is intuitive that, the single-cell tumor lineage tree constructed from input data resulting from biclustering is as reliable, if not more reliable, than the tree constructed from heuristically filtered data, since the former will be constructed from at least as much information from the unfiltered data as the latter. Improving the run time and memory efficiency of our ILP implementation and devising alternative strategies to perform biclustering are among our future objectives.

2.2.2 Likelihood-based sequencing error correction accounting for copy number estimates

Errors in sequencing may provide false evidence for an underlying allele methylation status different from the truth. A common strategy employed in prior works to mitigate the effect of such sequencing errors in downstream tasks is to employ the 90% rule, which assigns a CpG site in a cell the methylation status with support from 90% of the reads available for the site in the cell, and discards the site if no status can be assigned [42–44]. This approach has two drawbacks. First, the binarization

of methylation status disallows heterozygosity. Second, while there are high levels of copy number gains in some cancer types, this approach does not distinguish between having reads sampled from two underlying alleles and having reads sampled from many. For a CpG site in a cell, we would like a copy number-aware way to determine, if we observe both reads with the site methylated and reads with it unmethylated, whether they are truly sampled from heterozygous alleles, or they are sampled from homozygous alleles and the methylation statuses on the disagreeing reads need to be corrected.

To this end, we present a maximum log-likelihood-based approach to correct likely errors in methylation reads, incorporating site copy number estimates in each cell (Algorithm 1). For a CpG site in a cell, given $n \geq 0$ reads with a methylated status, $m \geq 0$ reads with an unmethylated status, its copy number c, and sequencing error probability $0 \leq \epsilon < 1$ (which we set to .01 in our experiments to proxy the per-base sequencing error rate in Illumina sequencing instruments), we can enumerate the likelihood of having drawn the observed reads from all possible underlying allele statuses, letting $0 \leq \gamma \leq c$ be the number of alleles with the CpG site methylated. Note that in Algorithm 1, 1^c and 0^c respectively denote the events that all c alleles are methylated and unmethylated; $1^\gamma 0^{c-\gamma}$ denotes the event where γ alleles are methylated and $c - \gamma$ are not.

Sequencing error correction takes place when we have $n > 0$ and $m > 0$, but the allele status with the highest log-likelihood is homozygous. In that case, we correct the reads with the minority methylation status to the majority one. When $c = 1$ the site will implicitly be identified as homozygous and the methylation status of any read that does not conform to the majority allele will be corrected. In the case where sequencing error is needed, yet $n = m$, which will likely happen for $c = 1$, the CpG site is discarded. In case copy number information is not available, one can assume an appropriate (uninformative) copy number for all sites in all cells (e.g. $c = 2$ for diploid). Sex chromosomes are always excluded from our analysis. We use the sequence error-corrected reads for pairwise distance

Algorithm 1 Sequencing error correction accounting for copy number

$n, m \leftarrow$ number of methylated and unmethylated reads for a CpG site in a cell, respectively
$c \leftarrow$ copy number for the CpG site in the cell
$\epsilon \leftarrow$ sequencing error probability

function SEQUENCINGERRORCORRECTION(n, m, c, ϵ)
  if $c = 0$ **then return** $None$ ▷ site is discarded in case of deletion event
  $\mathcal{L}[1^c], \mathcal{L}[0^c] \leftarrow \ln[(1-\epsilon)^n \epsilon^m], \ln[\epsilon^n (1-\epsilon)^m]$
  for $0 < \gamma < c$ **do**
    $\mathcal{L}[1^\gamma 0^{c-\gamma}] \leftarrow \ln[\frac{\gamma}{c}^n (1 - \frac{\gamma}{c})^m]$
  end for
  $status \leftarrow \arg\max \mathcal{L}$
  if [$status = 1^c$ OR $status = 0^c$] AND [$n \neq 0$ AND $m \neq 0$] **then**
    if $n > m$ **then** $n, m \leftarrow n + m, 0$
    else if $n < m$ **then** $m, n \leftarrow n + m, 0$
    else return $None$ ▷ site is discarded if correction is impossible
  end if
  return n, m
end function

estimation between cells, as described in the following section.

2.2.3 Expected distance calculation between cell pairs for tree construction

In this section, we present a formulation for computing the expected distance between two cells given their respective copy number and (sequencing error-corrected) reads. We consider such a formulation with the goal of correcting for the potential bias contributed by low coverage CpG sites. Intuitively, this distance measure serves as an estimate for the expected proportion of methylation statuses altered across all CpG sites between a cell pair.

Given a cell, let the copy number at a CpG site be c, and we have $0 \leq \gamma \leq c$ alleles with the CpG site methylated, and $c - \gamma$ alleles with the CpG site unmethylated. To allow modeling preferential allele sampling based on site methylation status as previously observed [74], we additionally introduce parameter p, probability of drawing a read from the allele with the CpG site methylated given a pair of

alleles heterozygous for the site. We set p to the uninformative .5 in our experiments (see Supplemental Section A.8 for when to choose alternative values). Then, we can compute the probability of drawing from an allele with the CpG site methylated by normalizing over all alleles (Equation 2.1):

$$p_{c,\gamma} = \frac{\gamma p}{\gamma p + (c - \gamma)(1 - p)} \tag{2.1}$$

It follows that the probability of drawing from an allele with the CpG site unmethylated is $1 - p_{c,\gamma}$.

Consider the case where we observe n reads at a CpG site for a cell, and all n reads are methylated for the site. Let the copy number at the site for the cell be c, the probability of sampling all n methylated reads from c alleles with the CpG site methylated (i.e. $\gamma = c$) is: P(reads $\mid 1^c$) = 1; here 1^c is the event that in all c copies the CpG site is methylated, and reads is the event that all n sampled reads are methylated. The probability of drawing all n methylated reads from c alleles with the CpG site unmethylated (i.e. $\gamma = 0$) is: P(reads $\mid 0^c$) = 0. To compute the probability of drawing n methylated reads from c alleles with mixed methylation status for the site, we need to sum the probabilities over all other possible values of γ, the number of alleles with the site methylated, assuming that all other possible values of γ are equally likely; and when there is a copy number loss (i.e. $c = 1$), the formulation assigns only non-zero probability to a homozygous combined allele status:

$$\text{P(reads} \mid \text{mixed)} = \begin{cases} \frac{1}{c-1} \sum_{\gamma=1}^{c-1} p_{c,\gamma}^n & , c \geq 2 \\ 0 & , c = 1 \end{cases}$$

Given user-defined prior probabilities P(1^c), P(0^c), and P(mixed), which we all set to the uninformative .33 in our experiments (see Supplemental Section A.8 for when to choose alternative values),

let $a = \text{P(reads} \mid 1^c)\text{P}(1^c) + \text{P(reads} \mid \text{mixed)P(mixed)} + \text{P(reads} \mid 0^c)\text{P}(0^c)$, then we apply Bayes' Theorem to get the likelihood of any allele status given the observed reads (Equations 2.2):

$$\begin{aligned}
\text{P}(1^c \mid \text{reads}) &= \frac{\text{P(reads} \mid 1^c)\text{P}(1^c)}{a} = \frac{\text{P}(1^c)}{\text{P}(1^c) + \frac{\text{P(mixed)}}{c-1}\sum_{\gamma=1}^{c-1} p_{c,\gamma}^n} \\
\text{P(mixed} \mid \text{reads}) &= \frac{\text{P(reads} \mid \text{mixed)P(mixed)}}{a} = \frac{\frac{\text{P(mixed)}}{c-1}\sum_{\gamma=1}^{c-1} p_{c,\gamma}^n}{\text{P}(1^c) + \frac{\text{P(mixed)}}{c-1}\sum_{\gamma=1}^{c-1} p_{c,\gamma}^n} \\
\text{P}(0^c \mid \text{reads}) &= \frac{\text{P(reads} \mid 0^c)\text{P}(0^c)}{a} = 0
\end{aligned} \tag{2.2}$$

In the case where the n reads all have the CpG site unmethylated, the three values can be computed similarly with P(reads $\mid 1^c$) = 0 and P(reads $\mid 0^c$) = 1. In the case we observe among the n reads both ones with the CpG site methylated and ones with the site unmethylated, we have P(mixed | reads)=1, and P(1^c| reads) = P(0^c| reads) = 0. Then, for a given CpG site s in Cell A and Cell B, the respective copy numbers $c_{\text{A},s}$, $c_{\text{B},s}$, and respective reads, we can compute the expected distance over the possible combinations of allele status between the two cells at s (Equation 2.3):

$$\text{dist}(\text{reads}_{\text{A},s}, c_{\text{A},s}, \text{reads}_{\text{B},s}, c_{\text{B},s}) = \sum_{\substack{\text{status}_{\text{A},s} \\ \in\{1^{c_\text{A}}, \\ \text{mixed}, \\ 0^{c_\text{A}}\}}} \sum_{\substack{\text{status}_{\text{B},s} \\ \in\{1^{c_\text{B}}, \\ \text{mixed}, \\ 0^{c_\text{B}}\}}} \text{P}(\text{status}_{\text{A},s}|\text{reads}_{\text{A},s})\text{P}(\text{status}_{\text{B},s}|\text{reads}_{\text{B},s})\text{dist}(\text{status}_{\text{A},s}, \text{status}_{\text{B},s}) \tag{2.3}$$

where dist(11,11)=dist(10,10)=dist(00,00)=0, dist(11,10)=dist(10,11)=dist(00,10)=dist(10,00)=0.5, and dist(11,00)=dist(00,11)=1.

The total expected distance between cells A and B can now be computed with some distance function over the vector of expected distances over all shared sites. The L_1 norm normalized by the

number of shared sites is computed via Equation 2.4. The use of shared CpG sites between pairs of single cells to estimate their distances was established by pdclust:

$$\text{dist(A,B)} = \frac{\sum_{s \in \text{sites}_\text{A} \cap \text{sites}_\text{B}} \text{dist}(\text{reads}_{\text{A},s}, c_{\text{A},s}, \text{reads}_{\text{B},s}, c_{\text{B},s})}{|\text{sites}_\text{A} \cap \text{sites}_\text{B}|} \tag{2.4}$$

A comparison of our distance formulation against a baseline distance measure commonly used in prior studies! [42,44] on simulated data can be found in Supplemental Section A.2.1 (Supplemental Figures A.4,A.5,A.6). After computing the distance between each pair of cells, one can incorporate any distance-based tree reconstruction method into `Sgootr` to obtain the lineage tree and arbitrarily choose a single cell from the normal tissue as the root. We use the `scikit-bio` implementation of the NJ algorithm [75] for our main analysis. It is possible to replace NJ with an alternative distance-based tree construction method in `Sgootr`. We compare `Sgootr`'s performance while using FastME 2.0 [76], a popular alternative, against that using NJ in Supplemental Section A.5 and show that they lead to similar results (Supplemental Figures A.15,A.16, A.17).

2.2.4 Pruning of CpG sites according to a tree-based methylation status persistence measure

The main body of `Sgootr` consists of an iterative procedure: at each iteration, it (i) computes the pairwise distances among single cells to form the tumor lineage tree using a distance-based algorithm, then (ii) measures the methylation persistence score of each CpG site along the tumor lineage tree and prunes away a fraction of CpG sites that have the least scores before continuing onto the next iteration, and (iii) outputs the tumor lineage tree of the particular iteration where distance from the tumor lineage trees obtained in consecutive iterations are minimum possible.

We have described how to compute pairwise distance among single cells in the previous section, and we leverage the widely used RF distance [77, 78] to measure the differences among the tumor lineage trees constructed on the same set of cells in consecutive iterations. In this section, we focus on how we measure the methylation persistence of CpG sites at each iteration given the tumor lineage tree obtained for these CpG sites. In particular, we define the *methylation persistence score* for a CpG site given a tumor lineage tree (Equation 2.6): the higher the methylation persistence score for a CpG site, the more stably maintained its methylation status change along a tumor lineage tree.

To facilitate the analysis, we assign to each CpG site covered in each cell its most likely methylation status: given sequencing error-corrected reads from component (ii) of `Sgootr`, we call a site homozygously methylated if all of its reads have the site methylated, homozygously unmethylated if all unmethylated, and heterozygous if the read status for the site is mixed. The status remains unknown if there are no reads covering the CpG site. After this step, we measure the persistence of each CpG site independently: first at each branch of the tree, then for the overall tree.

2.2.4.1 Methylation persistence score for a CpG site at a particular branch in the lineage tree

Each branch in the methylation tumor lineage tree induces a bipartition of the tree and subsequently of the leaf nodes, which represent single cells. In other words, let TN denote the full set of leaf nodes in a methylation tumor lineage tree t, cutting a branch b in the tree creates disjoint subsets of leaf nodes TN_b and $\overline{TN_b}$. Let $Q_{s,tn}$ denote the probability distribution across three possible methylation statuses - homozygously methylated, heterozygously methylated, and homozygously unmethylated - at CpG site s across a subset of cells tn, and let I_s represent the set of cells with status information at site

s. Then, we define $\mathcal{MP}_{s,b}$, the methylation persistence score of CpG site s at branch b (Equation 2.5):

$$\mathcal{MP}_{s,b} = \sqrt{JSD\big(Q_{s,TN_b \cap I_s}, Q_{s,\overline{TN}_b \cap I_s}\big)}. \tag{2.5}$$

Here $JSD(X, Y)$ denotes the Jensen-Shannon divergence [79, 80] for a pair of distributions X, Y, and is defined as $JSD(X, Y) = \frac{1}{2}D(X\|Z) + \frac{1}{2}D(Y\|Z)$ with $Z = \frac{1}{2}(X + Y)$ and $D(K\|L)$ denoting the Kullback-Leibler divergence measure for any arbitrary distributions K and L [81]. It is worth noting that the square root of the Jensen-Shannon divergence measure, which computes $\mathcal{MP}_{s,b}$ (Equation 2.5), is metric [82, 83] and is commonly referred to as Jensen-Shannon distance in popular implementations.

The intuition is that, if the change in methylation status of CpG site s has an observed persistent effect in tumor progression - namely, s is lineage-informative - there exists a branch b^* in the methylation tumor lineage tree such that the leaf nodes in the two subtrees induced by b^* show very different distributions of methylation statuses. In other words, $\mathcal{MP}_{s,b^*}$ will be large for such b^*. In contrast, suppose the bipartitions contain the same distributions for methylation statuses for s - either that most cells share the same status, or that both bipartitions are similarly heterogeneous - the score will be low.

2.2.4.2 Overall methylation persistence score for a CpG site in the lineage tree

We define the overall methylation persistence score of CpG site s in methylation tumor lineage tree t as the maximum methylation persistence score the site s has across all valid bipartitions (Equation 2.6). We recognize that (i) an extreme difference in the number of cells between the bipartitions, or (ii) a severe lack of cells with status information could both lead to meaningless divergence measurements. Therefore, we only consider bipartitions induced by branch b such that (i) both partitions

contain no fewer than a user-defined fraction δ of total number of leaves, and (ii) there is read information in no fewer than a user-defined fraction ω of cells in both partitions. In our experiments, we set $\delta = .05$ and $\omega = .5$.

$$\mathcal{MP}_{s,T} = \max_{\substack{b \in \{b || TN_b| \geq \delta |TN| \wedge |\overline{TN_b}| \geq \delta |TN| \\ \wedge |TN_b \cap I_s| \geq \omega |TN_b| \wedge |\overline{TN_b} \cap I_s| \geq \omega |\overline{TN_b}|\}}} \mathcal{MP}_{s,b} \tag{2.6}$$

2.2.4.3 Iterative joint tumor lineage tree reconstruction and lineage-informative site identification

In a given iteration i, `Sgootr` first computes the tumor lineage tree t_i with the distances between pairs of cells based on the persistent sites identified in iteration $i-1$ (for iteration $i = 0$ the entire set of sites remaining after component (i) and (ii) of `Sgootr` are used). Then, for each CpG site s used in computing t_i, `Sgootr` calculates its overall methylation persistence score $\mathcal{MP}_{s,t_i}$. Among the CpG sites with overall methylation persistence scores, `Sgootr` prunes away a user-defined fraction κ of the CpG sites with the lowest scores, along with those with tying scores at the threshold. It outputs the remaining CpG sites to be used in iteration $i+1$. The process continues for a user-defined maximum number of iterations, and `Sgootr` outputs $t^* = t_i$ where i is the last iteration with $RF(t_i, t_{i-1})$ equals the global minimum across iterations.

Intuitively, we would like to detect an approximate point in the iterative procedure where most non-lineage-informative CpG sites have been pruned out (leading to the initial roughly decreasing trend of RF distance) but most lineage-informative CpG sites still remain (whose further elimination will lead to increasingly inaccurate distance measurements between cell pairs and therefore increasingly unstable tree topologies, which is reflected in a once-again increasing trend of RF distance). We

demonstrate empirical observations corroborate with such intuition (Figure 2.2C., Supplemental Figures A.11,A.17) and provide practical recommendations for choosing κ and the maximum number of iterations in Supplemental Section A.9.

2.2.5 Inference of migration history from lesion-labeled tumor lineage tree leveraging prior knowledge on migration patterns

Given a rooted tumor lineage tree t^* produced by `Sgootr` through the procedures described above, provided that the single cells in the tree are labeled by their lesion of origin, we can infer the underlying *tumor migration history*, which we represent as a directed multigraph (without self loops) where each vertex represents a distinct lesion and each edge represents a distinct migration event from the source lesion to the target lesion. This intuitive representation of metastatic migration events was introduced as a "migration graph" by El-Kebir and colleagues [29]. We accomplish this by first adapting the well-known Fitch-Hartigan algorithm [84,85] with slight modification to obtain a unique, maximally parsimonious labeling of the internal nodes of t^*, then identifying the migration events.

The Fitch-Hartigan algorithm was originally intended to solve the Small Parsimony Problem, namely, labeling the internal nodes of a character-based evolutionary tree in which every leaf is labeled with a single character and the objective is to minimize the total number of changes (i.e., mutations from parent node to child node) [84, 85]. In our reuse, the bottom-up phase of the Fitch-Hatigan algorithm computes in a tree the number of migration events - in other words, the change of lesion of origin labels from a parent node to a child - in the most parsimonious labeling of its internal nodes. This phase also produces a set of possible labels for each internal node in the tree. The top-down phase, then, assigns each internal node a label from its set, as described in Algorithm 2, starting with

Algorithm 2 Partial order-guided modification to the Fitch-Hartigan top-down node label assignment

$label_set(node)$ ← maps a node to the set of labels it is given by the bottom-up phase of Fitch-Hartigan algorithm fitch,Hartigan1973
$parent(node), children(node)$ ← returns the parent and the two children of $node$, respectively
$poset_minimal(labels)$ ← returns the minimal element in subset $labels$ of a poset L, breaking ties arbitrarily
$assignment[node]$ ← records the final label assignments of all nodes
$node$ ← a node in the tree

function LABELASSIGNMENT($node$)
  $candidate_labels \leftarrow label_set(node)$
  if $node$ is not root AND $assignment[parent(node)]$ is in $candidate_labels$ **then**
    $assignment[node] \leftarrow assignment[parent(node)]$
  else
    $assignment[node] \leftarrow poset_minimal(candidate_labels)$
  end if
  if $node$ is not a leaf **then**
    $child1, child2 \leftarrow children(node)$
    LABELASSIGNMENT($child1$)
    LABELASSIGNMENT($child2$)
  end if
end function

root node of t^*.

A general formulation of Fitch-Hartigan algorithm may produce multiple optimal labelings of the internal nodes, and there are existing tools to capture the space of all optimal labelings: for example, MACHINA by El-Kebir and colleagues enumerates optimal solutions allowing restrictions on the types of transition events [29], and FitchCount by [30] generates a stochastic matrix summarizing all transition events in the optimal solution space. However, some of these optimal labelings may adhere to our prior understanding of the sequential-progression (a.k.a. metastatic-cascade) model of tumor evolution [86] better than others. One can encode such prior understanding as a partial order over the lesion of origin labels of single cells. For example, in this work, we leverage the following partial order:

$$\text{Normal Tissue} \prec \text{Primary Tumor} \prec \text{Lymph Nodes} \prec \text{Distant Metastases}.$$

In our modification to the Fitch-Hartigan algorithm, during the top-down phase, when a node cannot inherit the label of its parent, we assign it the minimal element in its candidate label set according to the partial order, breaking ties arbitrarily when the elements in the label set cannot be compared (Algorithm 2). This returns a maximally parsimonious labeling of the internal nodes in the tree that also acknowledges the sequential-progression model.

With a fully labeled tree, we produce a directed multigraph that corresponds to the tumor migration history as follows. First, we generate a vertex for each distinct lesion represented in the tree. Then, as we traverse down the tree from the root, for each label change from a parent to a child in the tree, we add one directed edge to the multigraph, from the lesion vertex corresponding to the parent label to that corresponding to the child label. Parallel edges between a pair of vertices correspond to a polyclonal seeding event, and any back edge indicates a reseeding event.

Note that our proposed method resolves equally optimal labelings of the internal nodes with prior knowledge on the likely migration patterns among sampled sites. We have that prior knowledge for the primary dataset we analyze in this work, and the method provides a basis on which we compare single-cell lineage trees with lesion labeled leaves constructed with `Sgootr` or alternative methods. In the case where no prior knowledge is available and the goal is to survey the whole space of parsimonious migration histories possibly suggested by a lesion-labeled lineage tree, users may instead use tools like MACHINA [29] or FitchCount [30]. It may also be intuitive to more directly encode such prior knowledge by assigning different weights to migration events between distinct lesions, and solve the weighted version of the Small Parsimony Problem with Sankoff's algorithm [87]; however, the output of Sankoff's algorithm is highly sensitive to the input weights, and it is not immediately clear which

set of weights would be most appropriate for our purpose.

2.3 Results

The performances of `Sgootr`'s distinct components on simulated data are described in the Supplemental Sections A.1 (biclustering), A.2.1 (distance calculation), and A.2.2 (iterative pruning; Supplemental Figure A.7). Each of these components improves upon the respective baseline approach on a wide range of settings, and the results of the iterative pruning component demonstrate that `Sgootr` can capture the migration history of a tumor accurately. Furthermore, we applied `Sgootr` to scRRBS data of single cells from an U87MG glioblastoma cell line *ex vivo* clone tree experimentally generated by [88]. We show in Supplemental Section A.3 that, even though the original study fails to construct an accurate lineage tree from microsatellite allelotype data (see Supplemental Figure S10 of the RETrace study), `Sgootr` is able to reconstruct a CpG methylation-based lineage tree that accurately captures the ground truth clonal relationships (Supplemental Figure A.9).

In the remainder of the chapter we focus on the application of `Sgootr` to the scBS-seq dataset generated by Bian and colleagues [43] on 9 metastatic CRC patients[1], among whom 2 (CRC01 and CRC04) have matching scRNA-seq data (via the scTrio-seq2 protocol) from both normal and tumor cells with which copy number calling can be performed. We highlight results from patient CRC01 since in addition to having copy number calls available, they have the largest number of distinct tissue types, sampling locations, and treatment conditions. Full details and results for the other 8 patients can be found in Supplemental Section A.4 (Supplemental Table A.3, Supplemental Figure A.10). We additionally present in Supplemental Section A.7 `Sgootr`'s results on a GBM patient by Chaligne and

[1]Out of 12 total patients in the Bian cohort, we excluded those without scBS-seq data (CRC03 and CRC06), and those without metastasis information (CRC09).

colleagues [44], which demonstrate the applicability of our approach to MscRRBS data (Supplemental Figure A.19). To the best of our knowledge, this is the only other single-cell bisulfite sequencing dataset with multi-regional tumor sampling that is publicly available.

2.3.1 `Sgootr` obtains a simple tumor migration history with scBS-seq and scRNA-seq data from patient CRC01.

The single cells of CRC01 are sampled from 4 distinct lesions - PT, LN, ML, and MP - as well as NC adjacent to the primary lesion; furthermore, there are multiple sampling locations within each lesion (Figure 2.2A.). We only include cells with both scBS-seq and scRNA-seq data, and employ `inferCNV`[2] to call copy number from scRNA-seq data. We apply `Sgootr` with $\kappa = .1$, terminating the iterative procedure after 40 rounds. A unique global minimum among the RF distances is identified at iteration 8 (Figure 2.2C.), hence $t* = t_8$ (Figure 2.2B.) is output as the lineage tree. The tumor migration history inferred by `Sgootr` (Figure 2.2D.) is simpler than the CNA-based results previously reported (see Supplemental Figure S7 of the work by Bian and colleagues): NC grows into PT, followed by a polyclonal migration to LN, which appears to proceed to seed both ML and MP. While the migration history also suggests a polyclonal reseeding from ML to LN, both edges in the graph have low support: they are both due to a subtree of a particular cellular origin harboring a (potentially mislabeled) singular cell of a different origin.

For comparison, we also present in Supplemental Section A.6 the tumor lineage tree we constructed on mutations called from the matching scRNA-seq data from patient CRC01 using the `Trisicell` toolkit [91]. Due to the high level of sparsity in the mutation calls, we clustered single cells prior to constructing the mutation tree. We observed a high level of heterogeneity in terms of lesion of origin

[2] from the Trinity CTAT Project `https://github.com/broadinstitute/inferCNV`

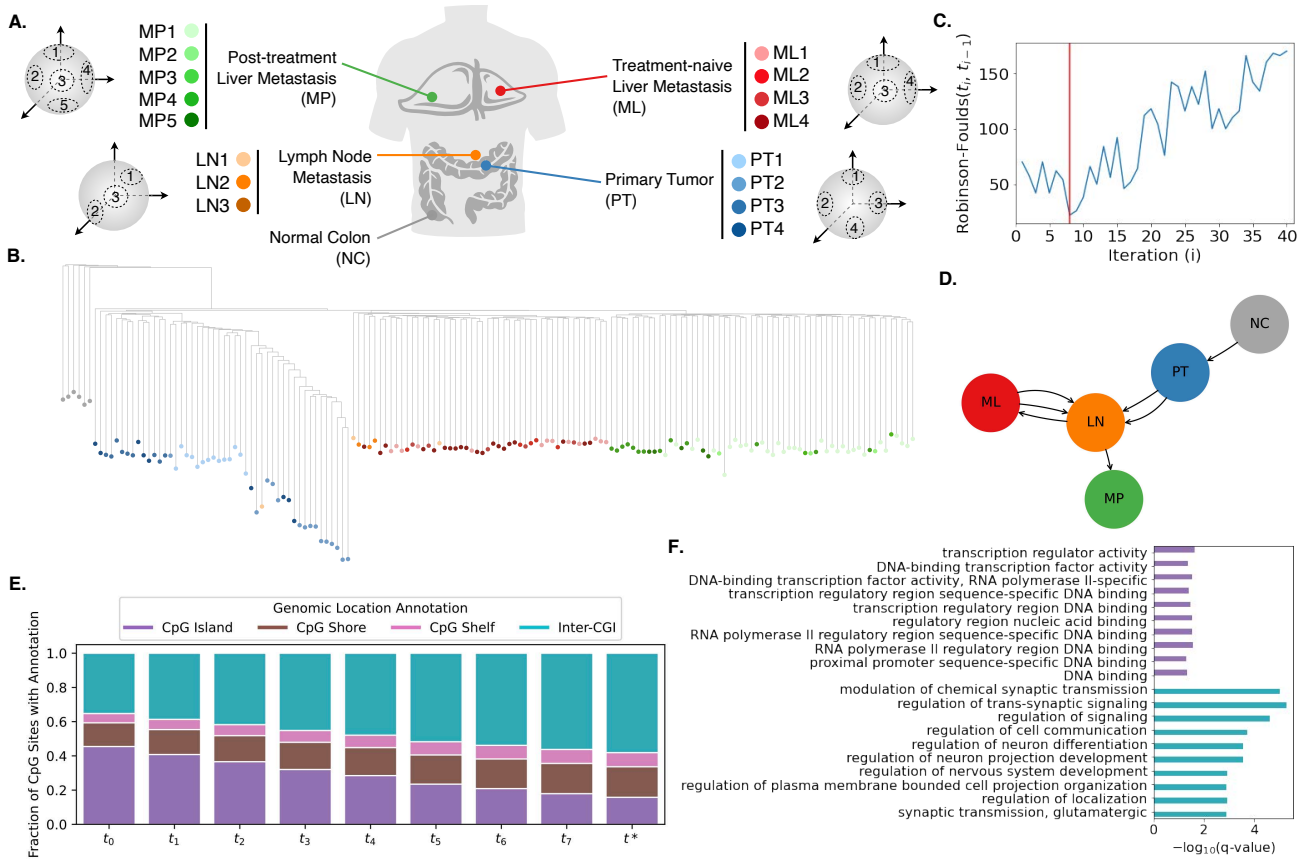


Figure 2.2: **Application of `Sgootr` to patient CRC01 scTrio-seq data by Bian and colleagues. A.** The multi-regional patient data consists of single cells sampled from 4 distinct lesions (in addition to normal tissue) and multiple sampling locations within each lesion. **B.** Tumor lineage tree constructed by `Sgootr` with $\kappa = .1$. Each single cell is represented as a leaf, colored by its sampling location. **C.** RF distances between the trees of consecutive iterations of the pruning procedure, with the global minimum occurring at $t* = t_8$. **D.** Tumor migration history inferred by `Sgootr` for patient CRC01. **E.** Fraction of CpG sites located in CGI (regions with over-representation of CpG sites) [89], CpG shore (2kb-long regions flanking both sides of CGIs) [90], CpG shelf (2kb-long regions adjacent to CpG shores on the side away from CGIs) [90], and inter-CGI regions in the CRC01 trees at each iteration of the pruning procedure of `Sgootr`, from t_0 to $t*$. **F.** Top 10 GO terms with significant ($< .05$) q-values in enrichment analysis of non-pseudo, protein-coding genes spanning lineage-informative sites in intra-CGI and inter-CGI regions respectively in CRC01. See Supplemental Figure A.12A. for a full list of enriched terms.

and sampling locations in the cell clusters (Supplemental Figure A.18), which makes inferring simple tumor migration history from trees constructed on them unlikely. Together with the CNA-based results(see Supplemental Figure S7 of the work by Bian and colleagues), this mutation-based result further highlights the usefulness of CpG methylation in reconstructing single cell lineage trees when sparse SNV and CNA data fail to provide sufficient signal for lineage reconstruction.

2.3.2 `Sgootr` infers migration histories simpler than alternative methods for the Bian CRC cohort.

We benchmark `Sgootr` against (i) a naïve baseline distance-based tree construction method (column "baseline" in Figure 2.3A.), (ii) IQ-TREE [92] with the two-state general time reversible model (`GTR2`), an instance of the popular maximum-likelihood-based tool used for inferring lineage trees from single-cell bisulfite sequencing data in prior studies [42, 44] (column "IQ-TREE" in Figure 2.3A.), and (iii) IQ-TREE (with `GTR2` model) preceded by FG, a lineage-informative site-selection method previously described by Gaiti and colleagues [42], which returns a subset of input CpG sites with lower than expected epimutation rate (column "FG+IQ-TREE" in Figure 2.3A.). Note that it is IQ-TREE preceded by FG that was employed exactly in previous studies [42, 44]. However, as discussed in more detail below, the FG step is not only very slow, it also leads to highly complex migration histories. These observations lead us to additionally benchmark against IQ-TREE without the FG step, even though it had not been directly applied to single-cell bisulfite sequencing data in prior studies. See Supplemental Section A.10 for further experiment details.

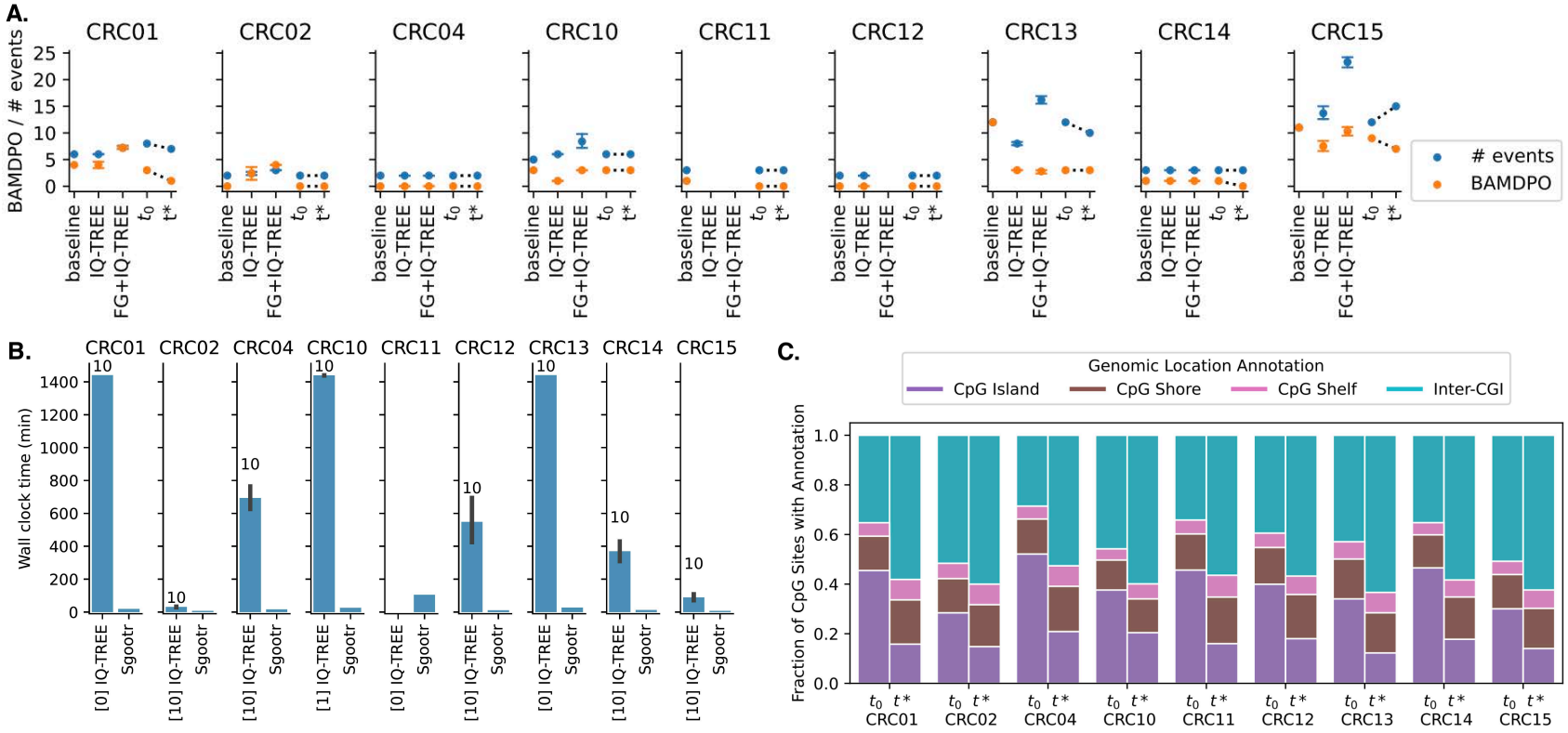


Figure 2.3: **Overview of results from the Bian metastatic CRC cohort. A.** Number of migration events and BAMDPO values of migration histories obtained via (i) naïve baseline method, (ii) IQ-TREE, (iii) IQ-TREE on FG-selected CpG sites, (iv) `Sgootr` intermediate tree prior to iterative procedure (t_0), and (v) final tree output by `Sgootr` ($t*$). IQ-TREE result is represented by the mean across 10 runs with different random seeds, with the error bars denoting the 95% confidence interval of the mean. The lack of data points in "IQ-TREE" columns means IQ-TREE fail to finish model optimization to start tree search after 24 hours, and that in "FG+IQ-TREE" columns means FG analysis fail to terminate after 100 hours. **B.** Run time comparison between `Sgootr` and IQ-TREE. Each experiment is performed on 4 cores. The number in brackets next to "IQ-TREE" on the x-axis represents the number of IQ-TREE runs (out of 10) that have converged within 24 hours. The number above each IQ-TREE bar represents the total number of runs (out of 10) with any tree search output (converged or intermediate after 24 hours), and hence contributes to the run time and migration history comparison. **C.** Fraction of CpG sites located in CpG island, CpG shore, CpG shelf, and inter-CGI regions, before (t_0) and after ($t*$) the iterative pruning component of `Sgootr`.

For each benchmarking experiment, we take as input the cell-by-site read count matrices resulting from first-pass heuristic filtering step of `Sgootr`, namely, removing low-quality cells, CpG sites with coverage in less than $\frac{2}{3}$ of remaining cells, sites on sex chromosomes, and sites within Chromosome 21 peri-centromeric regions. Then, for each CpG site covered in each cell, call whether it is methylated or not by the 90% rule [42–44]. This binarization step is necessary in generating input for the `GTR2` model. For the FG+IQ-TREE experiments, we further perform FG (Supplemental Section A.10, Supplemental Figures A.20, A.21) to select CpG sites to input to IQ-TREE. Additionally, we further filter the input to the benchmarking experiments so that all experiments (Figure 2.3A.) on a particular patient act on the same set of input cells. This calibration is to ensure fair comparisons among results from different methods using the two performance measures we will introduce below. To facilitate fair runtime comparison between IQ-TREE and `Sgootr`, we also filter to ensure the input to IQ-TREE for each patient is the same set as those to `Sgootr`, so that the runtime measurements in Figure 2.3B. are taken on the same set of cells and CpG sites for each patient.

For each patient in the Bian cohort, we generate a baseline lineage tree by first computing the Hamming distance between every pair of cells over their binarized CpG site methylation status vectors, averaging across the sites covered in both cells, then applying NJ to the distance matrix. As IQ-TREE is a stochastic method, we perform 10 separate runs of each IQ-TREE experiment with different random seeds, and for each instance report the intermediate tree at convergence or after 24 hours if convergence has yet to be achieved by then. If none of the 10 runs of IQ-TREE for a particular experiment have finished model optimization, started tree search, and produced some intermediate result, we do not report IQ-TREE result for that experiment. Similarly, if FG does not terminate in 100 hours, we do not report FG+IQ-TREE result for that experiment. We root the baseline and IQ-TREE trees with the same normal cells we used to root the trees by `Sgootr`, and we infer the tumor

migration histories from them via `Sgootr`'s modified Fitch-Hartigan algorithm.

We compare the migration histories inferred via `Sgootr` and the three alternative approaches by two measures: (i) the total number of migration events, namely the number of edges in our multigraph representation, and (ii) one we name BAMDPO, which is the Hamming distance between the binary adjacency matrix of the simple directed acyclic graph induced by the lesion partial order and that obtained by collapsing all parallel edges in the migration history multigraph. Intuitively, the first measure (output from Fitch-Hartigan algorithm) captures the conventional definition of parsimony, and the latter measures the degree the inferred history deviates from the sequential-progression model of tumor evolution. Together, they offer complementary perspectives on the complexity of the inferred histories. For both measures, a lower value would correspond to a simpler history.

Figure 2.3A. compares the tree $t*$ obtained by `Sgootr` against the alternatives with respect to the two measures. The figure also includes results of t_0 for each patient, the tree obtained by `Sgootr` prior to its iterative site pruning procedure. Comparisons of the measures for t_0 against those for $t*$ show `Sgootr`'s iterative procedure almost always reduces the complexity of the inferred migration history, provided such a reduction is at all possible. In the case of patient CRC15, even though the total number of migration events increases from t_0 to $t*$, BAMDPO decreases; this suggests what first appeared to be deviations from the simple sequential-progression model could potentially be resolved as polyclonal seeding events more in concordance with the model, once the tree structure becomes more refined through the iterative procedure. Compared to naïve baseline and the two IQ-TREE-based methods, `Sgootr` typically infers as simple, if not simpler, migration histories; the only exception is patient CRC10, where IQ-TREE (without FG site-selection) infers a migration history with the lowest BAMDPO.

Note that all patients except CRC01 and CRC15 have single cells sampled from only 2 lesions

in addition to the normal tissue (Supplemental Table A.3); among them, patients CRC02, CRC04, CRC12, and CRC14 have low number of input cells. For that reason, one may expect a lower number of migration events detected from the constructed lineage trees. Furthermore, `Sgootr` has achieved the minimal possible values in terms of BAMDPO prior to the iterative procedure in CRC02, CRC04, CRC11, CRC12; for CRC02, CRC04, and CRC12, `Sgootr` has also achieved the minimal possible value for the number of events prior to the iterative procedure. In those cases, there are no changes in terms of those two measures between the initialized tree (t_0) and the optimal tree ($t*$).

2.3.3 `Sgootr`-identified lineage-informative CpG sites construct lineages suggesting simpler migration histories than those identified via FG.

Specifically, comparisons between the measures from running IQ-TREE on FG-selected sites and those from `Sgootr` show that `Sgootr` consistently outperforms the former in terms of the number of events and BAMDPO where further improvements were possible. This suggests the sites selected by `Sgootr` may be more meaningful for the purpose of lineage reconstruction than those by FG. FG is also time-consuming: for patient CRC11 and CRC12, FG failed to complete after 100 hours, and no experimental result was reported (Figure 2.3A.).

2.3.4 `Sgootr` is orders-of-magnitude faster than IQ-TREE.

While `Sgootr` produces as simple, if not simpler migration history than IQ-TREE (without FG site-selection) in terms of the two measures in all but one case, `Sgootr` obtains the results in a fraction of time IQ-TREE took given the same input dimensions. Each IQ-TREE run was allotted 24 hours. If the program converged within 24 hours, we report the wall clock time elapsed; if the program

did not converge within 24 hours but produced some intermediate output from tree search, we report the program run time as 24 hours; and if IQ-TREE failed to produce any intermediate result from tree search within 24 hours, we do not include the elapsed time from that run. Within the allotted time, IQ-TREE produces converged results in 6 out of 9 patients, some (intermediate) results for 8 out of 9 patients. IQ-TREE is not able to finish model optimization and start tree search within 24 hours for any of the 10 runs for CRC11 (Figure 2.3B.). `Sgootr` is deterministic and obtains lineage trees suggesting simpler histories in a mere fraction of the time when given the same number of cores and memory as IQ-TREE, positioning itself as the better option in more time-frugal settings.

2.3.5 `Sgootr`-identified lineage-informative CpG sites are enriched in inter-CGI regions.

In all 9 CRC patients in the Bian cohort, as well as the multi-regional GBM patient MGH105 in the [44] cohort, the set of lineage-informative sites identified by `Sgootr` show enrichment in inter-CGI regions (Figures 2.2E., Supplemental Figures 2.3C., A.19F.). The median distances between CpG sites and their closest CGIs also increase as the iterations approach $t*$ across all patients (Supplemental Figure A.14).

2.3.6 Genes associated with `Sgootr`-identified lineage-informative CpG sites are enriched for pan-cancer and CRC-related terms.

To examine whether the lineage-informative sites discovered by `Sgootr` in the Bian cohort have potential functional implications, we stratified lineage-informative sites by their genomic annotations with respect to CGIs, for each group finding a set of non-pseudo, protein-coding genes whose

gene bodies and 1000 bp upstream promoter regions contain the sites, and carried out GO [93, 94] enrichment analyses respectively using GOrilla [95, 96]. All annotated non-pseudo, protein-coding genes in the genome were used as the background set. A hypergeometric test was used to measure the over-representation of the genes in associated GO terms, and the Benjamini-Hochberg false discovery rate-adjusted p-values (q-values) are presented. In 5 out of 9 CRC patients, genes mapped to lineage-informative sites within CGI are strongly associated with transcription regulation and DNA binding (Figure 2.2F., Supplemental Figure A.12), processes known to be linked to cancer onset [97]. Among informative sites mapped to inter-CGI regions, prominent enrichment in neural activities including ion transport and gated channel activities were observed across all 9 CRC patients (Figure 2.2F., Supplemental Figure A.12), concurring with recent studies which support the striking cross-talk between the neural system and CRC development [98, 99]. Enrichment results using an alternative tool, DAVID [100, 101], are also included for comparison (Supplemental Figure A.13). While the GO terms discovered are not identical between the tools, the results highly concur as indicated by the overlap coefficients of the terms (Supplemental Table A.4).

2.4 Discussion

In this chapter, we present `Sgootr`, the first distance-based tool to jointly infer tumor lineage using single-cell CpG methylation data and select for lineage-informative CpG sites. `Sgootr` concomitantly tackles two key challenges in lineage reconstruction from single-cell methylation data: sparsity and diverse inheritance dynamics of CpG methylation. We show `Sgootr` is able to construct lineage trees that suggest simpler tumor migration histories, perform CpG site selection, and obtain results orders-of-magnitude faster than alternative methods in applications to real multi-regionally-

sampled single-cell methylation data in CRC and GBM.

In addition to tumor lineage inference, `Sgootr` captures meaningful biological insights of identified lineage-informative CpG sites. In both CRC and GBM, the two cancer types included in our study, we discovered the enrichment of lineage-informative sites in inter-CGI regions across patients. We also observed that the variability of the methylation status of a CpG site among sequenced cells seems to increase with its distance from the closet CGI (Supplemental Table A.5). While the underlying mechanisms of these phenomena are yet to be elucidated and beyond the scope of this study, a plausible explanation is that CpG sites in islands are protected by the reduced chromatin accessibility in the region marked by histone modifications [102], whereas inter-CGI sites harbors more stochastic changes due to malfunctioning of DNA methyltransferases and TET enzymes [103]. While CGIs have attracted much attention in the studies of tumor epigenetic changes and more frequently covered in methylation arrays, our findings suggest that inter-CGI regions should not be overlooked especially in the context of tumor lineage tracing. We are looking to further study the relations between the methylation status of lineage-informative inter-CGI CpG sites and gene expression and explore their prognostic potentials in our subsequent studies.

Chapter 3: Studying melanoma evolution with Nanopore long-read sequencing data

3.1 Motivation

Short-read genome sequencing (e.g., Illumina) has been a workhorse of cancer genomics and has advanced our understanding of cancer heterogeneity and evolution [104]. However, these methods struggle to accurately detect structural rearrangements due to limitations in read mappability [105] or genomic coverage [106], and they require error-prone bisulfite conversion to detect epigenetic alterations [107]. As a result, short-read methods provide an incomplete view of the underlying mechanisms of tumor evolution.

Emerging long-read sequencing technologies, such as ONT and PacBio, overcome the limitations of short-read sequencing by offering improved read mappability, direct variant phasing, and simultaneous capture of DNA methylation [108]. Several recent studies have used long-read sequencing to study complex patterns of somatic variation and methylation in various cancers [105, 109, 110], but so far mostly at bulk resolution.

To study tumor evolution on sub-clonal level, single-cell genome sequencing has recently been used to profile variants in individual tumor cells, however the technology is limited by sampling sparsity [106, 111]. Alternatively, whole genome sequencing of single-cell derived subclones have been used to precisely trace subclonal differentiation with high sequencing depth [112–114]. So far all such studies were based on short-read results, thus providing only partial view of genomic and epigenomic evolutionary landscape.

To address this issue, in this chapter, we introduce an approach based on long-read sequencing of single-cell derived subclones; and apply it to analyze 23 melanoma subclones, each derived

from a single cell of the B2905 cell line [115]—a murine model of UV-induced, RAS-mutant, and differentiation-transitory subtype of human melanoma. These subclones were previously characterized by differential therapeutic responses [116]; however, the contributions of SV and methylation changes to the genomic and phenotypic heterogeneity of these subclones remained unclear. We therefore generated high-quality SNV, SV, CNA, and DMR calls in each subclone and developed new algorithms to infer their relative timing in distinct lineages through phylogenetic analysis.

Our long-read approach revealed the complex and multifaceted evolution of this heterogeneous tumor. We profiled the dynamic involvement of distinct mutational processes, including UV-induced damage, oxidative stress, and impaired DNA repair, offering deeper insight into their roles across the evolutionary stages of melanoma progression. We also observed recurrent amplifications of oncogenes occurring independently across multiple lineages and driven by distinct rearrangements, suggesting parallel adaptation under evolutionary pressures. Furthermore, we identified lineage-specific methylation trajectories associated with increased aggressiveness in distinct lineages, linking epigenetic changes with gradual phenotypic differentiation.

3.2 Results

3.2.1 Long-read sequencing of single-cell derived subclones of B2905 mouse melanoma model

To showcase our long-read approach to profile tumor evolution at subclonal level, we used B2905 cell line, derived from M4 cutaneous melanoma model, previously characterized with high intratumoral heterogeneity and in preclinical studies showed a diverse response to ICB therapy [115]. Here we performed long-read sequencing of 23 single-cell derived subclones of the B2905 cell line, previ-

ously isolated for expansion to become individual clonal sublines, and implanted into C57BL/6 mice to observe distinct survival times [116]. While previous studies also analyzed bWES, whole transcriptome, and scRNA-seq for each of the B2905 sublines [117–119], the contributions of SVs, CNAs and methylation to the evolution of this model remained unexplored.

We sequenced each subline using a single ONT R10.4 PromethION flow cell. On average, this yielded 69.77Gbp of data per subline with a mean read N50 of 29.1Kbp and an average of 82.95% of base pairs with a quality score of 20+. As a matching normal sample, we sequenced the splenocytes of a HGF-transgenic C57BL/6N mouse, the genetic background that was used to derive the B2905 model.

Following basecallling, we aligned the reads from each subline against the GRCm38 reference genome and to generate somatic SNVs, we used `DeepVariant` [120], and subtracted germline variants from the normal sample. For SV calls, we used `Severus` [105] in multi-sample mode. Additionally, since the mouse strain that derived B2905 cell line was generated by out-crossing HGF-transgenic FVB mice with C57BL/6 mice for over 10 generations, for SNV and SV variants, we subtracted known germline variants of the FVB mouse strain. Finally, we used `Wakhan` [121] for CNA calling, which uses SV breakpoint as the boundaries of CNA segments and fits a model integer copy number profile. Additionally, 5mC base modifications were profiled directly from the sequencing data, enabling the detection of DMRs using `BSmooth` [122], which we will discuss in further detail in a subsequent section. Together, these identified variants serve as the basis of the multifaceted analyses we will describe in the following sections (Figure 3.1).

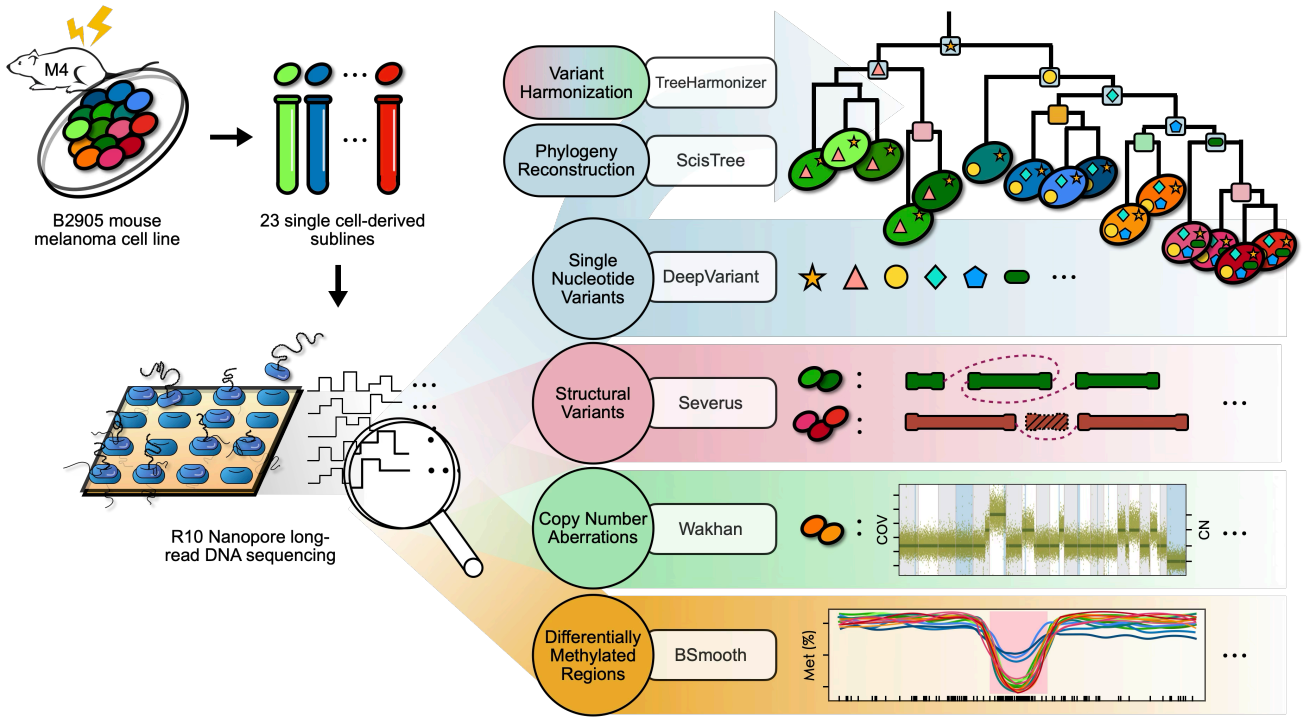


Figure 3.1: **An overview of the long-read approach and analysis of 23 subclones of a mouse melanoma tumor.** A tumor subline was cultured from each of the 23 single cells extracted from the B2905 cell line derived from the M4 mouse model of UV-induced melanoma, and the bulk samples were subjected to Nanopore long read sequencing. We studied melanoma tumor evolution integrating insights from mutation, structural variation, copy number variation, and differential methylation profiled from the sequencing data of the 23 sublines. Specifically, a tumor phylogeny of the sublines was constructed using SNVs, and then other types genomic and epigenomic events detected by various tools were placed onto the phylogeny. This provides an unprecedented look into melanoma evolution across time and various types of (epi-)genomic alteration events.

3.2.2 Constructed phylogeny reveals phenotypically-distinct major clades

To infer the evolutionary relationship between subclones, we constructed a phylogenetic tree with `ScisTree` [117, 123] using a subset of high-confidence SNVs (Supplementary Section B.1.1). From the initial set of 277,818 phylogenetically informative SNVs (61.75% of all SNVs used; excludes SNVs present in one or all sublines), we removed SNVs with low coverage (7.09% of informative SNVs) and SNVs that overlapped with deletions in any subline (89.00% of informative SNVs). This resulted in a final set of 10,862 high-confidence, informative SNVs. The constructed tree consisted of 4 clades that we will refer as "major clades" - colored in green, orange, blue, and red - representing 4 distinct subclonal populations in the B2905 cell line (Figure 3.2a.).

In a parallel study, we characterized the contrasting phenotypes of selected sublines from each clades according to their *in vivo* growth rates, expression of cell cycle genes, and response rates to anti-CTLA4 ICB treatment [116, 118]. Sublines with zero or low tumor take rate in immunocompetant C57BL/6 mice could grow tumors in immunocompromised nude mice, suggesting that immunogenicity impacts *in vivo* growth of these sublines. Red clade had an aggressive phenotype with rapid *in vivo* growth and lower response to ICB, resembling invasive melanoma subtype [124]; while orange sublines had slower *in vivo* growth and higher ICB response, resembling proliferative subtype. Green sublines showed mixed phenotypes of red and orange clades. Blue sublines had low tumor growth in immunocompetent C57BL/6 mice are referred as immunogenic [125]. (Supplementary Section B.1.2, Supplementary Table B.1) Overall, phylogenetic reconstruction reflected the observed phenotypic differences between sublines and enabled the analysis of variant timing and lineage phenotypes associations that we describe below.

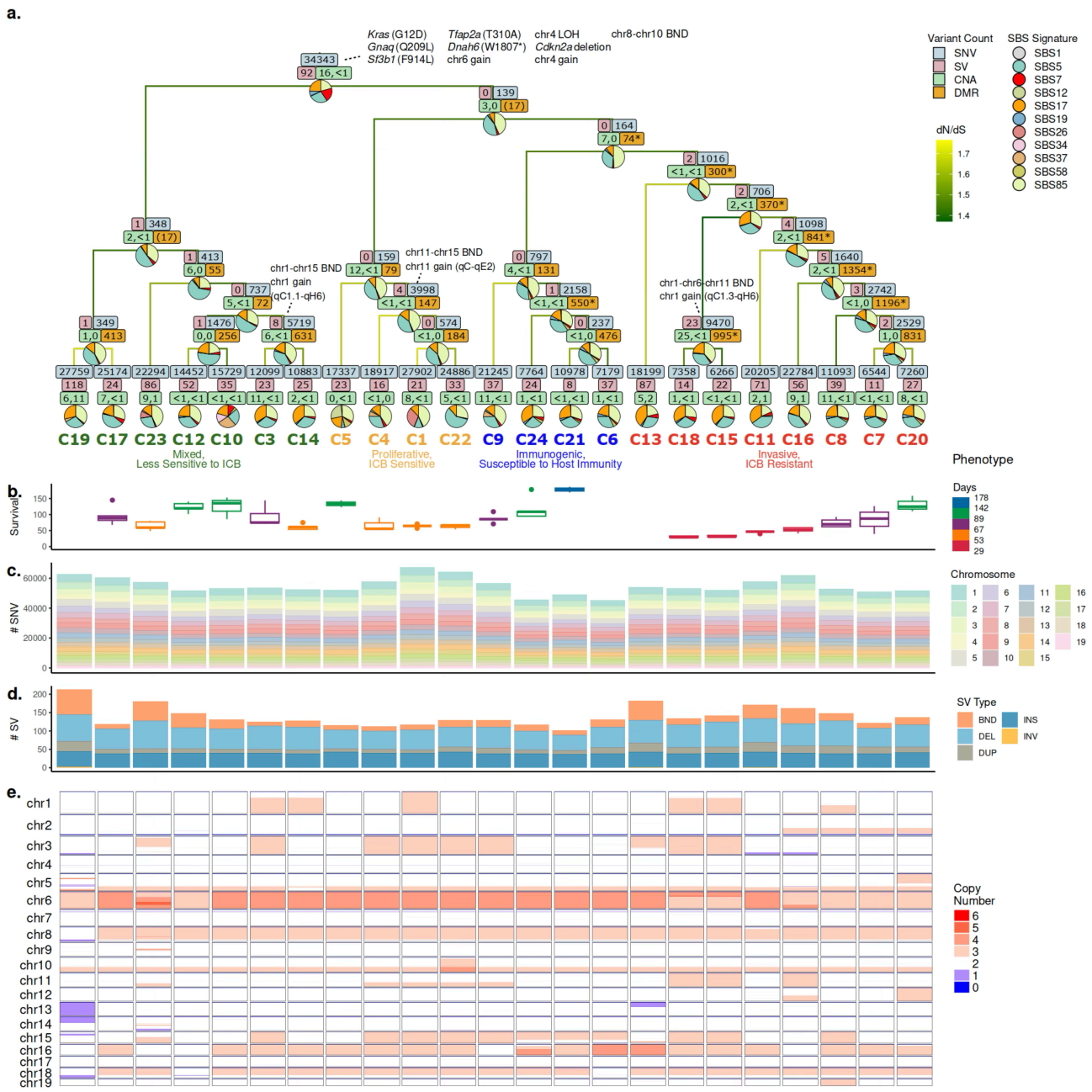


Figure 3.2: **Phylogenetic tree and mutational landscape of 23 single-cell derived sublines. a.** Phylogenetic tree with each node labeled with (i) counts of SNV, SV, CNA, and DMR assigned to the incoming branch of the node, (ii) exposure proportion of mutational signatures active along the incoming branch of the node, and (iii) key variants and events of note placed along the incoming branch of the node, if any. SNV and SV counts reflect those originating at that node. CNA numbers represent rounded percentages of genome collectively amplified and lost, respectively, among the subline below the branch, excluding those already accounted for in ancestral branches. DMR counts reflect the number of regions detected between the two groups of sublines defined by the bipartition of the tree at the node: the parenthesized numbers at the two immediate children of the root node reflect the same set of regions, and the numbers marked with an asterisk * denote those that are statistically significant per permutation tests (see Methods). The sublines are organized into 4 major clades, denoted by the distinct colors of the leaf labels. Sublines with previously determined phenotypes [118] also marked. **b.** Growth kinetic of tumors derived from individual sublines [116]. **c.** Per subline total SNV count per autosome. **d.** Per subline total SV count per SV type. **e.** Per subline copy number profiles.

3.2.3 Phylogeny-guided harmonization with `TreeHarmonizer` provides evolutionary timing of genomic variants

While the tumor phylogeny described above accurately reflects evolutionary relationship between the sublines, it excludes a large portion of SNVs that potentially overlap with deletions or have low coverage, and it does not include SVs or CNAs. This is typical for popular tumor phylogeny reconstruction methods that rely on SNVs [126]. Inclusion of different variant types is complicated by possible overlaps (e.g. SNVs within subclonal deletions) or erronous variant calls. We thus developed a tool called `TreeHarmonizer` that (i) places variants on the branches of a phylogenetic tree and (ii) identifies variants in disagreement with tree.

Briefly, `TreeHarmonizer` takes as input (i) a phylogenetic tree and (ii) individual SNV, SV, and CNA calls for every subline. For each unique variant, it evaluates whether the variant is consistent with the input phylogeny based on (i) the MRCA of all sublines carrying the variant, (ii) the variant status in all sublines, and (iii) the presence of nearby deletions. We denote a variant as clonal, subclonal, or private, if they were placed at the trunk, internal branch, or terminal branch of the tree, respectively (Figure 3.2).

The variant allele frequency (VAF) of placed clonal and subclonal variants was distributed around 50%, as expected for heterozygous somatic variants fixed in a given lineage (Supplementary Figure B.6c.,d.). Overall, high proportion of placed variants confirms the agreement between the phylogenetic tree and genomic landscape of the sublines. It also suggests that sublines were free of contamination, as it would result in an increase in unplaced variants.

To further evaluate the correctness of variant placement, we examined the set of SNVs that were placed by `TreeHarmonizer` differently, compared to the initial phylogenetic assignment by

`ScisTree`, which typically relies on more stringent filters for input variant calls in searching for the correct tree structure (Supplementary Section B.1.3). We compared the partition function values [119] that estimates the probability that, given a set of calls in individual sublines, a mutation is placed at the MRCA branch of a clade. Indeed, `TreeHarmonizer` variant placement led to a significant increase (one-tailed paired-sample t-test $\mathrm{t}(26) = 11.99, p = 2.133e-12$) in their partition function values, confirming accurate variant placement (Supplementary Figure B.5g.).

3.2.4 Evolutionary dynamics of somatic SNVs and mutational processes

Overall, the cumulative number of variants (all variants acquired along the course of evolution at a given tree node) was variable between sublines and major clades; in particular, sublines in the blue clade tend to have below-median number of both SNVs and SVs (Figure 3.2c.,d., Supplementary Figure B.6a.). The number of SNVs and SVs within individual branches in the tree were strongly correlated ($r(43) = .798$, $p = 5.158e-11 < .001$; Supplementary Figure B.6b.), suggesting that the rate of SNV acquisition relative to that of SV is roughly constant over time.

Given the number of variants placed by `TreeHarmonizer`, we sought to explore the timing of mutational processes, commonly explained through the patterns of somatic SNVs. We estimated activities of COSMIC [127] mutational signatures at each branch of the tumor phylogeny (Figure 3.2a.; Supplementary Section B.1.4). Notably, SBS7 (a UV damage signature) was prevalent at the trunk (Figure 3.2a., Supplementary Figure B.2), consistent with the fact that the tumor model was induced by UV radiation.

Some signatures were present throughout the phylogeny, such as clock-like signatures (SBS1 and SBS5), and SBS85 (attributed to AID) (Figure 3.2a., Supplementary Figure B.2). While clock-

like signatures presence is expected, SBS85 has been typically described in the context of lymphoid cancers [128]. However, it has also been shown that inflammation-induced AID expression in non-lymphoid cells promotes skin cancer development independently of UV damage [129].

Other signatures were more specific with timing or lineages. For example, SBS26 (defective DNA mismatch repair), had increased activity solely at the terminal branch attached to subline C1, which had the highest cumulative number of SNVs among all sublines (Figure 3.2a.,c., Supplementary Figure B.2). SBS17, a signature potentially associated with damages inflicted by ROS, shows significant difference in activity both between internal and terminal branches (adjusted $p = 9.08\text{e-}05$; Supplementary Section B.1.4), and between branches in the orange and red clade (adjusted $p = 9.68\text{e-}03$; Supplementary Section B.1.4). While the differences between internal and terminal branches could be explained by the additional oxidative stress caused by cell culture [130], the difference between the orange and red clade may reflect the difference in the metabolism between them, as we have shown previously that the expression of mitochondrial genes was selected in the former [118](Figure 3.2a., Supplementary Figure B.2).

To identify putative initiating events and variants that contributed to lineage differentiation, we performed functional annotation and enrichment analysis of SNVs (Supplementary Section B.1.6, Supplementary Figure B.3). Interestingly, dN/dS values at internal branches were significantly lower than terminal branches (p=5.9e-07, Supplementary Figure B.3a., Supplementary Section B.1.5), suggesting an increasing positive selection pressure over the evolutionary timeline (Figure 3.2a.). Among the clonal SNVs, we observe hotspot coding mutations, such as *Kras* (G12D) and *Gnaq* (Q209L) that are well-known cancer drivers, and variants in *Sf3b1* and *Tfap2a*, which are frequently mutated genes in melanoma [131, 132]. (Figure 3.2a.). Overall, impacted genes were enriched in pathways related to neural system development and differentiation as well as differentiation of neural crest to melanocytes

(Supplementary Figure B.3b.).

At the same time, individual clades exhibited enrichment in functionally distinct and phenotypically relevant pathways, such as DNA damage checkpoint signaling in the green clade (Supplementary Figures B.3c.), cell cycle and metabolism in the orange clade (Supplementary Figures B.3d.) and cell migration and methylation in the red clade (Supplementary Figures B.3f.). Overall, this analysis revealed the dynamics of mutational processes and putative functional effects of lineage-specific mutations, that we associate with other genomic, epigenomic or phenotypic changes below.

3.2.5 Somatic SVs dynamics suggest early drivers and late genomic instability

Long-read sequencing enables accurate detection of somatic SVs, which may be driven by orthogonal mutational processes and have substantial impact on lineage phenotypes, as we illustrate below.

Overall, small deletions (<1kb) were the most common SV; insertions were also common at the clonal level, but less common as private; and (interchromosomal) breakend junctions were more common at the private level (Figure 3.3a.-d.). `Severus` further groups breakend junctions into clusters if they are likely to result from a single mutational event. We identified 65 such clusters involving a total of 298 SVs: of which four were composed of clonal SVs 3.3e.), one of subclonal SVs (e.g. Figure 3.3f.), and 60 of private SVs (e.g. Figure 3.3g.). Private SV clusters were the most complex: while clonal clusters consisted only of two SVs each, 14 private clusters had at least five SVs, and largest cluster had 28 SVs (in C13). The green clade was overall enriched with clustered SVs, compared to the other clades (150/298 clutered SVs were in the green clade; Fisher's exact test, p = 3.055e-07), primarily driven by C19 that had the highest number of clustered SVs (n = 68) in nine clusters (Sup-

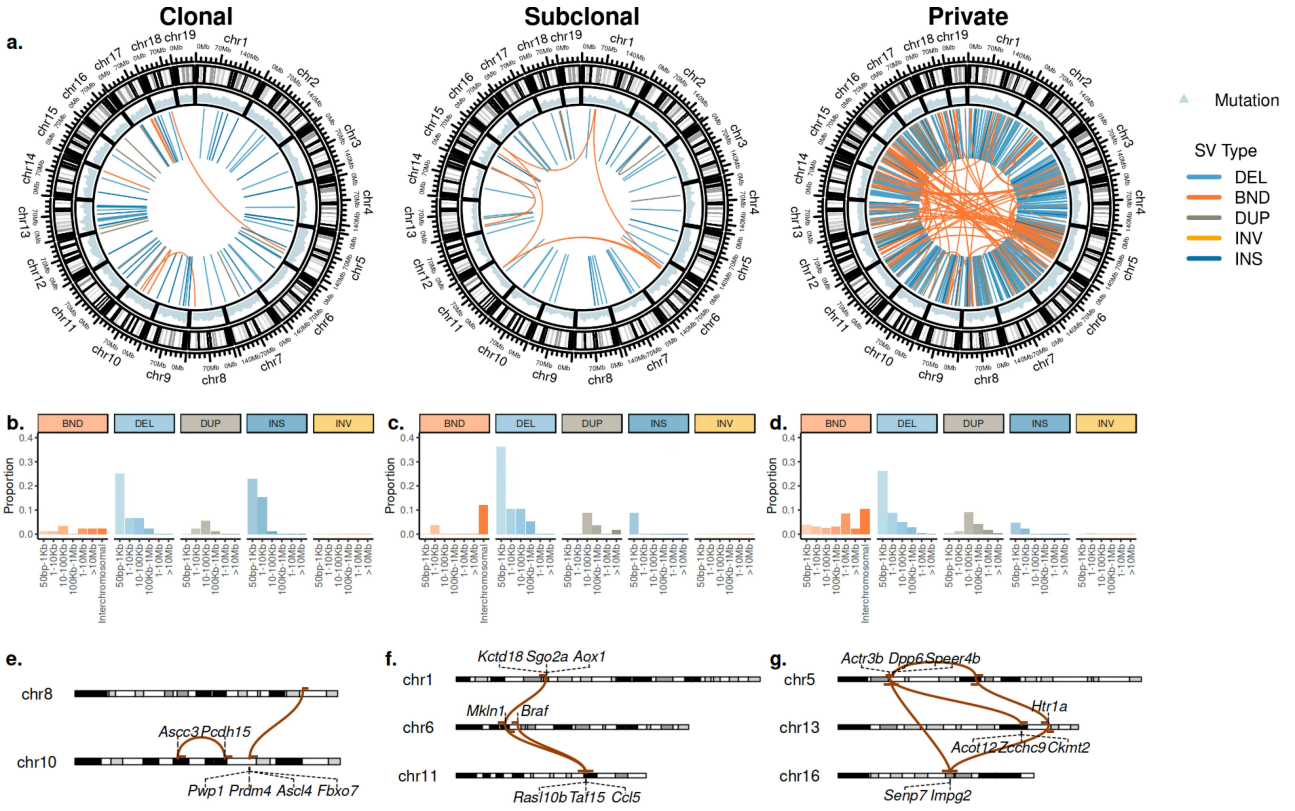


Figure 3.3: **Patterns and dynamics of somatic SVs. a.** Chromosomal distribution of clonal, subclonal, and private SNV and SV events. **b-d.** Distribution of clonal (b.), subclonal (c.), and private (d.) SVs across categories defined by SV type and size. **e-g.** Examples of notable clonal (e.), complex subclonal (f., shared by sublines C15 and C18), and complex private (g., private to C10) interchromosomal SVs. Genes near breakpoints are annotated.

plementary Figure B.7). In general, the vast majority of clustered SVs were private, suggesting that in this tumor complex rearrangements were late events, potentially driven by an increased genomic instability.

Using `Padfoot` (https://github.com/KolmogorovLab/Padfoot), we identified 564 SVs with breakpoints occurring within gene boundaries (Supplementary Section B.1.7). Of these, 330 affected coding regions through exon deletions or disruptions involving exonic sequences. Notably, we observed clonal deletions affecting exon 4-9 of *Nav3* (Supplementary Figure B.10a.), which included the first coiled-coil domain, the mutations in which have been shown to result in a loss of function [133]. In

accordance, deletion of *NAV3* gene has been linked to poor prognosis in cancer [134].

Additionally, in sublines C3 and C14, we identified a subclonal deletion in *Met* resulting in the loss of exon 4-11 (Supplementary Figure B.10b.), which included extracellular and regulatory domains. Such truncated *MET* has been shown to be self-activated and promoting immortalization of normal cells [135].

We also detected 24 candidate gene fusions, of which 23 were private. Among them, a subclonal fusion between *Map4* and *Smarcc1* was identified in C3 and C14 and validated using scRNA-seq (Supplementary Figure B.10c.).

Beyond disrupting genes and generating oncogenic fusions, SVs also cause CNAs that alter gene expressions, which will be the focus of the following subsection.

3.2.6 Recurrent CNAs in independent lineages suggest parallel adaptation process

SVs often cause CNAs potentially affecting large contigous regions of the genome. Overall, a high proportion of the genome was impacted by amplifications (26.37% on average per subline), and smaller fraction impacted by losses (1.24%) (Figure 3.2a.,e.). To understand functional implications of these alterations, we focused on genes with one-to-one human homologs curated in the COSMIC Gene Census [127] and labeled them as ONC, TSG), or ND; Supplementary Section B.1.8, Figure 3.4, Supplementary Figures B.11-B.16).

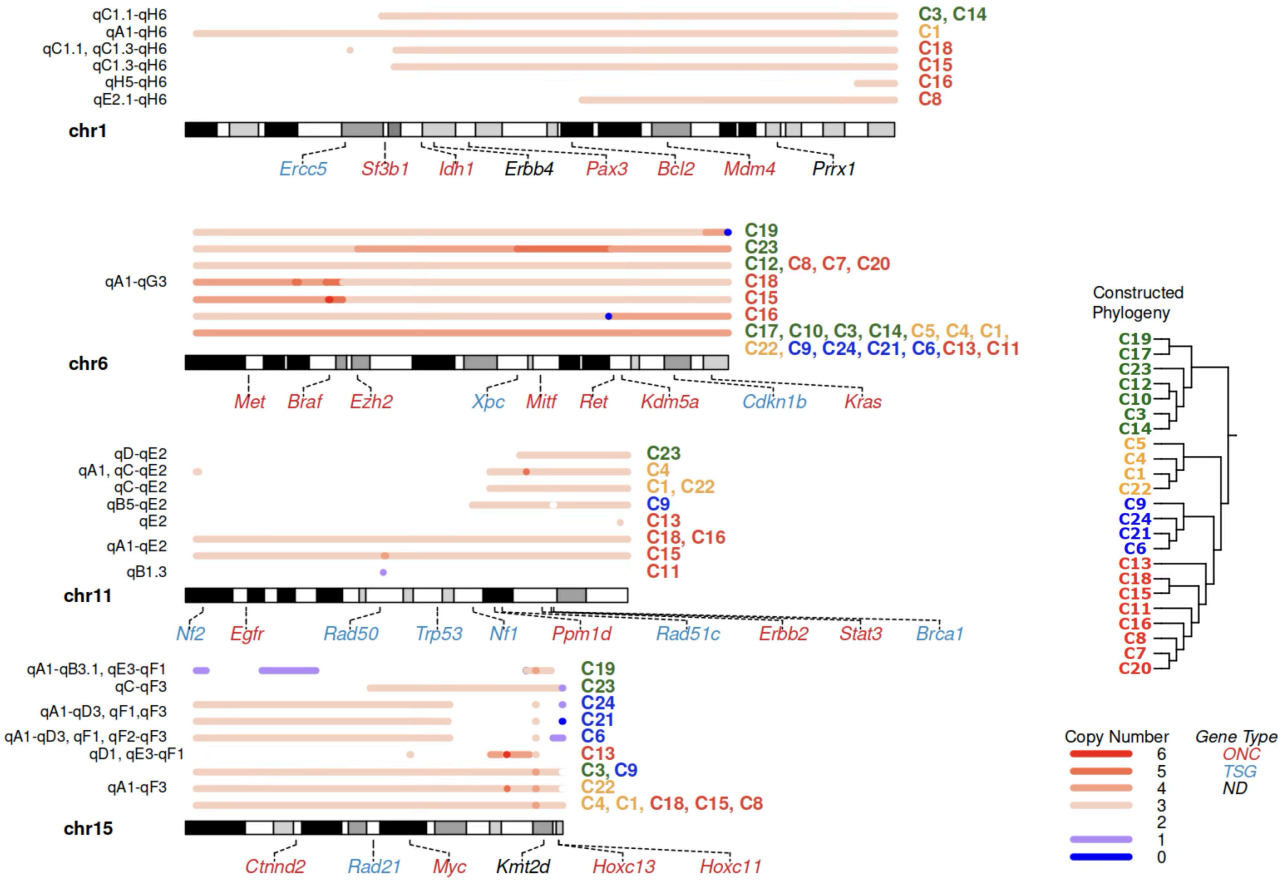


Figure 3.4: **Genomic regions recurrently affected by CNAs suggest parallel evolution.** Examples of regions on chr1, chr6, chr11, and chr15, recurrently harboring CNAs, possibly due to independent genomic events that impact the same region. Within each chromosome, sublines harboring CNAs are grouped by the similarity of their copy number profiles, and the respective copy number profiles are annotated by the chromosome bands they span. Select genes are annotated along the chromosomes and colored by whether their human homologs are designated as oncogene (ONC) or tumor suppressor (TSG) according to COSMIC Gene Census [127].

We first focused on clonal CNAs, such as whole-chromosome amplification of chr6 in all sublines, although some sublines acquired additional copies in later events (Figure 3.4). These CNAs affected key ONCs, such as *Met*, *Braf*, *Mitf*, and clonally mutated *Kras* (Figure 3.4, Supplementary Figure B.12). We also observed a focal loss of of *Cdkn2a* in both copies of chr4 with the same

breakpoint boundaries. This was likely a result of focal deletion, followed by copy-neutral loss of heterozygosity (LOH) in a larger region, happened approximately half way through the phylogeny trunk based on variant timing analysis (Supplementary Section B.1.9). These overall suggests a substantial number of early CNA events could have contributed to tumor initiation.

Next, we analyzed the patterns of subclonal and private CNAs. While many of them were attributed to internal tree branches by `TreeHarmonizer`, others remained unplaced. In unplaced CNAs, similar regions were amplified by different SV breakpoints in independent parts of the tree (Supplementary Section B.1.8). This suggests that these amplifications were independent recurrent events in different lineages, representing examples of parallel evolution. In one case, five distinct genomic rearrangement events resulted in independent amplifications of a large portion of chr1 in seven sublines (Figure 3.4): two SVs were subclonal, three were private; most of them were products of intra-chromosomal fusions, with one exception of whole-chromosome amplification in C1.

In another example of parallel CNA, a chr11 gain in C4, C1, and C22 is associated with the same subclonal interchromosomal SV involving chr15, the one in C9 involves chr18 and results in the amplification of TSG *Nf1*, which also carries a private, non-synonymous SNV in C9 (Figure 3.4, Supplementary Figure B.14).

This analysis suggests that subclonal CNAs substantially contributed to differentiation of tumor clades, and indicates convergent evolution in independent lineages.

Indeed, among COSMIC genes impacted by CNAs, ONCs were preferentially gained ($p = 3.832\text{e-}02$, Supplementary Section B.1.8) and TSGs were preferentially lost ($p = 7.294\text{e-}03$, Supplementary Section B.1.8) in this tumor, as expected. This suggests that those changes may be independently selected for in a highly heterogeneous tumor [136]. For example, parallel amplification of *Myc* was recurrently present in a majority of sublines (Figure 3.4, Supplementary Figure B.15); likewise,

Myc homolog is commonly amplified in human melanoma [137]. In C13, *Myc* is duplicated via a focal SV spanning merely 1.6kb, likely reflecting targeted selection.

Furthermore, a closer examination of ONC/TSG COSMIC gene homolog pairs adjacent to copy number change boundaries shows that, more often than expected by chance, ONC falls on the side of the boundary with the higher copy number state while TSG on the lower side (p=3.74e-02, Supplementary Section B.1.8). This provides evidence that relative positioning of ONC and TSG genes on chromosome may influence the selection of CNA boundaries.

3.2.7 Methylation changes contribute to differentiation of an aggressive clonal lineage

In addition to genomic variants, epigenomic changes may shape the observed distinct lineage phenotypes. Long-read sequencing, combined with phylogenetic approach, enabled us to identify genomic regions that are differentially methylated at subclonal resolution (Figure 3.2a).

For each internal branch in the phylogeny, a set of DMRs between the bipartitions of the sublines induced by the corresponding tree split was identified using `BSmooth` [122, 138](Supplementary Section B.1.10). We then evaluated significance of the number of DMRs obtained via a permutation test for branches ancestral to a particular number of leaves (Supplementary Section B.1.11, Supplementary Figure B.17).

We found a strong positive correlation between the number of cumulative SNVs and the number of DMRs at the branch ($r(5) = .842, p = 1.758$e-2, Supplementary Section B.1.12; Figure 3.5a.), suggesting that the rate of differential methylation to some extent mirrors that of SNVs in the evolution of this tumor.

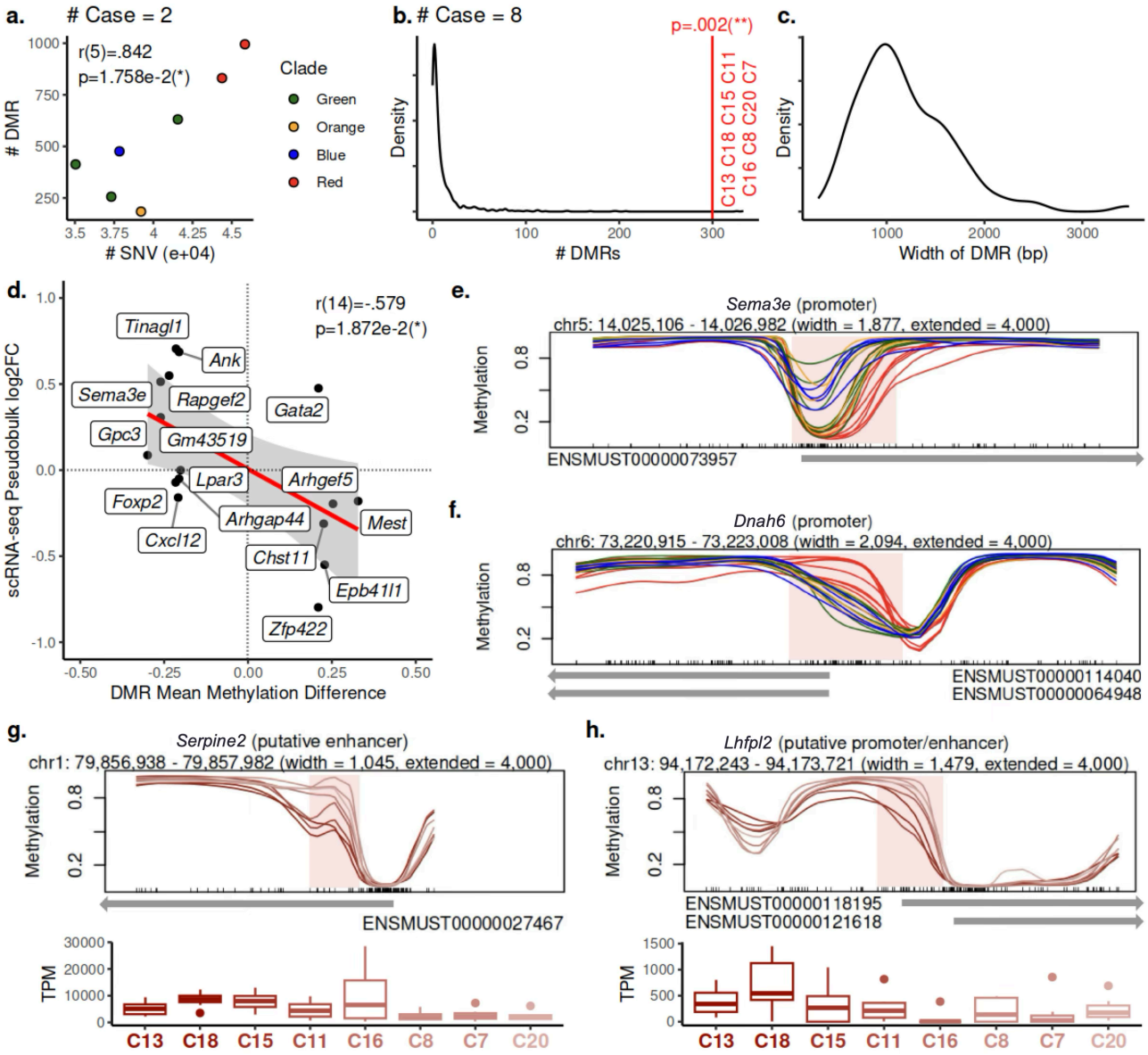


Figure 3.5: **DMRs detected at subclonal resolution. a.** The number of SNVs cumulative from trunk to the branch and the number of DMRs detected at the branch, for each branch parental to 2 leaves. **b.** The number of DMRs detected at the branch rooting the red clade and the corresponding null distribution of the number of DMRs generated during the permutation test for case groups of size 8. **c.** The distribution of the widths of the 300 DMRs detected at the branch rooting the red clade. **d.** Mean methylation difference within DMRs between sublines in the red clade and others and the corresponding log2 fold change in gene expression in genes whose (putative) promoter or enhancer the DMRs overlap as measured by scRNA-seq. **e.** DMR overlapping the promoter of *Sema3e* where sublines in the red clade are hypomethylated. **f.** DMR overlapping the promoter of Dnah6 where sublines in the red clade are hypermethylated. **g.** Genomic region overlapping a putative enhancer of *Serpine2* where it is monotonically becoming hypermethylated in the red clade while the gene expression of *Serpine2* follows a monotonicly decreasing trend. **h.** Genomic region overlapping a putative promoter of *Lhfpl2* where it is monotonically becoming hypermethylated in the red clade while the gene expression of *Lhfpl2* follows a monotonicly decreasing trend. In **e.** and **f.**, the methylation level in each subline is plotted with the color corresponding to the clade of the subline in the tumor phylogeny. In **g.** and **h.**, the methylation level and TPM values as profiled by scRNA-seq in each subline are plotted with the color hue corresponding to its phylogenetic order in the red clade.

We then focused on epigenomic analysis of the red clade since it was characterized with significantly worse survival, compared to other clades (Kaplan-Meier, $p < .0005$; Figure 3.2a.). For the root branch of the red clade, `BSmooth` [122] obtained 300 DMRs, which is significantly more than what is expected for any group of eight sublines ($p = .002$, Supplementary Section B.1.11, Figure 3.5b.). The median length of the DMR regions was 1,081bp (Figure 3.5c.). Among all 300 DMRs, 17 overlap with the putative promoter regions in GENCODE transcription start sites [139, 140], and nine were annotated in the EPDnew (v.003) database of experimentally validated promoters [141, 142]. Another nine DMRs displayed enhancer-like signature and were within 2kb of an annotated transcription start site [143–146]. (Table B.2).

Notably, the EPDnew-validated promoter of *Sema3e* was found to be hypomethylated (Figure 3.5e.), and its human homolog has been shown to drive cancer cell invasiveness [147]. Additionally, the EPDnew-validated promoter of *Dnah6*, previously implicated in ICB response in human melanoma [148], was found to be hypermethylated in the red clade (Figure 3.5f.). Coincidentally, *Dnah6* also harbors a clonal truncating mutation p.W1807* (Figure 3.2a.). Upon closer examination, we found evidence for haplotype-specific methylation of *Dnah6* promoter on the mutated copy in green, orange and blue clades, and the hypermethylation in the red clade was likely associated with the loss of the wild-type copy (Supplementary Figure B.18).

While the red clade was distinct from the rest of the phylogeny in terms of survival, there are also gradual changes of genotype and phenotype within the clade. The 8 red sublines followed a linear differentiation pattern, instead of a branching pattern in other clades. Further, survival monotonically improved along the evolutionary trajectory, $\tau(5)=1$, $p < .001$ (Supplementary Section B.1.14, Figure 3.2a.). To evaluate if methylation was driving gradual change in phenotype, we analyzed DMRs within the red clade along the trajectory and focused on (putative) promoter or enhancers of known

genes.

We found 146 genes whose (putative) promoter or enhancer overlap with regions displaying monotonic change in methylation rate along the evolutionary trajectory, which is found to be significantly more than what is expected by chance via a permutation test on the order of the sublines within the red clade ($p = .021$, Supplementary Section B.1.14). Some of these methylation changes were correlated with changes in scRNA-seq expression. In particular, a putative enhancer of *Serpine2* and the putative promoter and enhancer of *Lhfpl2* are gradually hypermethylated in the red clade, and their expression correspondingly decreases (Figure 3.5g.,h.). In human melanoma, SerpinE2 expression has previously been associated with high invasive potential of slow-cycling cells [149], and *LHFPL2* expression was associated with mTORC1 signaling [150], which promotes mesenchymal-epithelial transition in cancer [151].

3.2.8 Systematic analysis of multiple variation types provides insights into B2905 lineages differentiation

Overall, our long-read approach offers us a unique lens to observe the complete spectrum of genomic and epigenomic events at the subclonal level and provided insights into the evolution of the B2905 melanoma (Figure 3.6).

We detected clonal mutations of all types that affected genes known to contribute to melanoma initiation, including hotspot SNVs in *Kras* and *Gnaq*, SNVs in frequently mutated sites of *Sf3b1* and *Tfap2a*, homozygous deletion of *Cdkn2a*, and amplifications of *Met*, *Braf*, *Mitf*, *Kras*, *Cdk4*, and *Erbb3*. We observed a substantial proportion of UV damage signature on the clonal level, but not on sublonal levels; confirming UV damage as a key initiation event [152], but unlikely the driver of

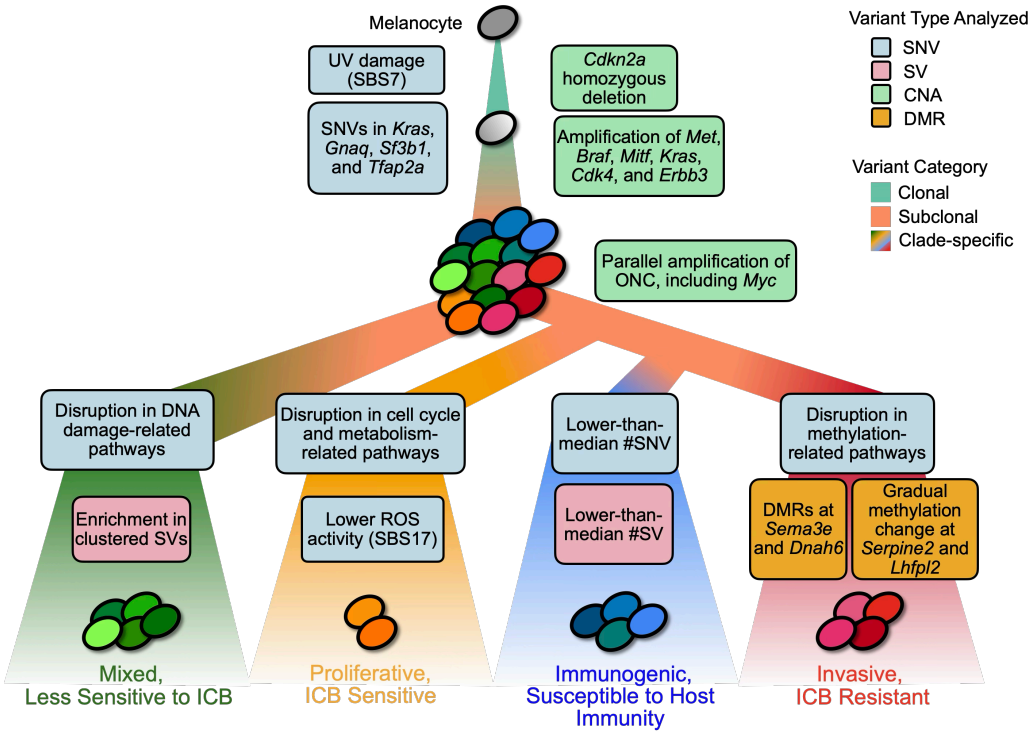


Figure 3.6: **A model of B2905 melanoma evolution informed by analyses of long-read sequencing data on the 23 sublines.** Key clonal, subclonal, and clade-specific observations from analyses of the SNVs, SVs, CNAs, and DMRs are labeled along the tumor evolutionary trajectories.

lineage differentiation.

On the subclonal level, accurate and precise SV breakpoints detection, enabled by long-read sequencing, allowed us to resolve parallel CNAs events that affected overlapping genomic regions. For example, *Myc* was independently amplified across different lineages via distinct genomic rearrangements, ranging from whole-chromosome duplications and a small, focal amplification. Overall, the positioning of CNA breakpoints relative to neighboring ONC and TSG genes suggested that these events occurred randomly, and then were independently selected.

Furthermore, our phylogenetic approach enabled the comparisons among different evolutionary trajectories of major clades that resulted in observed phenotypic heterogeneity. In the green clade, which had a significantly higher number of clustered SVs, we observed a disproportional disruption of pathways related to DNA damage signaling and response (Supplementary Figure B.3c.), possibly causing an increase in genomic instability. In orange clade, the disrupted pathways related to cell cycle and metabolism and lower ROS activity cohered with its proliferative phenotype (Supplementary Figure B.3c.,B.2). Blue clade sublines had significantly fewer number of clade specific SNVs (one-tailed t-test, $p = 3.31e-02$) and SVs (one-tailed t-test, $p = 1.65e-02$); which may reflect lower selection pressure on the blue sublines, but requires further investigation.

The red clade trajectory – showing faster growth in immunocompetent hosts and higher ICB resistance – stood out in terms of increased number of characteristic DMRs, including promoters of *Sema3e* and *Dnah6* that could drive the aggressive phenotype. Interestingly, methylation-related pathways in the red clade were disproportionally impacted by non-synonymous SNVs, which may explain the increase in observed methylation changes, compared to other clades (Supplementary Figure B.3f.). Moreover, we observed that a significant number of genes in the red clade, such as *Serpine2* and *Lhfpl2*, displayed a monotonic change in methylation at (putative) promoters or enhancers, often cor-

related with aggressiveness and scRNA-seq expression changes. This suggests that DMRs may drive a gradual phenotypic trajectory change.

Overall, this showcases how joint analysis of SNVs, SVs, CNAs and DMRs provides explanation of lineage-specific phenotypes and genotypes.

3.2.9 Curated set of subclonal variant calls enables benchmarking of long-read tools

In addition to enabling a highly detailed view of melanoma evolution, our approach provides a unique resource for development of new cancer genomics methods. Variant placement using our tool `TreeHarmonizer` resulted in a highly accurate set of SNV, SV and CNA calls with established phylogenetic relationship. This enables modeling a heterogeneous tumor sample by mixing subsets of sublines sequencing data, as for each such subset a benchmarking set of variants could be easily derived. As a proof of concept, we generated 5 pseudobulk samples by merging reads from different clade combinations (Figure 3.7a.), and then compared several recently developed state-of-the-art tools for long-read variant calling in terms of their precision and recall on somatic variants (Figure 3.7b.).

While precision was generally high for all callers, recall varied highly between callers and across different pseudobulk samples (Figure 3.7b.). Specifically, as compared to clonal SNVs, recall for subclonal and especially private SNVs was substantially lower, likely due to the decreased allelic fraction of such variants (Figure 3.7a.,c.). Overall, we found that recent tools that were developed for somatic analysis (`ClairS` and `DeepSomatic`) performed better than general-purpose variant callers, with `DeepSomatic` performing the best, maintaining the high precision as other callers while notably outperforming in recall (Figure 3.7b.).

This analysis underscores the difficulty of detecting rare variants from heterogeneous cancer

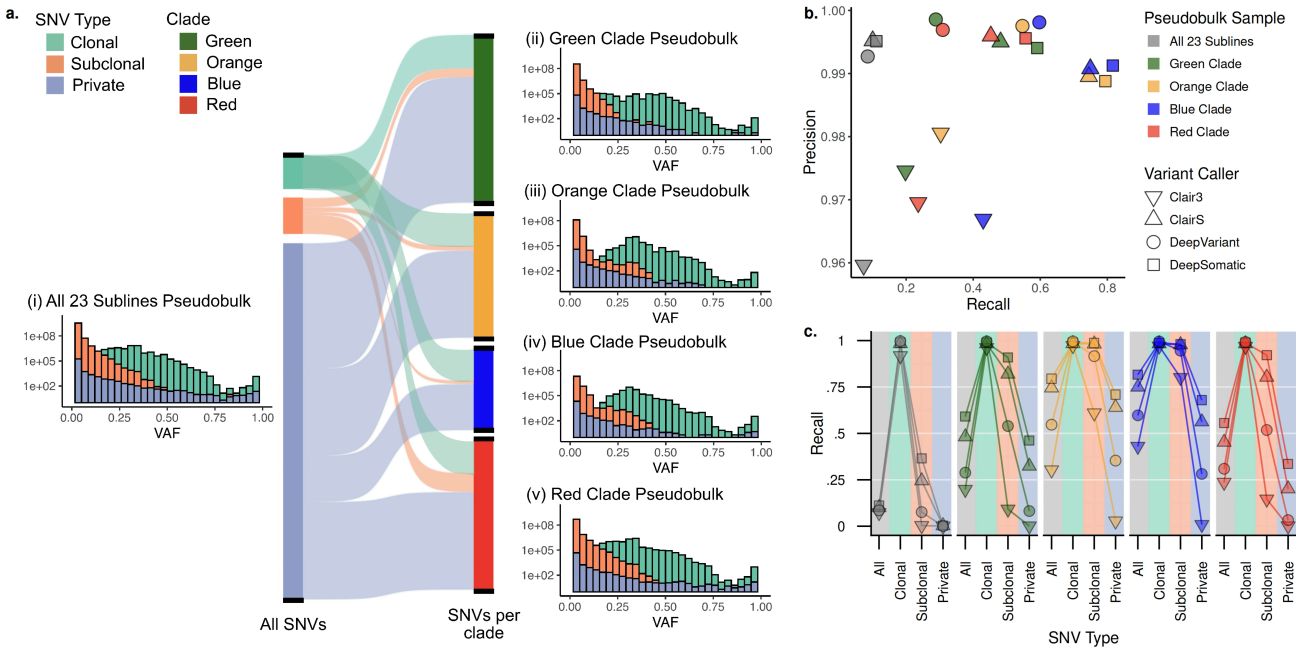


Figure 3.7: **Subline-derived pseudobulk samples enables benchmarking the performances of long-read variant callers at various clonality. a.** Five pseudobulk samples generated by merging reads from (i) all 23 sublines, and sublines in (ii) green clade, (iii) orange clade, (iv) blue clade, and (v) red clade. For each pseudobulk sample, the makeup in terms of clonal, subclonal, and private variants is shown via the Sankey diagram, and the VAF distribution of the corresponding ground truth set of variants is shown respectively. **b.** Precision and recall performance of `Clair3`, `ClairS`, `DeepVariant`, and `DeepSomatic` on the five pseudobulk samples. **c.** Recall performance of variant callers on the five pseudobulk samples, stratified by the clonality of the target variants in the ground truth set.

samples – one of the key challenges in precision oncology. Even though somatic variant callers demonstrated an improvement over germline callers, even the best performers left many variants undetected in the most complex pseudobulk dataset. We publicly release the data and variant calls to encourage the development and continuous improvement of new cancer genomics methods.

3.3 Discussion

In this study, we present an approach to comprehensively map genomic and epigenomic alterations that drive the initiation, progression and clonal differentiation of a mouse melanoma tumor model. We used Nanopore sequencing of clonal sublines – each derived from a single cell of the parental tumor – to systematically catalog somatic SNVs, SVs, CNAs, and DMRs. We developed an algorithmic framework to place all variant types in the context of the sublines phylogeny, representing dynamics of various mutational processes along the evolutionary timescale. This approach provided an highly detailed and comprehensive view of melanoma evolution, illuminating the complex interplay between mutational processes that drive phenotypic differentiation of clonal lineages. We also demonstrated that our data can serve as a useful resource for new cancer genomics developments.

Long-read sequencing of individual sublines enabled precise resolution of parallel evolution events [153], which has been challenging to resolve using conventional approaches. Prior studies have reported examples of parallel CNAs and SVs using bulk sequencing of multiple matching samples, for example, in ovarian [154], breast [155], and osteosarcoma [136] cancers. However, these analyses were typically constrained by heterogeneity of individual samples and limited spatial or temporal resolution, making it difficult to confidently identify independent events affecting the same genomic regions. More recent efforts using single-cell DNA sequencing reported large parallel CNAs in cell

line models [156] and in acute myeloid leukemia with complex karyotype [157]. Yet, due to data sparsity, these approaches often rely on pseudo-bulk inference of structural breakpoints, which prevents assignment of CNAs and SVs to individual cells. In contrast, our approach provides a highly resolved view of parallel clonal evolution thanks to the high purity of sublines, higher number of samples and the advantages of long-read sequencing.

In addition to genetic variation, Nanopore sequencing enabled the detection of lineage-specific DMRs, offering insights into the epigenetic landscape of tumor evolution. DNA methylation has increasingly been recognized as both a marker of clonal lineage [158–161] and a regulatory mechanism driving phenotypic diversity [162–165]. Our work focuses on the latter, extending epigenetic resolution to the subclonal level—beyond traditional comparisons of tumor versus normal tissue or across tumor subtypes.

Recent long-read studies have explored ASM, for example in brain tissue [166] or medulloblastoma [167]. While genome-wide ASM analysis was difficult due to the low heterozygosity of the B2905 mouse model, we illustrated that local analysis was possible in the presence of nearby somatic SNVs; furthermore, the analytical framework presented here could be readily adapted to genome-wide allele-specific analyses, for example in human.

Overall, our study highlights the advantage of long-read sequencing in resolving the the complete variational landscape of tumors, potentially benefiting precision oncology studies, as for various cancers it has been shown that variants other than SNVs, including SVs and DMRs, are important biomarkers for prognosis or therapeutic response [168, 169]. Our analytical framework, including the `TreeHarmonizer` tool, could be adopted to multi-site or longitudinal human tumor datasets. Finally, by publicly releasing all sequencing data and curated variant call sets, we provide a unique benchmarking resource to accelerate the development and validation of new cancer genomics tools.

Chapter 4: Studying melanoma subclones at spatial resolution

4.1 Spatial transcriptomics profiling of the control and anti-CTLA-4 treated tumors.

To obtain the data we use in this study, heterogeneous mixtures of cells from the B2905 cell line of the M4 mouse melanoma model (representative of the RAS-mutant TCGA melanoma, as previously discussed in Chapter 3) were injected into pure bread male mice [116]. Among them, 6 were assigned to the control group, and 11 to the treated group. Mice in the treated group were administered anti-CTLA-4 ICB treatment on Day 13, 16, 20, and 23 after tumor implantation. The tumors were resected at different time points: starting from Day 16 post-implantation for the control group, and starting from Day 27 (after the last ICB treatment in the course) for the treated group. The volume of each tumor was assessed across different time points up until resection, and the treated tumors were further classified into non-responder and responder based on whether its volume has exceeded 250 mm^3 at the time of resection (Figure 4.1). A tissue slice from the fresh frozen sample of each of the 17 tumors was subjected to Visium spatial transcriptomics sequencing.

4.2 Identifying tumor subclones from Smart-seq2 full-length scRNA-seq data of the control and anti-CTLA-4 treated tumors.

In addition to the 17 tumors subjected to ST profiling, 6 similarly derived tumors - 2 from additional mice in the control group, and 4 from additional mice in the treated group - were subjected to Smart-seq2 full-length scRNA-seq [170].

Mutations were profiled from the single cells in either group by a previous study [91], and 6 tu-

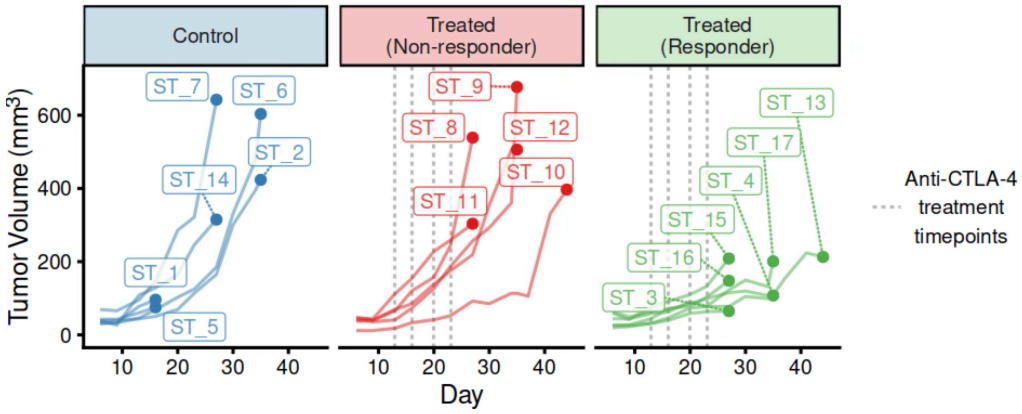


Figure 4.1: **Growth dynamics of the 17 tumors from which the ST slides were derived.** Mice in the treated group received anti-CTLA-4 treatment on Day 13, 16, 20, and 23. The treated tumors were further classified as non-responder and responder by tumor volume at the time of resection.

mor subclones were identified using `SBMClone` [171] (Figure 4.2a.), a stochastic block model that obtains cell-mutation joint clusters that characterize subclonal structures in sparse single-cell data [172]. The presence and absence of mutations in the inferred clusters suggest a clonal relationship shown in Figure 4.2b.

It is worth noting that, the `SBMClone` joint clustering was performed on the set of mutations with coverage in cells from both the control and treated tumors to avoid control-treated batch effect being falsely identified as a meaningful block structure. That said, when identifying the set of mutations specific to each clone, we retain mutations covered among either the control or treated cells but not both, as long as they satisfy the criterion that it is only present among cells in one particular subclone but not others.

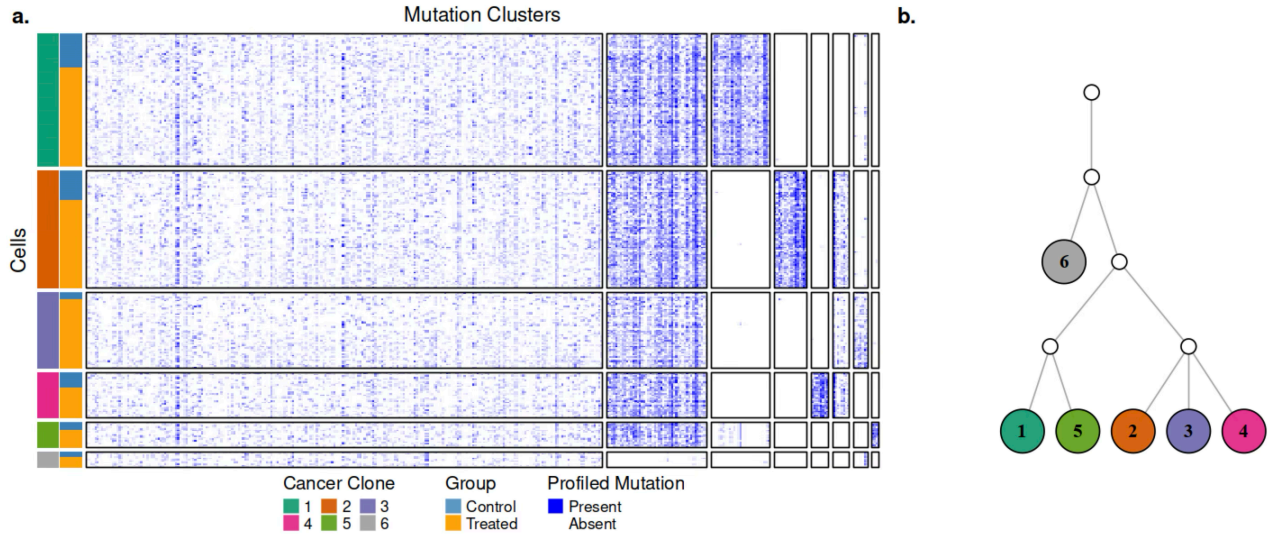


Figure 4.2: **Mutations profiled from Smart-seq2 full-length scRNA-seq data of control and treated tumors reveal 6 major tumor subclones. a.** cell-mutation joint clusters identified by `SBMClone`. **b.** Clonal relationship suggested by the clustering.

4.3 Estimating tumor subclone proportions on the ST slides via deconvolution using SpaCET

Now that we have obtained (i) 17 Visium ST slides from control or ICB-treated tumors and (ii) 6 tumor subclones identified using mutations profiled from Smart-seq2 full-length scRNA-seq data of similarly derived control and treated tumors, we set out to identify the presence of the tumor subclones across spatial coordinates.

Over 50 computational tools have been developed for determining cell identities profiled by scRNA-seq among spatial transcriptomics data, and as Gaspard-Boulinc and colleagues summarized in their review article, the deconvolution tools adopt various types of frameworks: Bayesian, regression, optimal transport, dimensionality reduction, deep-learning, and or ensemble approach [173]. While many such tools can serve as an alternative, for our purpose, we choose a non-negative least squares regression-based tool, `SpaCET` [174], which accommodates the use of a hierarchical reference panel

in deconvolution and produces an accurate estimation of the abundance of cell types in both the major and minor lineages across spatial spots. Our analysis can particularly benefit from such a framework because (i) the tumor clonal structure is intrinsically hierarchical and can fully exploit the unique approach, and (ii) the accurate distinction of immune sublineages offers finer resolution in studying subclone-immune interactions.

Including (i) scRNA-seq data of non-malignant stromal and immune cell types in the mouse skin profiled in previous works by Davidson and colleagues [175] and Hunter and colleagues [176], as well as (ii) scRNA-seq data of our 6 identified tumor subclones in the reference panel, we deconvolved the 17 ST slides using `SpaCET` [174]. For the full deconvolution result for each of the 17 ST slides, see Supplementary Section C.1.

4.4 Validation of SpaCET-deconvolved tumor subclone proportions

Prior to proceeding with analysis of the `SpaCET` deconvolution results, we first validated the tumor subclone proportions. While the number of mutations one is able to profile directly from Visium data is limited, as the probablility of detection is inversely related to the mutation's distance to the 3' end of the transcript (measured as expected distance given gene isoform abundance estimated by `RSEM` [177] from the SmartSeq2 data; Figure 4.3a.), a limited number of them are detectable given sufficient expression of the associated gene at each spatial coordinate. We ask the question: are the ST spots with mutations detected for a particular subclone consistent with where the subclones are according to SpaCET?

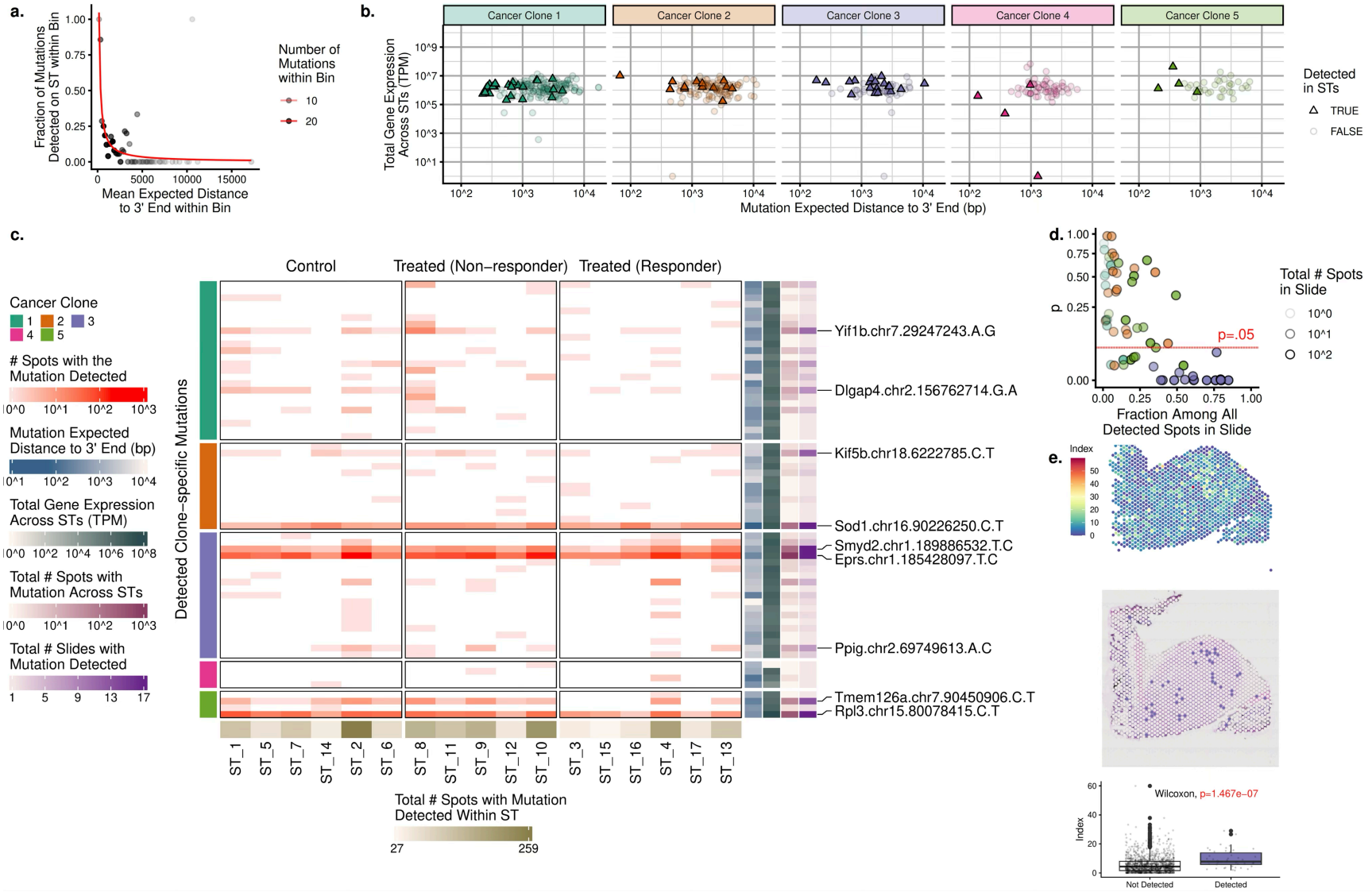


Figure 4.3: **Validation of tumor subclone proportions estimated by SpaCET using subclone-specific mutations detected on ST slides. a.** Among mutations profiled from Smart-seq2 data, the probability of a mutation being also detected on an ST slide is inversely related to its distance from the 3' end of a transcript. **b.** Distribution of detected subclone-specific mutations in terms of its distance from 3' end and the total expression of the associated gene across ST slides. Note that as Cancer Clone 6 harbors only truncal mutations, it lacks marker mutations specific to itself and therefore it is not considered.**c.** Distribution of detectable subclone-specific mutations across the 17 ST slides. **d.** Distribution of p values from one-tailed Wilcoxon rank sum tests that the ST spots in an ST slide with detected clone-specific mutation has higher detectability index for that clone than those without. **e.** An example: (top) the distribution of the detectability index for Cancer Clone 3 across ST_1, (middle) spots on ST_1 with Cancer Clone 3 marker mutations detected, and (bottom) one-tailed Wilcoxon rank sum test results.

As Cancer Clone 6 mainly harbor truncal mutations (Figure 4.2), there are no subclone-specific mutations to distinguish it from the others, and it is hence not considered in this analysis. The chance of detecting a subclone-specific mutation also varies across the 5 remaining subclones, depending on (i) the distance each of such mutation is from 3' end, and (ii) the overall expression levels of the associated genes. Indeed, the clone-specific mutations detected in any ST spot are typically closer to 3' end and have overall sufficient expression across ST slides (Figure 4.3b.). In particular, the few mutations specific to Cancer Clone 4 that are close to the 3' end have lower levels of overall expression compared to the other subclones, consistent with the observation that detecting a Cancer Clone 4-specific mutation is more rare (Figure 4.3b.,c.).

For each of the subclone-specific mutations at all detected across the ST slides, the number of detection across ST slides vary (Figure 4.3c.). A vast majority of the mutations was detected in very few total spots across very few slides; however, with the exception of Cancer Clone 4, each of the cancer clones has a couple of mutations that are both more frequently ($\geq$ 10 total spots) and consistently ($>$ 5 ST slides) detected across ST slides (labeled mutations on the right, Figure 4.3). Leveraging those marker mutations, we derive an index measuring the expectation of detecting any marker mutation m_c with an expected distance d_{m_c} bp from 3' end) specific to subclone c at an ST spot p_s in ST slide s given (i) the SpaCET-estimated subclone proportion $f_{s,c}$ at the spot, (ii) the expression (TPM) level of the associated gene e_{g_m}at the spot, and (iii) the discrete read length distribution $L_s(l)$ of the ST slide, as follows:

$$\text{Index}_{p_s,c} = \Sigma_{m_c} \left[f_{s,c} \cdot e_{g_m} \cdot \Sigma_{l \geq d_{m_c}} L_s(l) \right]$$

This model follows the simple intuition that a clone-specific mutation is more likely to be de-

tected in a spot that is (i) high in the clone proportion, (ii) with the associated gene highly expressed, (iii) and the read lengths distribution covers the distance of the mutation from 3' end.

For each of the 4 tumor subclones with subclone-specific mutation markers c in an ST slide s, we compute $\text{Index}_{p_s,c}$ across all spots p_s. Then, we examined whether the index values at spots with a subclone-specific mutation marker detected are significantly higher than those without by performing a one-tailed Wilcoxon rank-sum test. An example of results from Cancer Clone 3 in ST_1 is shown in Figure 4.3e., where the spots with Cancer Clone 3-specific mutation markers exhibit significantly higher index values. See Supplementary Section C.2 for the full result across all 17 ST slides.

Examining the p values from the Wilcoxon rank-sum tests, we see that, the differences are largely significant for clones with sufficiently high number of spots with clone-specific mutations detected (Figure 4.3d.). When the number of spots with detected subclone-specific mutation is low, the low sampling size unfortunately lends little statistical power. That said, the results from clones in ST slides with high number of subclone-specific mutations detected suggest confidence in the subclonal proportions obtained by SpaCET.

We would like to acknowledge that the validation of deconvolution results from empirical ST data is difficult with the lack of ground truth. In deriving the model above, we have found various aspects to present challenges. First, we make the assumption that the isoform abundance estimated from the Smart-seq2 data is accurate, and the expected distance measure captures some information about how likely a mutation is at all detectable from the ST data. While the overall trend is roughly as expected (Figure 4.3a.), we make the inherent assumption that the isoform relative abundance is similar among the ST slides, which is more difficult to verify. Furthermore, the data used in this study were derived from pure bred mice, and therefore we cannot achieve allele-specific resolution. Ideally, we would derive an explicit model for the observed read counts mapped to mutation loci across ST

spots; however, as we lack allele-specific information, reference and altered read counts cannot be fully modeled. Further validation using subclone-specific copy number information profiled from Smart-seq2 scRNA-seq and ST data may be attempted, with the understanding that expression-based copy number calls are noisy, and is only largely related to and not a complete reflection of the underlying genomic copy number.

4.5 Analysis of SpaCET-deconvolved tumor subclone proportions

Having validated the tumor subclone proportions estimated by SpaCET [174], we examine how such results may provide insight on the difference in the responses to anti-CTLA-4 treatment from the 17 tumors.

4.5.1 SpaCET-deconvolved proportions reveal differences in subclone growth dynamics between treatment responders and non-responders

Immediately, summing over the spatial axis of the deconvolution results, the aggregate proportions of each major cell type, as well as that for each of the 6 tumor subclones, already provide ample information regarding the growth dynamics of the tumors at subclone resolution. While tumor volume measured on the day of resection is one important feature of the tumor, as the various ST samples were collected on different days, we calibrate the growth phenotypes of various slides by measuring their tumor growth rates.

While we do not have information about the initial cell type compositions for every tumor, as ST_1 and ST_5 were control tumors collected early on at 16 days post-implantation (Figure 4.1) with similar cell type and tumor subclone composition (Figure 4.4a.), assuming the initial cell type compo-

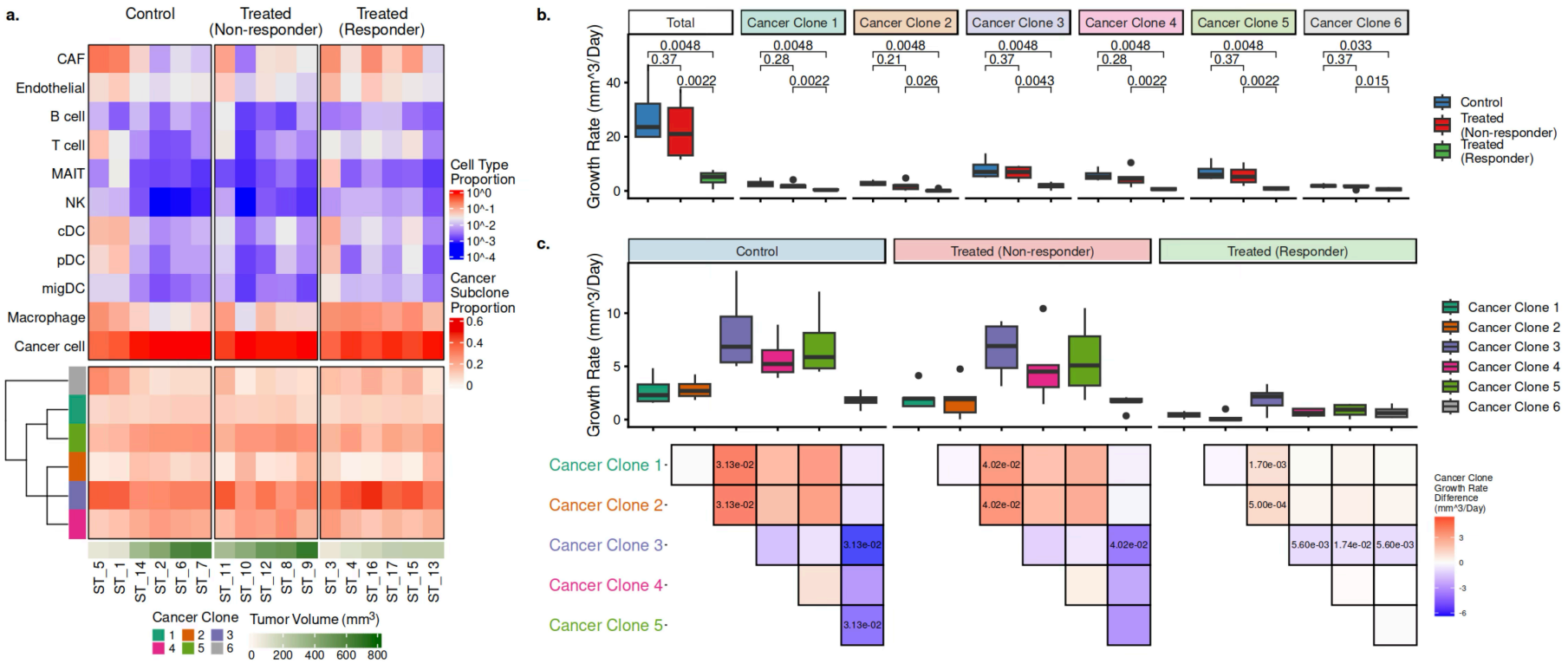


Figure 4.4: **Deconvolved subclone proportions inform differences in tumor growth dynamics. a.** Proportions of major stromal, immune cell types, and cancer cells estimated by SpaCET [174] across 17 ST slides. A break down of the proportions of the 6 tumor subclones relative to the cancer cell fraction is further displayed. **b.** Growth rates (mm^3/day) of the whole tumor, as well as that broken down by tumor subclones, across the control, treated (non-responder), and treated (responder) tumors. One-tailed Wilcoxon rank-sum test p-values for pairwise comparisons are displayed. **c.** Growth rates (mm^3/day) of tumor subclones across control, treated (non-responder), and treated (responder) conditions. Unpaired t test p-values - adjusted via the Benjamini-Hochberg procedure - for pairwise comparisons are displayed, the differences in the mean growth rates of the row subclone and the column subclone are displayed beneath, and statistically significant (adjusted $p < .05$) pairs are marked by their respective adjusted p values.

sitions in all 17 tumors were similar, we use the average of the ST_1 and ST_5 cell type composition as an estimate for that of each of the 15 other ST tumors at Day 16. Then, the total cancer growth rate for an ST tumor can be computed by first computing the difference in cancer volume (i.e. tumor volume $\times$ cancer cell proportion) from Day 16 until tumor resection, then divide by the number of days that have elapsed. Subclone-specific growth rates can be similarly obtained via subclone volume (i.e. tumor volume $\times$ subclone proportion relative to whole tumor). For the purpose of this analysis, ST_1 and ST_5 are excluded.

In terms of overall cancer cell volume growth rate, responder tumors exhibit significantly lower growth rates than both control and non-responder tumors (one-tailed Wilcoxon rank-sum test, p=.0022), while there is no statistically significant difference between control and non-responder tumor groups (Figure 4.4b.). This trend holds true when we compared the growth rates of individual tumor subclones among the 3 treatment conditions (Figure 4.4b.).

Then, we asked whether, within each treatment condition, whether there are meaningful differences in the growth rates of the 6 cancer subclones. To that end, for each treatment condition, the significance of the difference in mean growth rates of any pair of tumor subclones was assessed via an independent samples t test. While Cancer Clone 3 has significantly higher growth rates than others among the responder tumors, its growth rates are not significantly different from those of Cancer Clone 4 and Cancer Clone 5, whose mean growth rates reach a similarly high level among control and non-responding tumors (Figure 4.4c.).

Chapter 5: Concluding remarks

Our understanding of how heterogeneous subpopulations arise in a tumor has come a long way since Nowell first formalized the foundation of theory of tumor evolution in 1976 [4], largely owing to the advances in sequencing technologies, which have allowed us to study somatic alterations at the base pair level. Novel sequencing technologies afford us with more, and new types of information about tumor samples, all the while with decreasing sequencing cost. This thesis has focused on how we leveraged the unique advantages of 3 sequencing technologies developed in the last decade, scBS-seq, ONT long-read sequencing, and 10x Visium spatial transcriptomics sequencing, respectively, to obtain novel insights into tumor evolution.

Conventionally, phylogenetic analyses are performed to elucidate key somatic alteration events that have given rise to tumor subclones. Phylogenetic reconstruction methods have been designed for genomic alterations profiled from multi-sampled bulk tumors [12–22,178–180], single-cell sequencing-profiled tumor samples [31–36], or a combination of both [181, 182]. While single-cell data can resolve some ambiguities in possible phylogenies inferred from bulk samples, shallow coverage and allele drop-out make inferring single-cell phylogenied using SNVs challenging. As there is increasing evidence that epigenomic changes, such as CpG methylation, are also partially heritable and may be used in phylogenetic reconstruction [45, 158, 159], and its higher mutation frequency helps overcome some challenges sparsity presents in single-cell data [40, 41], we introduced `Sgootr`, a computational tool that leverages scBS-seq data to jointly construct a tumor phylogeny and select a set of lineage-informative CpG sites. While prior works have constructed single-cell methylation phylogenies using maximum-likelihood-based methods such as `IQ-TREE`, `Sgootr` as a distance-based tool offers an alternative that obtains results as accurate orders-of-magnitude faster. This work was

accepted at 27th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2023), and appeared in the July issue of Genome Research the same year [160]. Since the publication of our tool, Chen and colleagues have introduced `MethylTree`, a computational tool to perform high-resolution lineage-tracing using epimutations in mice and humans [161], demonstrating continued interests in the community in leveraging (partially) heritable non-genetic as signals in phylogenetic reconstruction. As multi-omic sequencing technologies that can simultaneously profile genetic and non-genetic somatic alterations at increasingly finer resolution continue to mature, novel methods that involve both genetic and non-genetic somatic alterations simultaneously in phylogeny reconstruction will need to be developed to properly leverage both types of signals in refining our understanding of their roles in tumor evolution.

While it is not yet feasible to jointly leverage genetic and non-genetic signals in phylogeny reconstruction, one may already overlay information from one omic on a phylogenetic tree constructed on the signal from a different omic on the same set of samples. For example, Gaiti and colleagues performed joint RNA (Smart-seq2) and DNA methylation (MscRRBS) profiling on CLL and normal B cells, and the expression information available for the cells in the methylation-based phylogeny allows one to examine transcriptomic dynamics before and 1 month after ibrutinib treatment [158]. Using a similar technique, Chaligne and colleagues jointly performed methylation and expression profiling of GBM and IDH-MUT glioma samples, and quantified the heritability of transcriptomic cell states along the methylation-based phylogenies, which reveal differences in rates of cell differentiation and self-renewal between GBM and IDH-MUT samples [159]. Being able to resolve the timing of genetic and non-genetic somatic changes allows one to examine the evolution of a tumor from a more holistic standpoint, as both types of heritable changes result in changes in clonal fitness [38]. Taking a further proof-of-concept step towards this goal, we profiled 23 single-cell derived sublines of a

mouse melanoma model using ONT long-read sequencing and simultaneously resolved detected SVs, CNAs, as well as methylation changes along an SNV-based phylogeny [183]. The experimental design allowed us to profile genomic and epigenomic information from cell populations from the mouse model, at single-cell resolution and bulk coverage. Through analyses uniquely enabled by this data set, we were able to elucidate the interplay between genomic and epigenomic somatic alterations in shaping phenotypically distinct subclones during the evolution of this melanoma. This work was presented at Nanopore Community Meeting 2023, and the manuscript is in revision during the writing of this thesis. New collaborations are currently underway, aiming to extend the analyses in this work to human data. One unfortunate limitation of the current study is the infeasibility of high-quality haplotype phasing due to the use of pure-bred mice. With human data, one may now detect the variants at an allele-specific resolution, which will be instrumental in determining the associations between different types of variants.

As we make strides towards a holistic characterization of the heterogeneous populations within a tumor, it is important to be reminded that tumors do not grow in a vacuum, and how the subpopulations interact with the immune microenvironment may have critical implications in the phenotype of the tumor and how well it may respond to ICB therapies. As spatial transcriptomics sequencing technologies continue to mature, there have been efforts identifying tumor subclones on the spatial slides. One approach is to identify copy number clones directly on the spatial slides [184, 185], and others have attempted to model the read counts at somatic mutation loci identified from matched multi-sampled bulk data on the spatial slide while incorporating allele-specific copy number information [186]. While the former is only able to identify large events in aneuploid cancers, the latter heavily relies on the correct subclone identification from multi-sampled bulk data, which oftentimes can be difficult. As such, we derive a different strategy. Using Smart-seq2 full length scRNA-seq data, one can detect major

subclones using detected mutations, and one can identify the subclones on matched spatial slides via deconvolution using subclone-specific expression profiles. Along with the deconvolved immune cell proportions across spatial spots, one can examine the spatial correlations between phenotypically distinct subclones and, for example, T cell activities, revealing their differential response to the immune microenvironment. The association of tumor subclones with their distinct differentiation phenotypes and their response to their immune microenvironment provide pieces of evidence that, the differences in the spatial distribution of differentiation phenotypes within a tumor is not necessarily the result of adaptation as previously suggested [187], but may be due to the selection pressure from the immune microenvironment. The manuscript for this work is currently being prepared for journal submission. In future work, it would be important to validate our observations in human melanoma spatial transcriptomics data sets; moreover, understanding the underlying mechanisms for the subclonal differential response to immunity will be key to better treatments for such heterogeneous tumors.

In summary, new sequencing technologies have brought us new capabilities and have allowed us to ask new and interesting questions about tumor evolution. A combination of appropriately adapting existing methods and developing new methods will be needed to maximize the information we can learn from new types of data. The development of new sequencing technologies, new computational tools, creative data analyses, as well as rigorous biological validations will go hand-in-hand in advancing our holistic understanding of tumor evolution.

Appendix A: Supplemental Material for Chapter 2

A.1 Biclustering: benchmarking on simulated data and a case study on CRC01

As can be seen in Supplemental Figure A.1 (left plot) on the CRC patient CRC01 [43], the distribution of the number of sites covered by at least one read across cells is bimodal and a relatively small fraction of cells have very few sites covered. In contrast, the distribution of the number of cells with at least one read covering a site (right plot) is unimodal across the sites and only a relatively small fraction of sites are covered in a large number of cells. Provided the presence of those well-covered sites are distributed over the cells independently and identically (i.e. in i.i.d. fashion), the heuristic filtering described in the Methods section would work well. However, if the distribution is not i.i.d., fixing the output size, the submatrix obtained by ILP-based biclustering formulation would have higher density, improving the overall reliability of the tree constructed by `Sgootr`. Below we explore how the running time and memory usage of the ILP formulation vary with key problem parameters, and identify under what parameter settings of the problem instance it remains feasible.

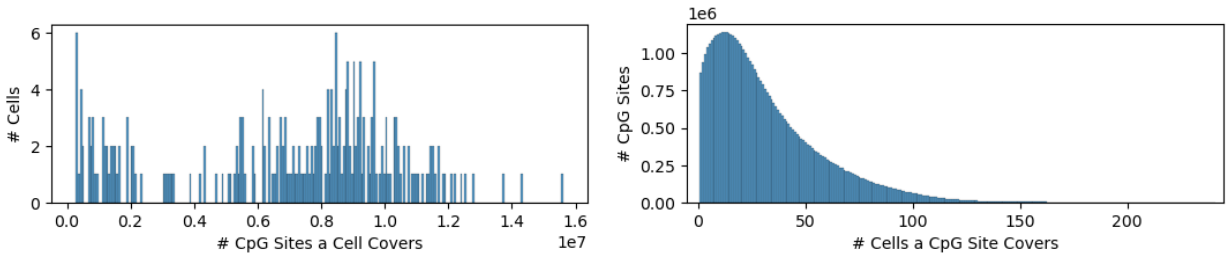


Figure A.1: **Patient CRC01 cell (left) and site (right) coverage level distribution.**

A.1.1 Running time and memory benchmark on simulated data

We first demonstrate how the expected running time (in other words, time the Gurobi solver takes to obtain an optimal solution) and memory behaviors of the Gurobi solver on our formulation vary as a function of the input size.

For the problem at hand, the input size is parameterized by v, the number of cells, and u, the number of CpG sites in the input matrix M. The number of variables (N_{var}), the number of constraints (N_{constr}), and the number of non-zero entries in the constraint matrix (N_{nz}) are some features of the Gurobi representation of the linear program that may influence its memory usage and running time; note that all three features are $\mathcal{O}(vu)$ (Supplemental Table A.1). Also note that Gurobi uses the sparse representation for the constraint matrix; as a result, even though the naive representation of the constraint matrix requires $\mathcal{O}(v^2u^2)$ space, the size of its representation by Gurobi is proportional to N_{nz}, which is $\mathcal{O}(vu)$.

Below we proceed to simulate datasets with varying sizes and study their running time and memory usage.

Feature	$f(v, u)$	**Asymptotics**
# Variables (N_{var})	$v \times u + v + u$	$\mathcal{O}(vu)$
# Constraints (N_{constr})	$3 \times v \times u + 2$	$\mathcal{O}(vu)$
Size of constraint matrix	$N_{var} \times N_{constr}$	$\mathcal{O}(v^2u^2)$
# Non-zeros (N_{nz})	$7 \times v \times u + v + u$	$\mathcal{O}(vu)$
Density of constraint matrix	$N_{nz}/(N_{var} \times N_{constr})$	$\mathcal{O}((vu)^{-1})$

Table A.1: **Feature sizes of the biclustering formulation model as a function of the number of cells, v, and the number of CpG sites, u in the input.**

A.1.1.1 Simulation details

For simplicity, we set $v = u$ in our simulations. For each (square) input binary matrix M_{sim}, we set a "dense" submatrix of size $.7v \times .7u$ such that each entry in the submatrix is set to 1 independently with probability q, while each entry outside of the submatrix is set to 1 with probability $1 - q$. An entry set to 1 indicates that the corresponding cell has at least one read covering the corresponding site. We vary the simulation parameters as follows:

$$v = u = \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$$

$$q = \{1, .9, .8, .7\}$$

For $q = 1$, we have a $.7v \times .7u$ submatrix with all 1's and 0's everywhere else. As $q > .5$ decreases, while we still have a dense $.7v \times .7u$ submatrix, its density difference with the rest of the matrix is not as pronounced, making it more difficult to achieve our objective. With each set of parameters (v, u, q), 5 matrices are simulated (for investigating the stochasticity introduced by $q < 1$), resulting in a total of $10 \times 4 \times 5 = 200$ simulated matrices.

A.1.1.2 Results

We perform biclustering on each of the 200 simulated binary matrices with $\alpha = \beta =$ {.6, .7, .8}, resulting in 600 trials total. Recall that α and β are respectively the proportion of the rows and columns of the densest submatrix we aim to identify. The trials are performed serially on the same machine with a single thread. As the dense submatrix in each simulated matrix has dimensions $.7v \times .7u$, performing biclustering with $\alpha = \beta = .6$ and $\alpha = \beta = .8$ will respectively correspond to an underestimation and overestimation of the two parameters. We profile (i) running time and (ii) peak memory usage for all

trials.

As a general pattern, both wall-clock running time (Supplemental Figure A.2A., where the y-axis is $\log_2$-transformed to better visualize the differences) and peak memory usage (Supplemental Figure A.2B.) grow linearly with $v \times u$. For any given input size, the running time increases with decreasing value of q, as the objective becomes more difficult to achieve (as demonstrated within each column in Supplemental Figure A.2A.). Notably, the running time also seems to be impacted by the choice of α and β: while both underestimation and overestimation lead to an increase of running time possibly due to the existence of several equally good solutions, underestimation seems to lead to a more substantial penalty (as can be observed across columns corresponding to the same input size in Supplemental Figure A.2A.). The peak memory usage seems to be similar for a given size input across choices of α and β, though an underestimation seems to occasionally lead to considerably higher usage and hence higher variance across trials (as can be observed across columns corresponding to the same input size in Supplemental Figure A.2B.).

To allow extrapolation of peak memory usage for any given input size, we perform linear regression using the ordinary least squares method on the data from the 600 trials, with $v \times u$ as the independent variable and peak memory usage in Mb as the dependent variable, obtaining a y-intercept of 65.59 and slope of $.0089$.

We do not extrapolate running time in the same manner, as it varies drastically with problem difficulty, which we proxy with q in our simulation study, but is hard to estimate in real data. Running time will also depend on the specific optimization strategies employed by Gurobi, a proprietary software, that may depend on other features of the problem instance, such as the density of the constraint matrix. Furthermore, one can always specify a time limit for a Gurobi run and obtain an intermediate output and its associated optimality gap, meanwhile Gurobi fails without output if memory limit is

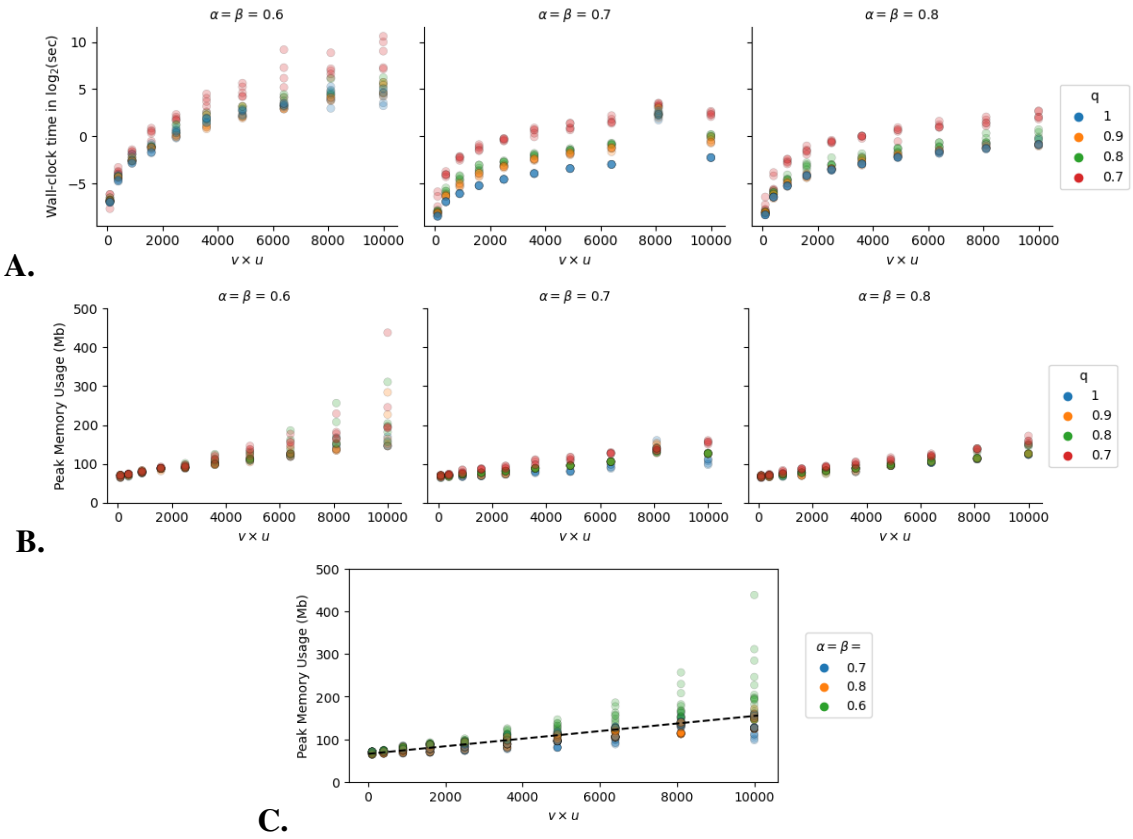


Figure A.2: **Running time and memory behavior on 200 simulated square binary matrices. A.** Wall-clock time in $\log_2$ of seconds. Measures were $\log_2$-transformed to better visualize the difference. **B.** Peak Memory consumption in Mb. **C.** Linear regression on peak memory usage across all 600 trials. Dashed line describes $y = .0089x + 65.69$.

reached; therefore, memory is more likely to present a bottleneck for large-scale problems than time in practice.

Results from simulated data lead us to two key conclusions: (i) memory usage may be a critical limiting factor to the feasibility on large input, and (ii) well-chosen α and β values will not only yield a more informative submatrix for the remainder of the `Sgootr` workflow, they will also (typically) lead to faster converge than ill-chosen parameters.

A.1.2 Biclustering case study on CRC01 from the Bian cohort

Next, we investigate the performance of our biclustering formulation on the CRC01 dataset from the Bian cohort as we vary α and β values.

A.1.2.1 Data descriptions

The dataset has in total 409 single cells with scBS-seq data from patient CRC01, among which 241 also has matching scRNA-seq data. As we will call copy number from scRNA-seq data to input to `Sgootr`, we only include those 241 cells with both scRNA-seq and scBS-seq data for the biclustering procedure. Collectively, the 241 cells cover 52,343,774 unique autosomal CpG sites. Supplemental Figure A.1 shows per-cell CpG site coverage distribution and per-site cell coverage distribution. The induced $241 \times 52,343,774$ binary coverage matrix has a density of 13.21%.

We aim for the resulting submatrix to have a density of at least 66%. The intuition here is that `Sgootr` estimates the distance between any pair of cells using the set of CpG sites which are covered by at least one read in each cell; provided $2/3$ of the picked sites are covered by a read in each cell, the distance between any pair of cells will be based on at least $1/3$ of the picked sites, i.e. at least half of the sites covered by a read in each cell. This ensures that the number of CpG sites employed in distance estimations involving one particular cell do not vary by more than a factor of 2. While we do not ensure that each row (or for that matter each column) has a density of at least $2/3$, an overall density of $2/3$ (or $\sim 66\%$) in the submatrix identified by our ILP formulation should provide a reasonable distance estimate between cell pairs.

A.1.2.2 Estimating α and β values

To estimate α and β, we generate 10 disjoint subsamples of the dataset, each containing 10% of the 241 cells and 0.1% of the 52,343,774 CpG sites. For each subsample, we generate the corresponding $24 \times 52,343$ binary matrix (Supplemental Table A.2), on which we perform biclustering with all possible combinations of $\alpha = \{.4, .5, .6, .7, .8, .9\}$ and $\beta = \{.001, .005, .01, .05, .1, .15\}$ in parallel, resulting in $10 \times 6 \times 6 = 360$ runs of biclustering in total. The choices for the ranges of α and β values are informed by the density of the original matrix, the desired density of roughly 66%, and the cell and site coverage distributions (Supplemental Figure A.1). We will determine α and β based on i) time-to-convergence, ii) optimality gap at time limit, and iii) density of resulting submatrix in the subsampled trials.

A.1.2.3 Results

A total of 360 runs of biclustering were performed across the 10 subsamples over a range of α and β values. Each run was given 4 threads and 20 hours. Runs with $\alpha = \{.8, .9\}$ were able to achieve negligible optimality gap in a matter of several hours, and smaller values of α leads to substantially higher time-to-convergence and optimality gap within the allotted time (Supplemental Figures A.3A.,B.). Fixing $\alpha = \{.8, .9\}$, looking at the other parameter, runs with $\beta = \{.001, .005\}$ all solved to convergence with no optimality gap (Supplemental Figures A.3A.,B.). With $\alpha = \{.8, .9\}$ and $\beta = \{.001, .005\}$, the resulting submatrices have all achieved a density within a neighborhood of 66%. We choose $\alpha = .8$ and $\beta = .005$ to i) avoid over-pruning, and ii) considering variance in the different subsamples, have some confidence that results from the full original binary matrix will be at least 66% dense. (Supplemental Figure A.3C.).

Original $(241 \times 52, 343, 774)$	**Subsamples** $(24 \times 52, 343)$											
	0	1	2	3	4	5	6	7	8	9	**mean**	**std**
13.21%	14.53%	12.22%	14.58%	11.90%	12.17%	12.74%	12.09%	15.11%	16.31%	10.34%	**13.20%**	**1.74%**

Table A.2: **Densities of the original binary matrix and the subsampled matrices.**

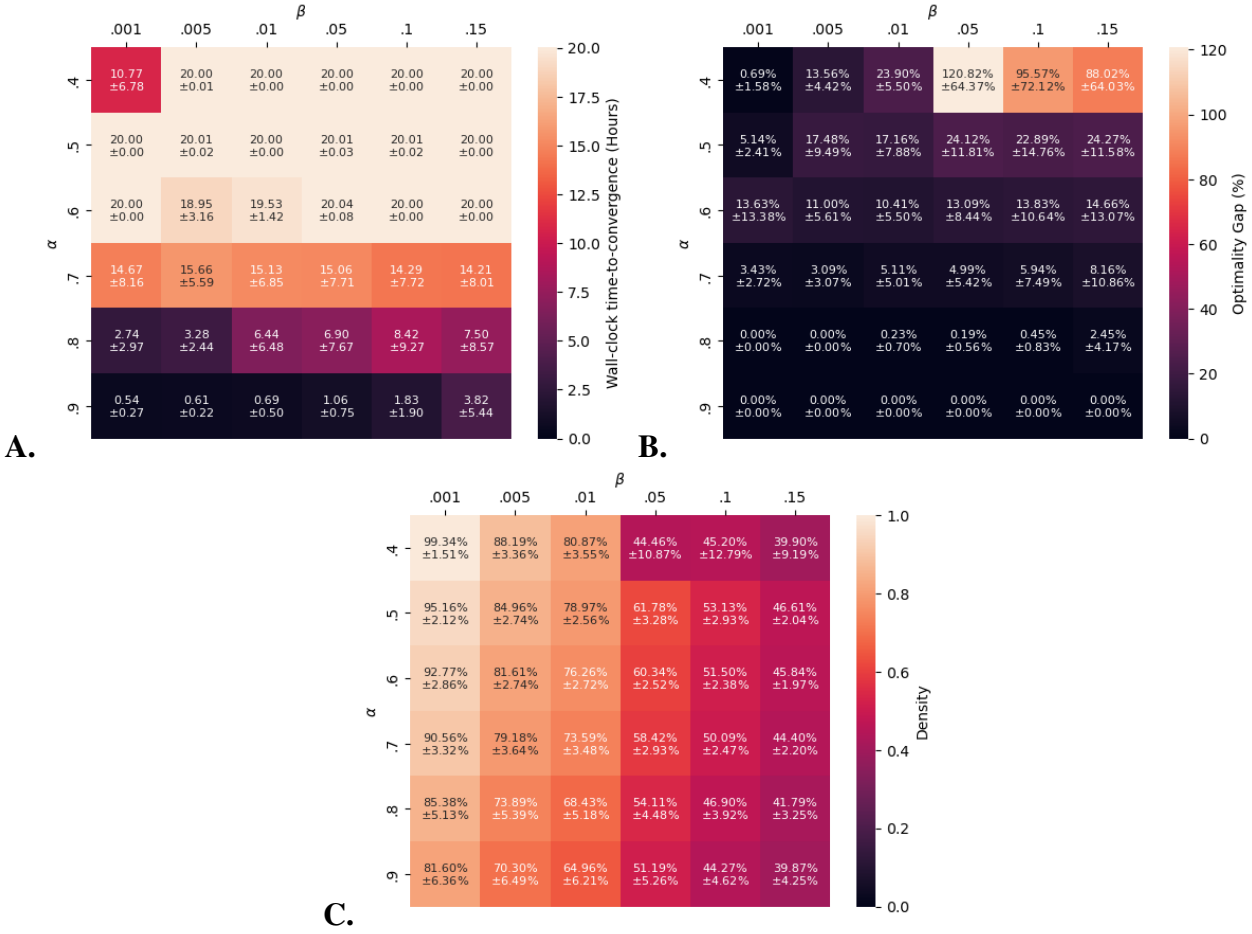


Figure A.3: **Mean and standard deviation of A. wall clock time-to-convergence in hours, B. optimality gap at 20 hours, and C. resulting submatrix density across the 10 subsamples, over a range of (α, β) value pairs.** Note that we set the time limit for the biclustering program to 20 hours, and therefore wall clock time-to-convergence is at most 20 hours.

The next step would be running biclustering on the original $241 \times 52,343,774$ binary matrix with the chosen parameters $\alpha = .8$ and $\beta = .005$. However, for input of this scale, the peak memory usage is predicted to be 112 Tb according to our linear regression result from simulated data, and is not feasible on the resources available to us and likely most researchers. Therefore, we opted for performing heuristic filtering in generating the results in the main text. Note that the submatrix we obtained with heuristic filtering on CRC01 is of a similar scale as that from performing biclustering on a $241 \times 52,343,774$ matrix with $\alpha = .8$ and $\beta = .005$.

For future iterations of the tool, we aim to optimize the implementation of the biclustering procedure to improve memory efficiency, and a combination of heuristic filtering and biclustering may be needed for best solving the problem at hand. In particular we aim to utilize the ILP formulation on the entire set of cells but only a subset of sites so as to stay within the limits of Gurobi's ability to solve the problem. Given the subset of cells identified by the ILP formulation, we will then be able to employ the heuristic filtering to identify the subset of sites with high coverage on these cells.

A.2 Benchmarking `Sgootr` on simulated data

In this section we evaluate the performance of distinct components of `Sgootr` on simulated data, in comparison to baseline approaches. All `Sgootr`-related parameter settings will be the default configuration. Furthermore, for simplicity, we will assume copy number = 2 in our simulations.

A.2.1 Expected distance calculation on simulated data

First we focus on our formulation for the expected distances between cell pairs and evaluate its performance against a simple baseline approach that was utilized in earlier studies [42, 44] which ignores sites that are potentially heterozygously methylated.

A.2.1.1 Simulation details

Our simulations generate pairs of diploid cells with length-normalized ground truth distance of $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$. For each value of π we simulate 50 cell pairs, producing in total $9 \times 50 = 450$ simulated cell pairs. Each cell starts with 2 (CpG) alleles of 1000 unmethylated CpG sites. For any simulated pair, given π, we uniformly sample the number of sites where one cell is

homozygously methylated but the other is homozygously unmethylated (each contributing 1 to the total distance) and the number of sites where one cell is homozygous but the other is heterozygously methylated (contributing .5 to the total distance) from all possible combinations. For example, with $\pi = .6$, by pigeonhole principle we have at least 200 and at most 600 CpG sites each of which contributing 1 to the total distance. After sampling those sites, the remainder of the ground truth distance need to be contributed by sites where one cell is heterozygous but the other is homozygous. After sampling this second set of sites, the ground truth distance is already achieved and the remaining sites will be of the same methylation status across the two cells.

Next, for each of the 450 cell pairs with known normalized ground truth distances, we simulate methylated and unmethylated read count vectors for either cell (i) for a range of possible depth of coverage to study how that may impact the measurements, and (ii) using the empirical per-site read coverage distributions from patient CRC01, the key patient we analyze in the main text. For (i), specifically, we model the number of reads covering a CpG site by a zero-truncated Poisson distribution (ZTPois) parameterized by $\lambda = \{1, 2, 3, 4, 5, 6\}$ (Supplemental Figure A.4A.). For (ii), we obtain the empirical per-site read coverage from patient CRC01 after first-pass filtering (Supplemental Figure A.4B.). The empirical distribution is truncated at 20, as probability substantially diminishes for larger values. For each simulated cell pair, we simulate 5 sets of methylated and unmethylated read count vectors for every depth of coverage distribution, obtaining $450 \times 7 \times 5 = 15,750$ sets of read count vectors.

We compare the distances between the read count vectors of a cell pair obtained by (i) a baseline distance measure, which we will describe in detail in the following section, and (ii) our expected distance formulation with default parameters. This comparison will help us better understand (a) how the depth of coverage affects the two distance measures, and (b) given the empirical read coverage

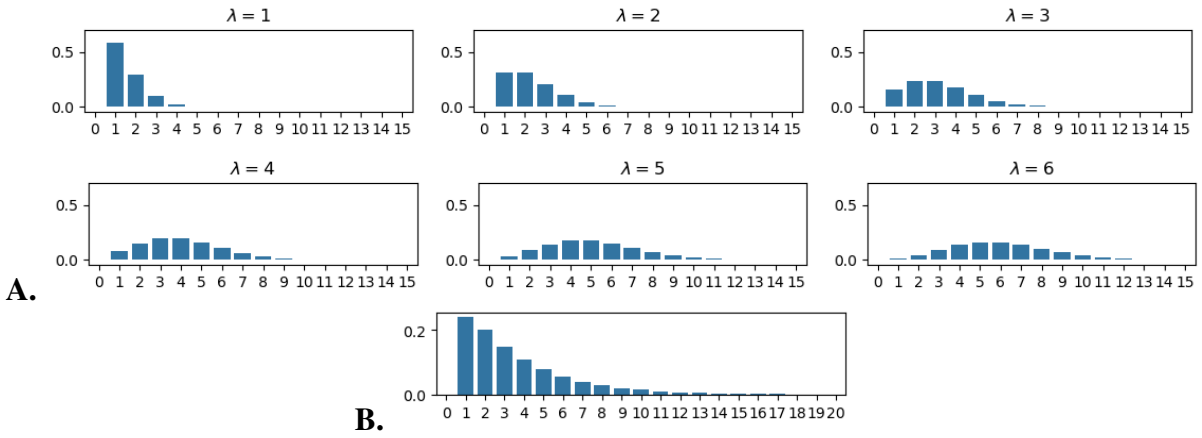


Figure A.4: **Probability mass functions of A. ZTPois(λ), where $\lambda = \{1, 2, 3, 4, 5, 6\}$ and B. the empirical per-site read coverage distribution from patient CRC01.**

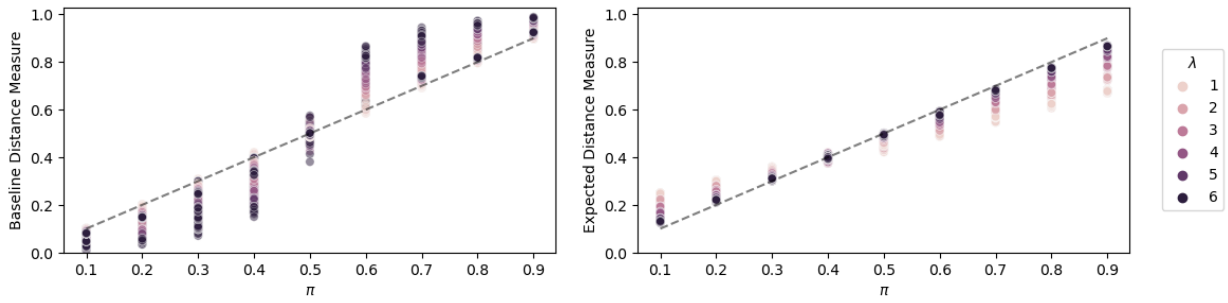
distribution, which measure better models the ground truth distance.

A.2.1.2 Results

The baseline distance measure we leverage in the comparison combines elements commonly used in prior studies [42, 44, 188]. For each CpG site in a single cell, this approach first calls its status as methylated or unmethylated based on the 90% rule. As a result, each cell will have a binary vector (likely with some missing entries due to failed binarization) indicating whether each site is methylated. Then, for each pair of cells, this approach computes the Hamming distance between the respective binary vectors over all sites at which both have a binarized status, then normalizes the distance by the number of such sites. This distance measure essentially excludes heterozygosity from consideration.

We first examine how the depth of coverage (the distribution of which is shown in Supplemental Figure A.4A.) affects the baseline in comparison to our expected distance measure. For each of the $450 \times 6 \times 5 = 13,500$ sets of read count vectors sampled according to ZTPois(λ), we measure both distances. The baseline distance measure tend to underestimate the ground truth distance when $\pi < .5$ and overestimate it when $\pi > .5$ (Supplemental Figure A.5A., left). This is because $\pi =$

A.



B.

λ	1	2	3	4	5	6
R^2(π, Baseline Distance Measure)	.998	.994	.987	.978	.970	.962
R^2(π, Expected Distance Measure)	.997	.997	.997	.998	.999	.999

Figure A.5: **Baseline distance measure and expected distance measure on** $13,500$ **sets of read count vectors, simulated from cell pairs with known normalized ground truth distance** $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$**, with level of read coverage depth modeled by ZTPois(**λ**), where** $\lambda = \{1, 2, 3, 4, 5, 6\}$**. A.** Measured distances across increasing value of π for various λ. The dashed lines describe $y = x$, and are drawn to highlight the direction of deviation in the measured distances. **B.** For a given level of read coverage depth parameterized by λ, the R^2 values of linear regressions with the base line vs. our expected distance measures as the independent variable and the true distance π as the dependent variable.

$\frac{\text{measured difference}}{\text{\# shared sites}}$ and the exclusion of each unbinarized (i.e. observed to be heterozygous) site subtracts .5 from the numerator, and subtracts 1 from the denominator. This directional deviation from true π becomes more substantial as read depth increases, witnessed by the decreasing R^2 values resulting from performing linear regression of π on the measured baseline distance (Supplemental Figure A.5B., top row). Intuitively, for small λ, likely only one allele at the heterozygous site will be sampled, and (assuming there is no preferential sampling of alleles of a particular status) its status will with 50% chance be called methylated and 50% unmethylated, and on expectation, a heterozygous status and a homozygous status will still contribute an expected distance of .5; however, for larger λ, heterozygosity is more likely to be reflected in the observed reads, therefore a larger portion of true heterozygous sites will be excluded, leading to a larger deviation of measured distance from π.

In contrast, for any given λ, our expected distance measurement has persistently high R^2 values

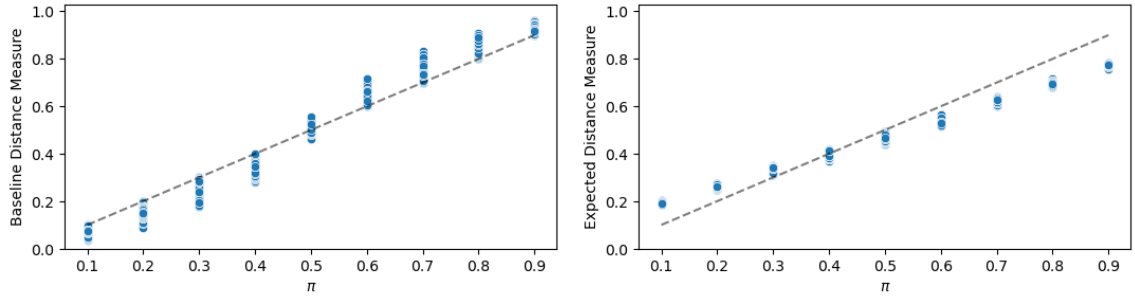
in a linear regression of π (Supplemental Figure A.5A., right; Supplemental Figure A.5B., bottom row). In fact, the R^2 values slightly improve with increasing level of coverage. We acknowledge that the expected distance measure, too, deviates from the ground truth π, especially at low coverage. As described by the intuition previously provided in the manuscript, with few reads at a CpG site, it is unlikely (quantified by the user-defined parameters) that they have sampled from all alleles present; therefore, if a CpG site in a cell only has reads of one methylation status, and the number of reads is low, our expected distance measure will assign some probability to it being heterozygous. This will lead to overestimation of the distance at a CpG site if two cells have the same homozygous status but are low-coverage for the site; similarly, this will lead to underestimation of the distance at a CpG site if two cells have the opposite homozygous statuses but are low-coverage for the site. There are proportionally more of the former type of sites for lower values of π and more of the latter type of sites for higher values of π, explaining the direction of deviation from $y = x$ (Supplemental Figure A.5A., right). That said, as the coverage level increases, the expected distance measure converges to the ground truth distance π.

Since it is reasonable to assume that the per-site read count distribution across all single cells sequenced with the same technology is close to being identical, our expected distance formulation is preferable to the baseline measure, as it tracks the true distance well for any given read coverage level. This is the case especially for higher coverage levels increases, which should be desirable if not destined as sequencing technology improves.

Next, we examine the performance of the two distance measures on data simulated using the empirical read coverage distribution (Supplemental Figure A.4B.). For each of the $450 \times 1 \times 5 = 2,250$ sets of read count vectors, we again measure distance using (i) the baseline measure and (ii) our expected distance measure. The trends previously observed with read counts sampled from ZTPois(λ)

still stand (Supplemental Figure A.6A.), and our expected distance measure again tracks the ground truth distance π better at this level of coverage, resulting a higher R^2 value (Supplemental Figure A.6B.).

A.



B.

R^2(π, Baseline Distance Measure)	.987
R^2(π, Expected Distance Measure)	.997

Figure A.6: **Baseline distance measure and expected distance measure on $2,250$ sets of read count vectors, simulated from cell pairs with known normalized ground truth distance $\pi = \{.1, .2, .3, .4, .5, .6, .7, .8, .9\}$, with level of read coverage depth modeled by the empirical read coverage distribution gathered from the CRC01 sample. A.** Measured distances across increasing value of π. The dashed lines describe $y = x$, and are drawn to highlight the direction of deviation in the measured distances. Note that the expected distance measure is based on the default configuration with the probability of a site being heterozygous set to $1/3$. **B.** R^2 values of linear regressions with the measured distances being independent variable and the true distance π being the dependent variable.

A.2.2 Iterative pruning on simulated data

Next, we evaluate our iterative pruning approach on simulated data against a baseline approach as described in the Results section.

A.2.2.1 Simulation details

We simulate 10 clone trees with different topologies, various degrees of branching, and various depths (Supplemental Figure A.7). Each clone tree has 1 normal clone (grey "ROOT" nodes in Supplemental Figure A.7) and 20 tumor clones (numbered nodes in Supplemental Figure A.7).

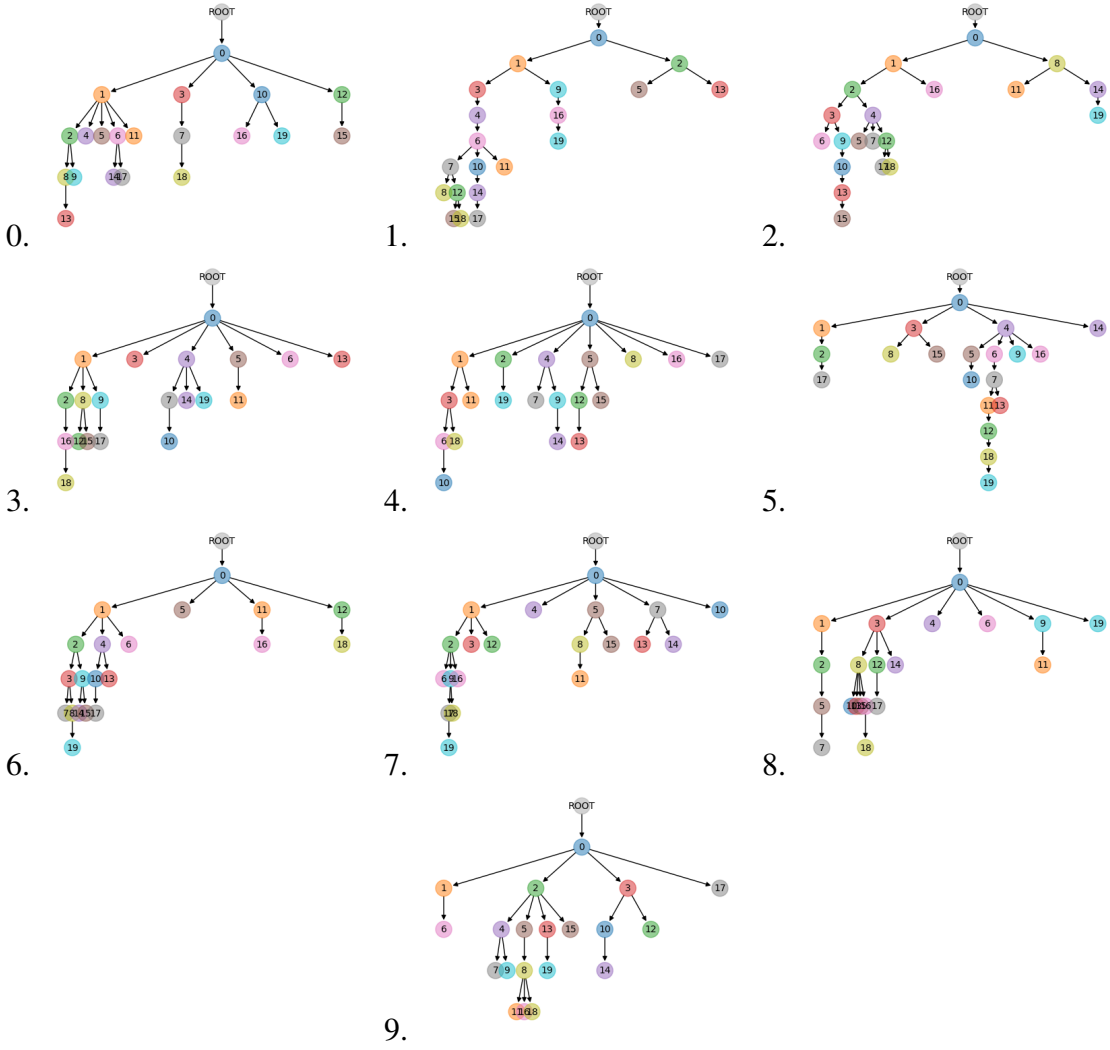


Figure A.7: **Topologies of the simulated clone trees.** A total of 3,000 CpG sites abiding by the infinite-sites assumption - and therefore may be lineage-informative - were simulated for each tree.

In a clone tree, each distinct clone is denoted by a distinct color and a distinct label, and it is represented by a pair of ground truth lineage-informative CpG-alleles with 9,000 CpG sites. 3000 of these are infinite-sites-abiding, another 3,000 "invariant", and the final 3,000 those of the "mixture" model [45]. For each clone tree, we first focus on the 3,000 infinite-sites-abiding CpG sites. At "ROOT", all of these sites are set to be homozygously unmethylated, and at each branch in the clone tree, a subset of these CpG sites that have not been previously selected are selected to have their

statuses change from homozygously unmethylated to (i) homozygously methylated with 50% chance and (ii) heterozygously methylated with 50% chance. The subsets are roughly equal in size. The remaining 6,000 CpG sites are simulated at the single cell level, as described below.

Given the clone tree with known (sub)clonal methylation status changes for the "infinite-sites-abiding" CpG sites, we simulate single-cell datasets with added noise and sparsity. To simulate one single-cell dataset from a clone tree, we simulate 10 diploid cells from each clone ($21 \times 10 = 210$ cells for each dataset), each with 6,000 additional CpG sites - the methylation status of the 3,000 "infinite-sites-abiding" sites are copied from their clone of origin. Among the additional CpG sites, the first 3,000 are "invariant" whose methylation status in each cell is with 80% chance homozygous (half of these homozygous sites are methylated by default and the other half are unmethylated), 10% chance heterozygous, and 10% chance the alternative homozygous. The remaining 3,000 sites are of the mixture model of methylation dynamics [45]: each is with 50% chance heterozygous, 25% chance homozygously methylated and 25% homozygously unmethylated.

Given such a set of 210 pairs of CpG-alleles, we sample the corresponding $210 \times 9,000$ methylated and unmethylated read count matrices according to the empirical read coverage distribution (Supplemental Figure A.4B.) as previously described in Supplemental Section A.2.1. Furthermore, we randomly mask out the same 33% of the coordinates in either $210 \times 9,000$ matrix, introducing sparsity observed in real datasets while mirroring the desired input matrix density (previously justified in Supplemental Section A.1.2).

We simulate 5 such single-cell datasets for each clone tree, resulting in $10 \times 5 = 50$ total pairs of (un)methylated read count matrices that are 66% dense. We leverage those as input to our simulation experiments.

A.2.2.2 Results

We compare the performance of `Sgootr` (no biclustering or error correction, $\kappa = .1$, for a maximum of 20 iterations) on simulated datasets against a baseline approach as described in the Results section with respect to the trees constructed. Briefly, given the read count matrices, the baseline approach computes the Hamming distance between the binarized methylation status vectors of each cell pair, normalizes by the number of sites covered in both cells, and applies NJ on the pairwise distance matrix to obtain a tree. We root the `Sgootr`- and baseline-generated trees with the same set of arbitrarily picked single-cell from the normal clone, and infer clonal evolution histories from the constructed trees using the Fitch-Hartigan-algorithm-based approach with the ground truth topologies (Supplemental Figure A.7) as the guiding partial orders. We use the same two evaluation metrics as in the Results section: (i) number of methylation events, and (ii) BAMDPO, which measures the difference between the inferred and the ground truth topologies.

In terms of the two performance measures, `Sgootr` outperforms the baseline approach in datasets generated from 9 out of 10 simulated clone trees, and performs comparably in datasets generated from simulated clone tree 4 (Supplemental Figure A.8, comparing between "baseline" columns and "$t*$" columns). Furthermore, despite that `Sgootr` many times already has superior performance compared to baseline prior to its iterative procedure, highlighting the benefits of using the expected distance formulation as opposed to the baseline distance measure (Supplemental Figure A.8, comparing between "baseline" columns and "t_0" columns), the iterative procedure of `Sgootr` seems to lead to additional boosts in performance by refining the signal (Supplemental Figure A.8, comparing between "t_0" columns and "$t*$" columns).

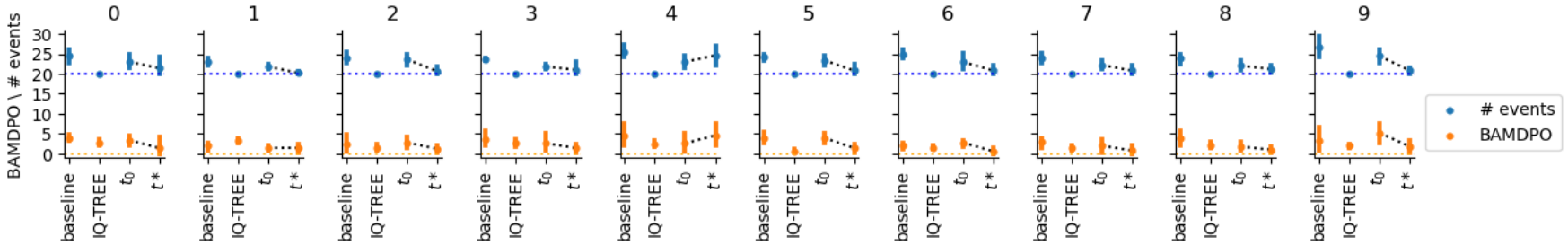


Figure A.8: **Results on 10 simulated clone trees from `Sgootr` and baseline method.** Number of migration events and BAMDPO values of migration histories obtained via (i) naïve baseline method, (ii) IQ-TREE, (iii) `Sgootr` intermediate tree prior to iterative procedure (t_0), and (iv) final tree output by `Sgootr` ($t*$). Results are represented by the mean across 5 single-cell methylation datasets generated from each clone tree, with the error bars denoting the 95% confidence interval of the mean. IQ-TREE result is represented by the mean across 10 runs with different random seeds, with the error bars denoting the 95% confidence interval of the mean. Dotted lines in blue and orange mark the ground truth (best possible) values for number of events and BAMDPO, respectively.

We also ran the maximum-likelihood-based IQ-TREE with the `GTR2` model (see Supplemental Section A.10 for details) on the simulated single-cell methylation datasets. For each simulated dataset, we performed 10 separate runs of IQ-TREE with different random seeds, terminating the program after 2 hours. Despite only 13 out of $50 \times 10 = 500$ IQ-TREE experiments converged after 2 hours, for these small simulated datasets, the intermediate outputs of IQ-TREE already exhibits high level of parsimony and accuracy in their clonal evolution history inferred (Supplemental Figure A.8). In fact, it has achieved the minimum number of events in all experiments (Supplemental Figure A.8, "# events" for "IQ-TREE" columns) and outperforms the baseline approach in terms of average BAMDPO in 9 out of 10 simulated clone trees (Supplemental Figure A.8, comparing "BAMDPO" between "baseline" columns and "IQ-TREE" columns). However, in terms of average BAMDPO, `Sgootr` slightly outperforms IQ-TREE in 8 out of 10 simulated clone trees (Supplemental Figure A.8, comparing "BAMDPO" between "IQ-TREE" columns and "$t*$" columns). It is also worth noting that, given the same computing resources, `Sgootr` produces results in a matter of minutes, as opposed to hours, which seems to be required for IQ-TREE convergence. This demonstrates that `Sgootr` is capable of achieving accuracy on average comparable to maximum-likelihood-based IQ-TREE on sparse single-cell methylation data in a fraction of the time required by the latter, suggesting `Sgootr` may be the more practical choice on large real datasets.

A.3 Application of `Sgootr` on single cells from U87MG *ex vivo* tree from the RETrace study

In the RETrace study, Wei and colleagues [88] presents a method to accurately capture microsatellites and simultaneously cytosine methylation (via scRRBS) from single cells. For validation,

Wei and colleagues experimentally cultured single cells from a cancer cell line for three generations, used the microsatellite allelotypes derived from the single cells from the resulting *ex vivo* clones to construct a lineage tree, and assessed the accuracy of the constructed tree with the experimentally-derived clonal relationships. While their lineage trees show high accuracy for microsatellite-unstable cells (See Figure 3 of the RETrace study), they demonstrate that for microsatellite-stable U87MG glioblastoma cell line, the accuracy of the microsatellite allelotype-based tree is only akin to that of a baseline random tree (See Supplemental Figure S10 of the RETrace study) [88].

We show that `Sgootr` can accurately reconstruct a single cell lineage tree using CpG methylation information from the microsatellite-stable U87MG *ex vivo* tree even though reconstruction using microsatellite allelotypes led to an inaccurate tree. We apply `Sgootr` to the scRRBS data of the 74 U87MG single cells from 7 distinct clones, whose clonal relationship is shown in Supplemental Figure S10a. of the work by Wei and colleagues [88]. While the quality control step in the original work ensures each of the 74 single cells has at least one read covering at least 30,000 unique CpG sites, both per cell and total CpG coverage in scRRBS datasets are lower compared to scBS-seq (Supplemental Figure A.9A.,B.). To accommodate such level of sparsity, we modify the parameters of `Sgootr` such that it: selects CpG sites covered in no less than 5% of the remaining cells, bypasses the post-filtering step, and relaxes the pruning parameters so that $\delta = .02, \omega = .05$. We root the tree with a single cell from clone 2-1-1D_3-1-10E, which is an out-group in the *ex vivo* tree.

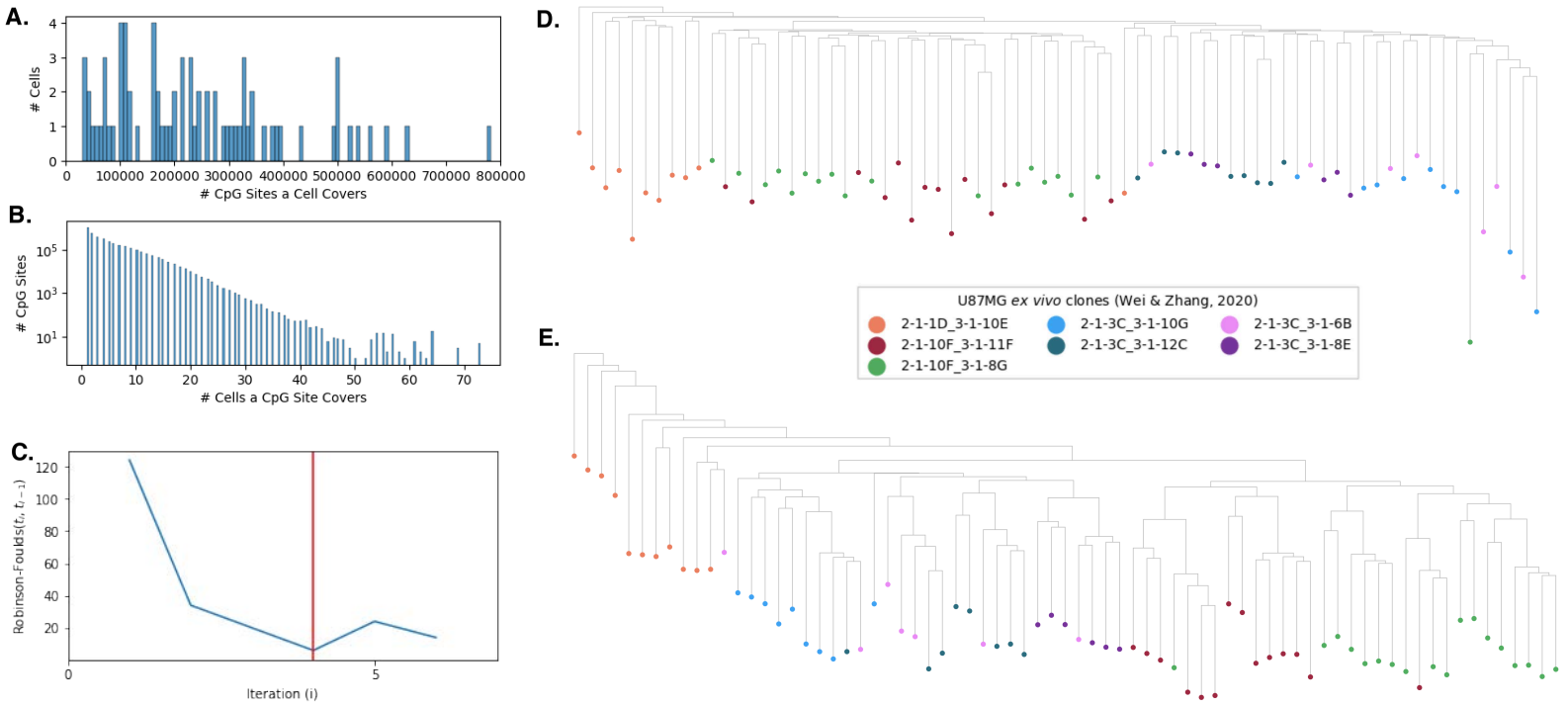


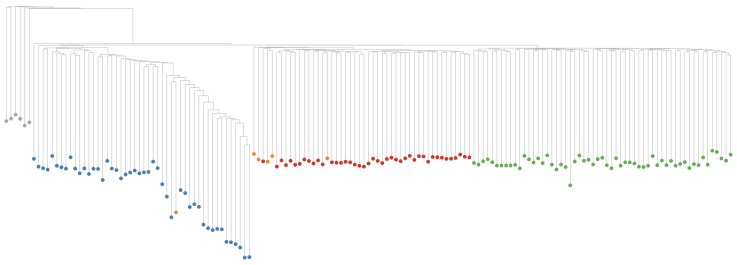
Figure A.9: **Application of `Sgootr` to scRRBS data from U87MG *ex vivo* tree by the RETrace study. A.** Cell coverage distribution among autosomal CpG sites. Cells with fewer than 30,000 unique CpG sites covered are excluded. **B.** Autosomal CpG site coverage distribution. Note that the y-axis is in log scale. **C.** RF distances across 6 iterations, with global minimum occurring at $i = 4$. **D.** Intermediate lineage tree constructed by `Sgootr` prior to the iterative procedure (t_0). **E.** Lineage tree output by `Sgootr` ($t*$).

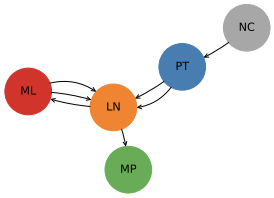
`Sgootr` obtains the optimal lineage tree $t* = t_4$ (Supplemental Figure A.9C.,E.). The final output $t*$ visibly improves from t_0, the intermediate tree `Sgootr` produces prior to the iterative pruning procedure (Supplemental Figure A.9D.), in terms of the clustering of single cells from the same clone. Furthermore, $t*$ captures the clonal relationships in the ground truth *ex vivo* tree: clone 2-1-1D_3-1-10E is the out-group, clones 2-1-10F_3-1-11F and 2-1-10F_3-1-8G are derived from the same parent clone, and clones 2-1-3C_3-1-10G, 2-1-3C_3-1-12C, 2-1-3C_3-1-6B, 2-1-3C_3-1-8E are derived from the same parent clone (See Supplemental Figure S10 of the RETrace study). This is in contrast to the microsatellite allelotype-based lineage tree constructed in the RETrace study, which shows little association with the clonal relationships exhibited in the *ex vivo* tree(See Supplemental Figure S10 of the RETrace study).

A.4 Application of `Sgootr` on the complete set of 9 metastatic CRC patients made available by Bian and colleagues

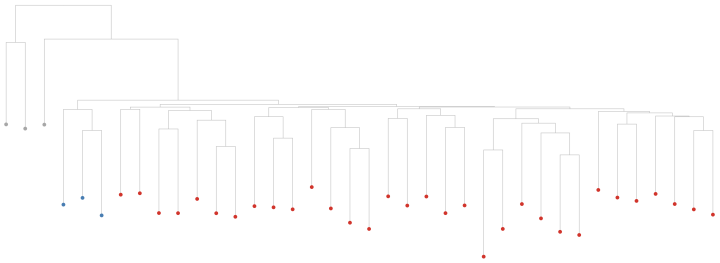
Patient	Primary site	Sampled lesions	Number of cells	Number of sites	CNA calls
CRC01	Left Colon	NC, PT, LN, ML, MP	160	140,801	✓
CRC02	Right Colon	NC, PT, ML	39	49,633	
CRC04	Right Colon	NC, PT, LN	65	303,849	✓
CRC10	Left Colon	NC, PT, LN	114	275,111	
CRC11	Left Colon	NC, PT, LN	228	579,272	
CRC12	Rectum	NC, PT, LN	22	650,140	
CRC13	Left Colon	NC, PT, LN	185	204,153	
CRC14	Rectum	NC, PT, LN	29	511,588	
CRC15	Right Colon	NC, PT, LN, ML, MO	50	42,824	

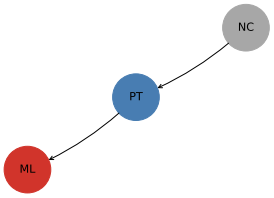
Table A.3: **Data summary of the Bian metastatic CRC cohort.** The number of cells and number of sites are those obtained after component (i) and (ii) of `Sgootr` before entering the iterative procedure.



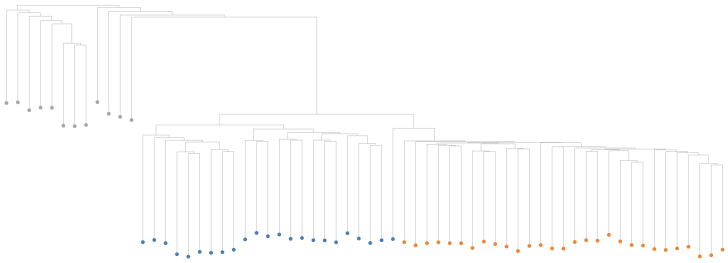


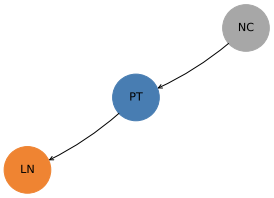
CRC01



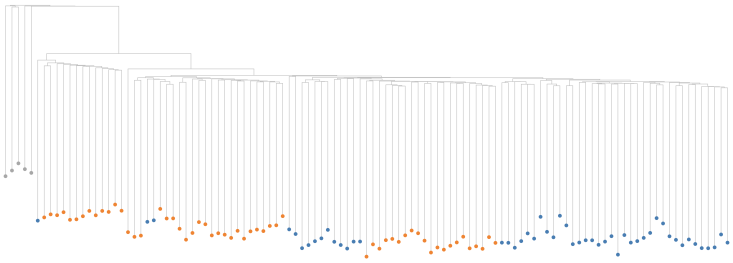


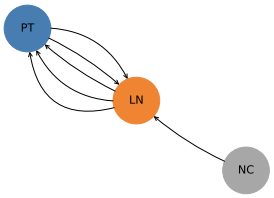
CRC02



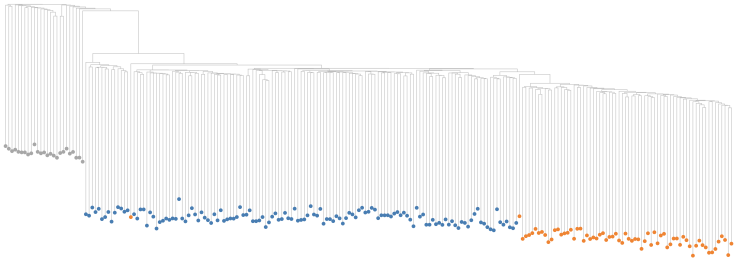


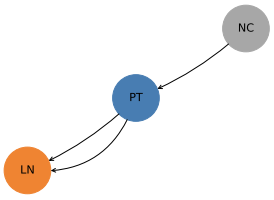
CRC04





CRC10





CRC11

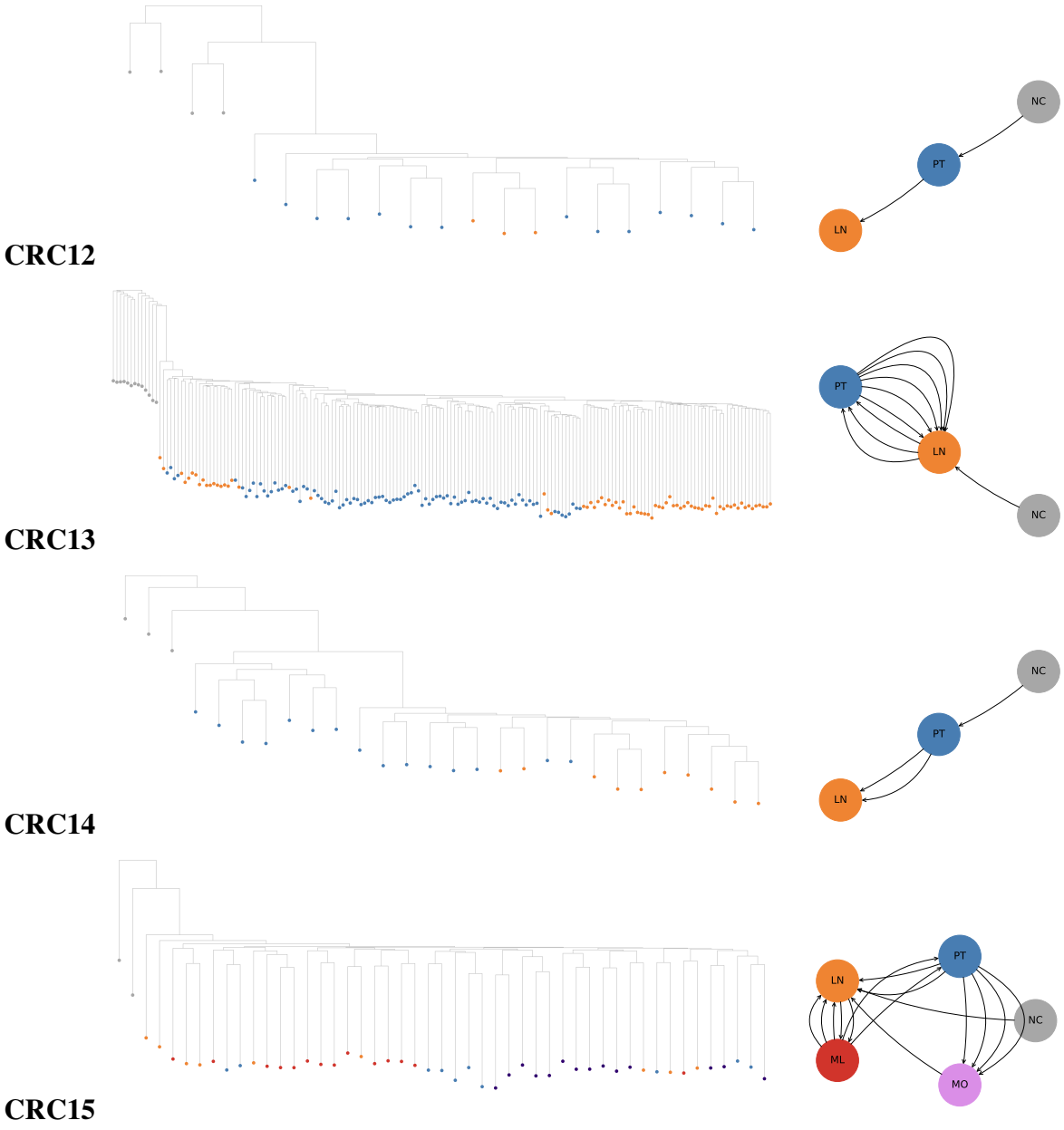


Figure A.10: **`Sgootr`-constructed ($\kappa = .1$) tumor lineage trees (t^*s) and `Sgootr`-inferred tumor migration histories for the 9 patients in the Bian cohort**

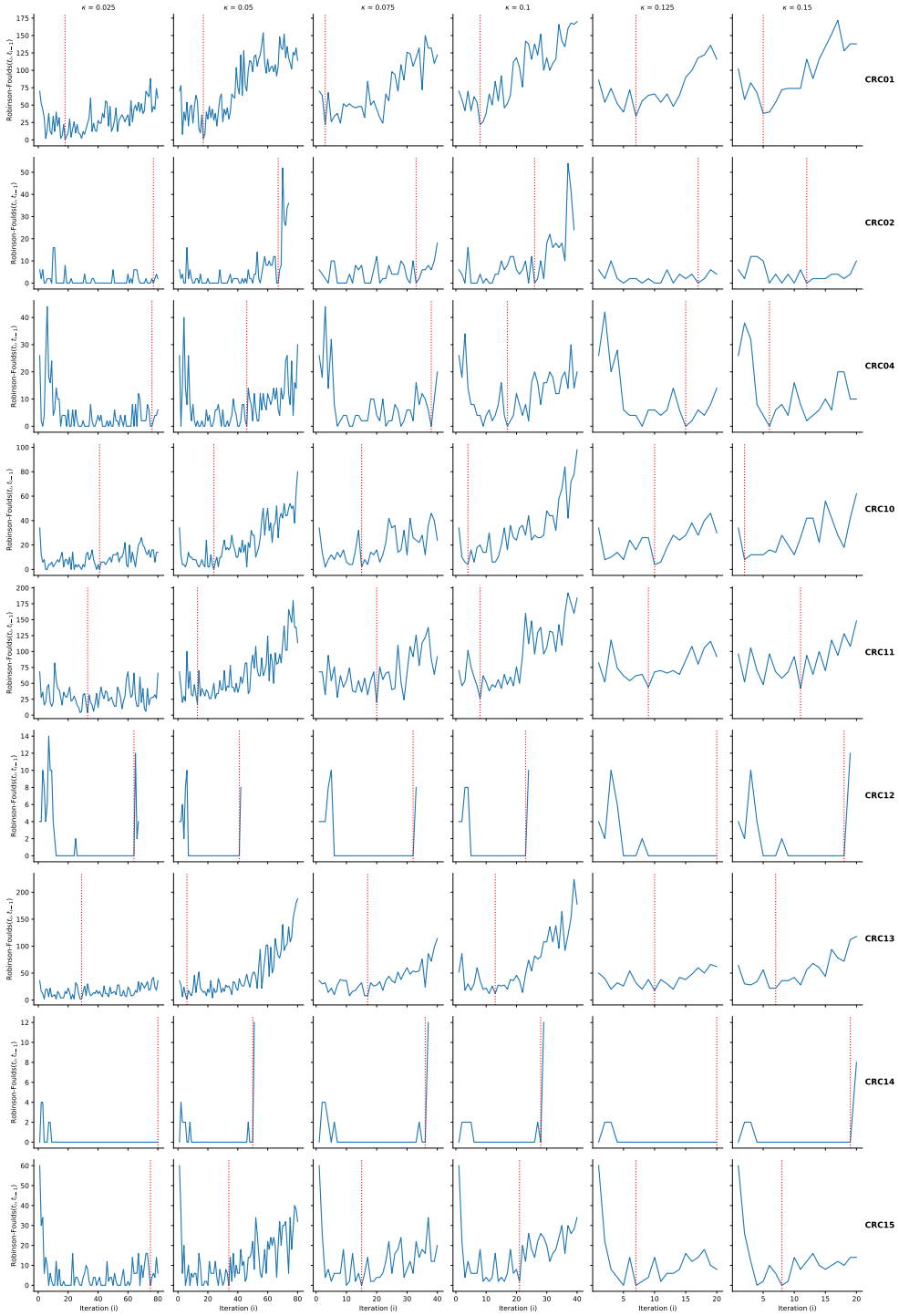
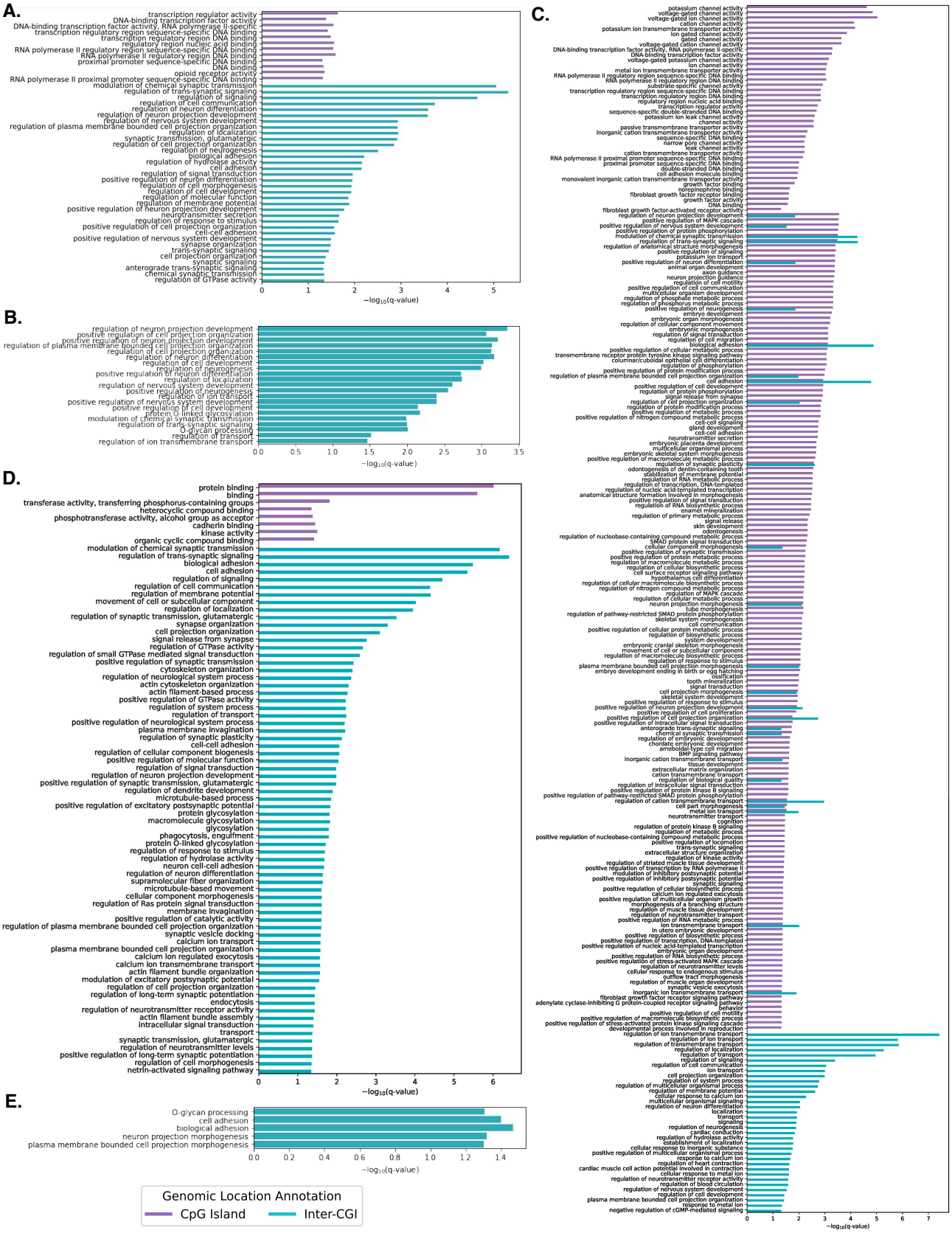


Figure A.11: **RF distances across iterations for each patient in the metastatic CRC cohort applying `Sgootr` with different κ values.** Each row represents each of the 9 patients, and each column represents κ=.025, .05, .075, .1, .125, and .15. For κ=.025, .05, the iterative procedure was run for a maximum of 80 rounds; for κ = .075, .1, maximum 40 rounds; and for κ=.125,.15, maximum 20 rounds. Each red vertical line marks the iteration `Sgootr` identifies as optimal. If the blue line ends before the maximum iteration, it means at the last iteration of pruning, the distance between some cells pairs can no longer be computed as all their shared CpG sites have been pruned out. Note that the maximum iteration is arbitrarily chosen, and `Sgootr` does not need to finish running the maximum iteration in order to detect whether the optimal condition has been met.


A.
transcription regulator activity
DNA-binding transcription factor activity
DNA-binding transcription factor activity, RNA polymerase II-specific
transcription regulatory region sequence-specific DNA binding
transcription regulatory region DNA binding
regulatory region nucleic acid binding
RNA polymerase II regulatory region sequence-specific DNA binding
RNA polymerase II regulatory region DNA binding
proximal promoter sequence-specific DNA binding
DNA binding
opioid receptor activity
RNA polymerase II proximal promoter sequence-specific DNA binding
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
regulation of signaling
regulation of cell communication
regulation of neuron differentiation
regulation of neuron projection development
regulation of nervous system development
regulation of plasma membrane bounded cell projection organization
regulation of localization
synaptic transmission, glutamatergic
regulation of cell projection organization
regulation of neurogenesis
biological adhesion
regulation of hydrolase activity
cell adhesion
regulation of signal transduction
positive regulation of neuron differentiation
regulation of cell morphogenesis
regulation of cell development
regulation of molecular function
regulation of membrane potential
positive regulation of neuron projection development
neurotransmitter secretion
regulation of response to stimulus
positive regulation of cell projection organization
cell-cell adhesion
positive regulation of nervous system development
synapse organization
trans-synaptic signaling
cell projection organization
synaptic signaling
anterograde trans-synaptic signaling
chemical synaptic transmission
regulation of GTPase activity
0 1 2 3 4 5
−log10(q-value)
B.
regulation of neuron projection development
positive regulation of cell projection organization
positive regulation of neuron projection development
regulation of plasma membrane bounded cell projection organization
regulation of cell projection organization
regulation of neuron differentiation
regulation of cell development
regulation of neurogenesis
positive regulation of neuron differentiation
regulation of localization
regulation of nervous system development
positive regulation of neurogenesis
regulation of ion transport
positive regulation of nervous system development
positive regulation of cell development
protein O-linked glycosylation
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
O-glycan processing
regulation of transport
regulation of ion transmembrane transport
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
−log10(q-value)
C.
potassium channel activity
voltage-gated channel activity
voltage-gated ion channel activity
cation channel activity
potassium ion transmembrane transporter activity
ion gated channel activity
gated channel activity
voltage-gated cation channel activity
DNA-binding transcription factor activity, RNA polymerase II-specific
DNA-binding transcription factor activity
voltage-gated potassium channel activity
ion channel activity
metal ion transmembrane transporter activity
RNA polymerase II regulatory region sequence-specific DNA binding
RNA polymerase II regulatory region DNA binding
substrate-specific channel activity
transcription regulatory region sequence-specific DNA binding
transcription regulatory region DNA binding
regulatory region nucleic acid binding
transcription regulator activity
sequence-specific double-stranded DNA binding
potassium ion leak channel activity
gated channel activity
passive transmembrane transporter activity
inorganic cation transmembrane transporter activity
sequence-specific DNA binding
narrow pore channel activity
leak channel activity
cation transmembrane transporter activity
RNA polymerase II proximal promoter sequence-specific DNA binding
proximal promoter sequence-specific DNA binding
double-stranded DNA binding
cell adhesion molecule binding
monovalent inorganic cation transmembrane transporter activity
growth factor binding
norepinephrine binding
fibroblast growth factor receptor binding
growth factor activity
DNA binding
fibroblast growth factor-activated receptor activity
regulation of neuron projection development
positive regulation of MAPK cascade
positive regulation of nervous system development
positive regulation of protein phosphorylation
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
regulation of anatomical structure morphogenesis
positive regulation of signaling
potassium ion transport
positive regulation of neuron differentiation
animal organ development
axon guidance
neuron projection guidance
regulation of cell motility
positive regulation of cell communication
multicellular organism development
regulation of phosphate metabolic process
regulation of phosphorus metabolic process
positive regulation of neurogenesis
embryo development
embryonic organ morphogenesis
regulation of cellular component movement
embryonic morphogenesis
regulation of signal transduction
regulation of cell migration
biological adhesion
positive regulation of cellular metabolic process
transmembrane receptor protein tyrosine kinase signaling pathway
columnar/cuboidal epithelial cell differentiation
regulation of phosphorylation
positive regulation of protein modification process
regulation of plasma membrane bounded cell projection organization
cell adhesion
positive regulation of cell development
regulation of protein phosphorylation
signal release from synapse
regulation of cell projection organization
regulation of protein modification process
positive regulation of metabolic process
positive regulation of nitrogen compound metabolic process
cell-cell signaling
gland development
cell-cell adhesion
neurotransmitter secretion
embryonic placenta development
multicellular organismal process
embryonic skeletal system morphogenesis
positive regulation of macromolecule metabolic process
regulation of synaptic plasticity
odontogenesis of dentin-containing tooth
stabilization of membrane potential
regulation of RNA metabolic process
regulation of transcription, DNA-templated
regulation of nucleic acid-templated transcription
anatomical structure formation involved in morphogenesis
positive regulation of signal transduction
regulation of RNA biosynthetic process
enamel mineralization
regulation of primary metabolic process
signal release
skin development
odontogenesis
regulation of nucleobase-containing compound metabolic process
SMAD protein signal transduction
cellular component morphogenesis
positive regulation of synaptic transmission
positive regulation of protein metabolic process
regulation of macromolecule metabolic process
regulation of cellular biosynthetic process
cell surface receptor signaling pathway
hypothalamus cell differentiation
regulation of cellular macromolecule biosynthetic process
regulation of nitrogen compound metabolic process
regulation of MAPK cascade
regulation of cellular metabolic process
neuron projection morphogenesis
tube morphogenesis
regulation of pathway-restricted SMAD protein phosphorylation
skeletal system morphogenesis
cell communication
positive regulation of cellular protein metabolic process
regulation of biosynthetic process
system development
embryonic cranial skeleton morphogenesis
movement of cell or subcellular component
regulation of macromolecule biosynthetic process
regulation of response to stimulus
plasma membrane bounded cell projection morphogenesis
embryo development ending in birth or egg hatching
ossification
tooth mineralization
signal transduction
cell projection morphogenesis
skeletal system development
positive regulation of response to stimulus
positive regulation of neuron projection development
positive regulation of cell proliferation
positive regulation of cell projection organization
positive regulation of intracellular signal transduction
anterograde trans-synaptic signaling
chemical synaptic transmission
regulation of embryonic development
chordate embryonic development
ameboidal-type cell migration
BMP signaling pathway
inorganic cation transmembrane transport
tissue development
extracellular matrix organization
cation transmembrane transport
regulation of biological quality
regulation of intracellular signal transduction
positive regulation of protein kinase B signaling
positive regulation of pathway-restricted SMAD protein phosphorylation
regulation of cation transmembrane transport
cell part morphogenesis
metal ion transport
neurotransmitter transport
cognition
regulation of protein kinase B signaling
regulation of metabolic process
positive regulation of nucleobase-containing compound metabolic process
positive regulation of locomotion
trans-synaptic signaling
extracellular structure organization
regulation of kinase activity
regulation of striated muscle tissue development
positive regulation of transcription by RNA polymerase II
modulation of inhibitory postsynaptic potential
positive regulation of inhibitory postsynaptic potential
synaptic signaling
positive regulation of cellular biosynthetic process
calcium ion regulated exocytosis
positive regulation of multicellular organism growth
morphogenesis of a branching structure
regulation of muscle tissue development
regulation of neurotransmitter transport
positive regulation of RNA metabolic process
ion transmembrane transport
in utero embryonic development
positive regulation of biosynthetic process
positive regulation of transcription, DNA-templated
positive regulation of nucleic acid-templated transcription
embryonic organ development
positive regulation of RNA biosynthetic process
positive regulation of stress-activated MAPK cascade
regulation of neurotransmitter levels
cellular response to endogenous stimulus
outflow tract morphogenesis
regulation of muscle organ development
synaptic vesicle exocytosis
inorganic ion transmembrane transport
fibroblast growth factor receptor signaling pathway
adenylate cyclase-inhibiting G protein-coupled receptor signaling pathway
behavior
positive regulation of cell motility
positive regulation of macromolecule biosynthetic process
positive regulation of stress-activated protein kinase signaling cascade
developmental process involved in reproduction
regulation of ion transmembrane transport
regulation of ion transport
regulation of transmembrane transport
regulation of localization
regulation of transport
regulation of signaling
regulation of cell communication
ion transport
cell projection organization
regulation of system process
regulation of multicellular organismal process
regulation of membrane potential
cellular response to calcium ion
multicellular organismal signaling
regulation of neuron differentiation
localization
transport
signaling
regulation of neurogenesis
cardiac conduction
regulation of hydrolase activity
establishment of localization
cellular response to inorganic substance
positive regulation of multicellular organismal process
response to calcium ion
regulation of heart contraction
cardiac muscle cell action potential involved in contraction
cellular response to metal ion
regulation of neurotransmitter receptor activity
regulation of blood circulation
regulation of nervous system development
regulation of cell development
plasma membrane bounded cell projection organization
response to metal ion
negative regulation of cGMP-mediated signaling
0 1 2 3 4 5 6 7
−log10(q-value)
D.
protein binding
binding
transferase activity, transferring phosphorus-containing groups
heterocyclic compound binding
phosphotransferase activity, alcohol group as acceptor
cadherin binding
activity
organic cyclic compound binding
modulation of chemical synaptic transmission
regulation of trans-synaptic signaling
biological adhesion
cell adhesion
regulation of signaling
regulation of cell communication
regulation of membrane potential
movement of cell or subcellular component
regulation of localization
regulation of synaptic transmission, glutamatergic
synapse organization
cell projection organization
signal release from synapse
regulation of GTPase activity
regulation of small GTPase mediated signal transduction
positive regulation of synaptic transmission
cytoskeleton organization
regulation of neurological system process
actin cytoskeleton organization
actin filament-based process
positive regulation of GTPase activity
regulation of system process
regulation of transport
positive regulation of neurological system process
plasma membrane invagination
regulation of synaptic plasticity
cell-cell adhesion
regulation of cellular component biogenesis
positive regulation of molecular function
regulation of signal transduction
regulation of neuron projection development
positive regulation of synaptic transmission, glutamatergic
regulation of dendrite development
microtubule-based process
positive regulation of excitatory postsynaptic potential
protein glycosylation
macromolecule glycosylation
glycosylation
phagocytosis, engulfment
protein O-linked glycosylation
regulation of response to stimulus
regulation of hydrolase activity
neuron cell-cell adhesion
regulation of neuron differentiation
supramolecular fiber organization
microtubule-based movement
cellular component morphogenesis
regulation of Ras protein signal transduction
membrane invagination
positive regulation of catalytic activity
regulation of plasma membrane bounded cell projection organization
synaptic vesicle docking
calcium ion transport
plasma membrane bounded cell projection organization
calcium ion regulated exocytosis
calcium ion transmembrane transport
actin filament bundle organization
modulation of excitatory postsynaptic potential
regulation of cell projection organization
regulation of long-term synaptic potentiation
endocytosis
regulation of neurotransmitter receptor activity
actin filament bundle assembly
intracellular signal transduction
transport
synaptic transmission, glutamatergic
regulation of neurotransmitter levels
positive regulation of long-term synaptic potentiation
regulation of cell morphogenesis
netrin-activated signaling pathway
0 1 2 3 4 5 6
−log10(q-value)
E.
O-glycan processing
cell adhesion
biological adhesion
neuron projection morphogenesis
plasma membrane bounded cell projection morphogenesis
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
−log10(q-value)
Genomic Location Annotation
CpG Island
Inter-CGI

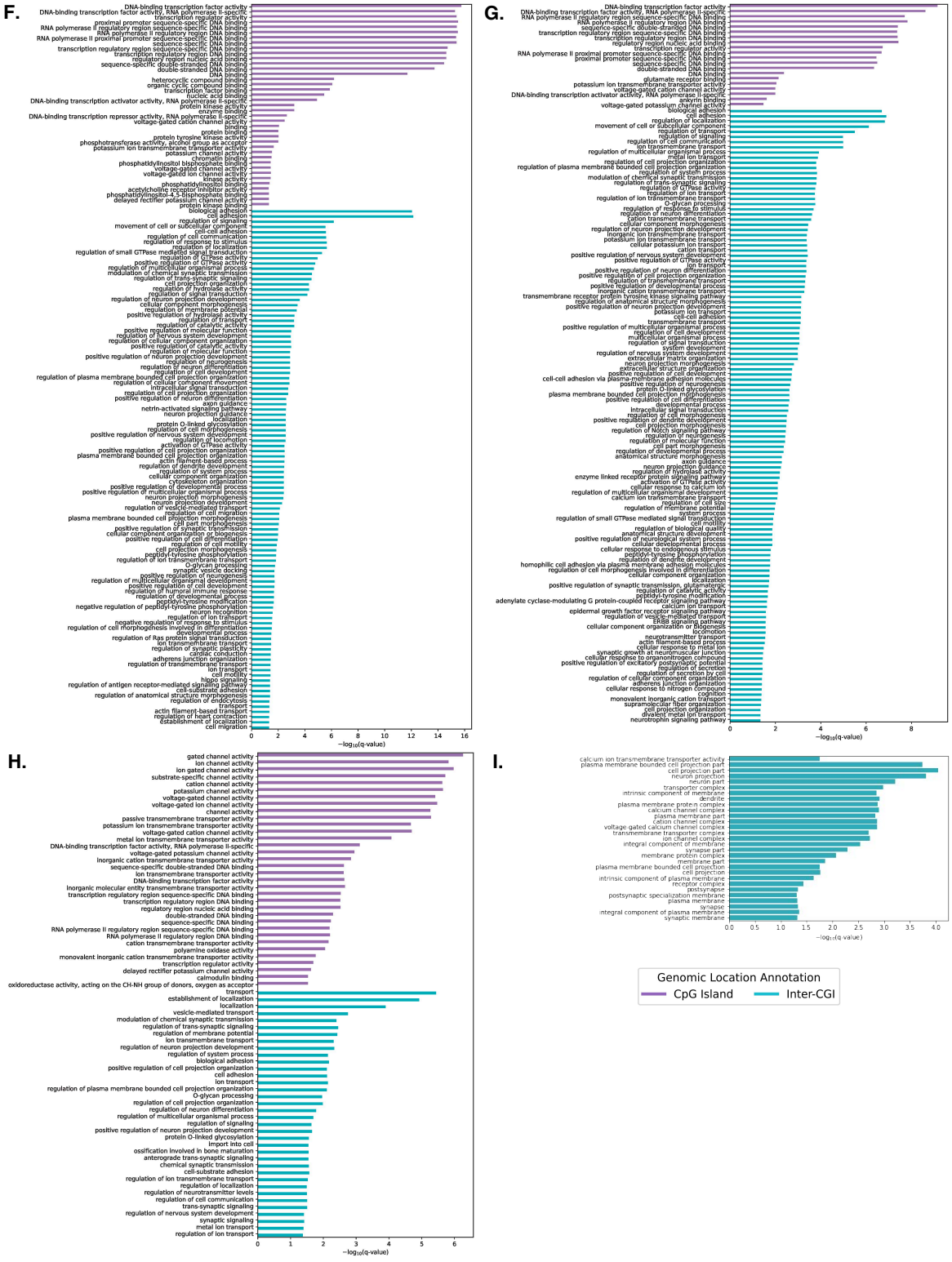
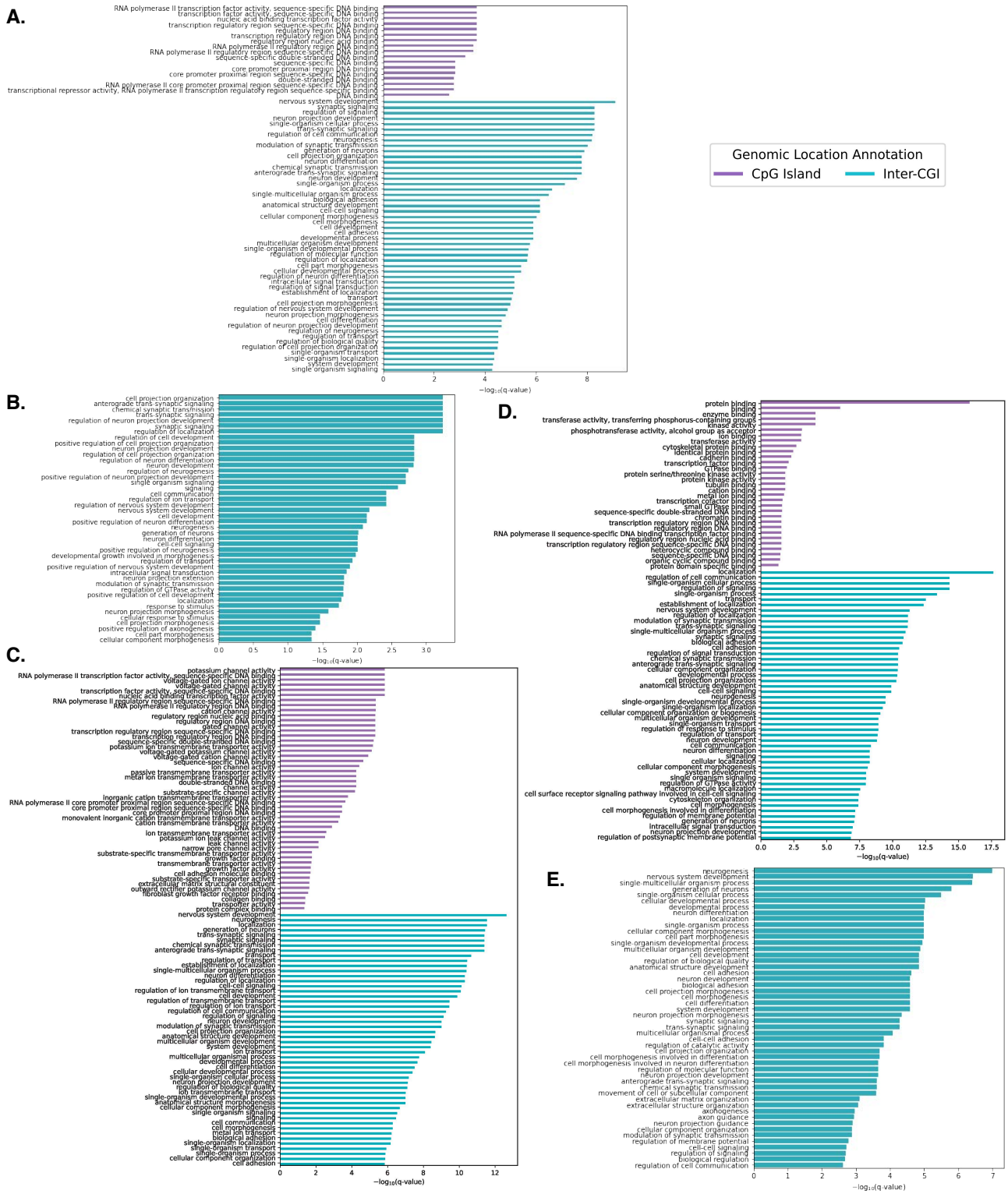


Figure A.12: **GOrilla [95] -identified GO terms with significant (<.05) q-values in enrichment analysis of non-pseudo, protein coding genes whose gene bodies and 1000bp upstream promoter regions contain the `Sgootr`-identified lineage-informative CpG sites, stratified by whether they locate in CpG island or inter-CGI regions for A. CRC01, B. CRC02, C. CRC04, D. CRC10, E. CRC13, F. CRC11, G. CRC12, H. CRC14, and I. CRC15.**


A.
B.
C.
D.
E.
Genomic Location Annotation
CpG Island
Inter-CGI
−log10(q-value)

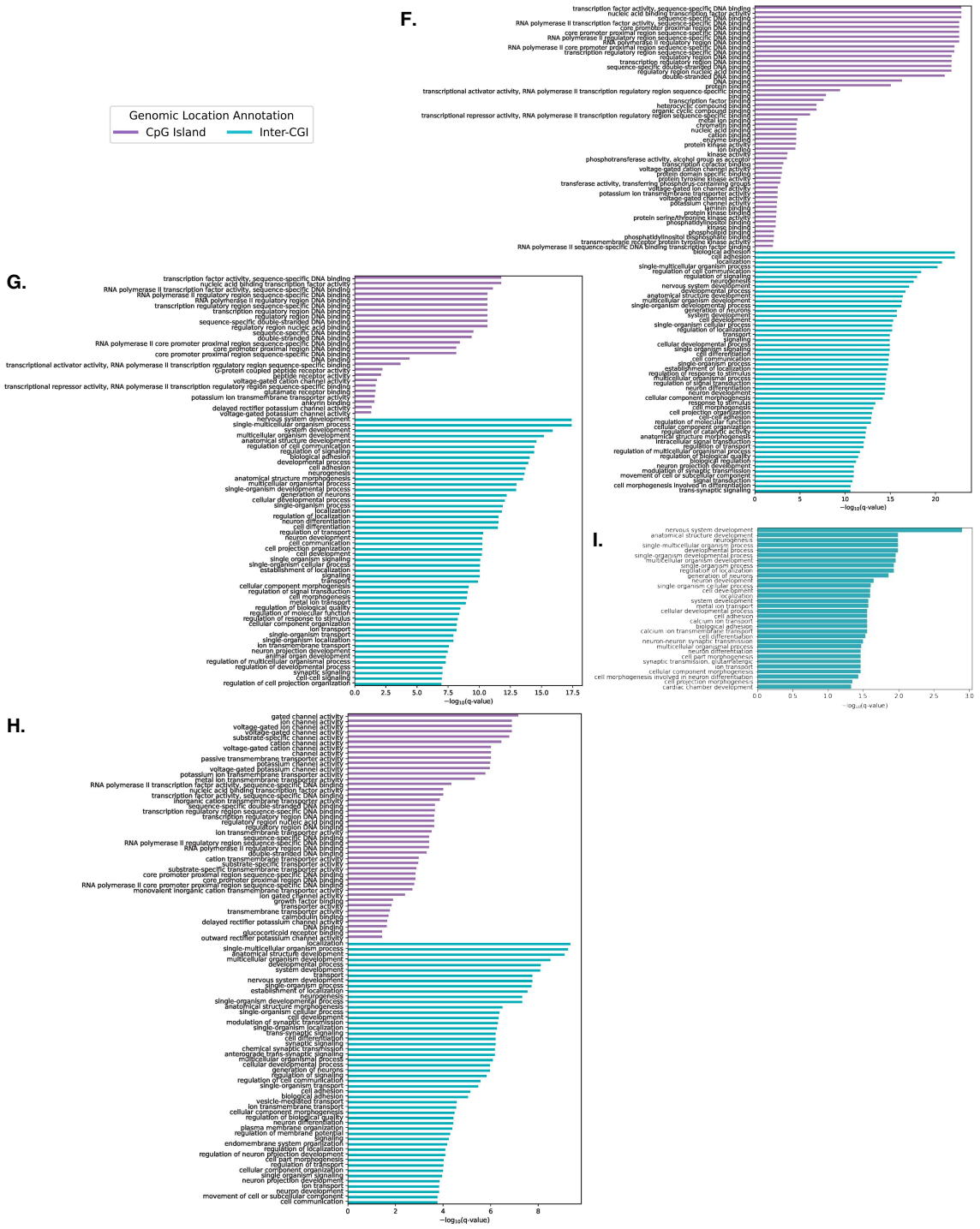


Figure A.13: **DAVID [100, 101] -identified GO terms with significant (<.05) q-values in enrichment analysis of non-pseudo, protein coding genes whose gene bodies and 1000bp upstream promoter regions contain the `Sgootr`-identified lineage-informative CpG sites, stratified by whether they locate in CpG island or inter-CGI regions for A. CRC01, B. CRC02, C. CRC04, D. CRC10, E. CRC13, F. CRC11, G. CRC12, H. CRC14, and I. CRC15.**

Patient	CpG Island	Inter-CGI
CRC01	0.8182	0.9412
CRC02	NA	0.7620
CRC04	0.7500	0.8689
CRC10	1.0000	0.8571
CRC11	0.9730	0.9200
CRC12	0.95	0.9160
CRC13	NA	0.6
CRC14	0.8824	0.8333
CRC15	NA	NA

Table A.4: **Overlap coefficients of GO terms associated with different genomic locations (CpG island and inter-CGI) discovered by GOrilla [95] and DAVID [100, 101] .** CRC02, CRC13, and CRC15 do not have any terms discovered by either GOrilla or DAVID in CpG island. The GO terms associated with inter-CGI regions in CRC15 discovered by GOrilla and DAVID belong to different sub-ontologies (molecular function and biological process, respectively), and are thus not comparable, despite having common themes in neural activities (Supplemental Figures A.12I.,A.13I.).

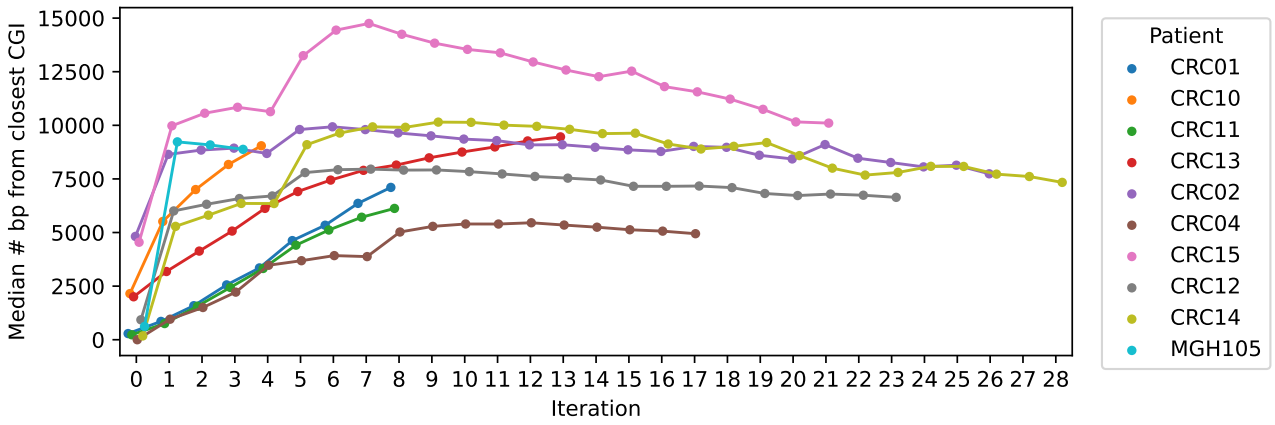
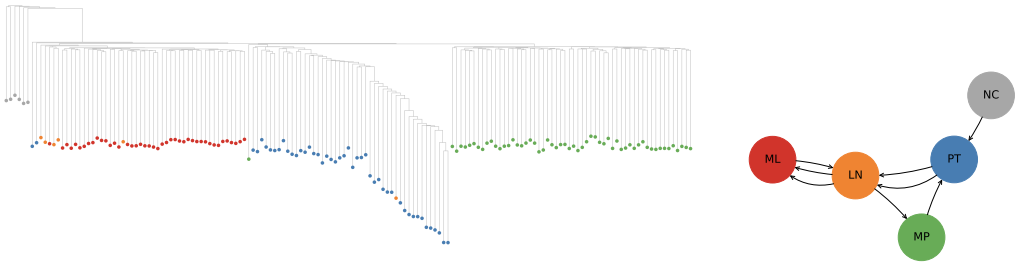


Figure A.14: **Median number of bp to the closest CGI from CpG sites used to construct the methylation lineage tree at each iteration for all CRC patients and GBM patient MGH105.** If a CpG site is within a CGI, it is 0 bp away from the closest CGI.

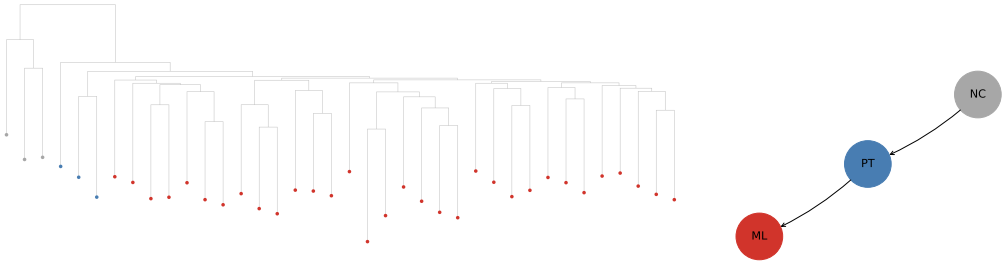
Patient	Spearman's Rho	p-value
CRC01	-0.5977	0.0
CRC02	-0.3006	0.0
CRC04	-0.6224	0.0
CRC10	-0.5827	0.0
CRC11	-0.6369	0.0
CRC12	-0.4700	0.0
CRC13	-0.5354	0.0
CRC14	-0.5922	0.0
CRC15	-0.4431	0.0
MGH105	-0.3907	0.0

Table A.5: **Spearman correlation of CpG site distance from closest CpG island and homogeneity.**

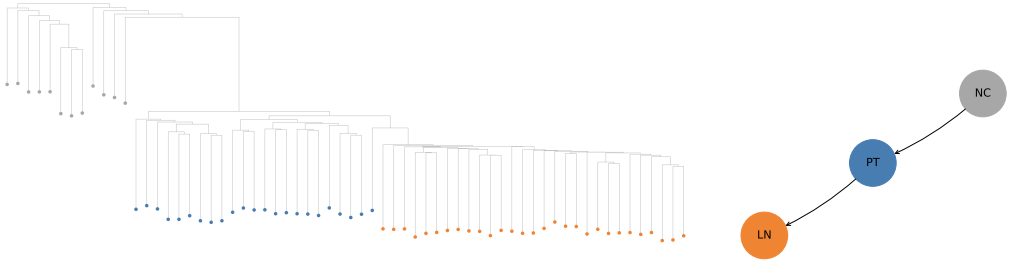
A.5 Application of `Sgootr` on the cohort, replacing NJ with FastME 2.0



CRC01

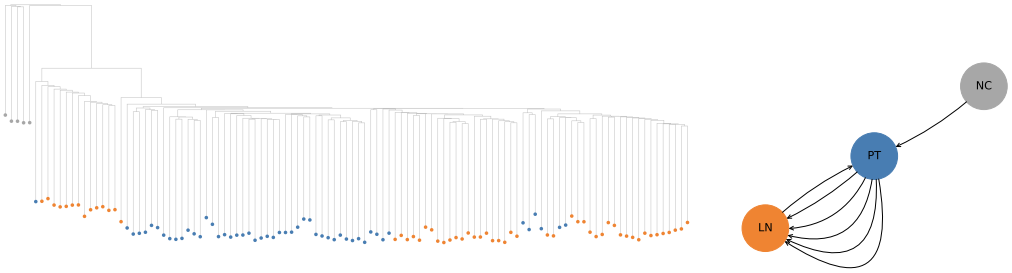


CRC02

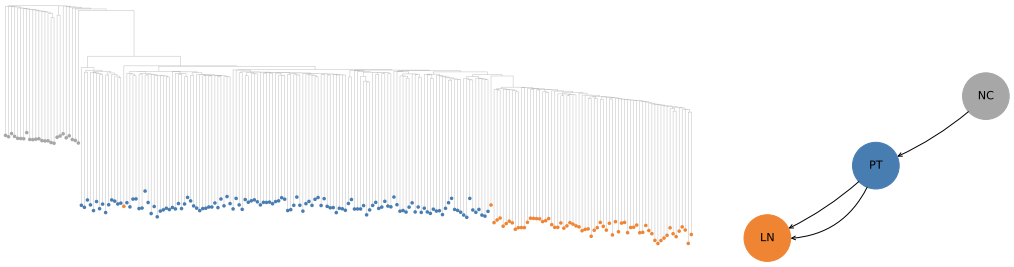


CRC04

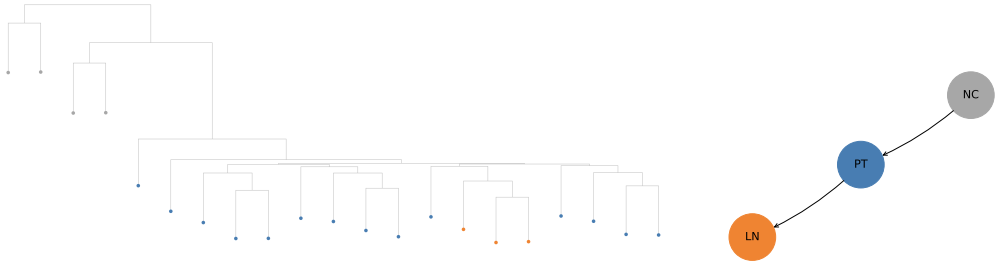
CRC10



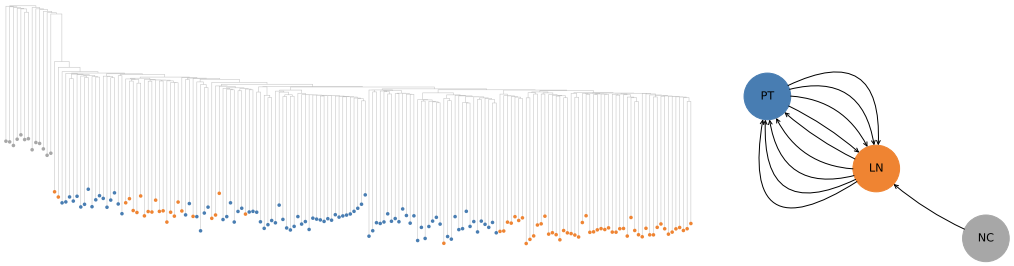
CRC11



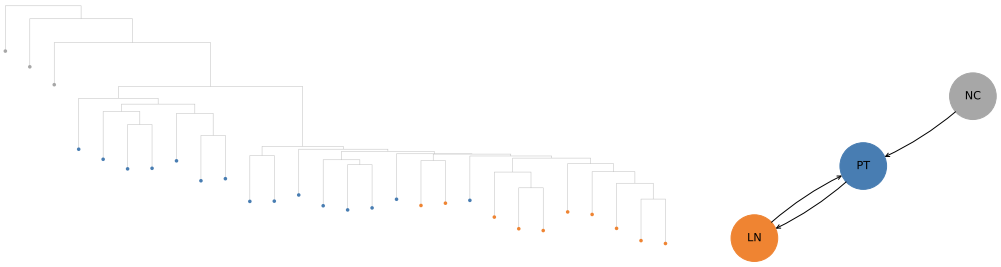
CRC12

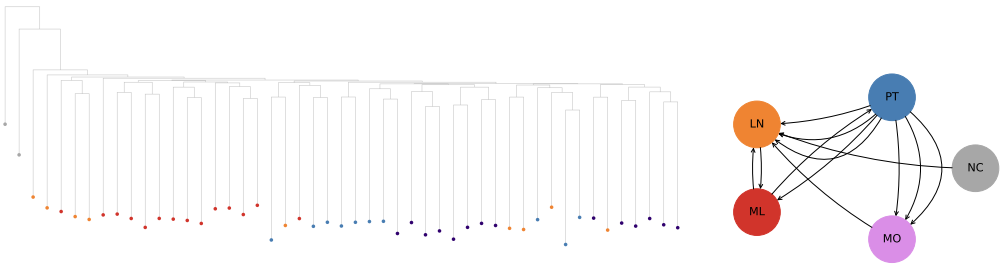


CRC13



CRC14





CRC15

Figure A.15: **`Sgootr`-constructed ($\kappa = .1$) tumor lineage trees (t^*s) and `Sgootr`-inferred tumor migration histories for the 9 patients in the Bian cohort, replacing NJ with FastME 2.0.**

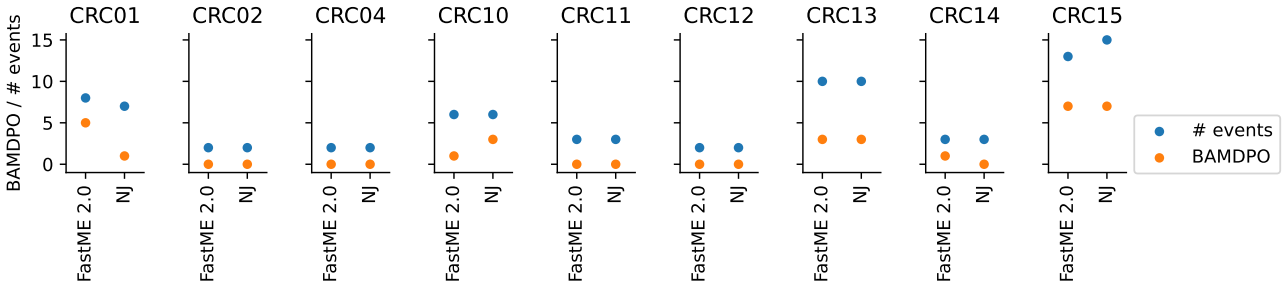


Figure A.16: **Number of migration events and BAMDPO of migration histories obtained via applying `Sgootr` with FastME 2.0 in comparison to NJ.**

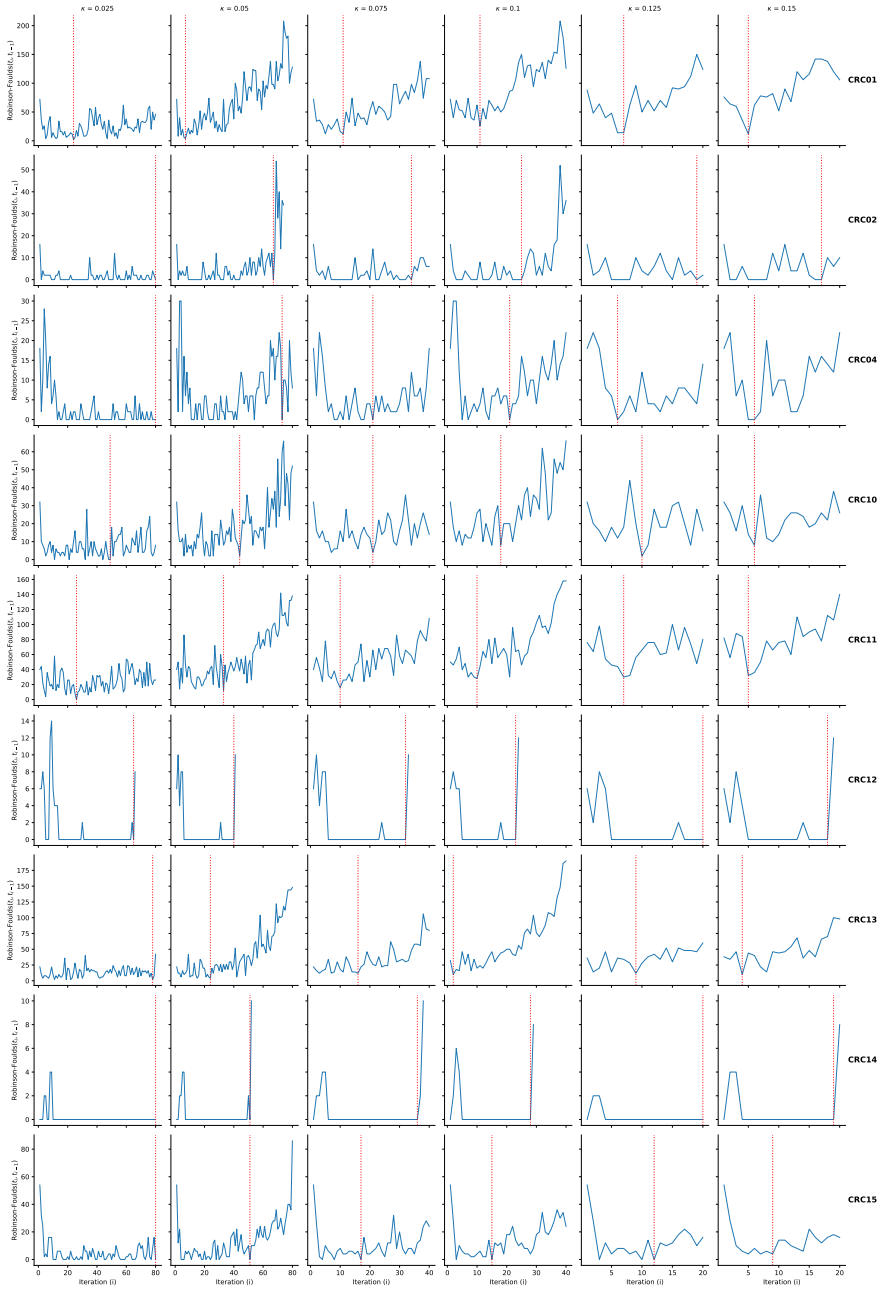


Figure A.17: **RF distances across iterations for each patient in the Bian metastatic CRC cohort applying `Sgootr` - substituting NJ with FastME 2.0 - with different κ values.** Each row represents each of the 9 patients, and each column represents κ=.025, .05, .075, .1, .125, and .15. For κ=.025, .05, the iterative procedure was run for a maximum of 80 rounds; for κ = .075, .1, maximum 40 rounds; and for κ=.125,.15, maximum 20 rounds. Red vertical lines marks the iteration `Sgootr` identifies as optimal.

A.6 Patient CRC01 tumor lineage tree constructed using mutation

Tumor lineage tree reconstruction is typically based on mutation calls in single cells, either through the use of genomic data or whole transcriptome data. However, calling mutations from single-cell transcriptomic data without whole exome data has many false negatives/positives, rendering the results less than reliable. Here, we construct a tumor lineage tree with SNV and insertion-deletion (indel) mutations derived from matching scRNA-seq data on patient CRC01 from the Bian cohort. We show the mutation-based tree suffers from sparsity and renders subsequent interpretation of tumor progression difficult, which motivates construction of tumor lineage trees based on richer signals from single-cell methylation data using tools such as `Sgootr` proposed in this work.

Constructing mutation-based lineage trees calls for an approach different from the methylation lineage tree reconstruction framework proposed in this work, because there are typically notably fewer observed mutation events than methylation events. The mutation lineage tree we construct is based on SNV and INDEL mutations derived from matching scRNA-seq data [43] through the use of GATK HaplotypeCaller [189] pipeline with default parameters followed by filtering germline variants (from matched bulk whole genome sequencing data), and SNPs from dbSNP (build 138, the reason we used this verison is first it was recommended by GATK best practices and second it mostly contains benign SNPs with high allele frequency to avoid including pathogenic SNPs with low allele frequency), and the 1000 Genomes database. We used ANNOVAR [190] to annotate genetic variants, and filtered those variants whose status was unknown (i.e. only synonymous and nonsynonymous variants were kept). In addition, all private mutations (i.e. mutations reported to be present in only one cell) were excluded as they are not informative for lineage tree reconstruction. Next, to reduce the artifacts and false positive errors of RNA sequencing, the mutations that were present in the bulk WGS of the same lesion were

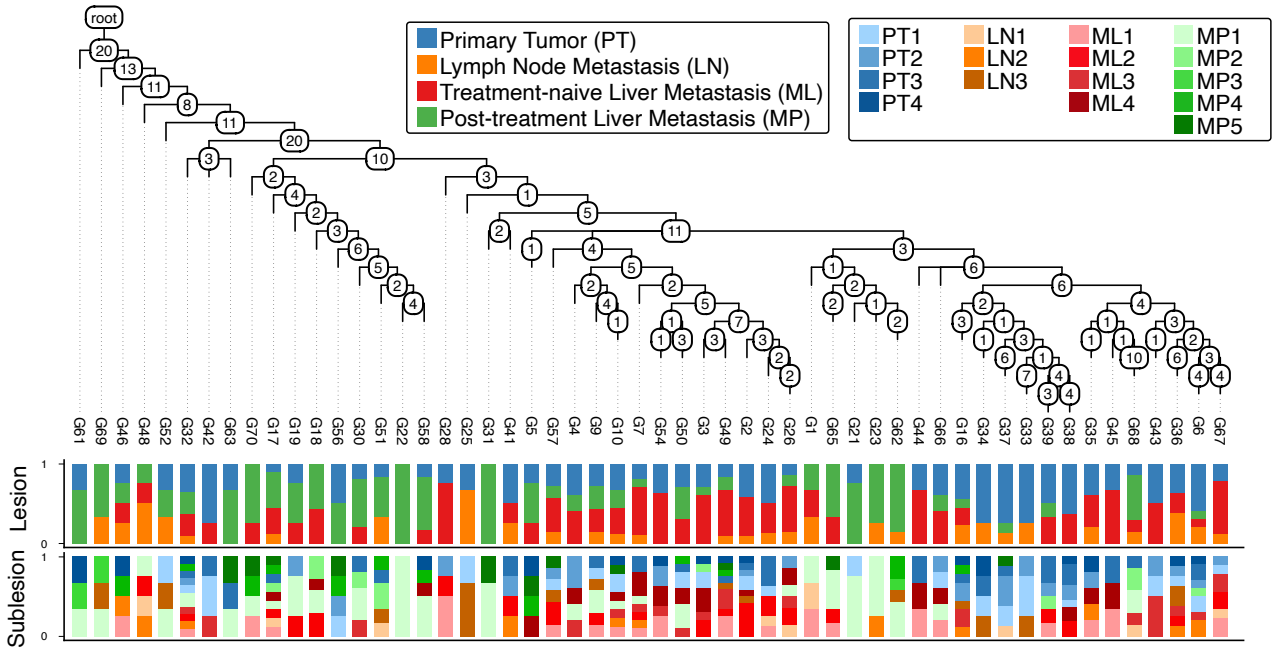


Figure A.18: **Tumor lineage tree based on SNVs/INDELs obtained from scRNA-seq data from patient CRC01.** Leaves are labeled by clusters of cells and numbers shown in rounded rectangular nodes represent number of *de novo* mutations shared only by the set of cells that are descendants of these nodes. The middle and bottom panels are stacked plots of the proportion of cells in leaves with respect to the tissue and sampling location, respectively. Note that the normal adjacent colon cells contain no mutations and therefore are part of the root in this lineage tree.

kept for downstream analysis resulting in 290 mutations in total. Finally, a local clustering of cells was performed for reducing the sparsity of the data - as a result each leaf depicted in the tree represents a combination of multiple cells with similar mutational profiles (Supplemental Figure A.18). The final lineage tree on these cell clusters was obtained by applying Trisicell [91].

A.7 Application of `Sgootr` to multi-regional GBM patient MGH105 MscRRBS data by Chaligne and collegues.

We apply `Sgootr` to the MscRRBS data from MGH105, the only multi-regionally-sampled GBM patient from the Chaligne cohort [44]. Single tumor cells are obtained from 4 sampling lo-

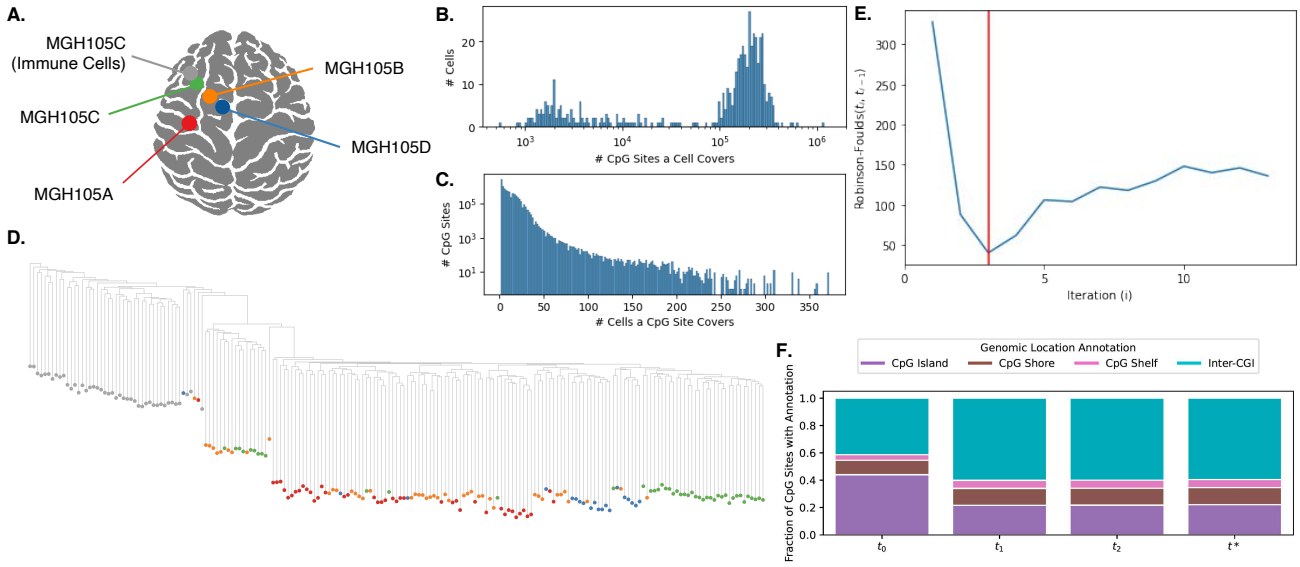


Figure A.19: **Application of `Sgootr` to patient MGH105 MscRRBS data by Chaligne and colleagues. A.** The multi-regional patient data include single cells from 4 distinct sampling locations (in addition to immune cells from sampling location MGH105C) **B.** Cell coverage distribution among autosomal CpG sites in patient MGH105. Note that the x-axis is in log scale. **C.** Autosomal CpG site coverage distribution in patient MGH105. Note that the y-axis is in log scale. **D.** Tumor lineage tree constructed by `Sgootr` with $\kappa = .1$ and each single cell at the leaves colored by its sampling location. **E.** RF distances across 14 iterations, with global minimum occurring at $i = 3$, the tree at which is shown above. **F.** Fraction of CpG sites located in CpG island, CpG shore, CpG shelf, and inter-CGI regions in the MGH105 trees at each iteration of the pruning procedure of `Sgootr`, from t_0 to $t*$.

cations within the same lesion: MGH105A, MGH105B, MGH105C, and MGH105D; additionally, non-tumor immune cells are sampled from MGH105C (Supplemental Figure A.19). Since both per cell and total CpG coverage in MscRRBS dataset are lower in comparison to scBS-seq (Supplemental Figure A.19B.,C.), we modify the parameters of `Sgootr` such that it: selects cells covering no fewer than 200,000 CpG sites, selects autosomal sites covered in no less than 5% of the remaining cells, bypasses the post-filtering step, and relaxes the pruning parameters so that $\delta = .02, \omega = .05$. We root the trees with an immune cell from sampling location MGH105C.

`Sgootr` obtains the optimal lineage tree $t* = t_3$ (Supplemental Figure A.19D.,E.), and the clustering implied by the tree topology seems to highly correspond to the sampling locations of the

single cells. We attempted to construct a lineage tree on the same set of cells and sites as we use here, with the set of methods Chaligne and colleagues leveraged in the original work, namely FG+IQ-TREE; however, FG fail to terminate within 100 hours hence no result is reported.

A.8 How to choose p, P(1^c), P(0^c), and P(mixed) for `Sgootr`

While in our experiments, we default to $p = .5$, P(1^c)=P(0^c)=P(mixed)=.33, it is possible for users to choose other values with additional pieces of prior knowledge. Here we offer examples of when the users may choose other values for those parameters.

The user may set $p > .5$ if there is reasonable belief that, given a pair of alleles truly heterozygous for a CpG site, the sequencing technology will sample from the methylated allele with higher probability than from the unmethylated one. This may manifest in a given dataset as a consistent observation of a higher number of methylated reads than unmethylated reads at diploid heterozygous sites. Similarly, the user may set $p < .5$ if there is reasonable belief that the unmethylated allele is more likely sampled given a heterozygous pair. If strong evidence for such sequencing bias is present, the user should adjust the value of p accordingly.

The user may have P(1^c)=P(0^c)>>P(mixed), if there is reasonable belief that true heterozygous CpG sites are rare. In this case, at low-coverage CpG sites, the observation of reads of a single status will give diminished probability that the site is truly heterozygous in the expected distance computation. In another potentially extreme example, the user may set P(1^c)>>P(0^c)=P(mixed), if there is reasonable belief that the homozygously methylated status is prevalent, and homozygously unmethylated and heterozygous statuses are equally rare. In this case, at a particular low-coverage CpG site, heterozygous status will only be given diminished probability if all observed reads are methylated;

however, if all observed reads are unmethylated, some larger-than-diminished probability will still be given to the heterozygous status.

A.9 How to choose κ for `Sgootr`

In the main text, we present `Sgootr` results with $\kappa = .1$. Having also experimented with a range of κ values, namely $\kappa = \{.025, .05, .075, .1, .125, .15\}$ (Supplemental Figures A.11,A.17), we provide some guidance on how to choose κ for future applications.

Recall that our goal is to detect a point in the iterative procedure where most non-lineage-informative sites have been pruned out but most lineage-informative sites still remain. Translating the objective into measurable terms, we return the last iteration where the global minimum RF distance occurs, after the initial decrease and before the ultimate increase.

When the RF distances across iterations appear flat, or have a region where RF distances have small fluctuations, and iterations where global minimum occur are far from unique - as one can observe across many patients for $\kappa = .025, .05, .075$ columns in Supplemental Figures A.11,A.17 - κ is likely too small. With small κ, `Sgootr` prunes out fewer sites at each iteration. Intuitively, RF distances between trees constructed from more similar sets of sites will be lower, possibly explaining the over all low RF distances in many instances. In those cases, the turning point in the RF distances becomes difficult, if not impossible, to observe, and a higher κ is called for.

While higher κ often lead to clear identification of a unique global minimum in RF distances across iterations - as one can observe across many patients for $\kappa = .125, .15$ columns in Supplemental Figures A.11,A.17 - it bares the risk of over-pruning. One may consider attempting lower values of κ while the V-shaped RF distance trend can still be observed.

Of course, determining the maximum number of iterations to run goes hand-in-hand with determining κ. If resource availability allows, we recommend choosing a number of iterations such that a consistent increase in RF distance towards the later iterations can be observed.

A.10 Experiment details

In this section, we provide details and parameters used to generate the results presented in this work.

A.10.1 `Sgootr`

While we specifically mention in main text the parameters used to generate the `Sgootr` results presented in this work, see `config.yaml` at the GitHub repository `https://github.com/algo-cancer/Sgootr` for the full set of default `Sgootr` configurations in one place.

A.10.2 FastME 2.0

In constructing a tumor lineage tree from a distance matrix using FastME 2.0 instead of NJ, we ran FastME 2.0 with the following command: `fastme -i {.mat distance matrix file} -o {.nwk output} -n -s`

A.10.3 IQ-TREE

We ran IQ-TREE with the following command, referencing what was done in prior work by Gaiti and colleagues [42] and Chaligne and colleagues [44] : `iqtree2 -s {.phy file} -st BIN -m TESTNEW -mset GTR2 -nt AUTO -bb 1000 -wbtl -abayes -alrt 1000 -nstop 500 -wsr -nm 2000 -v -alninfo -seed {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}`

A.10.4 FG

For FG, we follow the method descriptions by Gaiti and colleagues [42]. Given the cell-by-site matrices for methylated and unmethylated reads, we first obtain the binarized methylation status for each covered site in a cell via the procedure described in the Results section. We further prune out sites with coverage in less than 5 cells. For each remaining site s, we randomly downsample the other sites by 100x, and count the number of times s forms four gametes with sites in the downsampled set, disregarding the pair if they are within a read length of each other (150bp for Bian cohort, and 50bp for Chaligne cohort) or do not have at least 4 cells with coverage for both sites. For each site, we divide the number of occurrences of four gametes by the total number of pairings, obtaining the FG frequency.

The sites are binned into intervals of width 0.05 with regard to its methylation rate (number of cells with methylated status over total number of cells with read coverage for that site), disregarding bins with rates <0.1 or >0.9, and for each bin compute the median FG ratio. Within each bin, sites whose FG frequency is 1.5x median absolute deviation away from the bin median frequency were selected. The aggregate set of sites across all bins form the output of FG.

While random downsampling is necessary for problems of the scale in this work, it also intro-

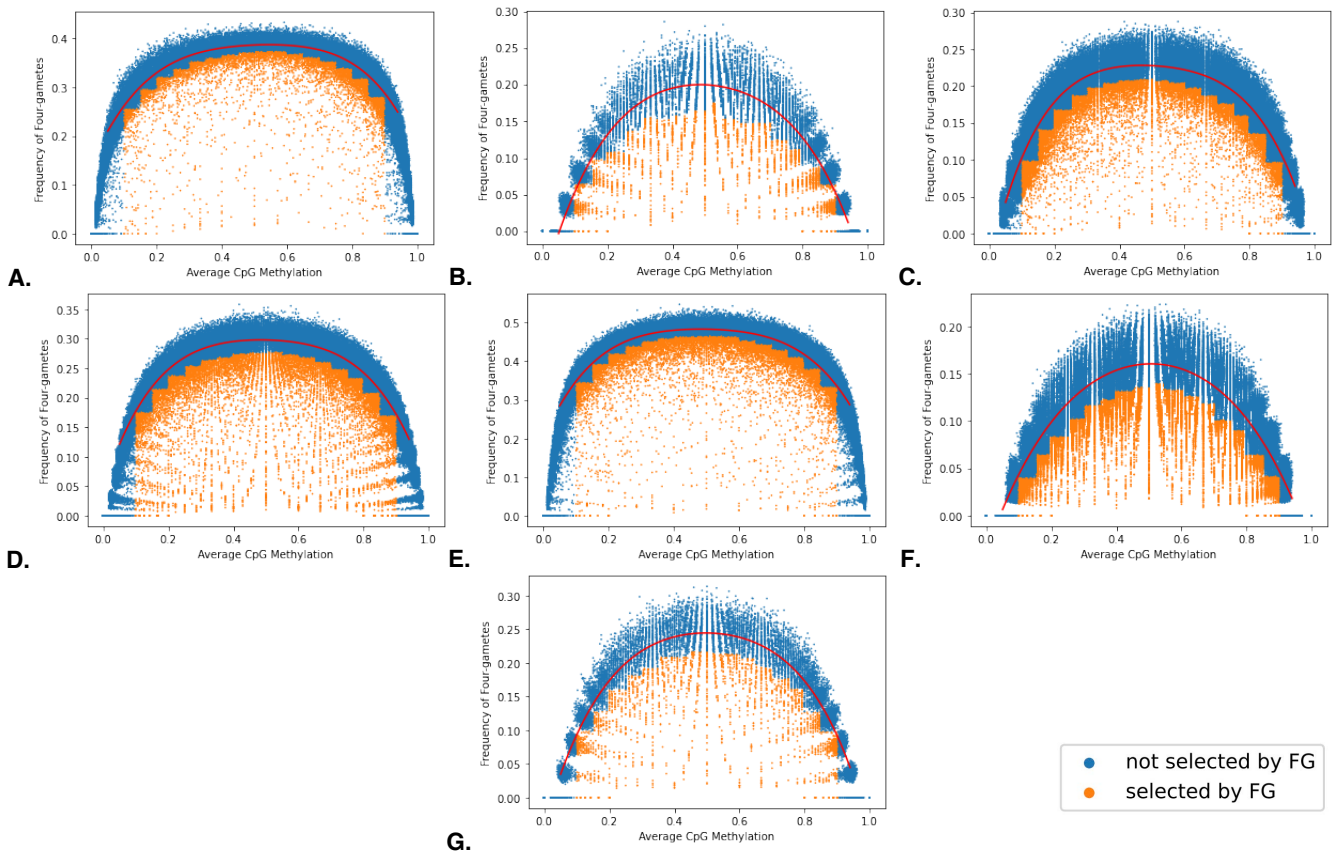


Figure A.20: **FG frequencies (averaged across 5 runs) for A. CRC01, B. CRC02, C. CRC04, D. CRC10, E. CRC13, F. CRC14, G. CRC15.** The red line denote the smoothed line connecting the median FG frequency within each bin.

duces stochasticity. Like Chaligne and colleagues [44] , we perform 5 separate runs of FG for each experiment, and found high concordance of the per site FG frequencies across the runs (Supplemental Figure A.21). For each site, we compute the mean FG frequency across 5 runs to use for the selection process described above (Supplemental Figure A.20).

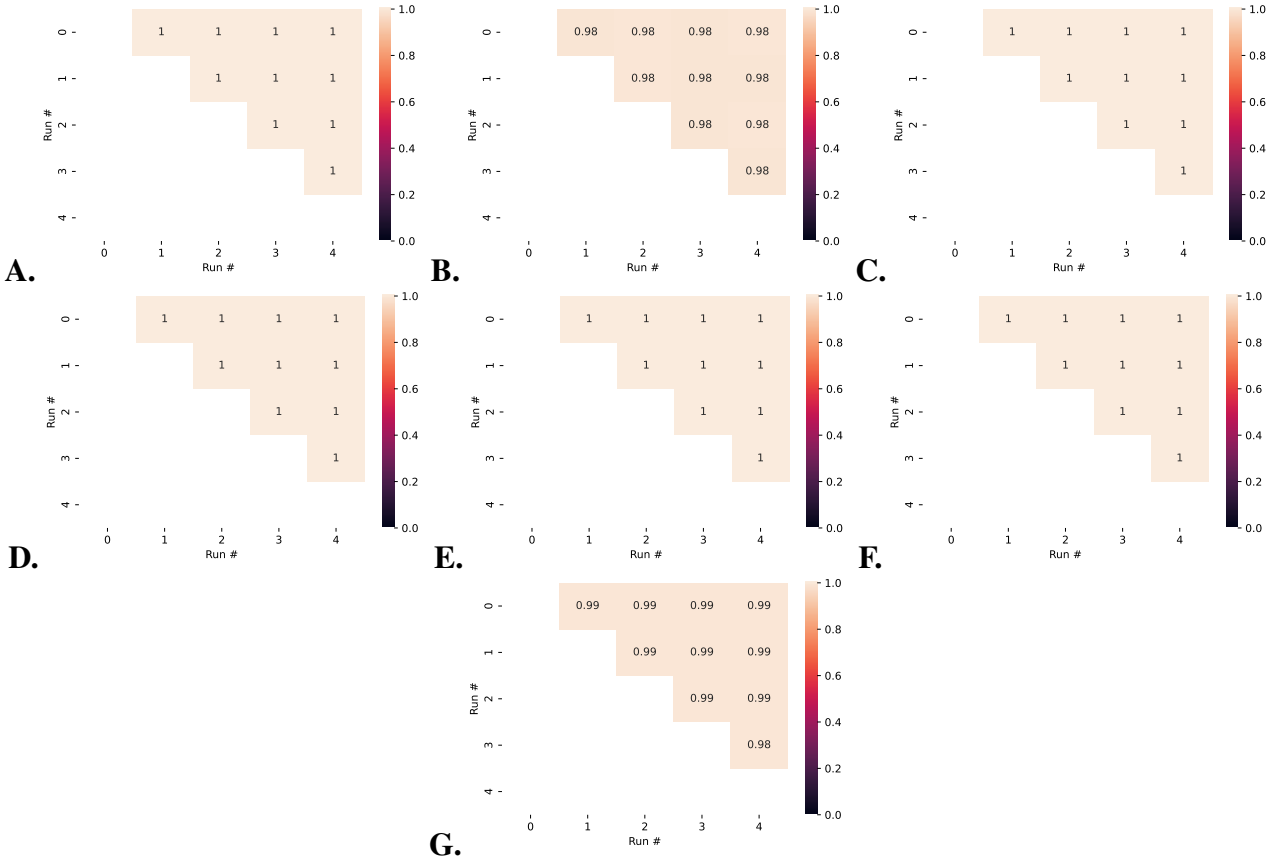


Figure A.21: **Pearson product-moment correlation coefficients between FG frequencies across sites between downsampled runs for A. CRC01, B. CRC02, C. CRC04, D. CRC10, E. CRC13, F. CRC14, G. CRC15.** Results from each pair of runs from a single experiment exhibit high concordance.

Appendix B: Supplemental Material for Chapter 3

B.1 Methods

B.1.1 Constructing tumor phylogeny using long read sequencing data from 23 sublines

To infer binary cell-linage tree of 23 sublines using `ScisTree` [123], we obtained a subset of highly reliable SNVs. For that, we first filtered out (i) SNVs in regions that were impacted by copy number losses in at least 1 subline, as such mutations are subject to the violations of the infinite-sites assumption and (ii) private (present in only one subline) or clonal (present in all sublines) SNVs, as these are not informative for phylogenetic analysis. We note that although some recently developed methods allow mutational losses [191–193], our filtered SNV set was large enough for reliable tree construciton using traditional methods. After that, for each mutation in each subline, we set the mutation status to "missing" if the coverage at the site is ≤ 10, deferring to the tree building algorithm to determine the presence-absence status of the mutation in the subline due to insufficient coverage. To avoid sites with consistently low coverage, as their mutation calls are unreliable across many sublines, we also filtered out sites with mean or median coverage across all sublines < 20. In addition, we also removed mutations with mean or median VAF across all sublines in which they were called as present below $1/3$, a value lower than the expected VAF for most non-private heterozygous SNVs in our dataset (which is typically 0.50 for mutations from diploid regions). Lastly, we observed that some SNVs had some, but low, read support for the alternate allele in a significant fraction of sublines for

which they were called as absent. Such SNVs are typically unreliable and require the tree building algorithm to make many false positive/negative corrections. We suspect that such ambiguous SNV calls are the result of various sequencing or mapping artifacts. To reduce the impact of such calls, we ignored sites for which there was read support for the alternate allele in at least half of the sublines in which the SNV was called absent. After the above filtering steps, we were left with 10,862 high-quality, phylogenetically informative somatic SNVs that were used for tree inference.

To construct the phylogenetic cell lineage tree from the filtered set of SNVs, we used `ScisTree` [123], a neighbor-joining based method that performs search in the space of cell-lineage trees. It is suitable for our dataset which has significantly lower number of cells compared to the number of SNVs. We performed a grid search over a range of false positive (10^{-2}, $5 \cdot 10^{-2}$, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}) and a range of false negative ($.05$, $.075$, $.1$, $.125$, $.15$, $.175$, $.2$) priors. In addition, as `ScisTree` performs a heuristic search, its output can depend on the order of the rows (cells) in the input. Therefore, we also repeated the tree building process for 100 different permutations of the input rows. As the final output, we selected the tree with the maximum likelihood among all the 8 (false positive priors) $\times$ 7 (false negative priors) $\times$ 100 (permutations of rows) $=$ 5600 trees inferred by `ScisTree`.

B.1.2 Determining the phenotypes of the 4 major clades

Here we provide a brief summary of the previous work that characterized the phenotypes of the B2905 subclones [116]. After generation from parental B2905 melanoma cell line, the work performed the following characterization studies: (1) each subline was subjected to exome sequencing and scRNA analysis; (2) each subline was tested for *in vivo* growth in C57BL/6 mice; (3) selected sublines were tested for response to ICB in C57BL/6 mice.

Feature \ Clade	**Green**	**Orange**	**Blue**	**Red**
Growth rate in C57BL/6 mice	++	+	+/-	+++
Response to host immunity (Delay of growth initiation in C57BL/6)	High	Medium	Very high	Low
ICB response (anti-CTLA4 vs. IgG)	Medium	High	(growing too slow for study)	Low
Cell cycle gene expression	Medium	High	Medium	Low
Wnt signaling gene expression	Mixed	Canonical		Non-canonical
Phenotype	**Mixed, Less Sensitive to ICB**	**Proliferative, Sensitive to ICB**	**Immunogenic, Susceptible to Host Immunity**	**Invasive, ICB Resistant**

Table B.1: **Phenotypes of the 4 major clades.** The phenotype for each of the major clades was determined by features defined by the behavior or characteristics of (select) sublines from the clade previously characterized [116].

As defined by the criteria by Hirsch et al. [118], invasive phenotype exhibits faster tumor growth when implanted in immunocompetent host, lower expression of cell cycle genes, and expression of non-canonical Wnt signaling genes; meanwhile, proliferative phenotype exhibits slower tumor growth when implanted in immunocompetent host, higher expression of cell cycle genes, and expression of canonical Wnt signaling signature. Given the relevant results from the characterization study [116], accordingly, the orange clade was determined to be proliferative, the red clade invasive, and the green clade exhibiting intermediate levels of these features was defined as having a mixed phenotype (Supplementary Table B.1).

The response to host immunity (immunogenicity) of each subline was estimated by DGI (the time from implantation of cancer cells to starting continuous growth without regression) in immunocompetent C57BL/6 mice implanted with each subline [116]. The longer period of DGI indicates stronger response to host immunity. DGI diminished when cancer cells were implanted to immunocompromised host, as compared to that in immunocompetent host, confirming the association of DGI with tumor response to host immunity (Supplementary Figure B.1).

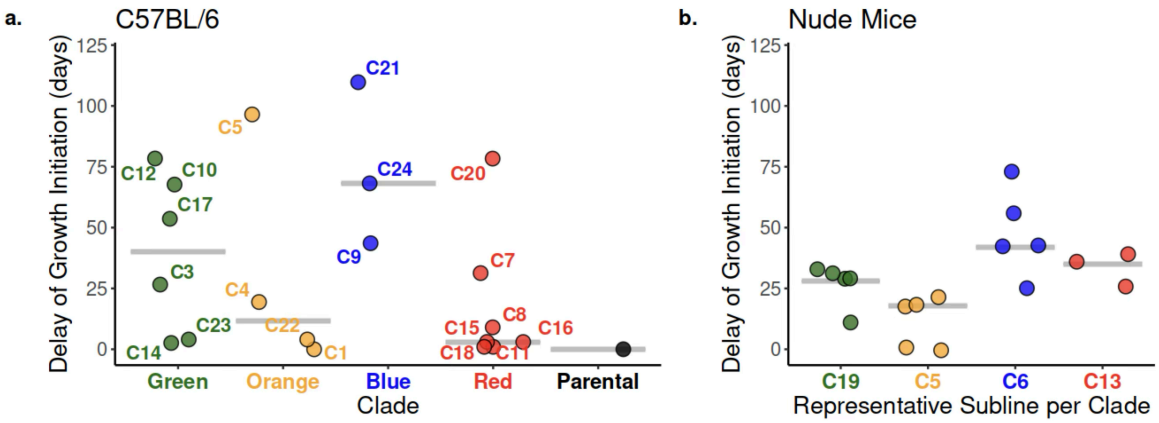


Figure B.1: **DGI of sublines in C57BL/6 and nude mice.** DGI is defined as the number of days between cell implantation in a mouse and the tumor started continuous growing without regression. **a.** One million cells of each subline were implanted into each of the five C57BL/6 mice (Day 1). The tumor size was measured twice a week. The mean DGI of the sublines in each clade, as well as the median across sublines means within a clade, was shown. No tumor was grown from any mice implanted with cells of C19, C6, or C13. Note that two mice implanted with C5 cells did not grow tumor, and they were not included in the calculation of mean DGI. These results suggested that, on average, sublines in blue clade induced the stronger host immunity than those in other clades. **b.** To investigate the association of the DGI with host immunity, the subline from each clade with the highest DGI following C57BL/6 implantation was tested in immunocompromised host. For each of C19 (green clade), C5 (orange clade), C6 (blue clade), and C13 (red clade), one million cells were implanted into each of the five immunocompromised nude mice (Day 1), and the DGI values measured from each mouse, as well as the median across mouse DGI values for a subline are shown. As expected, we observed the tumor grew in all mice with diminishing DGI as compared to the implantation in C57BL/6 mice. These results indicated that DGI is associated with subline response to host immunity.

To test the response to ICB, one or two sublines were selected from green, orange, and red clade [116]. The blue clade was not tested because of low tumor take rate (the odds that implanted cancer cells can grow tumor in the hosts). For each selected subline, its cells were implanted into ten C57BL/6, which were randomized to received anti-CTLA4 or isotype control antibody treatment. Taken together, these results of the characterization studies label each major clade with a distinct phenotype (Supplementary Table B.1).

B.1.3 Assessing the effectiveness of `TreeHarmonizer` in determining the placement of variants impacted by filtering during tree construction

The partition function algorithm [119] uses a tree sampling strategy to estimate the probability that, given a set of mutation calls, a mutation should be placed at the most recent common ancestor of a clade. To measure whether `TreeHarmonizer` effectively re-determines the placement of variants impacted by stringent filtering criteria during tree construction, we compute for those variants the respective partition function values (i) at their placement determined by `ScisTree` [123], and (ii) at their placement by `TreeHarmonizer`, with the trees sample from the space defined by the original set of variant calls prior to filtering.

We focused on the set of variants whose placement differed after applying `TreeHarmonizer`. By definition, if two variants are called in an identical set of sublines in the unfiltered data, their partition function values at a given placement in the tree will be the same if the values are estimated on the same set of sampled trees. Therefore, instead of computing partition function values per variant, we compute per group of variants that (i) were called in an identical set of sublines in the unfiltered data, (ii) shared the same placement by `ScisTree`, and (iii) shared the same altered placement by

`TreeHarmonizer`. We computed the partition function values with 1000 sampled trees, 0.001 false positive rate and 0.15 false negative rate, and the results on the 27 groups are reported in (Figure B.5g.). One-tailed paired-sample *t*-test was applied to the partition function values of the 27 groups.

B.1.4 Mutational signature analysis along the tumor phylogeny

The cumulative placed SNV profiles of the 23 sublines were used as input samples to the signature refitting tool `SigProfileAssignment (v0.1.9)` [194], a mutational signature fitting tool chosen for its superior overall performance and specifically for the number of mutations per sample at hand [195]. The 23 SNV profiles were translated into mutational spectra over 96 categories for joint refitting, and COSMICv3.4 [127] mm10 signatures were used as the reference catalog, as it has been shown that excluding sets of signatures from refitting often leads to poorer performance [195].

`SigProfilerAssignment` outputted signature activity estimates per subline, as well as signature probability per individual mutation per subline. Each mutation placed at a particular branch was assigned the most probable signature given the aggregated signature probabilities for that mutation in sublines in the subtree below the branch. The activity of a mutational signature at a branch was then estimated by the proportion of mutations placed at the branch assigned to the signature. For concision of presentation, activities for signatures SBS7a, SBS7b, SBS7c, and SBS7c were aggregated and referred to as signature "SBS7", and those for signatures SBS17a and SBS17b were aggregated and referred to as signature "SBS17".

For each active signature, to assess whether signature activities have significant differences according to their evolutionary timing or between clades, each branch in the tumor phylogeny is labelled by (i) truncal, internal, terminal, and (ii) the color of the clade it is in (the trunk and two branches

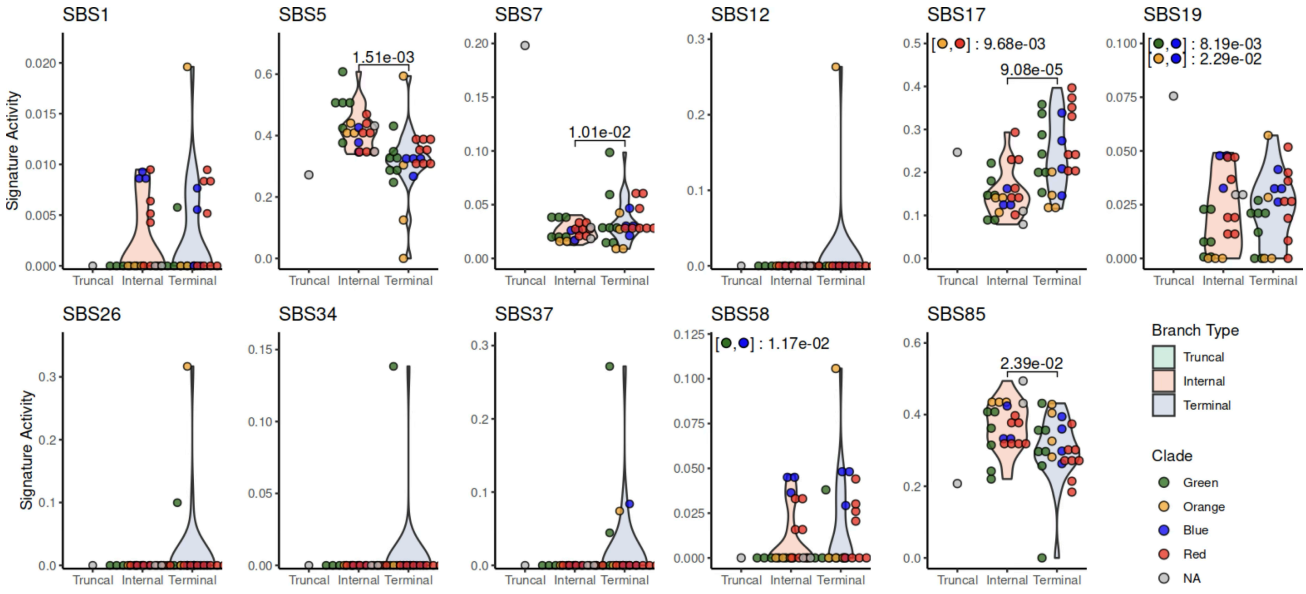


Figure B.2: **Mutational signature activities along the tumor phylogeny.** In each panel, a point represents a branch in the phylogeny. A branch can either be categorized into truncal, internal or terminal, or by the color of the clade it is in. ANOVA with post-hoc Tukey HSD Test was performed, and the adjusted p-values of statistically significant pairs are reported.

closest to the trunk have NA as clade assignment), and ANOVA with post-hoc Tukey HSD test was performed to see if there were significant differences between pairs along either independent variable. Since there is only 1 truncal branch, and the 3 branches closest to the root are only grouped together due to lack of clade labels, numerically significant results (adjusted p-value $< .05$) involving those as a group in the pair are not reported. In other words, only significant results between internal and terminal branches, or between pairs of clades are reported (Supplementary Figure B.2).

B.1.5 Computing dN/dS values along the tumor phylogeny

Placed SNVs were annotated using `Ensembl Variant Effect Predictor (v102)` [196]. For any given branch, we computed the dN/dS value as the ratio of cumulative non-synonymous SNVs placed along the path from the trunk and including the current branch to that of cumulative synonymous SNVs, normalized by the ratio of synonymous sites to non-synonymous sites [118]. dN/dS

values at internal branches were found to be significantly less than those at terminal branches via a one-tailed t-test (Supplementary Figure B.3a., $p < \alpha = .05$).

B.1.6 Pathway enrichment analysis

Given a set of genes of interest - either that they are impacted by a certain type of variant at a particular time during the tumor evolution (e.g. placed at the trunk of the phylogeny) or in a particular clade - we performed Gene Ontology over-representation analysis to examine their potential functional impact [197, 198]. Genes without a canonical name ("*Gm-*" and "*-Rik*" genes) were excluded from analysis. Using `clusterProfiler (v3.21)` [199], we tested whether the candidate gene set significantly over-represents a Gene Ontology Biological Process gene set between sizes of 3 and 500. The resulting p-values were adjusted using the Benjamini-Hochberg procedure, and terms with q-value $< .05$ were determined to be statistically significant.

B.1.7 Determining genes impacted by SVs

We used `Padfoot` (https://github.com/KolmogorovLab/Padfoot, commit ID:2d5237b) to annotate SV calls generated by `Severus` using the `--specie mouse` option and GRCm38.p4 GFF3 files with basic gene annotations. The *possible_fusion* and *exon_altering* events were manually validated using IGV and scRNA-seq alignments.

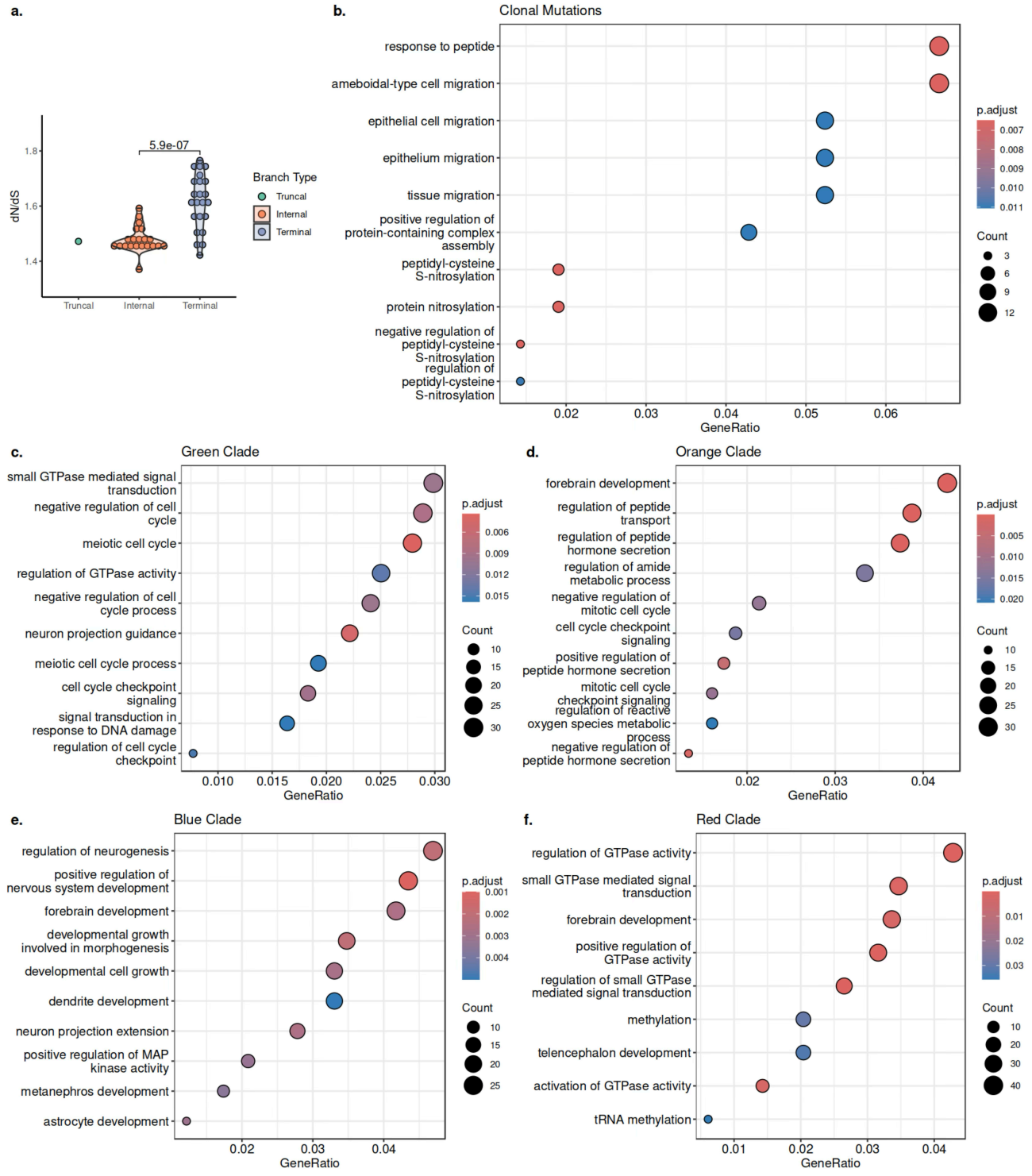


Figure B.3: **SNVs along the tumor phylogeny. a.** dN/dS value distributions among truncal, internal, and terminal branches. dN/dS values at internal branches are significantly lower than that at terminal branches by one-tailed t-test. **b.** Select enrichment of genes harboring non-synonymous clonal mutations among GO Biological Process term. **c.-f.** Select enrichment of genes uniquely harboring non-synonymous mutations in the green (c.), orange (d.), blue (e.), and red (f.) clade among GO Biological Process terms.

B.1.8 Determining the set of cancer-relevant mouse genes-of-interest according to the COSMIC Gene Census

To obtain a set of mouse genes that likely have functional impact in cancer, we selected those with one-to-one homology as genes curated in the COSMIC CGC (v101) [127]. Genes whose COSMIC homologs are uniquely labeled as ONC in CGC are labeled as ONC, and those whose COSMIC homologs are uniquely labeled as TSG are labeled as TSG. Those whose COSMIC homologs have no unique ONC/TSG labelings in CGC are labeled as ND. Among those without unique COSMIC ONC/TSG labelings, 4 genes (*Ezh2*: ONC [200]; *Rad21*: TSG [201]; *Tert*: ONC [202]; *Trp53*: TSG [203]) were given unique labels according to literature on melanoma or other cancer types. A gene is said to be impacted by CNA in a subline if it is entirely contained in a copy number gain or loss segment.

To determine if the gain or loss of a gene may be associated with its designation as ONC or TSG, we performed two χ^2 tests on the set of COSMIC ONC/TSG genes impacted by CNA in any subline. One χ^2 test evaluated enrichment of ONC that were gained in some subline but not lost in any sublines. The other performed the corresponding test of TSG lost in some subline but not gained in any subline.T he result was considered significant if $p < \alpha = .05$.

To see whether selection may be at play, we further examined the distribution of ONC/TSG around CNA boundaries. To that end, we specifically focused on COSMIC homolog pairs that are adjacent to each other along the genomic coordinate, where one is labelled ONC and the other TSG and have different absolute copy numbers in any subline, as they correspond to the nearest ONC/TSG COSMIC homologs upstream/downstream to a copy number change boundary. We counted in total 14 instances where ONC homolog is on the side of the boundary with higher copy number state, and TSG

homolog is on the other. To assess the statistical significance of the observation, a null distribution was generated by counting instances of such observation in 5,000 permutations of the ONC/TSG/ND labels of the COSMIC homologs. The p-value is the fraction of times the permutated labelings yield a greater or equal number of such observations than the actual labeling. We reject the null hypothesis if $p < \alpha$=.05.

B.1.9 Timing analysis of chr4 copy-neutral LOH

The copy number calls from `Wakhan` [121] suggest whole-chromosome copy-neutral LOH on chr4. Compared to other autosomes, there is a large proportion of clonal SNVs on chr4 that are homozygous, or in other words, have a median VAF of 1 across sublines evenly distributed across the chromosome (Supplementary Figure B.4a.,b.). The number of such chr4 clonal SNVs provide evidence for the approximate timing of the underlying chromosomal events. Copy-neutral LOH is likely the result of one of the following two scenarios: (i) LOH followed by whole-chromosome duplication of the remaining haplotype, and (ii) mitotic nondisjunction followed by LOH in the trisomic cell; nevertheless, the scenarios are indistinguishable given current data because neither a subline with one nor a subline with three copies of chr4 was observed.

In the first scenario, we assume the copy-neutral LOH is a result of LOH followed by duplication (Supplementary Figure B.4c.). Without additional information, we cannot distinguish SNVs obtained before the LOH event and those obtained after LOH and before duplication, as in either case they would appear homozygous; that said, the number of observed homozygous SNVs upperbounds the number of SNVs obtained prior to LOH. Assuming the rate of SNV acquisition per genomic length is constant, the total number of SNVs acquired prior to whole-chromosome duplication is at most two times the

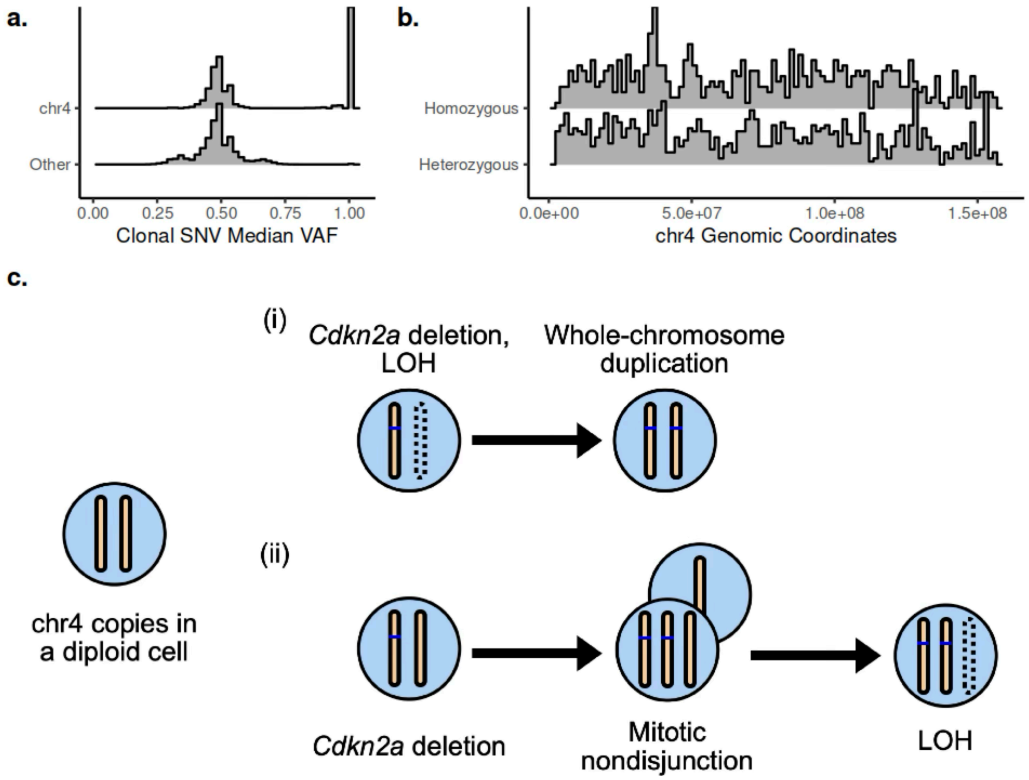


Figure B.4: **SNV evidence for the timing clonal CNA events. a.** Median (across sublines) VAF distribution for clonal SNVs on chr4 and other autosomes. **b.** Genomic distributions of homozygous (median VAF = 1) and heterozygous (median VAF < 1) chr4 clonal SNVs. **c.** Two indistinguishable scenarios that may have resulted in the copy-neutral LOH.

number of observed homozygous SNVs. Accounting for the unobserved SNVs in the lost copy in the total number of clonal SNVs, we also add the number of observed homozygous SNVs to the total number of observed chr4 clonal SNVs. The fraction of SNVs acquired prior to whole-chromosome duplication out of total clonal SNVs gives an overestimation of the timing of the duplication, which turns out to be .53, or around half way through the trunk of the tumor phylogeny.

In the second scenario, we assume the copy-neutral LOH is a result of LOH in the trisomic cell formed following mitotic nondisjunction (Supplementary Figure B.4c.). In this case, without additional information, we cannot resolve the timing of the LOH event from VAF distributions of observed clonal SNVs; however, the timing of mitotic nondijunction can be similarly estimated. The number of observed homozygous SNVs corresponds to the number of SNVs on the chr4 copy with Cdkn2a deletion in a diploid cell prior to mitotic nondisjunction. Again, assuming the rate of SNV acquisition per genomic length is constant, the total number of SNVs acquired prior to mitotic nondisjunction is roughly twice the number of observed homozygous SNVs. In the trisomic cell, the three copies of chr4 acquired SNVs independently. The number of observed number of chr4 clonal heterozygous SNVs, or in other words, those with median VAF < 1, corresponds to that acquired after mitotic nondisjunction, normalized such that the amount of genomic content is equivalent to two copies of chr4. The fraction of SNVs acquired prior to mitotic nondisjunction out of total SNVs acquired gives an overestimation of the timing of mitotic nondisjunction, which is again, .53.

The above analyses lead to the conclusion that, assuming the rate of SNV acquisition per genomic length is constant, the tumor cell first acquired two of the chr4 copy carrying *Cdkn2a* deletion half way through the trunk of the tumor phylogeny, at the latest.

B.1.10 Subclone-specific differential methylation analysis

Given modified base calls, modified base counts for each sample are aggregated using the tool `modbam2bed` [204] (`--aggregate --mod_base=5mC --cpg --mask`). The number of methylated reads (N_{mod}) and total read count (N_{mod} + N_{canon}) at each CpG site with methylation calls for the 23 subline samples are compiled for smoothing using `BSmooth` [122].

At a given branch in the phylogeny, the 23 sublines are bipartitioned into a case group, consisting of sublines in the subtree below the branch, and a control group, consisting of the others. With the case-control group assignment of sublines, we obtain a list of regions that are differentially methylated using `BSmooth` [122]. To reduce noise due to low coverage, only CpG sites covered by at least 3 reads in all sublines are considered. Signal-to-noise t-statistics were computed per CpG site using the smoothed methylation values, and a DMR is defined to be one that spans at least 10 consecutive CpG sites, with corrected t-statistics $< .025$ or $> .975$, and with absolute mean differences between groups at least .2. As replicates are required for DMR testing using `BSmooth`, DMRs were computed at branches where the bipartition of the sublines they induce both have a size of at least 2.

B.1.11 Permutation test on the number of DMRs obtained

Given a case group of a certain size, we used a permutation test to assess whether significantly more DMRs were identified in the case group than in the control group. A null distribution was generated by 1,000 (or, when the size of the case group is 2, all possible 253) random permutation of the case and control group labels over the sublines. We recorded the number of DMRs identified in each permutation using the permuted labels to implement the criteria described in the section above. The value of p is the fraction of permutations that yield a greater or equal number of DMRs than the

actual case-control groups of interest. We reject the null hypothesis if $p < \alpha = .05$.

B.1.12 Correlation between the number of DMRs and cumulative SNVs at a branch in the phylogeny

As the numbers of DMRs are not immediately comparable between tests involving case groups of different sizes, we focused on tests involving case groups of size 2 performed at branches parental to exactly 2 leaves, since there is a sufficient number (7) of such branches to assess correlation. The Pearson correlation was assessed between the cumulative numbers of SNVs from the trunk to the branch, and the number of DMRs detected between the 2 sublines below the branch and the rest of the 21 sublines. We reject the null hypothesis if $p < \alpha$=.05.

B.1.13 Differential gene expression analysis of subline scRNA-seq data

scRNA-seq data was obtained as previously described [118], and gene-level expected count values for the subline scRNA-seq data were obtained using `RSEM` [177]. As multiple cells were sequenced from each subline, the analysis treated them as biological replicates of the same subline. Genes with total count less than 2^{10} were removed, and the decimal expected count values were rounded to the closest digit. `DESeq2` [205] was used to assess differential gene expression, and the shrinkage estimator uses a normal prior (`coef=2`). The log2 fold change values for the genes of interest were obtained. Note that not all genes of interest have their expression values available in the scRNA-seq data set, and hence they are left out of the test for correlation.

B.1.14 Discovering key regions where changes in methylation is monotonic along the evolutionary trajectory

To obtain candidate DMRs to test for monotonicity, we collected regions that exhibited differential methylation among the sublines within the red clade that potentially exhibit such monotonic trend. As replicates are required for DMR testing with `BSmooth`, 4 DMR tests were performed with increasingly small subtrees as the case group (and the rest of the sublines in the red clade as the control group): first C11, C16, C8, C7, and C20; then C16, C8, C7, and C20; followed by C8, C7, and C20; and lastly C7 and C20. These groupings were chosen as they reflect cuts along the linear evolutionary trajectory in the red clade. The union of the 4 sets of DMRs were then taken, merging overlapping intervals. As this analysis is motivated by phenotypic observations, we focused our analysis on the regions overlapping (putative) promoters and enhancers of known genes, which we refer to as candidate DMRs.

For each of the candidate DMRs, the Mann-Kendall test was performed on four orderings of the sublines – the order in which the sublines were placed (C13, C18, C15, C11, C16, C8, C7, C20), as well as the three variations that may be formed by swapping immediate siblings (C18 and C15, C7, and C20).

Kendall's τ value with the greatest absolute value was recorded. We set a minimum threshold of .7 for the absolute value of Kendall's τ, as under assumptions it roughly corresponds to a Pearson correlation coefficient of .9 [206], which is commonly accepted to signify a strong magnitude of effect. Genes whose (putative) promoters or enhancers overlapped regions with an absolute value of Kendall's τ over the threshold .7 were kept. How the statistical significance of the discovery was assessed is discussed in the following section.

B.1.15 Permutation test on the number of genes associated with monotonic methylation changes along the evolutionary trajectory

We assessed the statistical significance of the number of genes we discover to be associated with monotonic methylation changes along the evolutionary trajectory in the red clade via a permutation test. Fixing the structure of the red subtree, we permuted the position of the 8 sublines in the structure. While there are 8! total permutations of 8 sublines, given the two sets of immediate siblings that are interchangeable in terms of order, and that there is symmetry in terms of the magnitude of correlation for the exact reverse of a particular order, for every permutation, there are 7 others that are equivalent for the purpose of our analysis. We obtain 1,000 permutations from the 7! equivalence classes, sampling no more than one permutation from each equivalence class.

For each permuted positioning of 8 sublines in the structure of the red subtree, we obtained the number of genes whose (putative) promoters or enhancers overlapped regions that show monotonicity in terms of methylation following the given order using the algorithm described in the previous section.

The p-value is the fraction of times the permuted orderings yield a greater or equal number of such genes than the actual ordering of the red clade suggested by Figure 3.2a. We reject the null hypothesis if $p < \alpha = .05$.

B.1.16 Calling copy number from bWES and Smart-Seq2 full-length scRNA-seq data on the 23 sublines

`CNVkit` [207] was used to obtain the CNA calls from bWES data. `CNVkit` was run with external bWES data from a normal spleen sample [152] as reference. The per-bin copy numbers were

shown in Supplementary Figure B.19, and the per-segment copy number calls were used to determine the copy number status (i.e. loss, neutral, gain) of available genes in a subline.

Full-length scRNA-seq data were obtained using the Smart-seq2 protocol [170] for several single cells per subline [118], and `inferCNV` [208] was used to obtain the CNA calls from the data. As Smart-seq2 data for normal melanocytes were not generated for the same study, `inferCNV` was run with external Smart-seq2 data from mature melanocytes [209] as reference (`cutoff=1`, `sd_amplifier=2`, `noise_logistic=TRUE` median filtering applied). The `inferCNV` i6 Hidden Markov Model was used to determine the copy number state, a posterior probability of .5 was used for filtering out likely normal observations, and a per-gene copy number status for each subline was derived.

B.2 Technology comparison: long-read, bWES, and scRNA-seq

Given the complex variational landscape of the B2905 evolution presented in Chapter 3, we sought to quantify the improvement in resolution enabled by long-read sequencing, as compared to the short-read bWES and scRNA-seq of B2905 subclones [118, 119].

Overall, long-read data had substantially better mappability and covered more variants (Figure B.5a.), consistent with other benchmarks [108]. For our purpose of studying tumor evolution, the advantage is two-fold. First, long-read sequencing offers signals across the whole genome instead of focusing on exonic, expressed regions. This leads to (i) a substantially higher number of profiled somatic SNVs (Figure B.5a.), (ii) identification of non-exonic SV breakpoints, and (iii) more contiguous estimation of copy number including non-exonic regions. Secondly, the reduction in mapping ambiguity avoids noisy signals due to mapping artifacts, amplifying real signals.

Using the available bWES and scRNA-seq data, we also validated the accuracies of the long-read

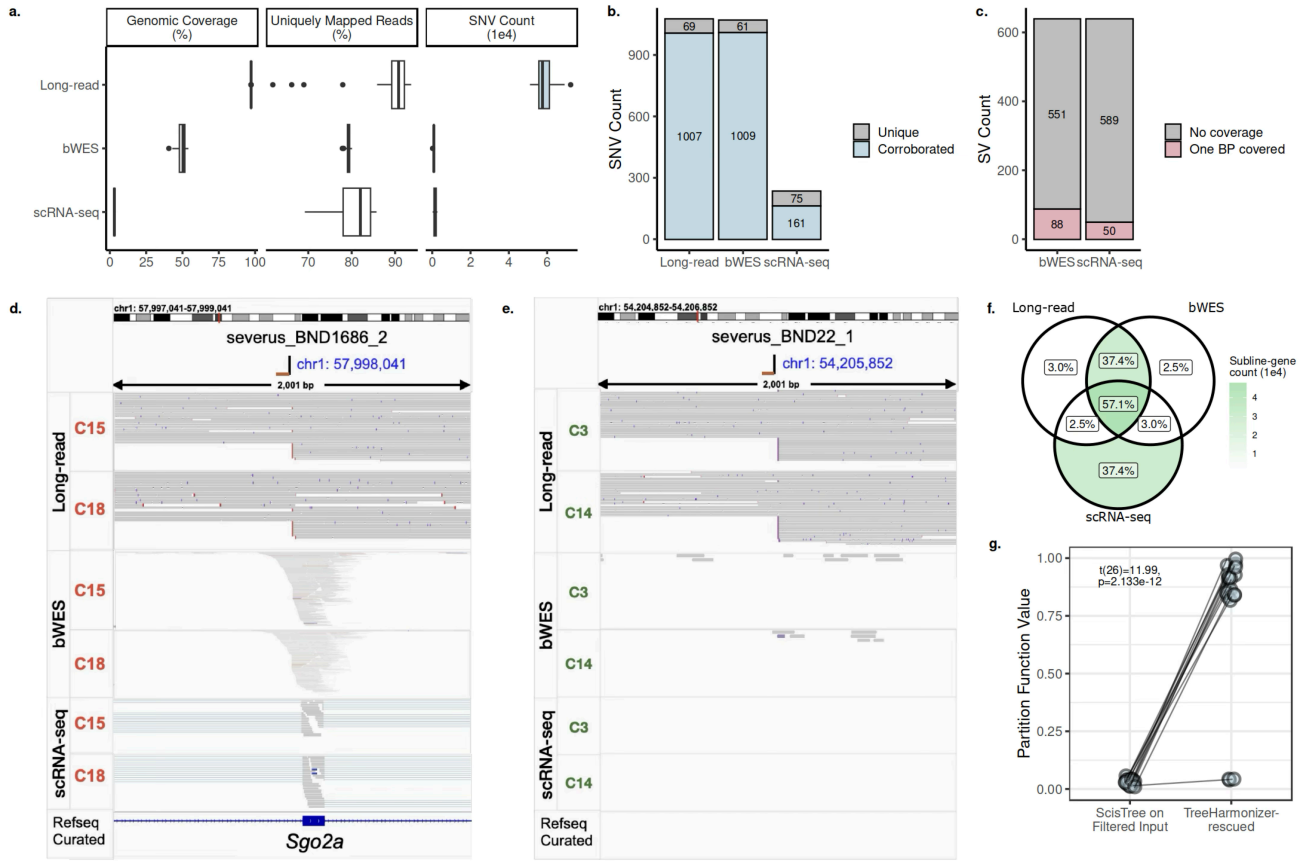


Figure B.5: **Comparison among long-read, bWES, and scRNA-seq data of the 23 sublines. a.** Per-subline autosomal genomic coverage (%), uniquely mapped reads (%), and count of unique somatic SNVs profiled. **b.** Counts of unique, non-private somatic SNVs called within bWES target regions across the 23 sublines for each technology. The counts are stratified by whether the variants' presence were corroborated by at least one other technology. **c.** Count of `Severus`-profiled SVs whose breakpoints have sufficient read coverage (200x after aggregating all reads from all sublines/cells) in bWES and scRNA-seq data. **d.** Example of genomic coverage at a near-exonic breakpoint associated with a subclonal SV across the 3 technologies. **e.** Example of genomic coverage at an intergenic breakpoint associated with a subclonal SV across the 3 technologies. **f.** Agreement on gene copy-number status (neutral, gain, loss) called from data from the 3 technologies across all sublines. For any subline, only genes with copy number calls available in all 3 technologies are included. **g.** Improvement in terms of partition function values [119] among variants impacted by filtering during tree construction, following variant placement by `TreeHarmonizer`.

profiled variants, quantifying the improvement in resolution when possible. Within bWES target regions, non-private somatic SNVs are $> 93\%$ consistent between long-read and bWES (Supplementary Figure B.5b.). Among those, $> 15\%$ are also corroborated by the scRNA-seq data. For each of the 639 `Severus`-profiled somatic SVs in long-read, we examined whether at least one of the breakends it is associated with has sufficient read coverage (200x after aggregating all reads from all sublines/cells) in bWES and scRNA-seq, as an overestimation of whether it can possibly be profiled in those technologies. Respectively, $> 86\%$ and $> 92\%$ of `Severus`-profiled somatic SVs have no breakends of sufficient coverage in bWES and scRNA-seq data (Supplementary Figure B.5c.), suggesting most SVs would have been missed in short-read data. We also profiled copy numbers from bWES and scRNA-seq data (Supplementary Section B.1.16) to gauge the accuracy of long-read calls. Looking across all sublines, among genes with copy number estimates available across all 3 technologies, $> 94\%$ had consistent copy number status (neutral, gain, loss) between long-read and bWES (Figure B.5f.). Among those, a majority also has its status corroborated in scRNA-seq calls, even though scRNA-seq calls suffer from noise and sparsity (Supplementary Figure B.5f., Supplementary Figure B.19).

Importantly, as the detection of parallel CNA events relies on the association of contiguous copy number changes with exact SV breakpoints, using long-read whole-genome sequencing is in practice necessary for that purpose. For example, the 2 key subclonal parallel CNAs observed on chr1 (Figure 3.4) in our work, distinguished by their respective breakpoints, both have insufficient support in bWES and scRNA-seq data at the breakpoint loci, regardless of whether it is near-exonic (Supplementary Figure B.5d.) or intergenic (Supplementary Figure B.5e.).

In addition, we validated the `TreeHarmonizer` placement of SNVs using the orthogonal partition function approach [119] that estimates the likelihood of mutation placement (Supplementary Section B.1.3). For the SNVs that were rescued by `TreeHarmonizer`, the partition function re-

ported significantly higher scores (one-tailed paired-sample t-test $\mathrm{t}(26) = 11.99, p = 2.133e-12$), indicating accurate placement in the phylogeny (Supplementary Figure B.5g).

B.3 Additional Supplementary Figures and Tables

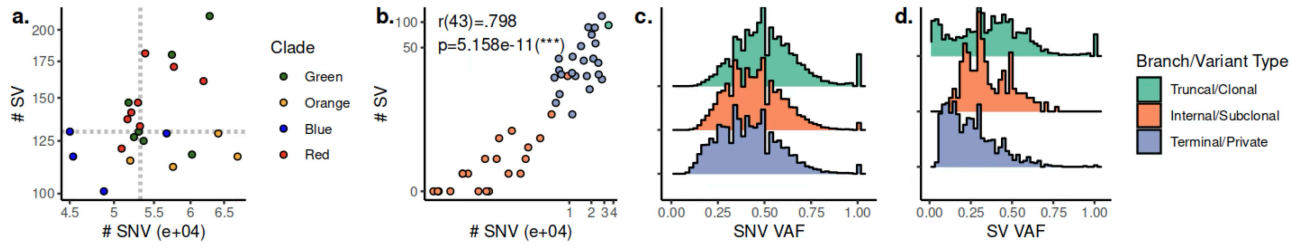


Figure B.6: **SNVs and SVs placed by `TreeHarmonizer`. a.** Cumulative number of placed SNVs and SVs per subline, colored by clade. Medians are marked by dashed lines. **b.** The number of SNVs and the number of SVs placed at each branch of the phylogeny strongly correlate with each other. **c.** Variant allele frequency distributions of placed SNVs. **d.** Variant allele frequency distributions of placed SVs.

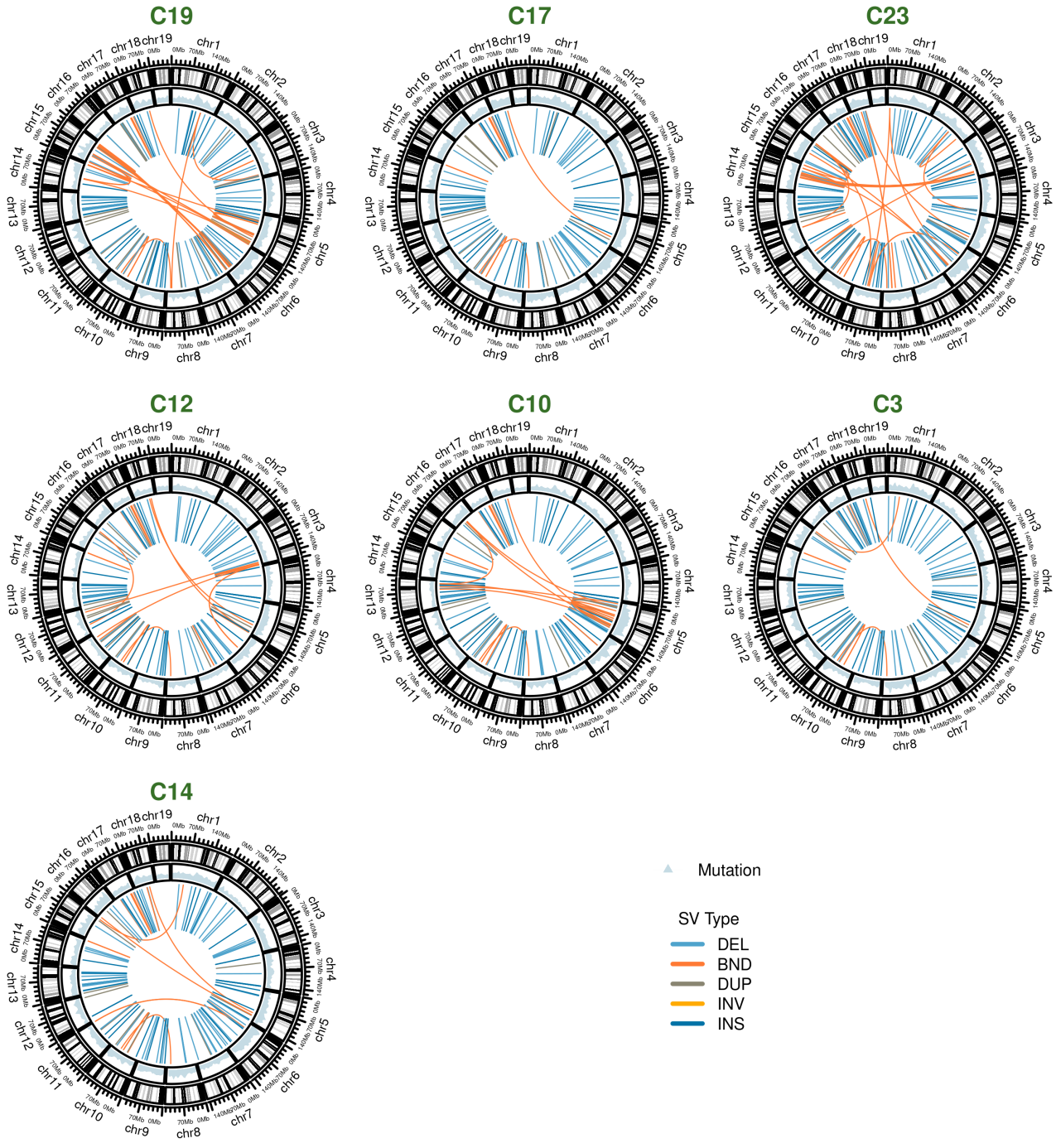


Figure B.7: **SNV density and SV events present in individual sublines in the green clade.**

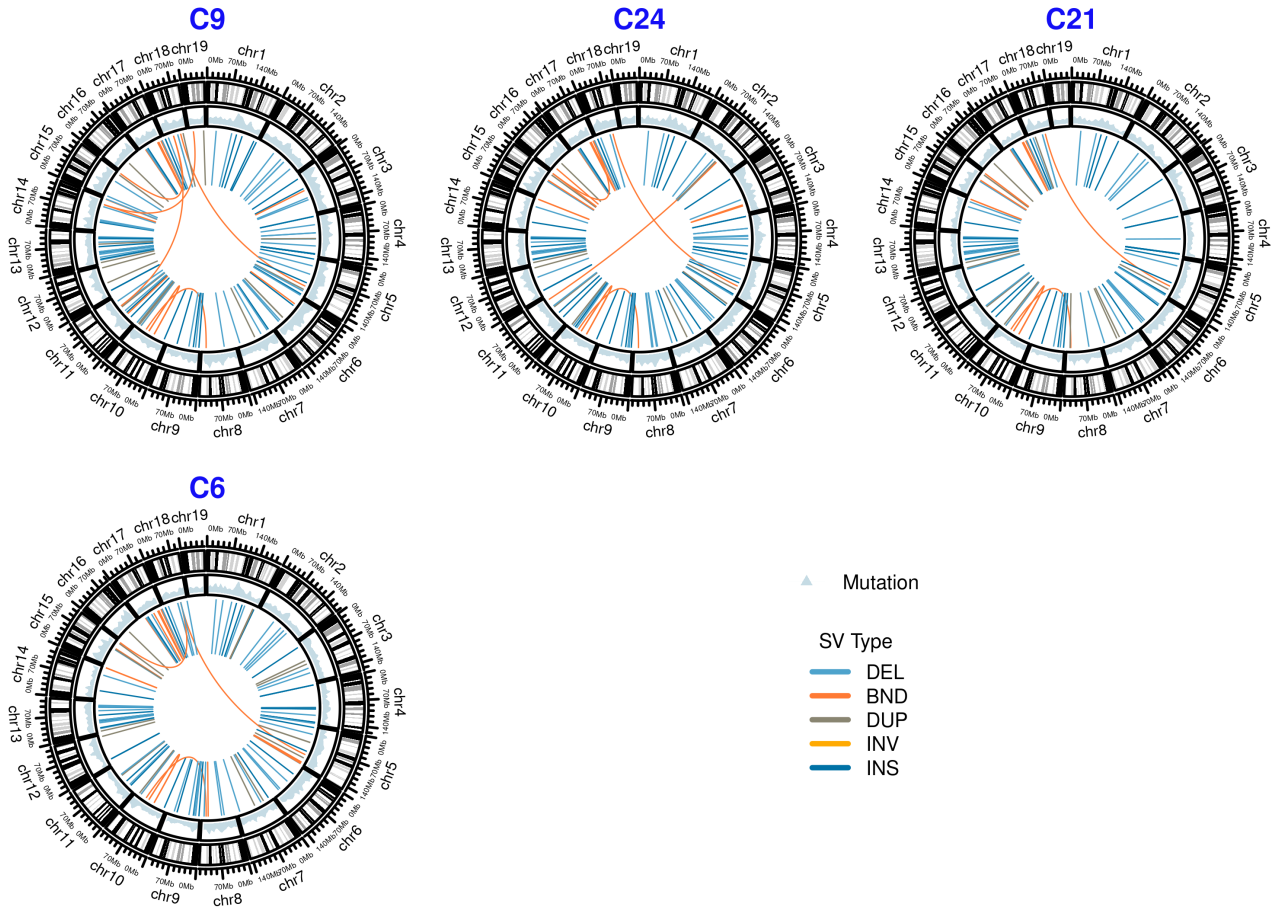


Figure B.8: **SNV density and SV events present in individual sublines in the blue clade.**

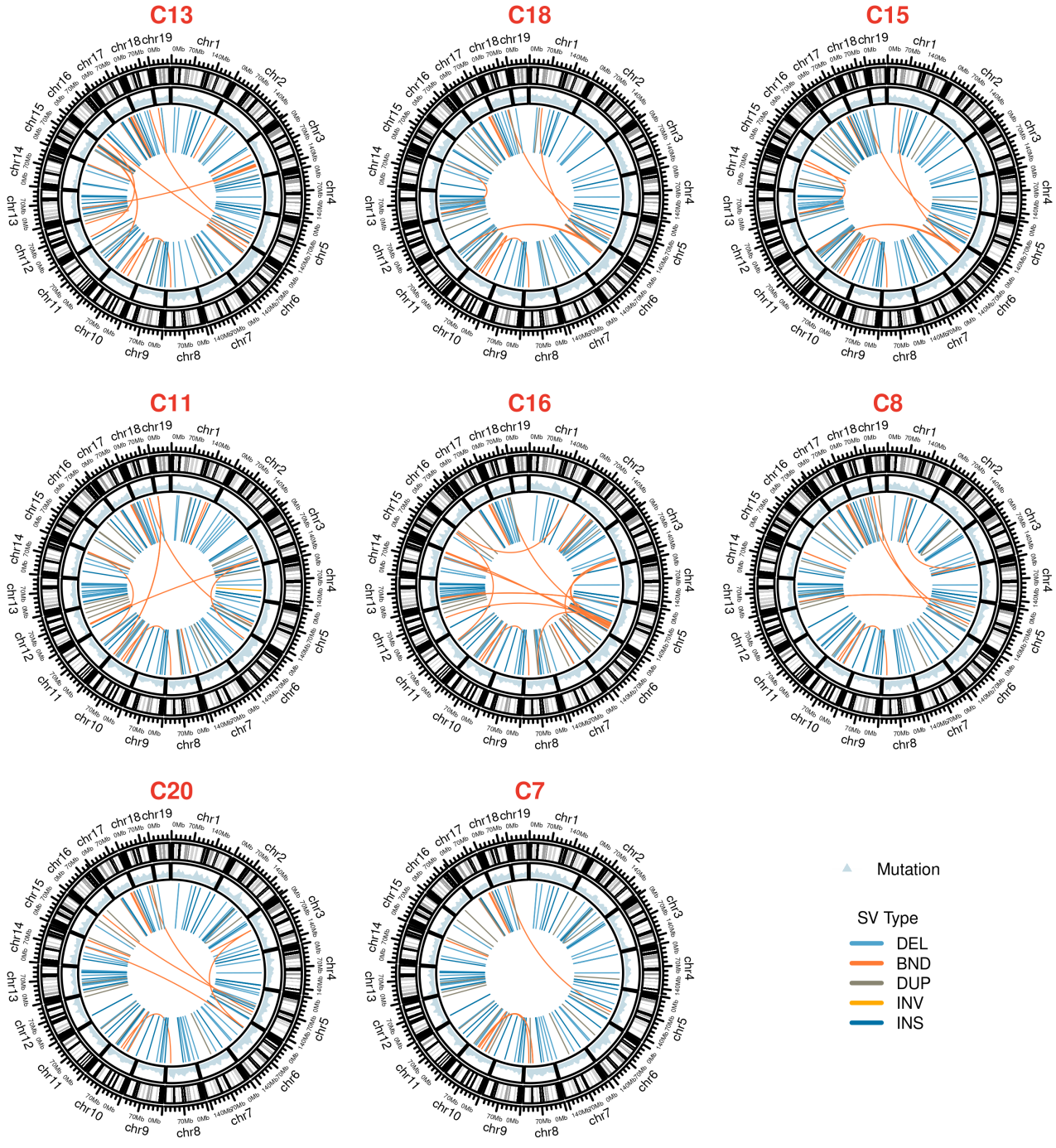


Figure B.9: **SNV density and SV events present in individual sublines in the red clade.**

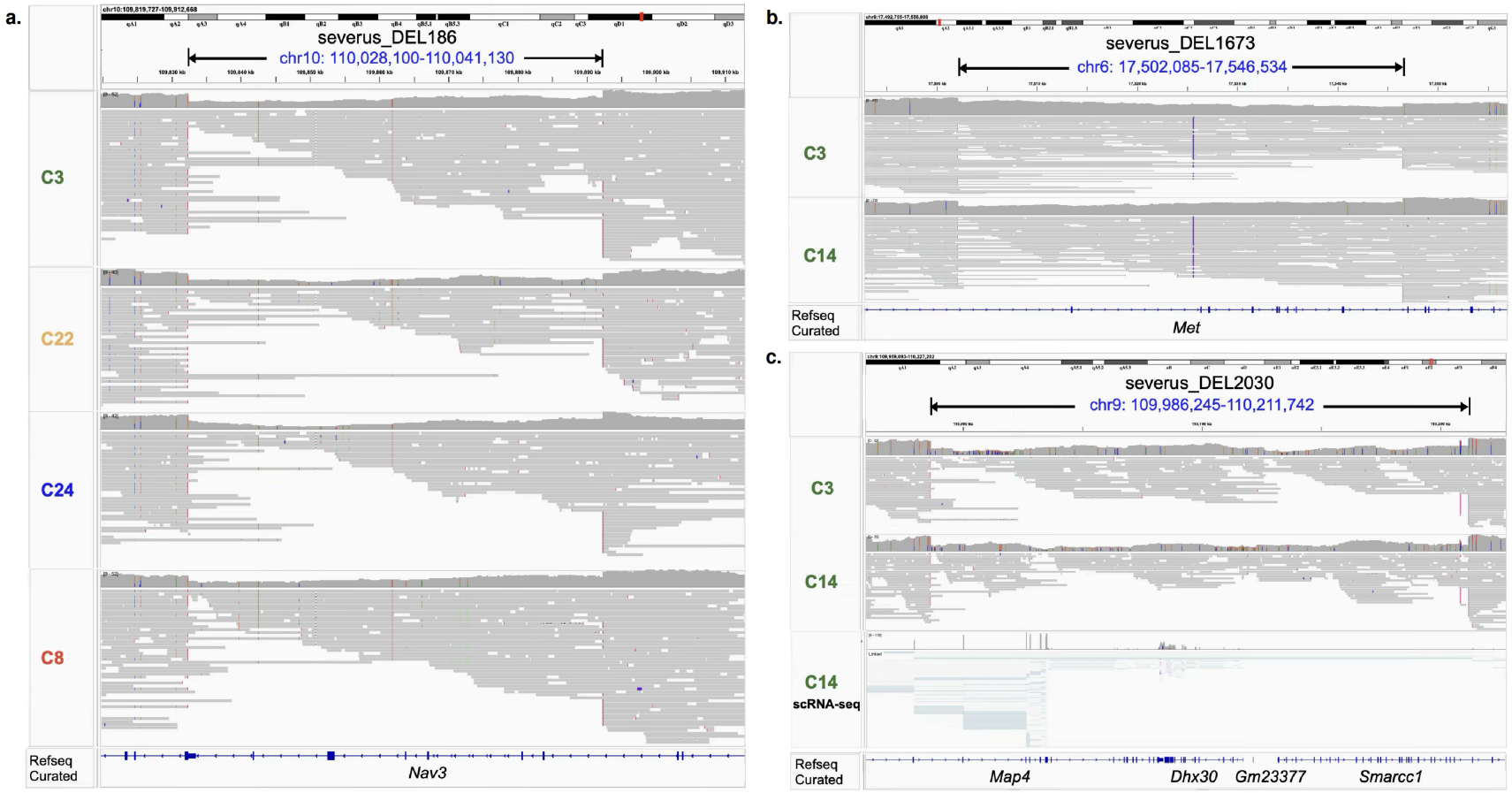


Figure B.10: **SV-disrupted genes and gene fusion annotated by `Padfoot`. a.** A clonal deletion affects exon 4-9 of *Nav3*. Long-read genomic reads of an representative subline from each major clade are shown. **b.** A subclonal deletion in C3 and C14 in *Met* resulted in the loss of exon 4-11. Long-read genomic reads from sublines C3 and C14 are shown. **c.** A subclonal deletion in C3 and C14 resulted in the fusion of *Map4* and *Smarcc1*. Long-read genomic reads from sublines C3 and C14, as well as scRNA-seq reads from a C14 cell corroborating the fusion event are shown.

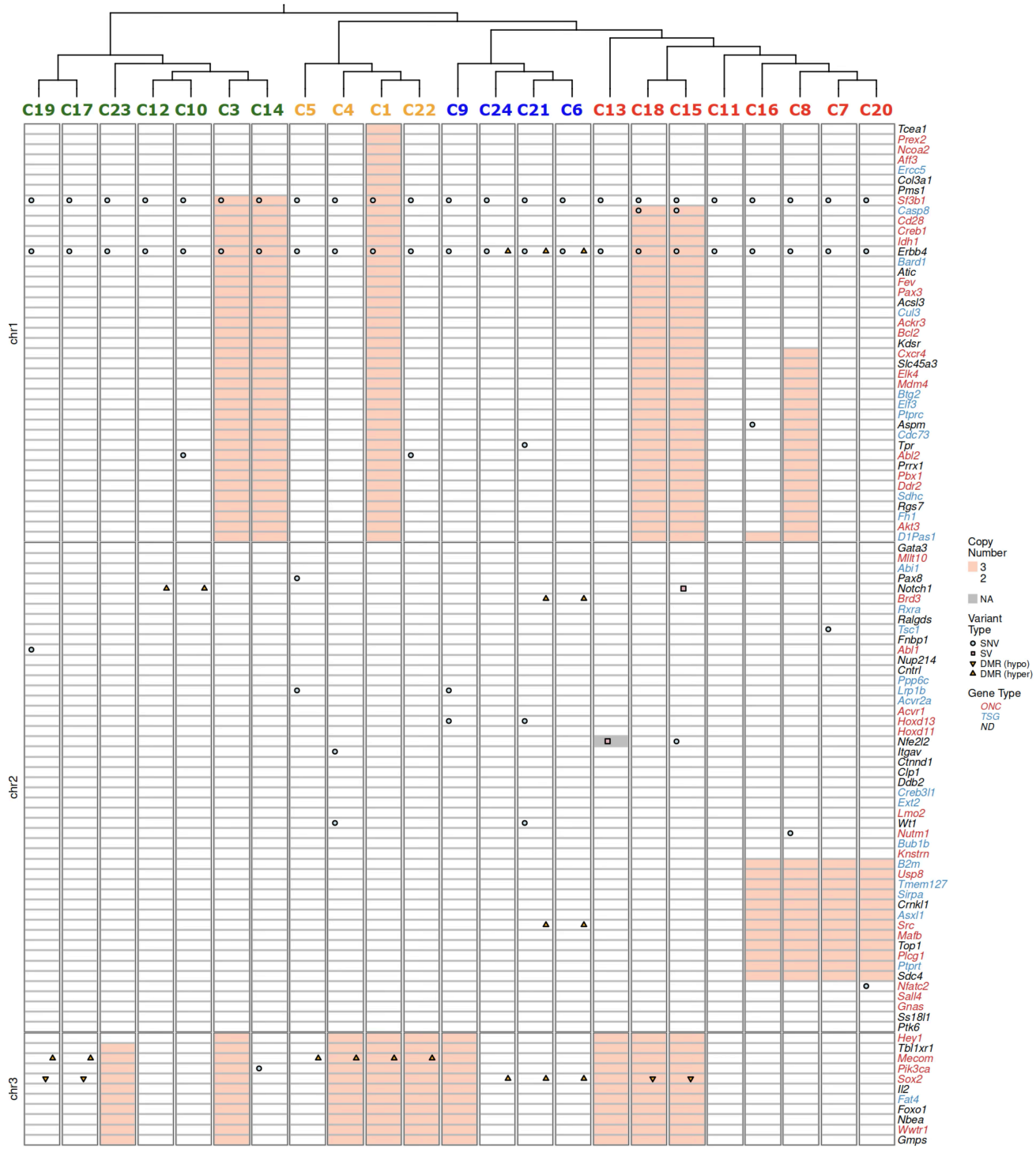
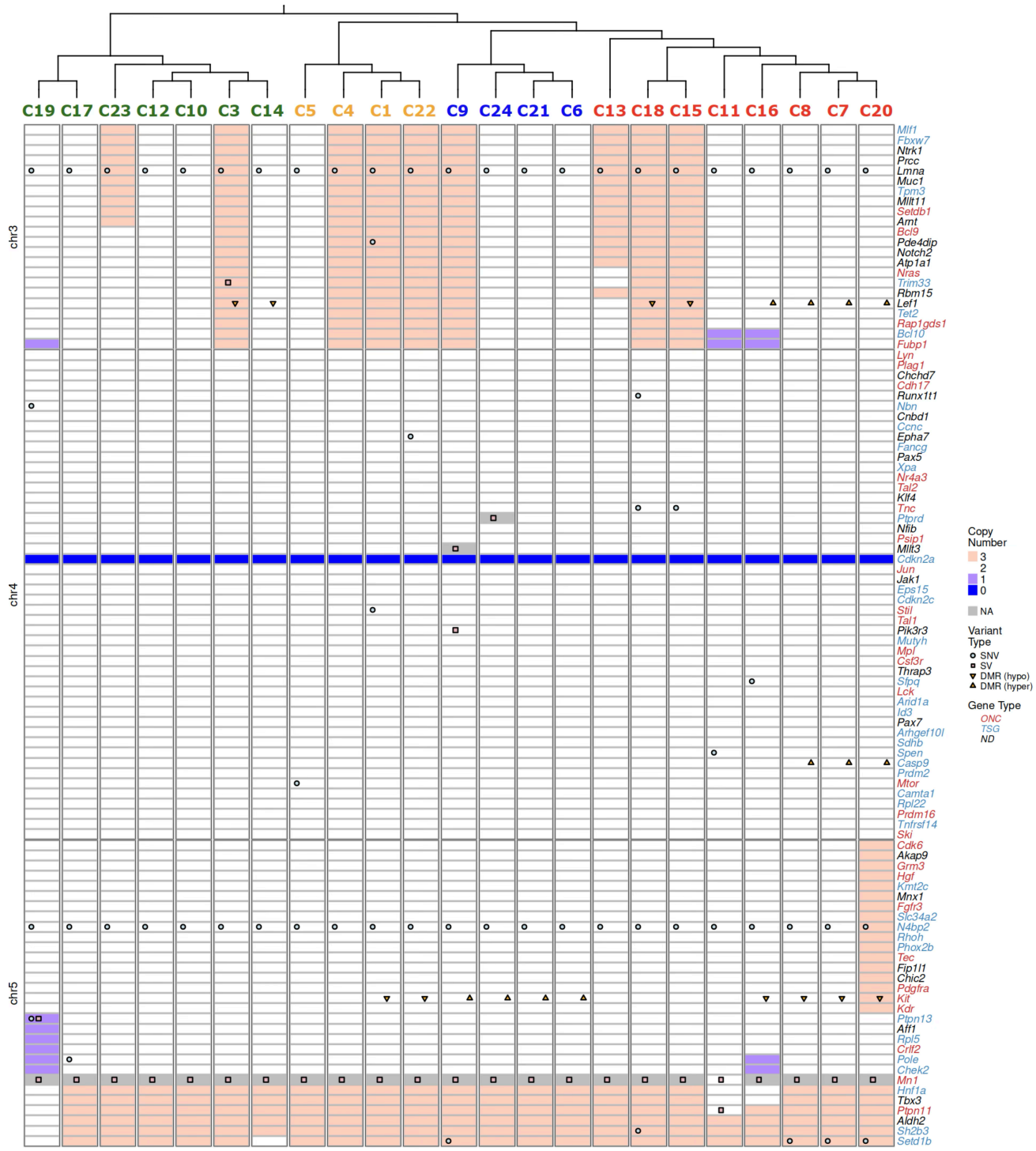


Figure B.11: **Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (1/7).** Genes with "NA" copy number in a subline span detected CNA boundaries and therefore cannot be assigned a unique copy number.



homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (2/7).**Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (2/7).** Genes with "NA" copy number in a subline span detected CNA boundaries and therefore cannot be assigned a unique copy number.

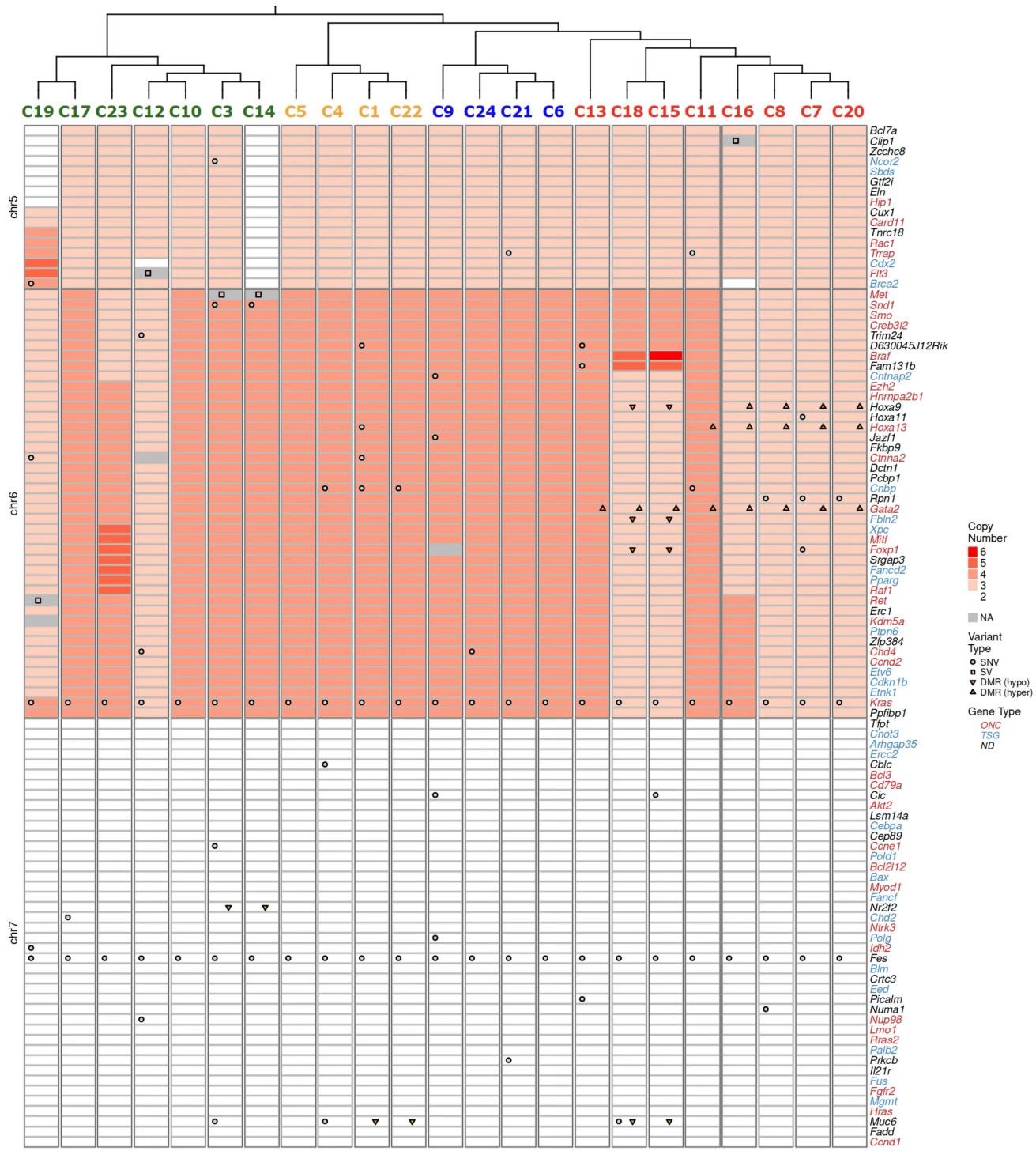


Figure B.12: **Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (3/7).** Genes with "NA" copy number in a subline span detected CNA boundaries and therefore cannot be assigned a unique copy number.

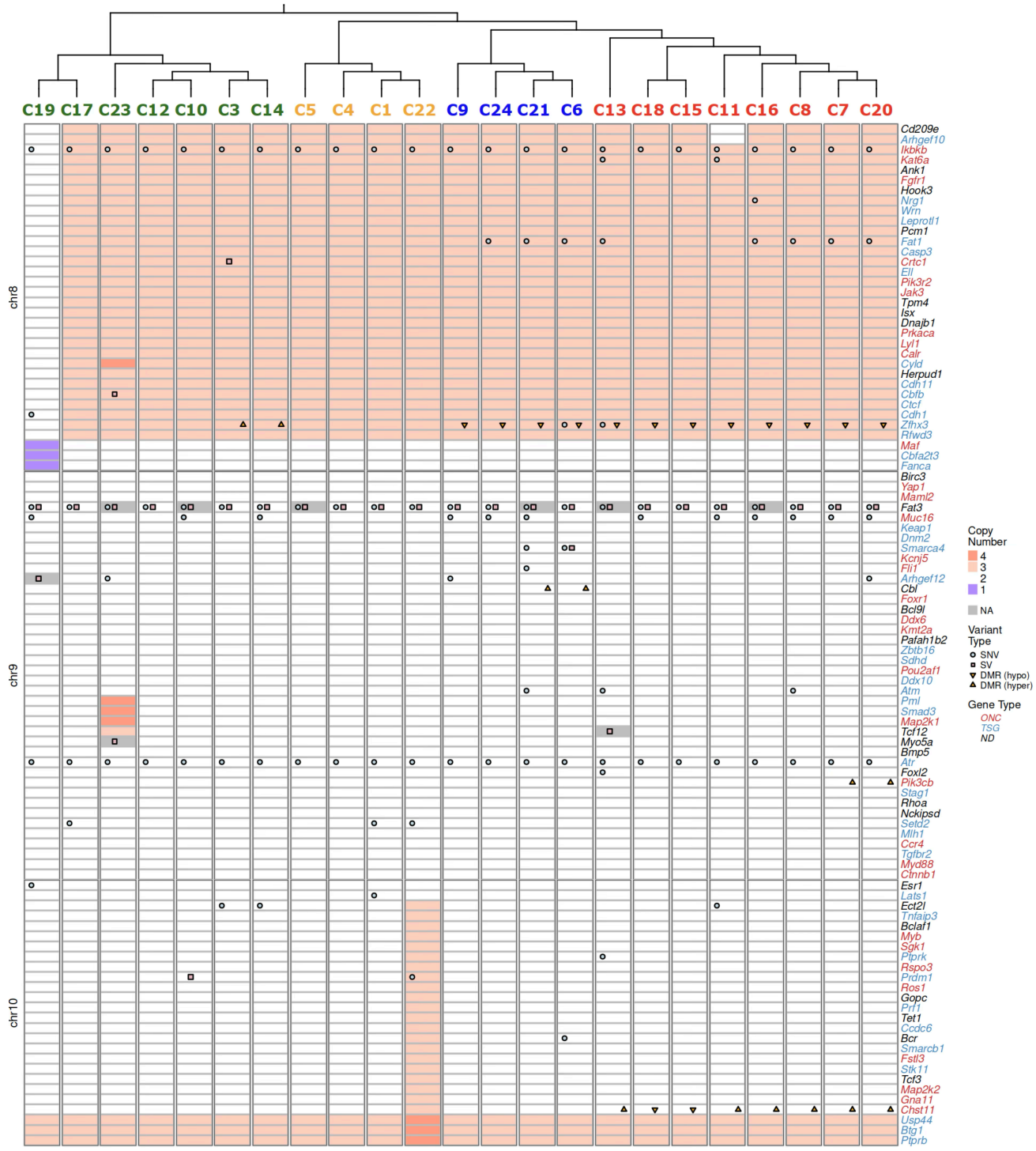


Figure B.13: **Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (4/7).** Genes with "NA" copy number in a subline span detected CNA boundaries and therefore cannot be assigned a unique copy number.

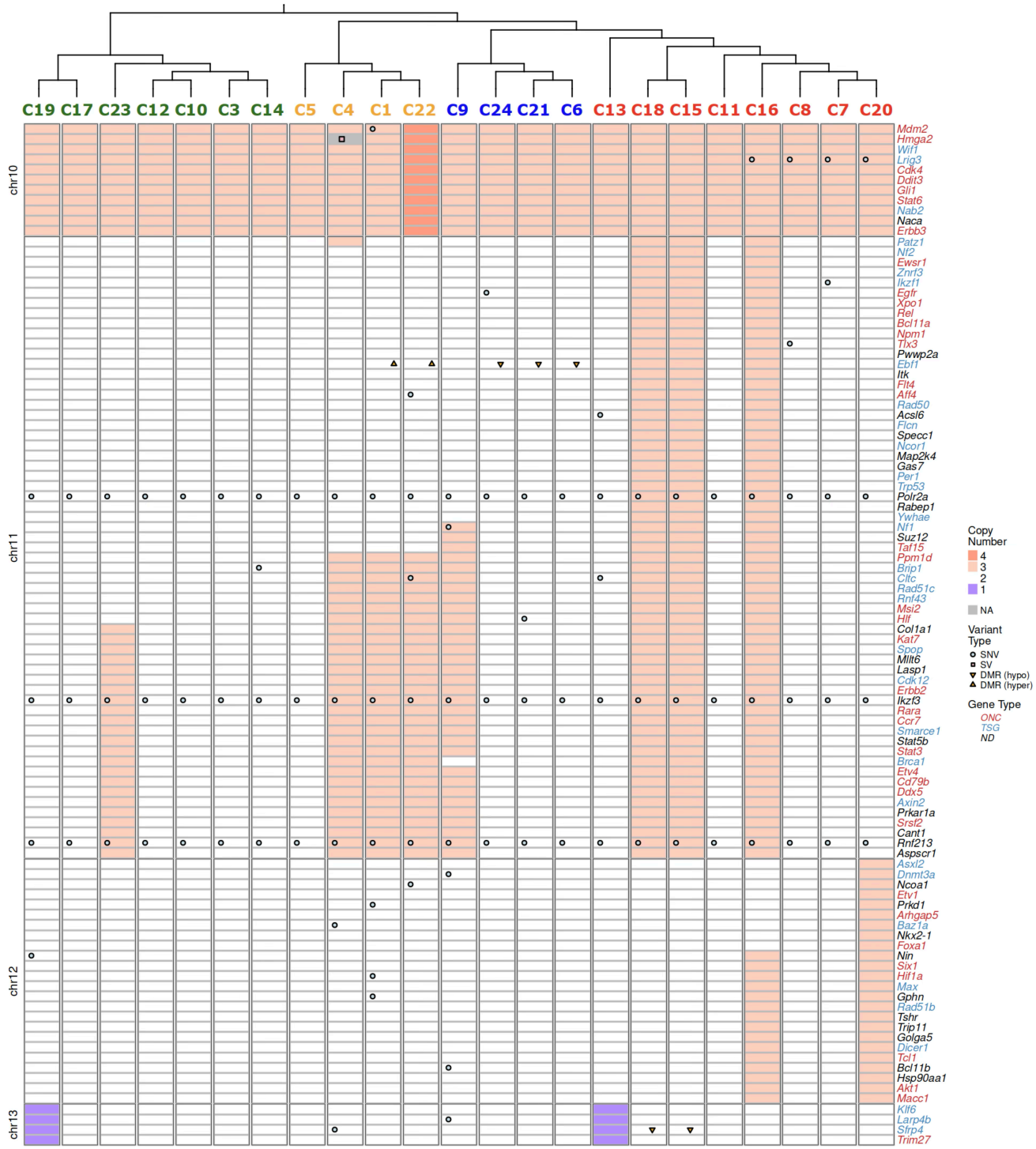


Figure B.14: **Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (5/7).** Genes with "NA" copy number in a subline span detected CNA boundaries and therefore cannot be assigned a unique copy number.

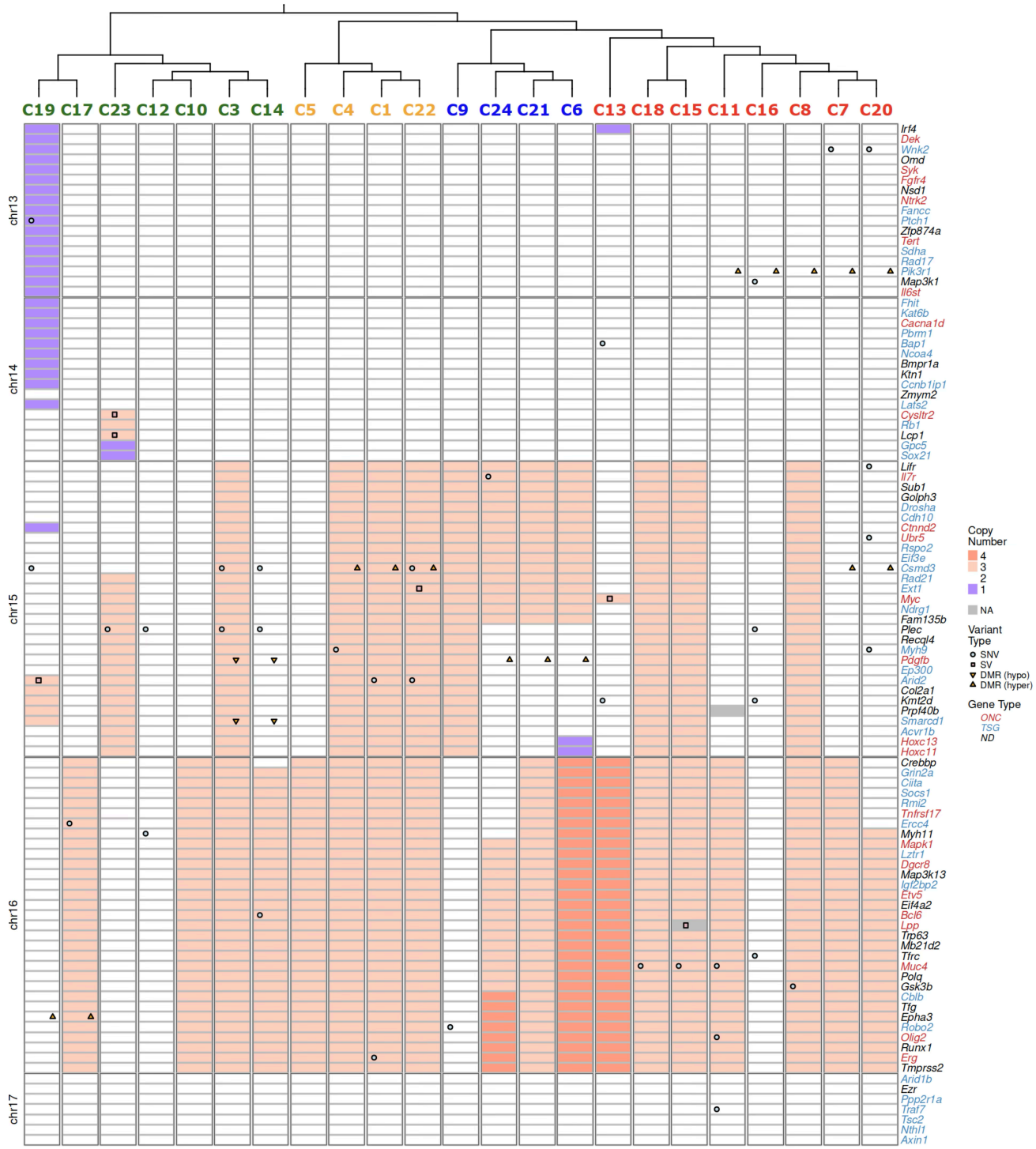


Figure B.15: **Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (6/7).** Genes with "NA" copy number in a subline span detected CNA boundaries and therefore cannot be assigned a unique copy number.

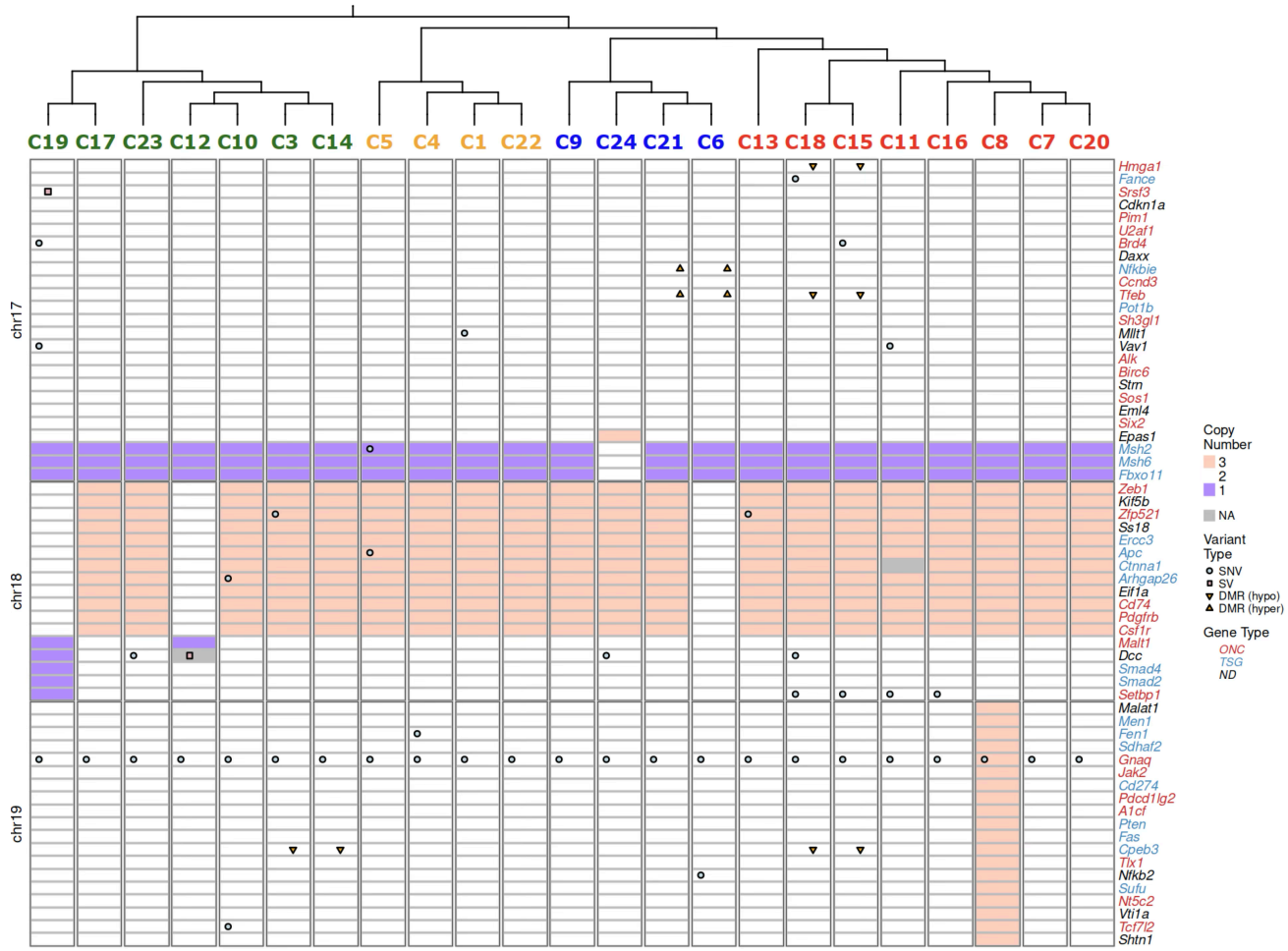


Figure B.16: **Mouse homologs of COSMIC genes impacted by (1) non-synonymous SNV, (2) SV breakpoint, (3) CNA, or (4) (hypo/hyper-) DMRs in their (putative) enhancers or promoters in 23 sublines (7/7).** Genes with "NA" copy number in a subline span detected CNA boundaries and therefore cannot be assigned a unique copy number.

Chr	Start	End	Direction	Putative Promoter	EPDnew	Proximal Enhancer
chr2	156447300	156448010	hyper	*Gm14225*	-	*Epb41l1*, *Gm14225*
chr6	30736515	30739964	hyper	*Mest*	*Mest*	*Mest*, *Mir335*, *Gm27522*, *Gm27568*, *Gm44296*
chr6	43264995	43267174	hyper	*Arhgef5*, *Gm44731*	*Arhgef5*	*Arhgef5*, *Gm44731*
chr6	52203136	52204841	hyper	*Hoxa5*	-	*Hoxa5*, *Hoxa6*, *Hoxaas3*, *Hoxa3*, *Gm28308*
chr6	73220915	73223008	hyper	*Dnah6*, *Gm40377*	*Dnah6*	*Dnah6*, *Gm40377*
chr6	88188695	88190058	hyper	-	-	*Gm38708*
chr6	88193674	88194653	hyper	*Gata2*	-	*Gata2*
chr6	116628225	116629385	hyper	*Zfp422*	*Zfp422*	*Zfp422*, *Gm44154*
chr7	23166848	23167923	hyper	*Gm8728*	-	-
chr10	83140358	83141188	hyper	-	-	*Chst11*, *1700025N21Rik*
chr11	28582204	28583082	hyper	-	-	*Ccdc85a*
chr3	79285435	79286044	hypo	-	-	*Rapgef2*, *6430573P05Rik*
chr3	146220813	146221659	hypo	*Lpar3*	*Lpar3*	*Lpar3*
chr4	130171758	130174329	hypo	-	-	*Tinagl1*
chr4	155774166	155775696	hypo	*Vwa1*	*Vwa1*	*Vwa1*
chr5	14025106	14026982	hypo	*Sema3e*	*Sema3e*	*Sema3e*, *Gm43519*
chr6	14897121	14900597	hypo	-	-	*Foxp2*
chr6	117168413	117169671	hypo	*Cxcl12*	*Cxcl12*	*Cxcl12*
chr8	34524973	34526019	hypo	-	-	*Gm33968*
chr9	54300511	54301797	hypo	*Gldnos*	-	-
chr11	65160117	65161795	hypo	*1700086D15Rik*	-	*Arhgap44*, *1700086D15Rik*
chr11	96279456	96280673	hypo	-	-	*Hoxb8*
chr12	82586939	82588254	hypo	*Rgs6*	-	-
chr15	27467528	27468632	hypo	-	-	*Ank*
chr17	31299560	31300475	hypo	*Slc37a1*	-	-
chrX	52613765	52614700	hypo	*Gpc3*	*Gpc3*	*Gpc3*, *A630012P03Rik*

Table B.2: **DMRs associated with the red clade that overlap putative promoter or putative proximal enhancers of known genes.** A region is said to be a putative promoter for a gene if it is within -350bp to +50bp of the gene's annotated transcriptional start site (Ensembl Release 102) [139, 140]. Of the 17 putative promoters, 9 are also annotated in the EPDnew (v.003) database of experimentally validated promoters [141, 142]. A region is said to be a putative proximal enhancer for a gene if is has enhancer-like signature [146] and is within 2Kbp of the annotated transcriptional start site [143–145] of the gene. Note that *Gata2* was an outlier in that its putative promoter was noted to be hypermethylated while its expression has a positive log2 fold change (Figure 3.5d.); however, this may be due to the hypomethylation of a putative promoter of *Gata2* alternative isoforms, which went undetected as its mean methylation difference is not large enough to meet our criteria.

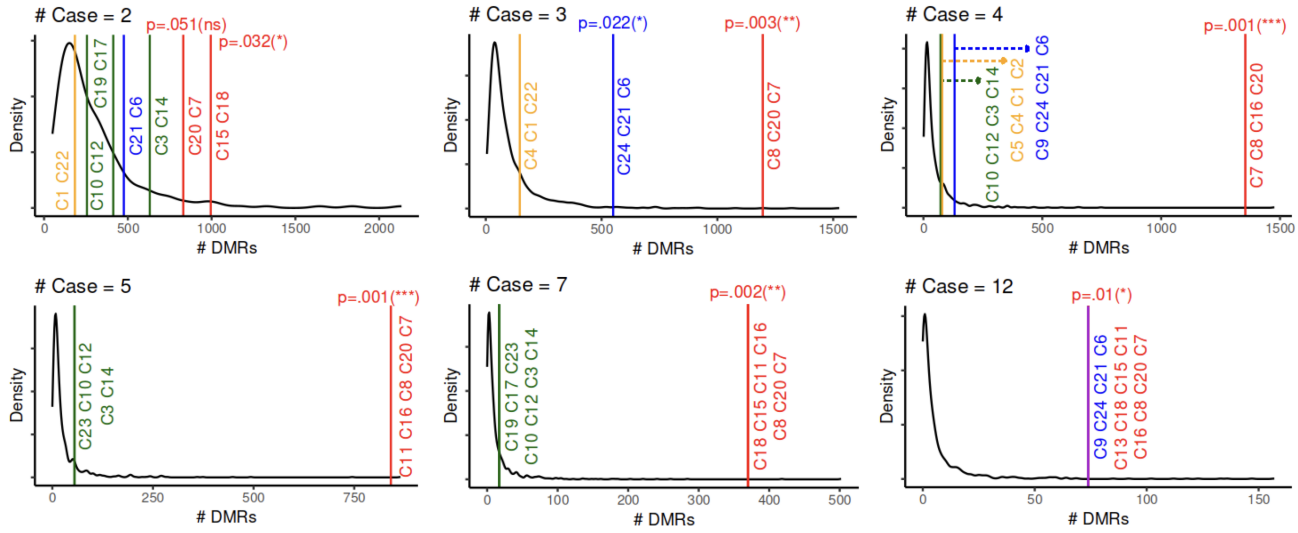


Figure B.17: **Permutation tests on the number of DMRs obtained for all subtree sizes of the phylogeny in Figure 3.2a.** In each panel, "# Case" indicates the number of leaves in the subtree below the branch at which the test was performed. See Figure 3.5b. for "# Case = 8". Note that due to symmetry, no additional permutation test is done for "# Case = 16" to account for the subtree consisting of the orange, blue, and red in Figure 3.2a., as the empirical null distribution is identical to that of "# Case = 7" and the number of DMRs obtained is identical to that of the subtree of size 7 colored in green in Figure 3.2a.

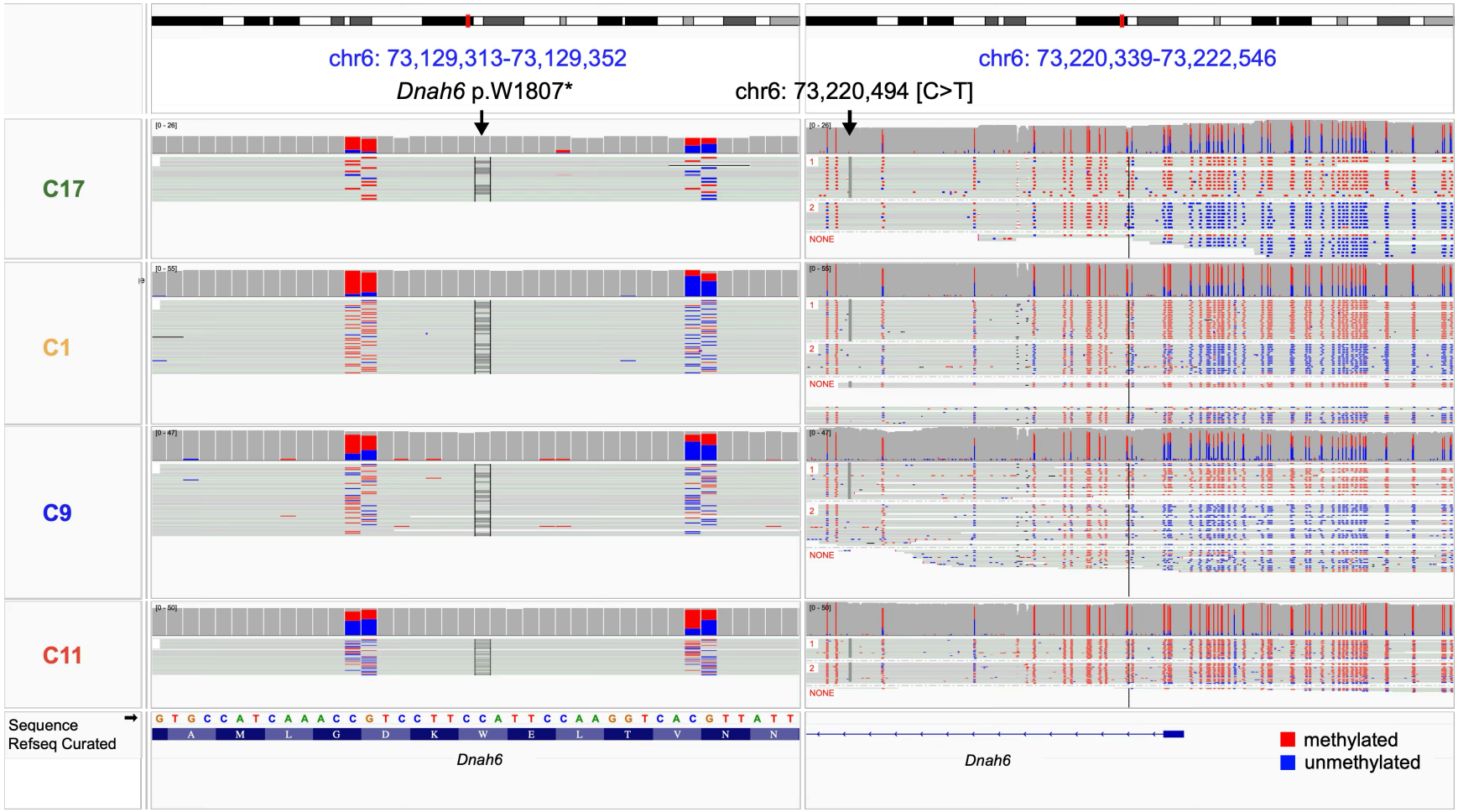


Figure B.18: **Associating truncating *Dnah6* mutation with (haplotype-specific) methylation of *Dnah6* promoter in major clades.** One representative subline from each major clade is shown. Clonal truncating mutation *Dnah6* p.W1807* is marked in the left panel displaying unphased reads, and a heterozygous variant chr6: 73,220,494 [C>T] is marked on the right panel with reads phased according to the variant. In C11, the high VAF of *Dnah6* p.W1807*, together with the homozygosity of chr6: 73,220,494 [C>T], not only indicates LOH in C11, but also provides evidence that the two variants exist on the same haplotype. As the phased reads on the right panel are also colored by DNA methylation, we observe that methylation at *Dnah6* promoter is specific to the haplotype harboring chr6: 73,220,494 [C>T], and hence *Dnah6* p.W1807*. We observe haplotype-specific methylation of the *Dnah6* promoter in the green, orange, and blue clade, and hypermethylation in the red clade. Note that while here we observe a few reads from the wild-type copy of *Dnah6* in C11, we see a complete loss of the wild-type copy of *Dnah6* in all other sublines in the red clade, and along with that a complete hypermethylation of the *Dnah6* promoter.

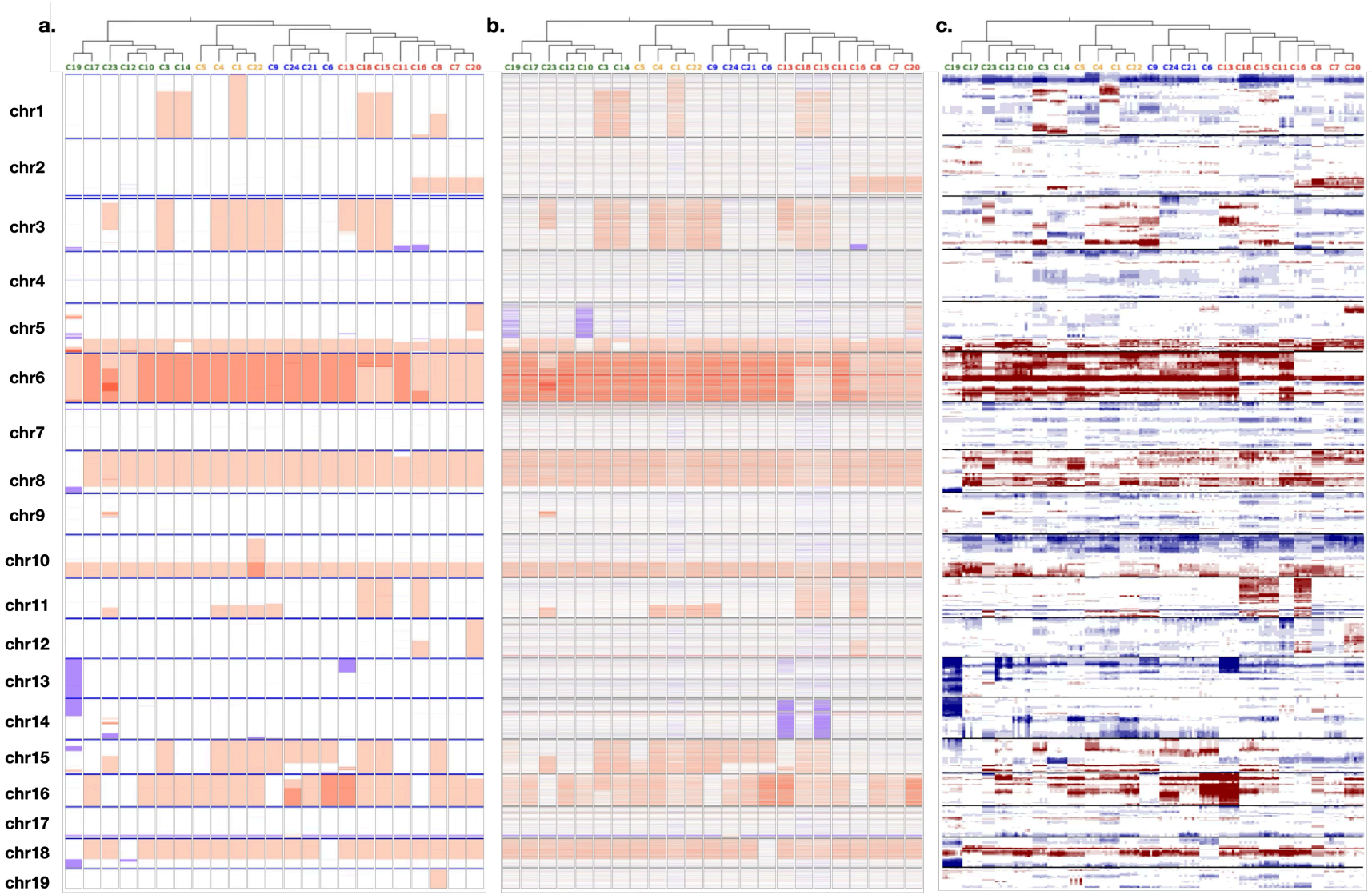
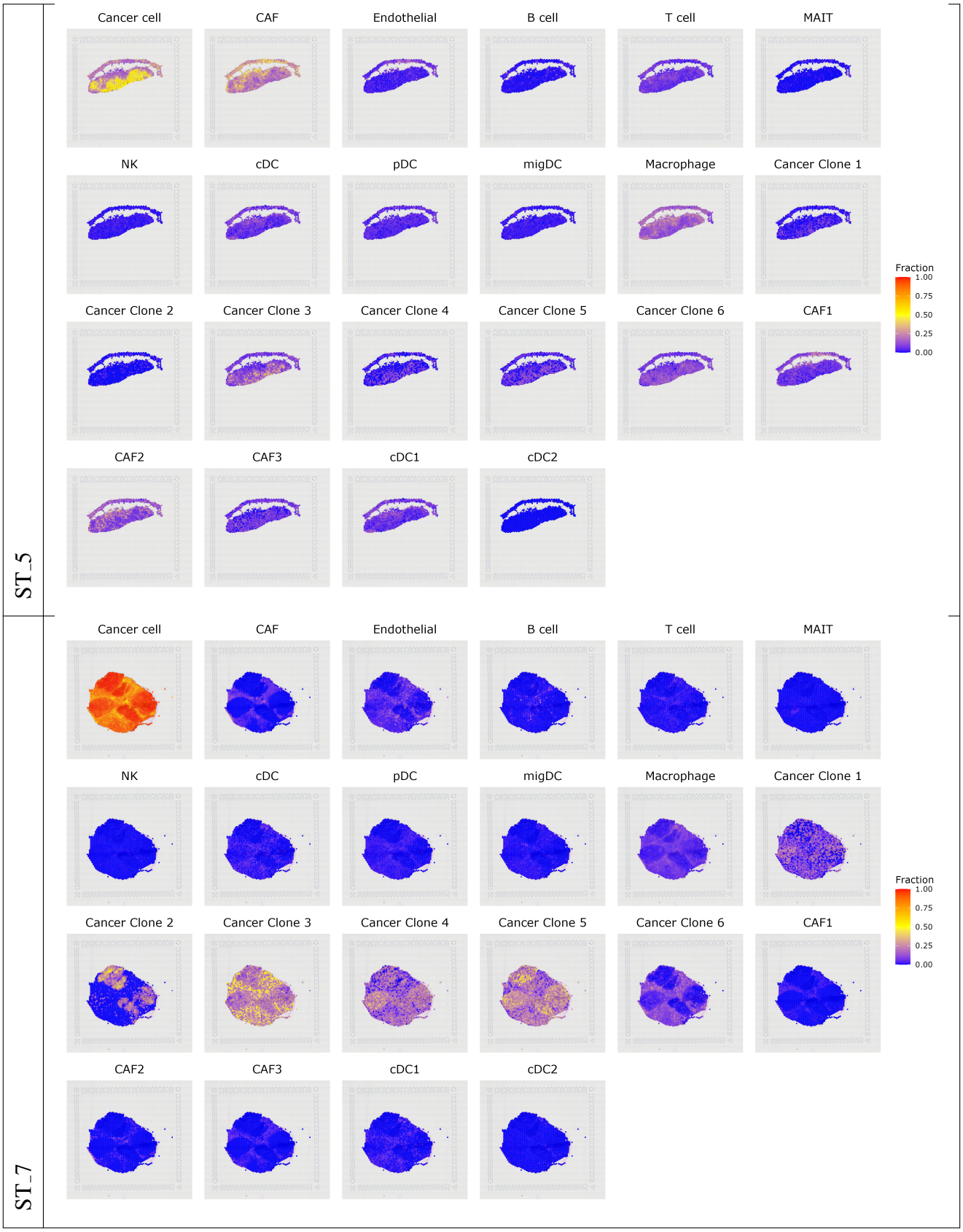


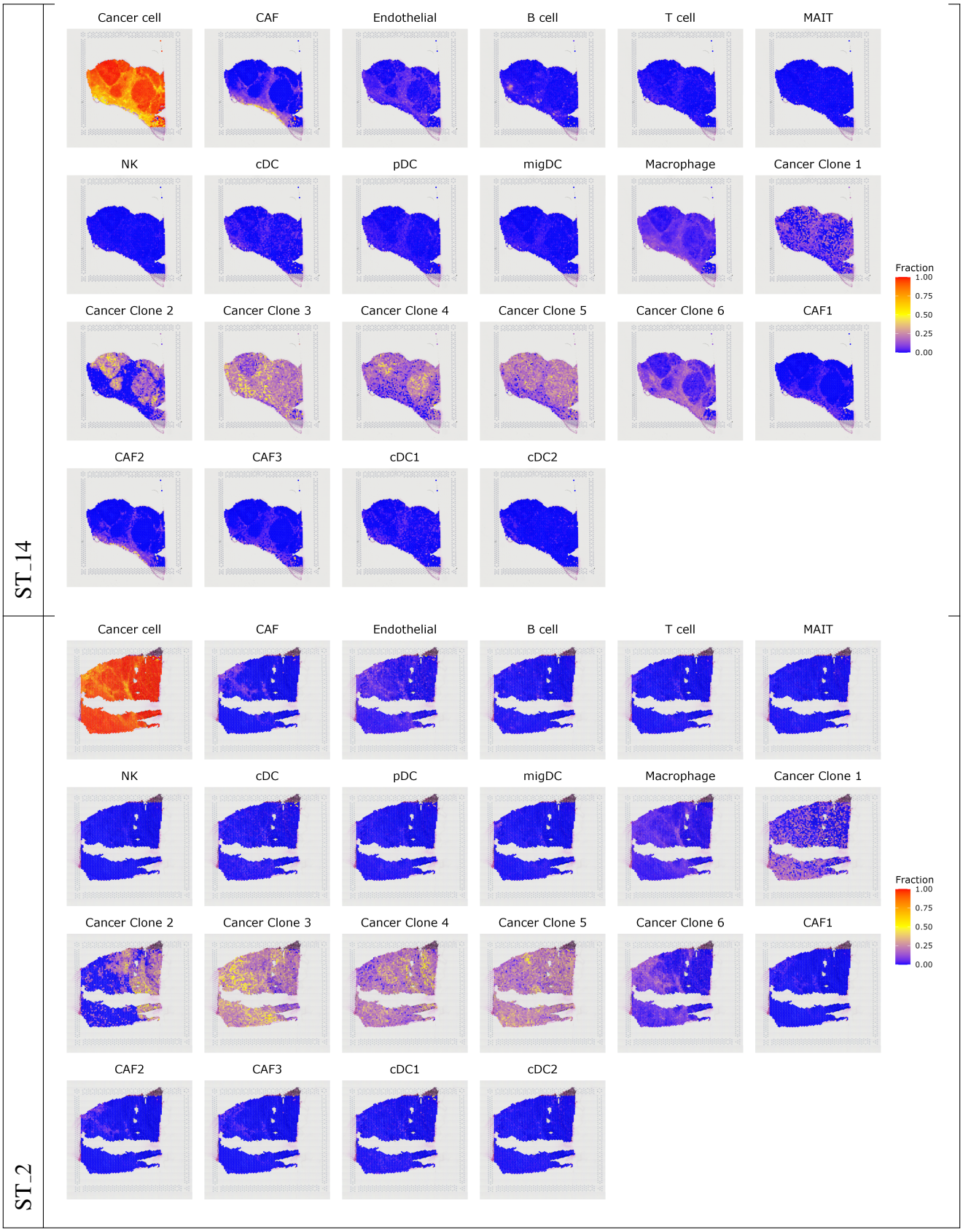
Figure B.19: **CNA profiles of 23 sublines obtained using long-read, bWES, and scRNA-seq data.** **a.** Long-read-derived CNA profiles using `Wakhan` [121] (Figure 3.2e.; Methods) **b.** CNA profiles derived by `CNVkit` [207] from bWES data [119]. **c.** CNA profiles derived by `inferCNV` [208] from full-length scRNA-seq data [118] obtained using the Smart-seq2 protocol [170].

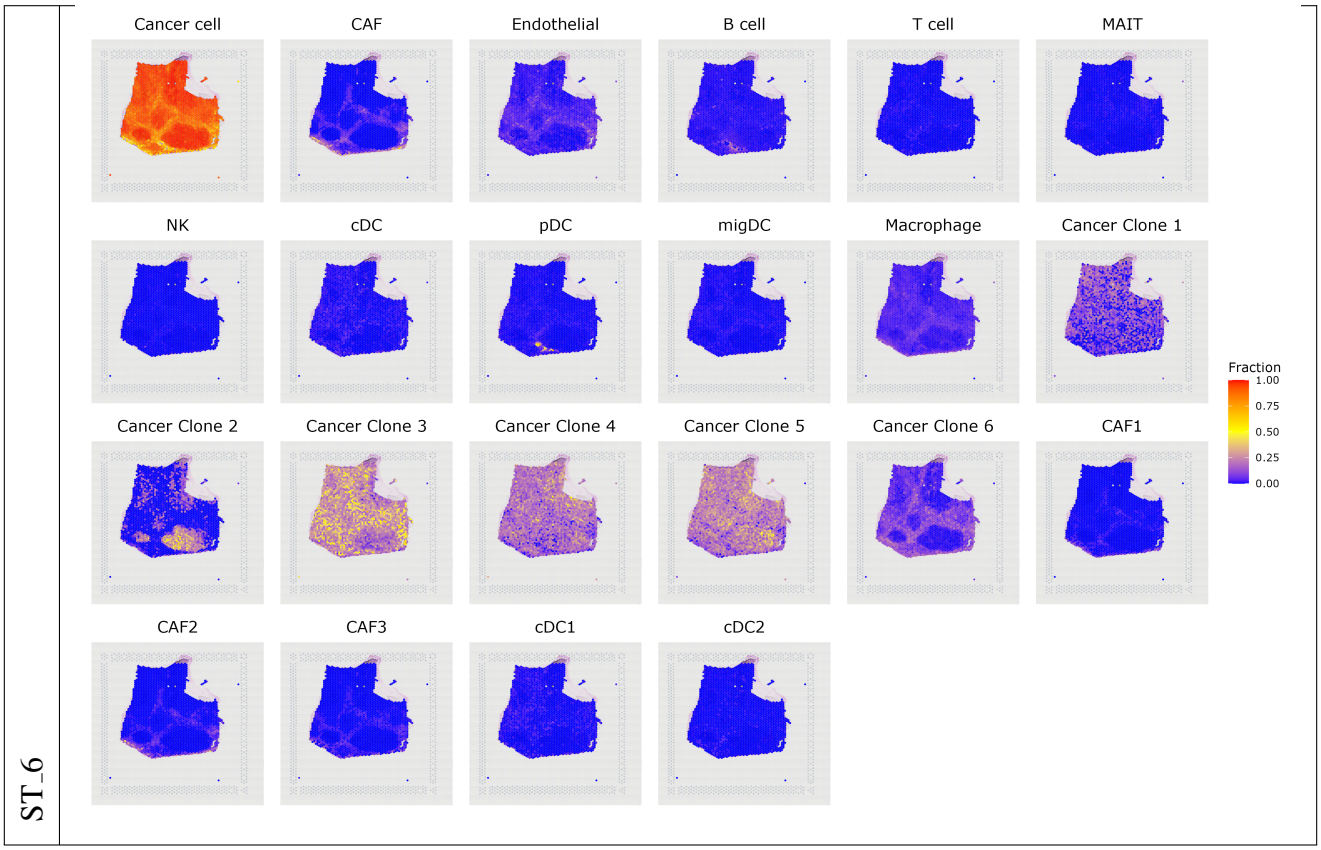
Appendix C: Supplemental Material for Chapter 4

C.1 SpaCET deconvolution results on 17 ST slides

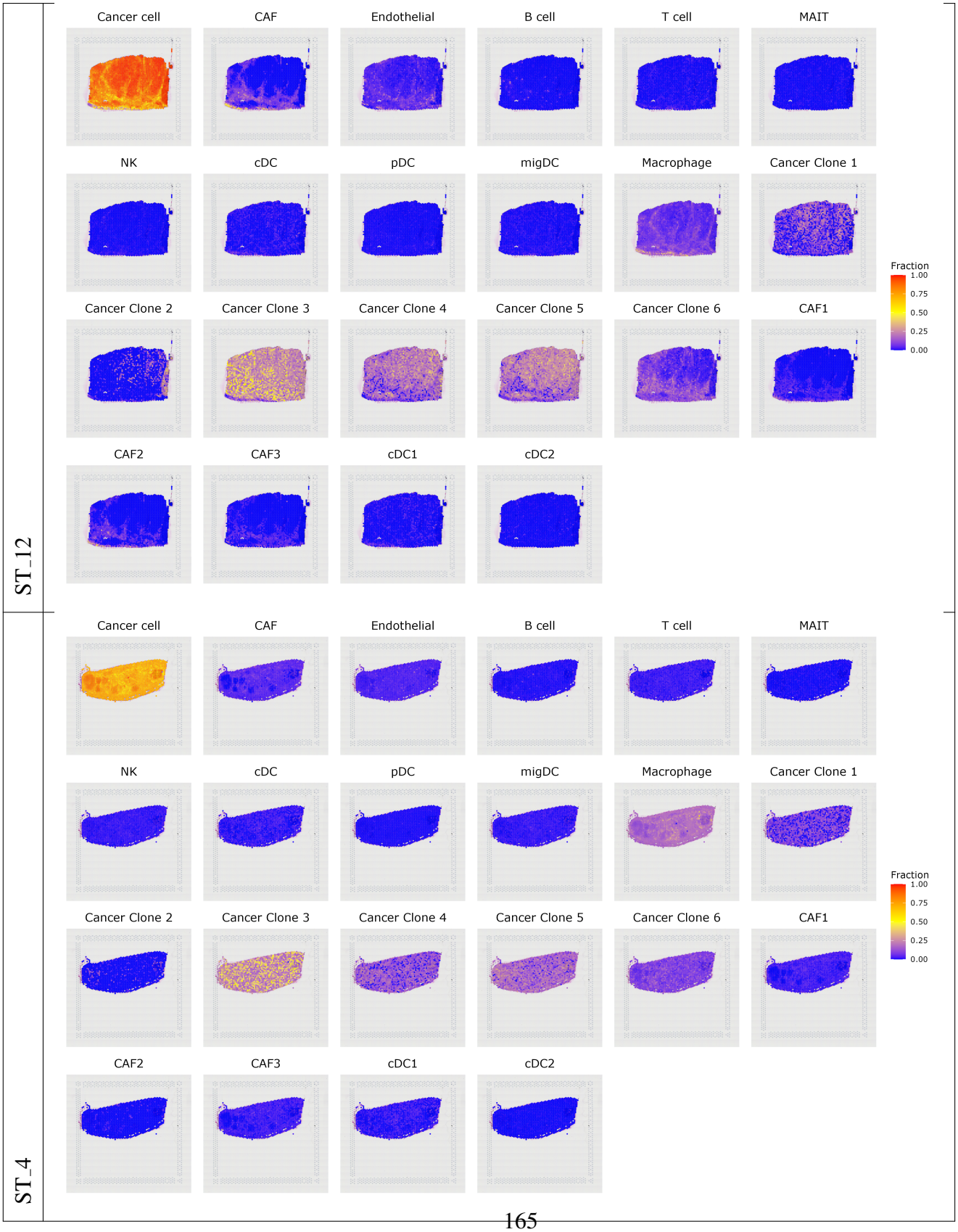
C.1.1 Control STs

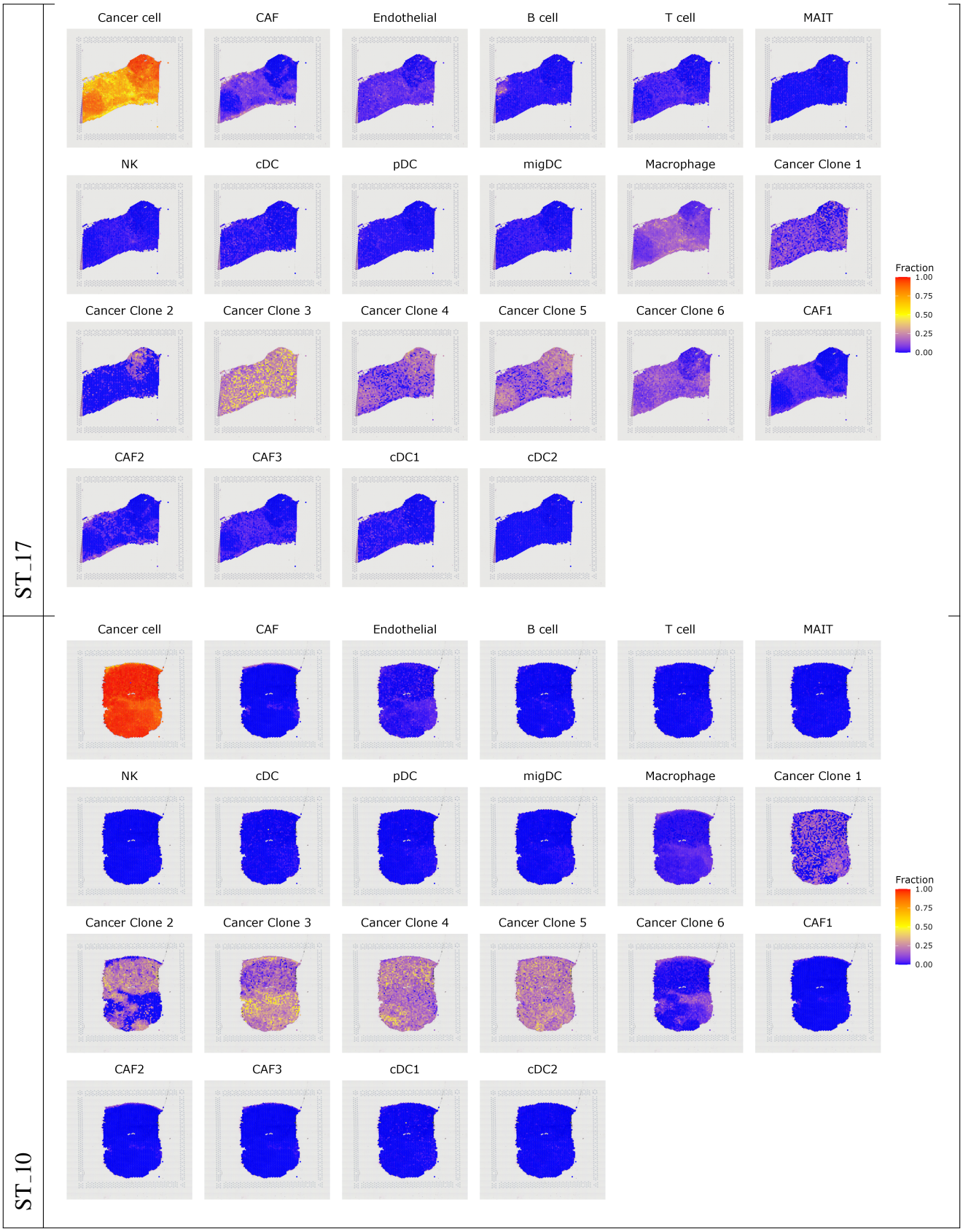

ST_5
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00
ST_7
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00

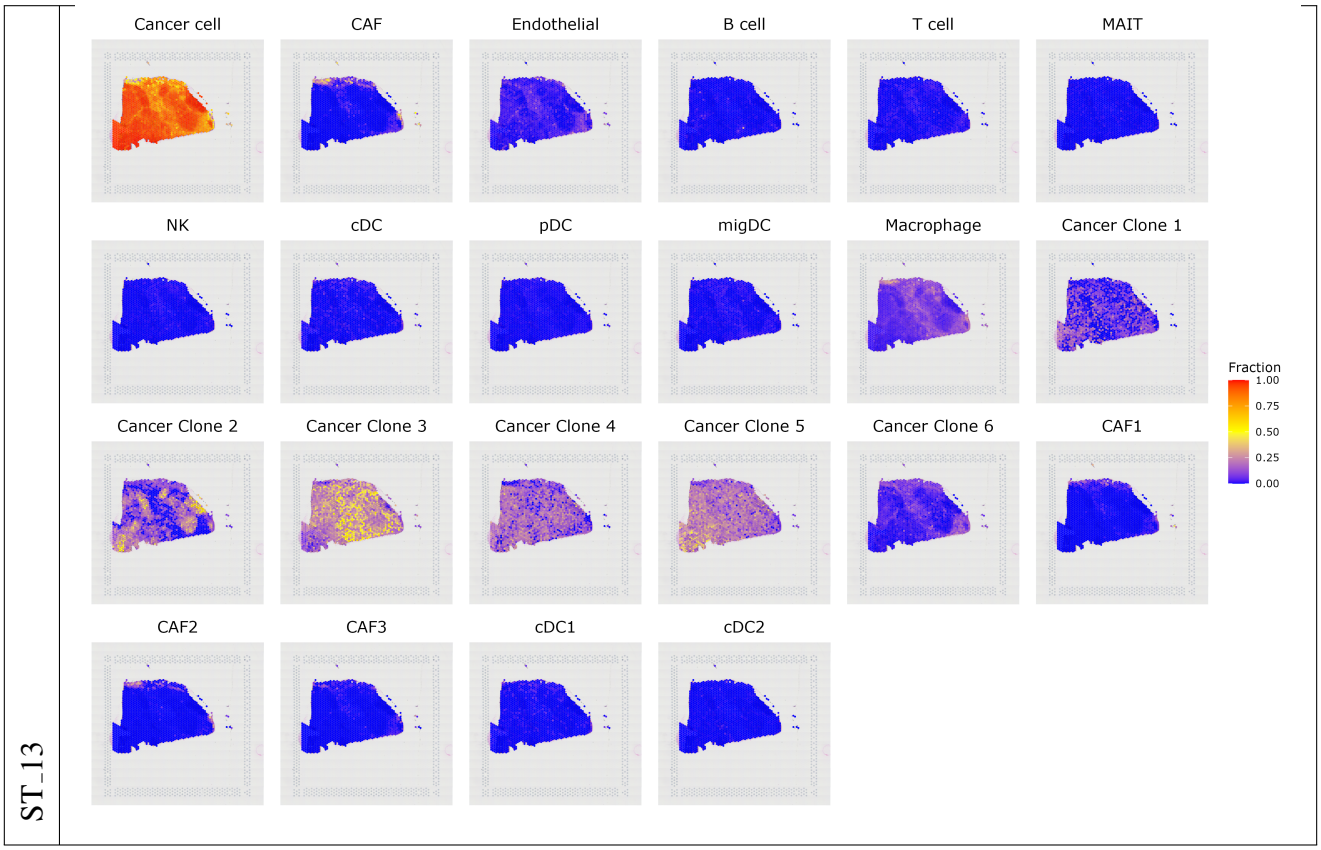

ST_14
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00
ST_2
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00


ST_6
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00

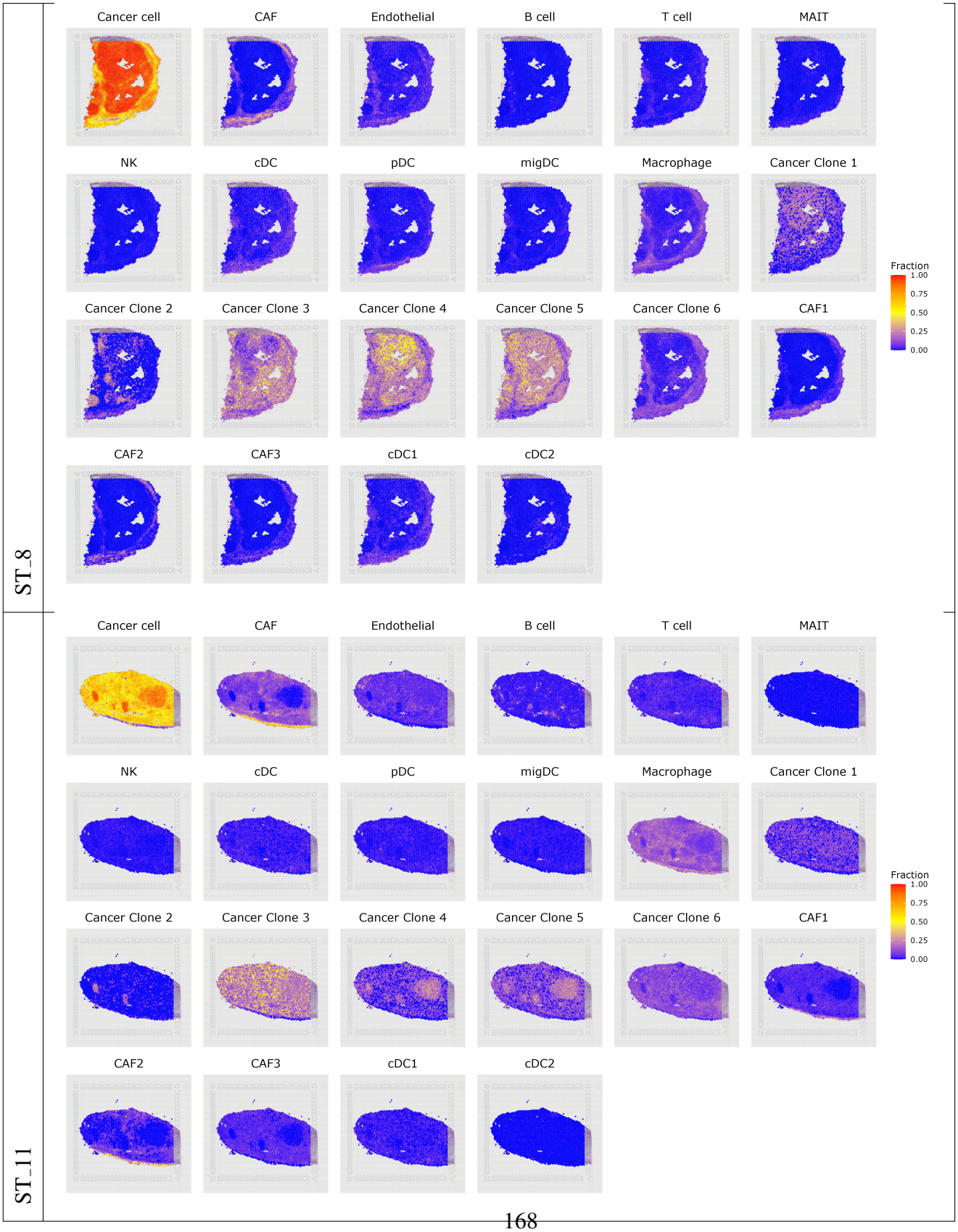
C.1.2 Treated STs (Non-responders)

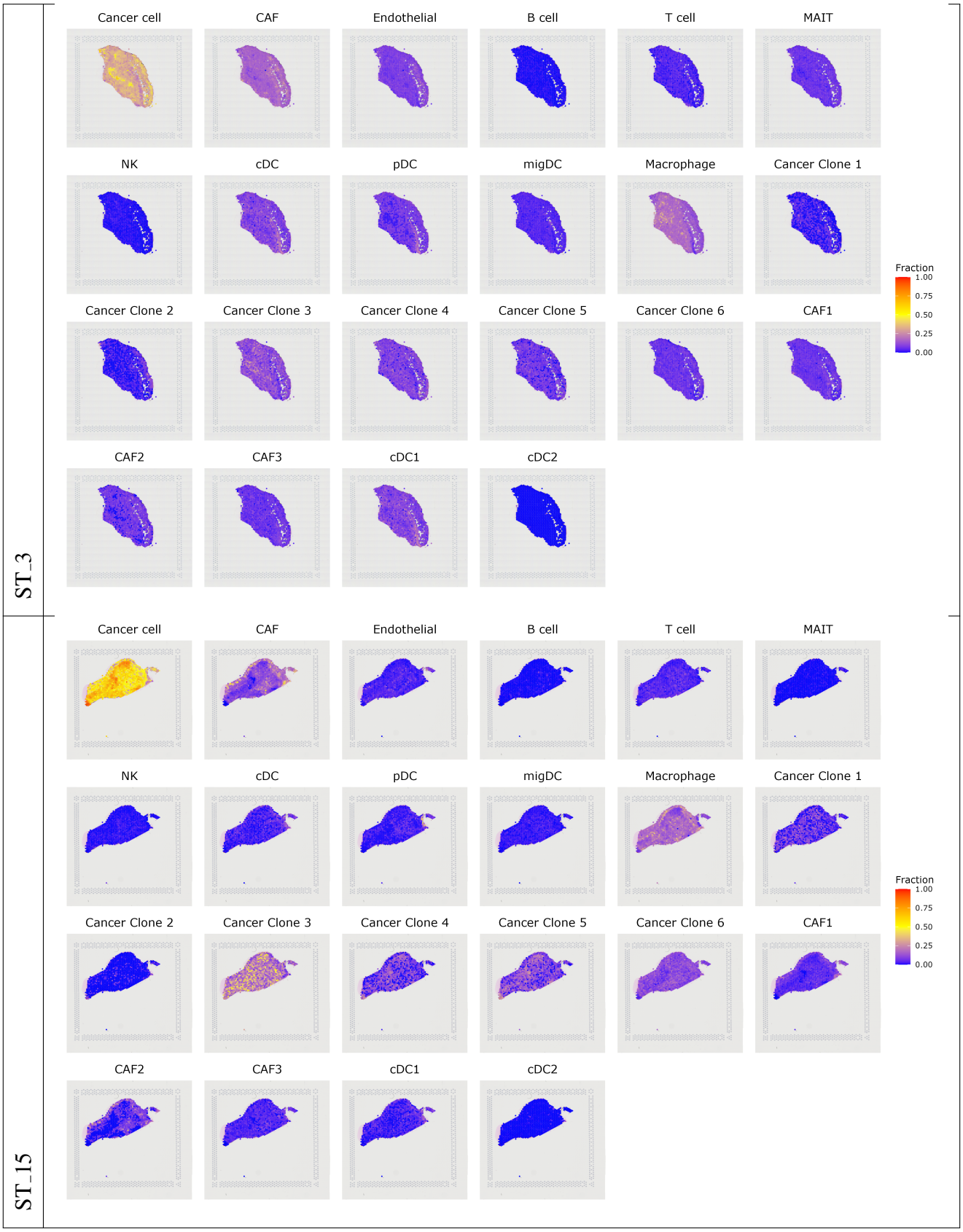


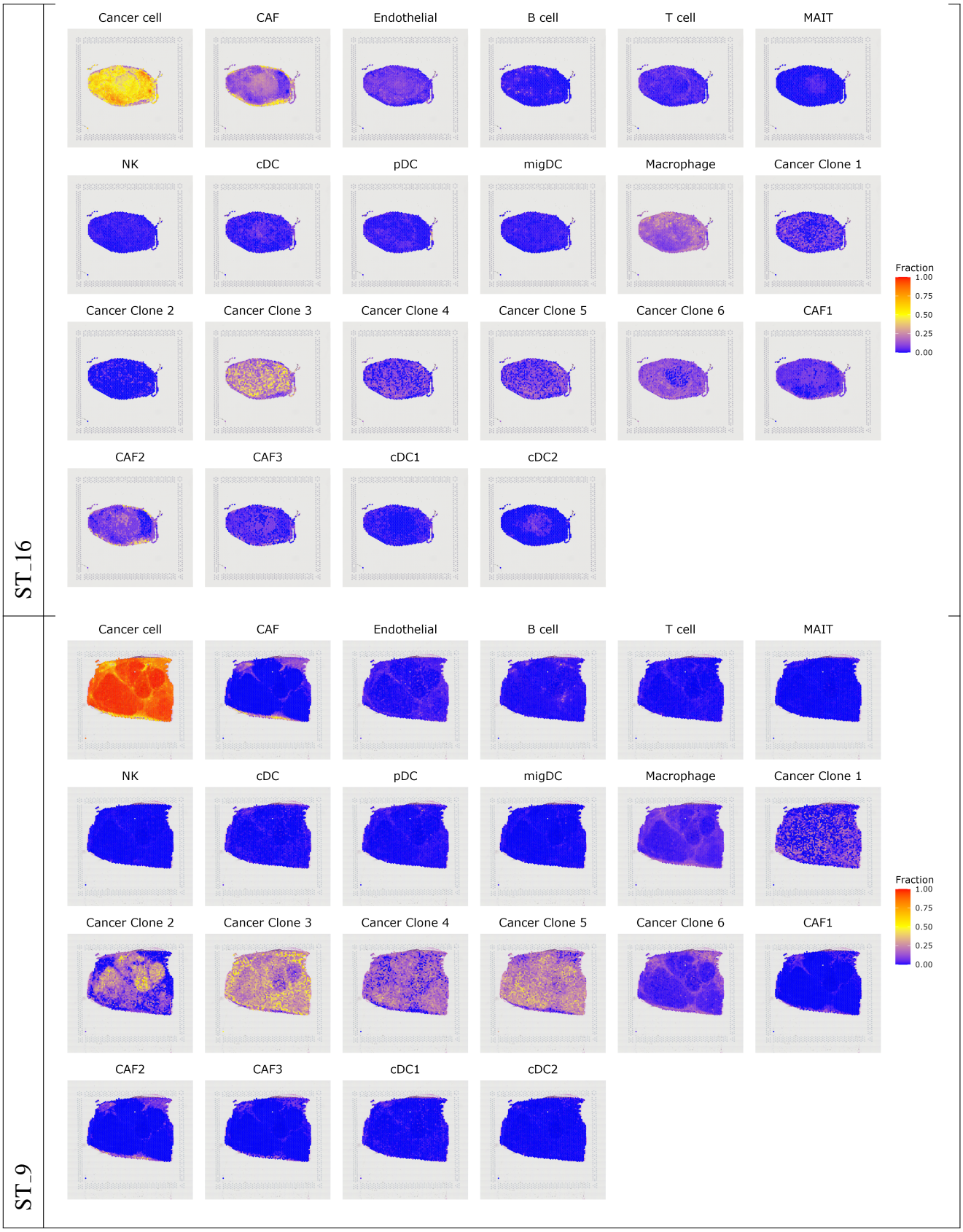

ST_17
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00
ST_10
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00


ST_13
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00

C.1.3 Treated STs (Responders)

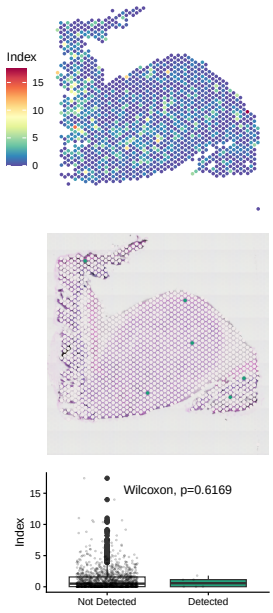
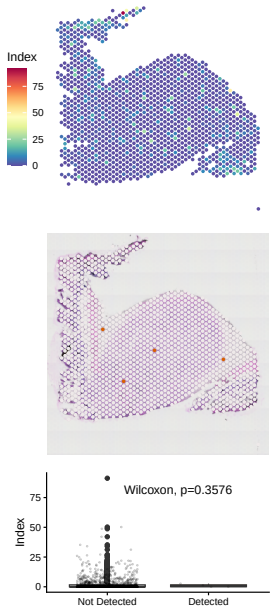
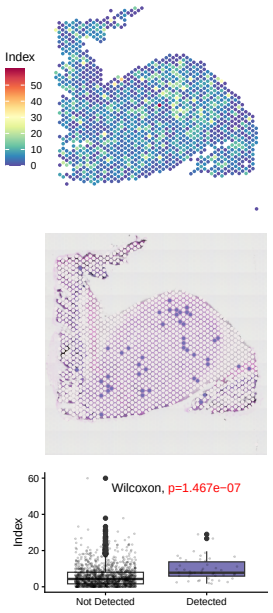
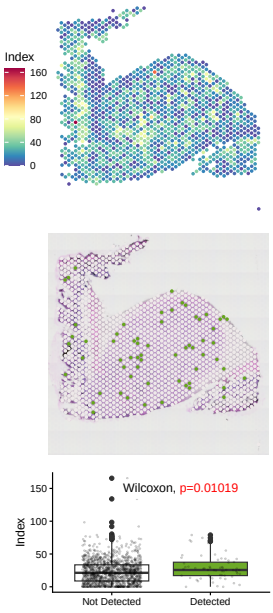
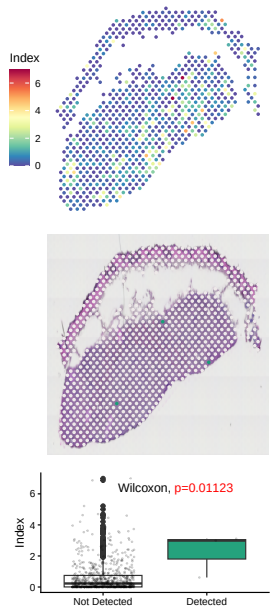
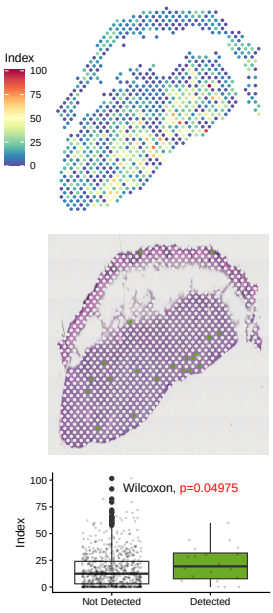


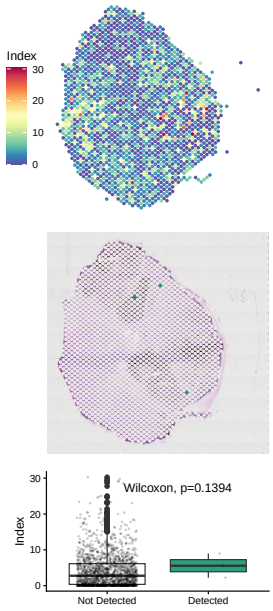
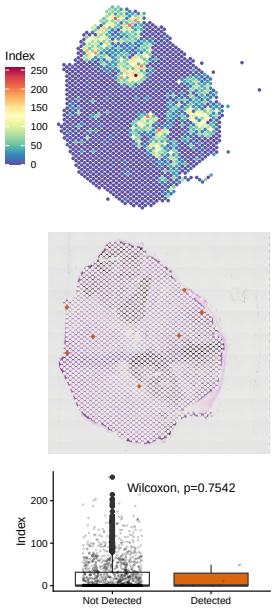
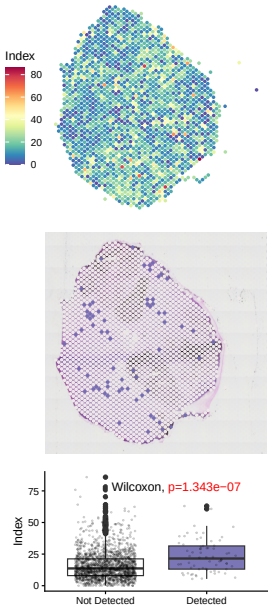
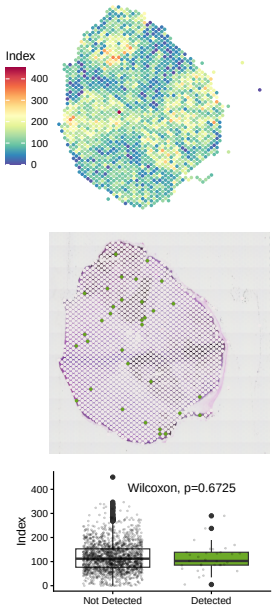
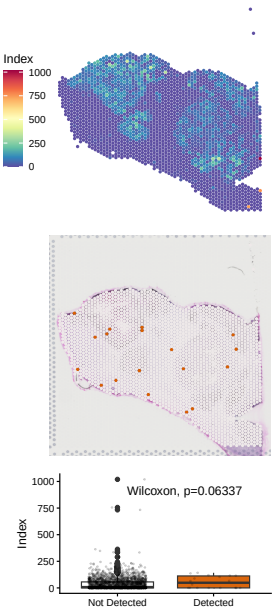
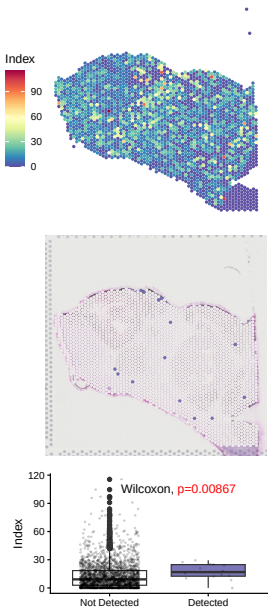

ST_3
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00
ST_15
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00

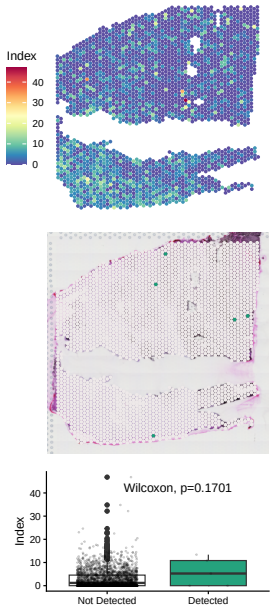
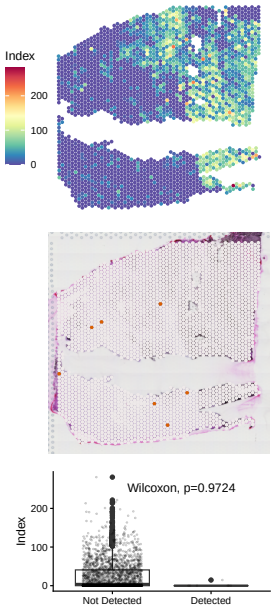
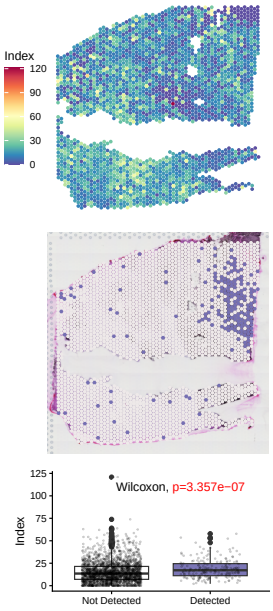
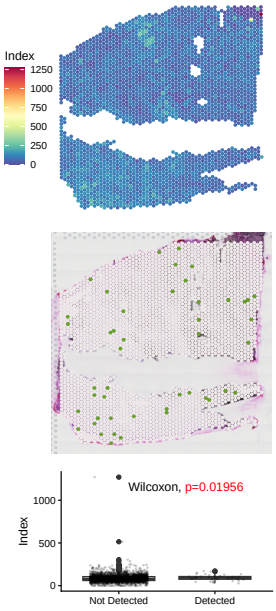

ST_16
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00
ST_9
Cancer cell
CAF
Endothelial
B cell
T cell
MAIT
NK
cDC
pDC
migDC
Macrophage
Cancer Clone 1
Cancer Clone 2
Cancer Clone 3
Cancer Clone 4
Cancer Clone 5
Cancer Clone 6
CAF1
CAF2
CAF3
cDC1
cDC2
Fraction
1.00
0.75
0.50
0.25
0.00

C.2 Validation of SpaCET-deconvolved tumor subclone proportions on 17 ST slides

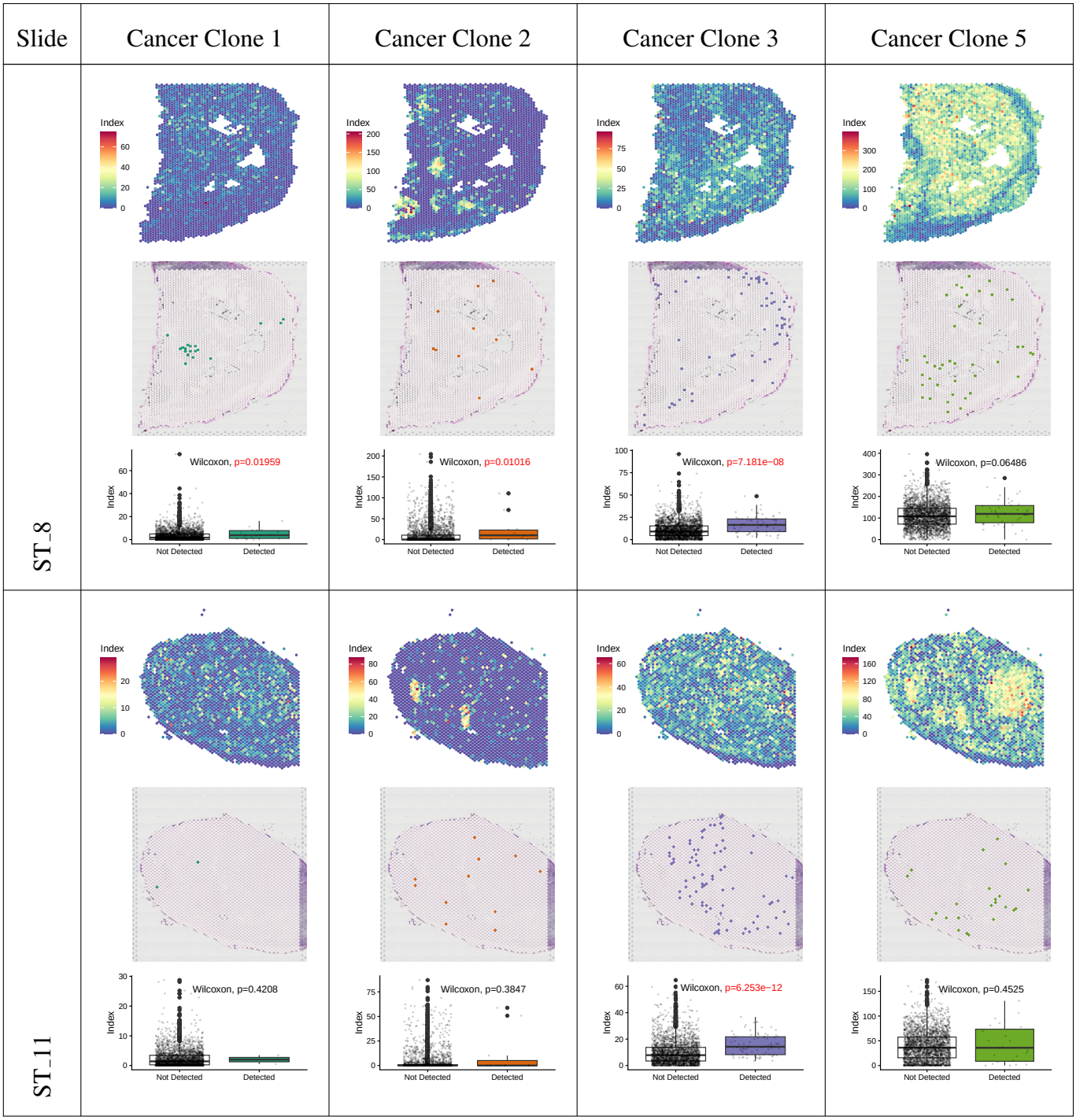
C.2.1 Control STs

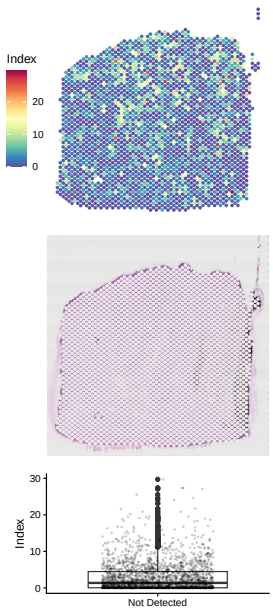
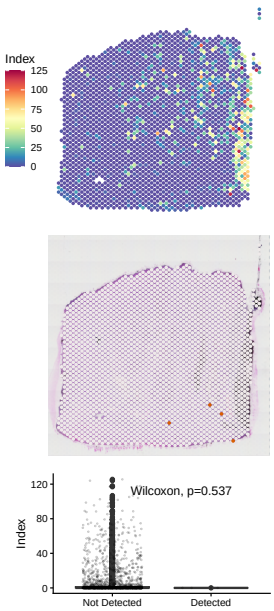
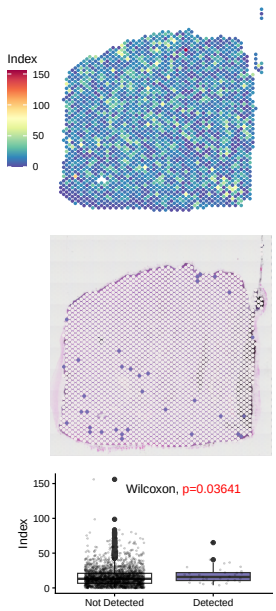
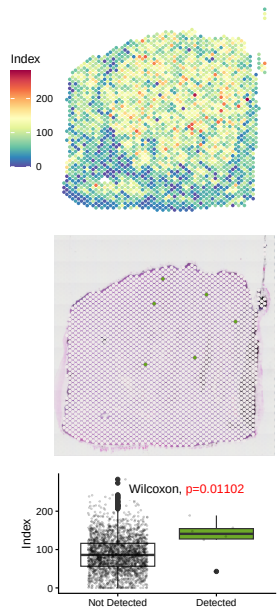
Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_1				
ST_5				

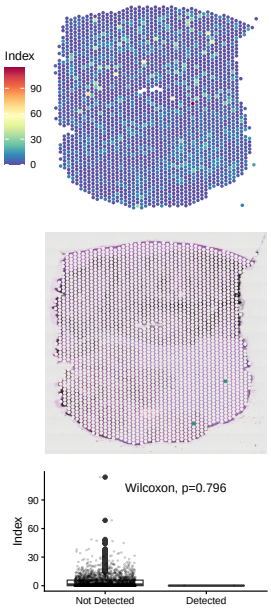
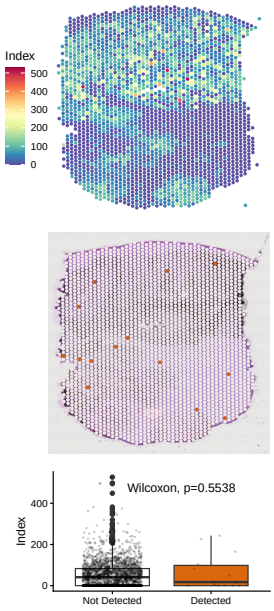
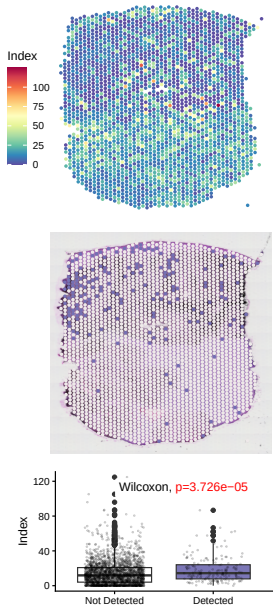
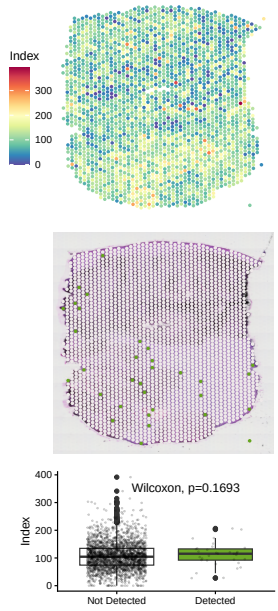
Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_7				
ST_14				

Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_2				
ST_6				

C.2.2 Treated STs (Non-responders)

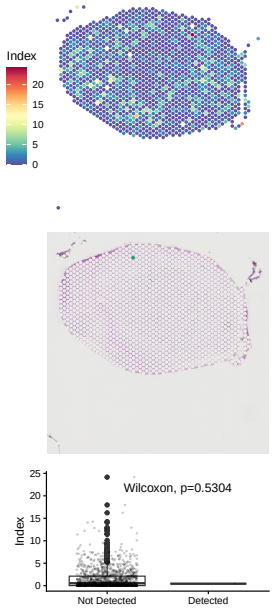
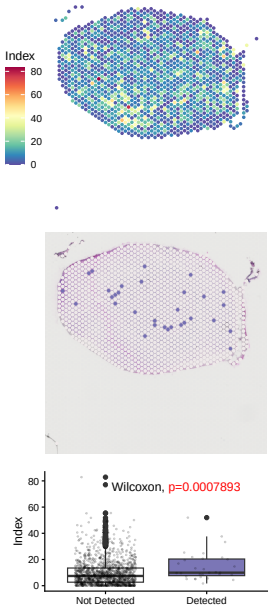
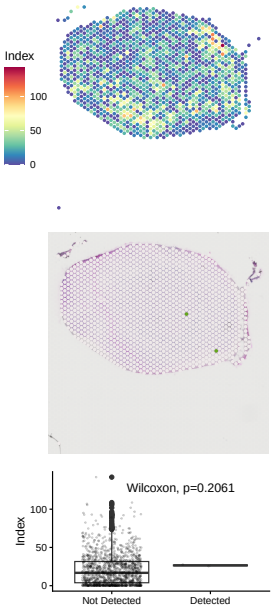
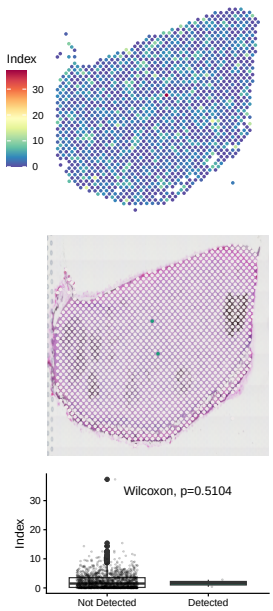
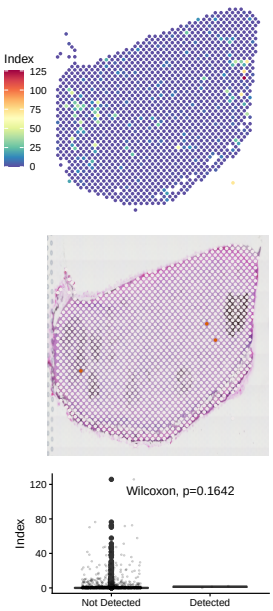
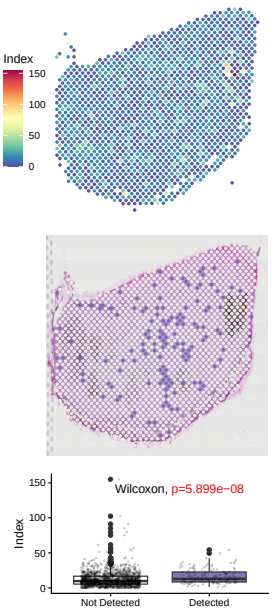
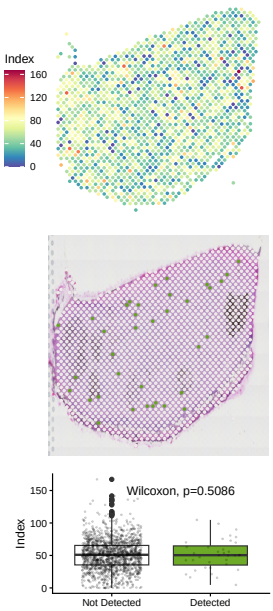


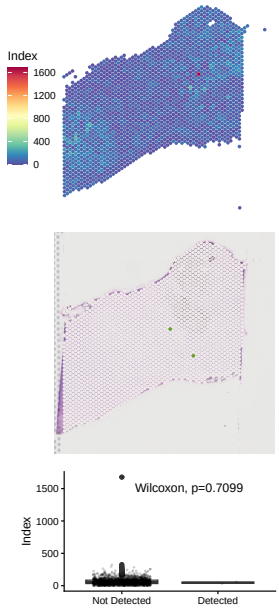
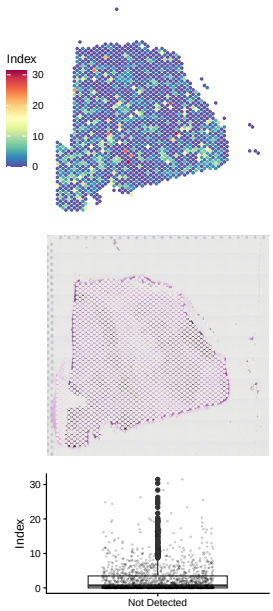
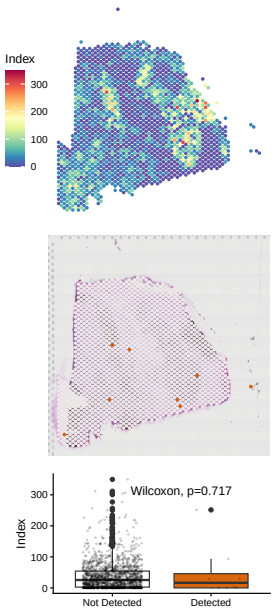
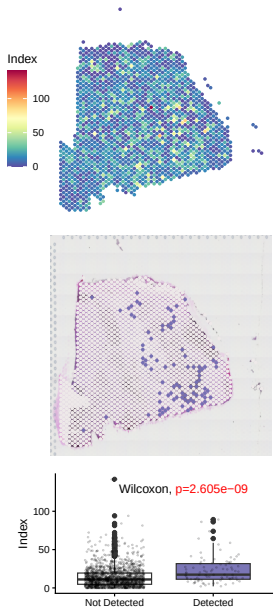
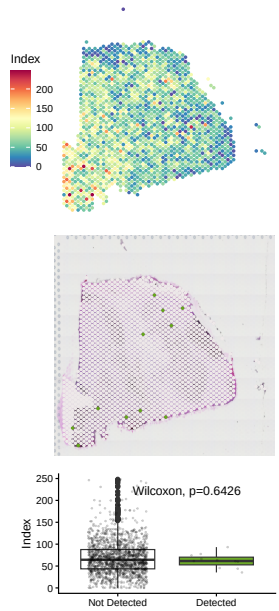
Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_9				
ST_12				

Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_10				

C.2.3 Treated STs (Responders)

Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_3				
ST_15				

Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_16				
ST_4				

Slide	Cancer Clone 1	Cancer Clone 2	Cancer Clone 3	Cancer Clone 5
ST_17				
ST_13				

Bibliography

[1] 1809-1882 Darwin, Charles. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. London : John Murray, 1859, 1859.

[2] P C NOWELL and D A HUNGERFORD. Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst*, 25:85–109, Jul 1960.

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

[4] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[5] Motoo Kimura. THE NUMBER OF HETEROZYGOUS NUCLEOTIDE SITES MAINTAINED IN A FINITE POPULATION DUE TO STEADY FLUX OF MUTATIONS. *Genetics*, 61(4):893–903, 04 1969.

[6] Philip J Stephens, Chris D Greenman, Beiyuan Fu, Fengtang Yang, Graham R Bignell, Laura J Mudie, Erin D Pleasance, King Wai Lau, David Beare, Lucy A Stebbings, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell*, 144(1):27–40, 2011.

[7] Yilong Li, Nicola D Roberts, Jeremiah A Wala, Ofer Shapira, Steven E Schumacher, Kiran Kumar, Ekta Khurana, Sebastian Waszak, Jan O Korbel, James E Haber, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*, 578(7793):112–121, 2020.

[8] Shahjehan A Wajed, Peter W Laird, and Tom R DeMeester. Dna methylation: an alternative pathway to cancer. *Annals of surgery*, 234(1):10–20, 2001.

[9] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature Methods*, 11(8):817–820, July 2014.

[10] Yu Hou, Huahu Guo, Chen Cao, Xianlong Li, Boqiang Hu, Ping Zhu, Xinglong Wu, Lu Wen, Fuchou Tang, Yanyi Huang, and Jirun Peng. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Research*, 26(3):304–319, 2016.

[11] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor

Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, 2016.

[12] E. Letouzé, Y. Allory, M. A. Bollet, F. Radvanyi, and F. Guyon. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome Biology*, 11(7):R76, 2010.

[13] W. Jiao, S. Vembu, A. G. Deshwar, L. Stein, and Q. Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15:35, 2014.

[14] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, Bouchard-Côté A., and S. P. Shah. Pyclone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014.

[15] C. A. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, Vij R., M. H. Tomasson, T. A. Graubert, M. J. Walter, Schierding W. Ellis, M. J., J. F. DiPersio, Ley T. J., E. R. Mardis, R. K. Wilson, and L. Ding. SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology*, 10(8):e1003665, 2014.

[16] R. F. Schwarz, A. Trinh, B. Sipos, J. D. Brenton, N. Goldman, and F. Markowetz. Phylogenetic quantification of intra-tumour heterogeneity. *PLoS Computational Biology*, 10(4):e1003535, 2014.

[17] Tom L. Kaufmann, Marina Petkovic, Thomas B. K. Watkins, Emma C. Colliver, Sofya Laskina, Nisha Thapa, Darlan C. Minussi, Nicholas Navin, Charles Swanton, Peter Van Loo, Kerstin Haase, Maxime Tarabichi, and Roland F. Schwarz. Medicc2: whole-genome doubling aware copy-number phylogenies for cancer evolution. *Genome Biology*, 23(1):241, 2022.

[18] A.G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein, and Q. Morris. Phylowgs: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16:35, 2015.

[19] M. El-Kebir, L. Oesper, H. Acheson-Field, and B.J. Raphael. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinformatics*, 31(12):i62–i70, 2015.

[20] S. Malikic, A. A. McPherson, N. Donmez, and C. S. Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 2015.

[21] Yuchao Jiang, Yu Qiu, Andy J Minn, and Nancy R Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences USA*, page 201522203, 2016.

[22] J.G. Reiter, A.P. Makohon-Moore, J.M. Gerold, I. Bozic, K. Chatterjee, C.A. Iacobuzio-Donahue, B. Vogelstein, and M.A. Nowak. Reconstructing metastatic seeding patterns of human cancers. *Nature Communications*, 8:14114, 2017.

[23] N. Beerenwinkel, R. F. Schwarz, M. Gerstung, and F. Markowetz. Cancer evolution: mathematical models and computational inference. *Systematic Biology*, 64(1):e1–e25, 2015.

[24] S. Turajlic and C. Swanton. Metastasis as an evolutionary process. *Science*, 352(6282):169–175, 2016.

[25] P. K. Brastianos, S. L. Carter, S. Santagata, D. P. Cahill, Taylor-Weiner A., R. T. Jones, E. M. Van Allen, M. S. Lawrence, P. M. Horowitz, K. Cibulskis, et al. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discovery*, 5(11):1164–1177, 2015.

[26] G. Gundem, P. Van Loo, B. Kremeyer, L. B. Alexandrov, J. M. Tubio, E. Papemmanuil, D. S. Brewer, H. M. Kallio, G. Högnás, M. Annala, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*, 520(7547):353–357, 2015.

[27] Z.-M. Zhao, B. Zhao, Y. Bai, A. Iamarino, S. G. Gaffney, J. Schlessinger, R. P. Lifton, D. L. Rimm, and J. P. Townsend. Early and multiple origins of metastatic lineages within primary tumors. *Proceedinsgs of the National Academy of Sciences USA*, 113(8):2140–2145, 2016.

[28] W.S. Hong, M. Shpak, and J.P. Townsend. Inferring the origin of metastases from cancer phylogenies. *Cancer Research*, 75(19):421–425, 2015.

[29] M. El-Kebir, G. Satas, and B.J Raphael. Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics*, 50(5):718–726, 2018.

[30] J.J. Quinn, M.G. Jones, R.A. Okimoto, S. Nanjo, M.M. Chan, N. Yosef, T. G. Bivona, and J. S. Weissman. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science*, 371(6532):eabc1944, 2021.

[31] Gryte Satas, Simone Zaccaria, Geoffrey Mon, and Benjamin J Raphael. Scarlet: Single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Syst*, 10(4):323–332, Apr 2020.

[32] Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. *Genome biology*, 17(1):86, 2016.

[33] Edith M Ross and Florian Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. *Genome biology*, 17(1):69, 2016.

[34] Jochen Singer, Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Single-cell mutation identification via phylogenetic inference. *Nature communications*, 9(1):5144, 2018.

[35] Xuan Liu, Jason I. Griffiths, Isaac Bishara, Jiayi Liu, Andrea H. Bild, and Jeffrey T. Chang. Phylogenetic inference from single-cell rna-seq data. *Scientific Reports*, 13(1):12854, 2023.

[36] Zilu Zhou, Bihui Xu, Andy Minn, and Nancy R. Zhang. Dendro: genetic heterogeneity profiling and subclone detection by single-cell rna sequencing. *Genome Biology*, 21(1):10, 2020.

[37] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M. Keizer, Indu Khatri, Szymon M. Kielbasa, Jan O. Korbel, Alexey M. Kozlov, Tzu-Hao Kuo, Boudewijn P. F. Lelieveldt, Ion I. Mandoiu, John C. Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J. Theis, Huan Yang, Alex Zelikovsky, Alice C. McHardy, Benjamin J. Raphael, Sohrab P. Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, 2020.

[38] Katie A Fennell, Dane Vassiliadis, Enid YN Lam, Luciano G Martelotto, Jesse J Balic, Sebastian Hollizeck, Tom S Weber, Timothy Semple, Qing Wang, Denise C Miles, et al. Non-genetic determinants of malignant clonal fitness at single-cell resolution. *Nature*, 601(7891):125–131, 2022.

[39] Mirang Kim and Joseph Costello. DNA methylation: an epigenetic mark of cellular memory. *Experimental & Molecular Medicine*, 49(4):e322–e322, April 2017.

[40] Toshikazu Ushijima, Naoko Watanabe, Eriko Okochi, Atsushi Kaneda, Takashi Sugimura, and Kazuaki Miyamoto. Fidelity of the methylation pattern and its variation in the genome. *Genome Res*, 13(5):868–874, May 2003.

[41] Tamir Biezuner, Adam Spiro, Ofir Raz, Shiran Amir, Lilach Milo, Rivka Adar, Noa Chapal-Ilani, Veronika Berman, Yael Fried, Elena Ainbinder, Galit Cohen, Haim M Barr, Ruth Halaban, and Ehud Shapiro. A generic, cost-effective, and scalable cell lineage analysis platform. *Genome Res*, 26(11):1588–1599, Nov 2016.

[42] Federico Gaiti, Ronan Chaligne, Hongcang Gu, Ryan M. Brand, Steven Kothen-Hill, Rafael C. Schulman, Kirill Grigorev, Davide Risso, Kyu-Tae Kim, Alessandro Pastore, Kevin Y. Huang, Alicia Alonso, Caroline Sheridan, Nathaniel D. Omans, Evan Biederstedt, Kendell Clement, Lili Wang, Joshua A. Felsenfeld, Erica B. Bhavsar, Martin J. Aryee, John N. Allan, Richard Furman, Andreas Gnirke, Catherine J. Wu, Alexander Meissner, and Dan A. Landau. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature*, 569(7757):576–580, May 2019.

[43] Shuhui Bian, Yu Hou, Xin Zhou, Xianlong Li, Jun Yong, Yicheng Wang, Wendong Wang, Jia Yan, Boqiang Hu, Hongshan Guo, Jilian Wang, Shuai Gao, Yunuo Mao, Ji Dong, Ping Zhu, Dianrong Xiu, Liying Yan, Lu Wen, Jie Qiao, Fuchou Tang, and Wei Fu. Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science*, 362(6418):1060–1063, November 2018.

[44] Ronan Chaligne, Federico Gaiti, Dana Silverbush, Joshua S. Schiffman, Hannah R. Weisman, Lloyd Kluegel, Simon Gritsch, Sunil D. Deochand, L. Nicolas Gonzalez Castro, Alyssa R. Richman, Johanna Klughammer, Tommaso Biancalani, Christoph Muus, Caroline Sheridan,

Alicia Alonso, Franco Izzo, Jane Park, Orit Rozenblatt-Rosen, Aviv Regev, Mario L. Suvà, and Dan A. Landau. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nature Genetics*, 53(10):1469–1479, 2021.

[45] Zohar Meir, Zohar Mukamel, Elad Chomsky, Aviezer Lifshitz, and Amos Tanay. Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. *Nature Genetics*, 52(7):709–718, Jul 2020.

[46] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Nae-Chyun Chen, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Ian T. Fiddes, Giulio Formenti, Robert S. Fulton, Arkarachai Fungtammasan, Erik Garrison, Patrick G. S. Grady, Tina A. Graves-Lindsay, Ira M. Hall, Nancy F. Hansen, Gabrielle A. Hartley, Marina Haukness, Kerstin Howe, Michael W. Hunkapiller, Chirag Jain, Miten Jain, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, Mikhail Kolmogorov, Jonas Korlach, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Ying Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sović, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Françoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz, Evan E. Eichler, Karen H. Miga, and Adam M. Phillippy. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022.

[47] Mario Fruzangohar, William A Timmins, Olena Kravchuk, and Julian Taylor. HaploMaker: An improved algorithm for rapid haplotype assembly of genomic sequences. *GigaScience*, 11, 05 2022. giac038.

[48] Peter Edge, Vineet Bafna, and Vikas Bansal. Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res*, 27(5):801–812, May 2017.

[49] Volodymyr Kuleshov. Probabilistic single-individual haplotyping. *Bioinformatics*, 30(17):i379–i385, 08 2014.

[50] Murray Patterson, Tobias Marschall, Nadia Pisanti, Leo van Iersel, Leen Stougie, Gunnar W. Klau, and Alexander Schönhuth. Whatshap: Weighted haplotype assembly for future-generation sequencing reads. *Journal of Computational Biology*, 22(6):498–509, 2015. PMID: 25658651.

[51] Jana Ebler, Marina Haukness, Trevor Pesout, Tobias Marschall, and Benedict Paten. Haplotype-aware diplotyping from noisy long reads. *Genome Biology*, 20(1):116, 2019.

[52] Kristen R. Maynard, Leonardo Collado-Torres, Lukas M. Weber, Cedric Uytingco, Brianna K. Barry, Stephen R. Williams, Joseph L. Catallini, Matthew N. Tran, Zachary Besich, Madhavi Tippani, Jennifer Chew, Yifeng Yin, Joel E. Kleinman, Thomas M. Hyde, Nikhil Rao, Stephanie C. Hicks, Keri Martinowich, and Andrew E. Jaffe. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature Neuroscience*, 24(3):425–436, 2021.

[53] Mirjana Efremova, Miquel Vento-Tormo, Sarah A Teichmann, and Roser Vento-Tormo. Cellphonedb: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature protocols*, 15(4):1484–1506, 2020.

[54] Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1088, 2021.

[55] Floriane Noël, Lucile Massenet-Regad, Irit Carmi-Levy, Antonio Cappuccio, Maximilien Grandclaudon, Coline Trichot, Yann Kieffer, Fatima Mechta-Grigoriou, and Vassili Soumelis. Dissection of intercellular communication using the transcriptome-based framework icellnet. *Nature communications*, 12(1):1089, 2021.

[56] Robin Browaeys, Wouter Saelens, and Yvan Saeys. Nichenet: modeling intercellular communication by linking ligands to target genes. *Nature methods*, 17(2):159–162, 2020.

[57] Simon Cabello-Aguilar, Mélissa Alame, Fabien Kon-Sun-Tack, Caroline Fau, Matthieu Lacroix, and Jacques Colinge. Singlecellsignalr: inference of intercellular networks from single-cell transcriptomics. *Nucleic acids research*, 48(10):e55–e55, 2020.

[58] Scott R Tyler, Pavana G Rotti, Xingshen Sun, Yaling Yi, Weiliang Xie, Michael C Winter, Miles J Flamme-Wiese, Budd A Tucker, Robert F Mullins, Andrew W Norris, et al. Pyminer finds gene and autocrine-paracrine networks from human islet scrna-seq. *Cell reports*, 26(7):1951–1964, 2019.

[59] Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, 22(1):78, 2021.

[60] Shila Ghazanfar, Yingxin Lin, Xianbin Su, David Ming Lin, Ellis Patrick, Ze-Guang Han, John C Marioni, and Jean Yee Hwa Yang. Investigating higher-order interactions in single-cell data with schot. *Nature methods*, 17(8):799–806, 2020.

[61] Matthew N Bernstein, Zijian Ni, Aman Prasad, Jared Brown, Chitrasen Mohanty, Ron Stewart, Michael A Newton, and Christina Kendziorski. Spatialcorr identifies gene sets with spatially varying correlation structure. *Cell Reports Methods*, 2(12), 2022.

[62] Damien Arnol, Denis Schapiro, Bernd Bodenmiller, Julio Saez-Rodriguez, and Oliver Stegle. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell reports*, 29(1):202–211, 2019.

[63] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):2084, 2020.

[64] Xin Shao, Chengyu Li, Haihong Yang, Xiaoyan Lu, Jie Liao, Jingyang Qian, Kai Wang, Junyun Cheng, Penghui Yang, Huajun Chen, Xiao Xu, and Xiaohui Fan. Knowledge-graph-based cell-cell communication inference for spatially resolved transcriptomic data with spatalk. *Nature Communications*, 13(1):4429, 2022.

[65] Zhuoxuan Li, Tianjie Wang, Pentao Liu, and Yuanhua Huang. Spatialdm for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns. *Nature Communications*, 14(1):3995, 2023.

[66] Zixuan Cang, Yanxiang Zhao, Axel A. Almet, Adam Stabell, Raul Ramos, Maksim V. Plikus, Scott X. Atwood, and Qing Nie. Screening cell–cell communication in spatial transcriptomics via collective optimal transport. *Nature Methods*, 20(2):218–228, 2023.

[67] Shiying Li, Ruohan Wang, Sitong Liu, and Shuai Cheng Li. Finding spatially variable ligand-receptor interactions with functional support from downstream genes. *Nature Communications*, 16(1):7784, 2025.

[68] Hongshan Guo, Ping Zhu, Fan Guo, Xianlong Li, Xinglong Wu, Xiaoying Fan, Lu Wen, and Fuchou Tang. Profiling dna methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nature Protocols*, 10(5):645–659, 2015.

[69] Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research*, 23(12):2126–2135, 2013.

[70] Jian Ma, Aakrosh Ratan, Brian J. Raney, Bernard B. Suh, Webb Miller, and David Haussler. The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, 105(38):14254–14261, 2008.

[71] Louis Dollo. The laws of evolution. *Bull. Soc. Bel. Geol. Paleontol*, 7:164–166, 1893.

[72] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. Sustainable data analysis with snakemake. *F1000Res*, 10:33, 2021.

[73] Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2022.

[74] Nelly Olova, Felix Krueger, Simon Andrews, David Oxley, Rebecca V. Berrens, Miguel R. Branco, and Wolf Reik. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting dna methylation data. *Genome Biology*, 19(1):33, 2018.

[75] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 07 1987.

[76] Vincent Lefort, Richard Desper, and Olivier Gascuel. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, 32(10):2798–2800, 06 2015.

[77] David F Robinson and Leslie R Foulds. Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147, 1981.

[78] Seung-Jin Sul and Tiffani L Williams. An experimental analysis of robinson-foulds distance matrix algorithms. In *European Symposium on Algorithms*, pages 793–804. Springer, 2008.

[79] Andrew K. C. Wong and Manlai You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7(5):599–609, 1985.

[80] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[81] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951.

[82] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55:639–653, 2003.

[83] Dominik Maria Endres and J E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information Theory*, 49(7):1858– 1860, July 2003.

[84] Walter M. Fitch. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology*, 20(4):406–416, 1971.

[85] J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29(1):53–65, 1973.

[86] Arthur W. Lambert, Diwakar R. Pattabiraman, and Robert A. Weinberg. Emerging biological principles of metastasis. *Cell*, 168(4):670–691, 2017.

[87] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics*, 28(1):35–42, 1975.

[88] Christopher Jen-Yue Wei and Kun Zhang. Retrace: simultaneous retrospective lineage tracing and methylation profiling of single cells. *Genome Res*, 30(4):602–610, Apr 2020.

[89] Sally H Cross and Adrian P Bird. Cpg islands and genes. *Current opinion in genetics & development*, 5(3):309–314, 1995.

[90] Rafael A Irizarry, Christine Ladd-Acosta, Bo Wen, Zhijin Wu, Carolina Montano, Patrick Onyango, Hengmi Cui, Kevin Gabo, Michael Rongione, Maree Webster, et al. The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nature genetics*, 41(2):178–186, 2009.

[91] Farid Rashidi Mehrabadi, Kerrie L. Marie, Eva Pérez-Guijarro, Salem Malikić, Erfan Sadeqi Azer, Howard H. Yang, Can Kızılkale, Charli Gruen, Welles Robinson, Huaitian Liu, Michael C. Kelly, Christina Marcelus, Sandra Burkett, Aydın Buluç, Funda Ergün, Maxwell P. Lee, Glenn Merlino, Chi-Ping Day, and S. Cenk Sahinalp. Profiles of expressed mutations in single cells reveal subclonal expansion patterns and therapeutic impact of intratumor heterogeneity. *bioRxiv*, 2021.

[92] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, November 2014.

[93] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[94] The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, 1 2021.

[95] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. Gorilla: a tool for discovery and visualization of enriched go terms in ranked gene lists. *BMC bioinformatics*, 10(1):1–7, 2009.

[96] Eran Eden, Doron Lipson, Sivan Yogev, and Zohar Yakhini. Discovering motifs in ranked lists of dna sequences. *PLoS computational biology*, 3(3):e39, 2007.

[97] M Ryan Corces, Jeffrey M Granja, Shadi Shams, Bryan H Louie, Jose A Seoane, Wanding Zhou, Tiago C Silva, Clarice Groeneveld, Christopher K Wong, Seung Woo Cho, et al. The chromatin accessibility landscape of primary human cancers. *Science*, 362(6413):eaav1898, 2018.

[98] Pingping Zhu, Tiankun Lu, Zhenzhen Chen, Benyu Liu, Dongdong Fan, Chong Li, Jiayi Wu, Luyun He, Xiaoxiao Zhu, Ying Du, et al. 5-hydroxytryptamine produced by enteric serotonergic neurons initiates colorectal cancer stem cell self-renewal and tumorigenesis. *Neuron*, 2022.

[99] Glenn Rademakers, Maartje Massen, Alexander Koch, Muriel X Draht, Nikkie Buekers, Kim AD Wouters, Nathalie Vaes, Tim De Meyer, Beatriz Carvalho, Gerrit A Meijer, et al. Identification of dna methylation markers for early detection of crc indicates a role for nervous system-related genes in crc. *Clinical epigenetics*, 13(1):1–12, 2021.

[100] Brad T Sherman, Ming Hao, Ju Qiu, Xiaoli Jiao, Michael W Baseler, H Clifford Lane, Tomozumi Imamichi, and Weizhong Chang. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*, 50(W1):W216–21, Mar 2022.

[101] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

[102] Howard Cedar and Yehudit Bergman. Linking dna methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5):295–304, 2009.

[103] Jiamu Du, Lianna M Johnson, Steven E Jacobsen, and Dinshaw J Patel. Dna methylation pathways and their crosstalk with histone methylation. *Nature reviews Molecular cell biology*, 16(9):519–532, 2015.

[104] Isidro Cortés-Ciriano, Doga C Gulhan, Jake June-Koo Lee, Giorgio EM Melloni, and Peter J Park. Computational analysis of cancer genome sequencing data. *Nature Reviews Genetics*, 23(5):298–314, 2022.

[105] Ayse Keskus, Asher Bryant, Tanveer Ahmad, Byunggil Yoo, Sergey Aganezov, Anton Goretsky, Ataberk Donmez, Lisa A Lansdon, Isabel Rodriguez, Jimin Park, Yuelin Liu, Xiwen Cui, Joshua Gardner, Brandy McNulty, Samuel Sacco, Jyoti Shetty, Yongmei Zhao, Bao Tran, Giuseppe Narzisi, Adrienne Helland, Daniel E Cook, Pi-Chuan Chang, Alexey Kolesnikov, Andrew Carroll, Erin K Molloy, Irina Pushel, Erin Guest, Tomi Pastinen, Kishwar Shafin, Karen H Miga, Salem Malikic, Chi-Ping Day, Nicolas Robine, Cenk Sahinalp, Michael Dean, Midhat S Farooqi, Benedict Paten, and Mikhail Kolmogorov. Severus: accurate detection and characterization of somatic structural variation in tumor genomes using long reads. *medRxiv*, Mar 2024.

[106] Simone Zaccaria and Benjamin J Raphael. Characterizing allele-and haplotype-specific copy numbers in single cells with chisel. *Nature biotechnology*, 39(2):207–214, 2021.

[107] Nelly Olova, Felix Krueger, Simon Andrews, David Oxley, Rebecca V Berrens, Miguel R Branco, and Wolf Reik. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting dna methylation data. *Genome biology*, 19:1–19, 2018.

[108] Qiuhui Li, Ayse G Keskus, Justin Wagner, Michal B Izydorczyk, Winston Timp, Fritz J Sedlazeck, Alison P Klein, Justin M Zook, Mikhail Kolmogorov, and Michael C Schatz. Unraveling the hidden complexity of cancer through long-read sequencing. *Genome research*, 35(4):599–620, 2025.

[109] Hillary Elrick, Carolin M Sauer, Jose Espejo Valle-Inclan, Katherine Trevers, Melanie Tanguy, Sonia Zumalave, Solange De Noon, Francesc Muyas, Rita Cascão, Angela Afonso, et al. Savana: reliable analysis of somatic structural variants and copy number aberrations using long-read sequencing. *Nature Methods*, pages 1–11, 2025.

[110] Kieran O'Neill, Erin Pleasance, Jeremy Fan, Vahid Akbari, Glenn Chang, Katherine Dixon, Veronika Csizmok, Signe MacLennan, Vanessa Porter, Andrew Galbraith, et al. Long-read sequencing of an advanced cancer cohort resolves rearrangements, unravels haplotypes, and reveals methylation landscapes. *Cell Genomics*, 4(11), 2024.

[111] Gilad D Evrony, Anjali Gupta Hinch, and Chongyuan Luo. Applications of single-cell dna sequencing. *Annual review of genomics and human genetics*, 22(1):171–197, 2021.

[112] Nicholas Williams, Joe Lee, Emily Mitchell, Luiza Moore, E Joanna Baxter, James Hewinson, Kevin J Dawson, Andrew Menzies, Anna L Godfrey, Anthony R Green, et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature*, 602(7895):162–168, 2022.

[113] Emily Mitchell, Michael Spencer Chapman, Nicholas Williams, Kevin J Dawson, Nicole Mende, Emily F Calderbank, Hyunchul Jung, Thomas Mitchell, Tim HH Coorens, David H Spencer, et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, 606(7913):343–350, 2022.

[114] Arcadi Cipponi, David L Goode, Justin Bedo, Mark J McCabe, Marina Pajic, David R Croucher, Alvaro Gonzalez Rajal, Simon R Junankar, Darren N Saunders, Pavel Lobachevsky, et al. Mtor signaling orchestrates stress-induced mutagenesis, facilitating adaptive evolution in cancer. *Science*, 368(6495):1127–1131, 2020.

[115] Eva Pérez-Guijarro, Howard H Yang, Romina E Araya, Rajaa El Meskini, Helen T Michael, Suman Kumar Vodnala, Kerrie L Marie, Cari Smith, Sung Chin, Khiem C Lam, et al. Multimodel preclinical platform predicts clinical response of melanoma to immunotherapy. *Nature medicine*, 26(5):781–791, 2020.

[116] Charli Gruen, Howard H Yang, Antonella Sassano, Emily Wu, Vishaka Gopalan, Kerrie L Marie, Andrea Castro, Farid Rashidi Mehrabadi, Chih Hao Wu, Isabella Church, et al. Melanoma clonal subline analysis uncovers heterogeneity-driven immunotherapy resistance mechanisms. *bioRxiv*, pages 2023–04, 2023.

[117] Farid Rashidi Mehrabadi, Kerrie L. Marie, Eva Pérez-Guijarro, Salem Malikić, Erfan Sadeqi Azer, Howard H. Yang, Can Kızılkale, Charli Gruen, Welles Robinson, Huaitian Liu, Michael C. Kelly, Christina Marcelus, Sandra Burkett, Aydın Buluç, Funda Ergün, Maxwell P. Lee, Glenn Merlino, Chi-Ping Day, and S. Cenk Sahinalp. Profiles of expressed mutations in single cells reveal subclonal expansion patterns and therapeutic impact of intratumor heterogeneity. *bioRxiv*, 2021.

[118] MG Hirsch, Soumitra Pal, Farid Rashidi Mehrabadi, Salem Malikic, Charli Gruen, Antonella Sassano, Eva Pérez-Guijarro, Glenn Merlino, S Cenk Sahinalp, Erin K Molloy, et al. Stochastic modeling of single-cell gene expression adaptation reveals non-genomic contribution to evolution of tumor subclones. *Cell Systems*, 16(1), 2025.

[119] Farid Rashidi Mehrabadi, Erfan Sadeqi Azer, John D. Bridgers, Eva Pérez-Guijarro, Kerrie L. Marie, Howard H. Yang, Charli Gruen, Chih Hao Wu, Welles Robinson, Huaitian Liu, Can Kizilkale, Michael C. Kelly, Cari Smith, Sung Chin, Jessica Ebersole, Sandra Burkett, Aydin Buluc, Maxwell P. Lee, Erin K. Molloy, Teresa M. Przytycka, Glenn Merlino, Chi-Ping Day, Salem Malikić, Funda Ergun, and S. Cenk Sahinalp. A partition function algorithm to evaluate inferred subclonal structures in single-cell sequencing data. In Sriram Sankararaman, editor, *Research in Computational Molecular Biology*, pages 409–413, Cham, 2025. Springer Nature Switzerland.

[120] Ryan Poplin, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, Jojo Dijamco, Nam Nguyen, Pegah T Afshar, Sam S Gross, Lizzie Dorfman, Cory Y McLean, and Mark A DePristo. A universal snp and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10):983–987, September 2018.

[121] Tanveer Ahmad, Mikhail Kolmogorov, and Ayse Gokce. Wakhan. `https://github.com/KolmogorovLab/Wakhan`, 2024.

[122] Kasper D. Hansen, Benjamin Langmead, and Rafael A. Irizarry. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10):R83, 2012.

[123] Yufeng Wu. Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics*, 36(3):742–750, 08 2019.

[124] Keith S Hoek, Ossia M Eichhoff, Natalie C Schlegel, Udo Dobbeling, Nikita Kobert, Leo Schaerer, Silvio Hemmi, and Reinhard Dummer. In vivo switching of human melanoma cells between proliferative and invasive states. *Cancer research*, 68(3):650–656, 2008.

[125] Chi-Ping Day, John Carter, Zoe Weaver Ohler, Carrie Bonomi, Rajaa El Meskini, Philip Martin, Cari Graff-Cherry, Lionel Feigenbaum, Thomas Tüting, Terry Van Dyke, et al. "glowing head" mice: a genetic tool enabling reliable preclinical image-based evaluation of cancers in immunocompetent allografts. *PloS one*, 9(11):e109956, 2014.

[126] Jessica A Rick, Chad D Brock, Alexander L Lewanski, Jimena Golcher-Benavides, and Catherine E Wagner. Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. *Systematic Biology*, 73(1):76–101, 10 2023.

[127] Zbyslaw Sondka, Nidhi Bindal Dhir, Denise Carvalho-Silva, Steven Jupe, Madhumita, Karen McLaren, Mike Starkey, Sari Ward, Jennifer Wilding, Madiha Ahmed, Joanna Argasinska, David Beare, Manpreet Singh Chawla, Stephen Duke, Ilaria Fasanella, Avirup Guha Neogi, Susan Haller, Balazs Hetenyi, Leonie Hodges, Alex Holmes, Rachel Lyne, Thomas Maurel, Sumodh Nair, Helder Pedro, Amaia Sangrador-Vegas, Helen Schuilenburg, Zoe Sheard, Siew Yit Yong, and Jon Teague. Cosmic: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Research*, 52(D1):D1210–D1217, 11 2023.

[128] S Kasar, J Kim, R Improgo, G Tiao, P Polak, N Haradhvala, MS Lawrence, A Kiezun, SM Fernandes, S Bahl, et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature communications*, 6(1):8866, 2015.

[129] Taichiro Nonaka, Yoshinobu Toda, Hiroshi Hiai, Munehiro Uemura, Motonobu Nakamura, Norio Yamamoto, Ryo Asato, Yukari Hattori, Kazuhisa Bessho, Nagahiro Minato, et al. Involvement of activation-induced cytidine deaminase in skin cancer development. *The Journal of clinical investigation*, 126(4):1367–1382, 2016.

[130] Barry Halliwell. Cell culture, oxidative stress, and antioxidants: avoiding pitfalls. *Biomed J*, 37(3):99–105, May-Jun 2014.

[131] Asher Mullard. The kras crowd targets its next cancer mutations. *Nat. Rev. Drug Discov*, 22:167–171, 2023.

[132] Douglas B Johnson, Jason Roszik, Alexander N Shoushtari, Zeynep Eroglu, Justin Balko, Catherine Higham, Igor Puzanov, Sapna P Patel, Jeffrey A Sosman, and Scott E Woodman. Comparative analysis of the gnaq, gna11, sf3b1, and eif1ax driver mutations in melanoma and across the cancer spectrum. *Pigment cell & melanoma research*, 29(4):470, 2016.

[133] Amama Ghaffar, Tehmeena Akhter, Petter Strømme, Doriana Misceo, Amjad Khan, Eirik Frengen, Muhammad Umair, Bertrand Isidor, Benjamin Cogné, Asma A Khan, et al. Variants of nav3, a neuronal morphogenesis protein, cause intellectual disability, developmental delay, and microcephaly. *Communications Biology*, 7(1):831, 2024.

[134] Olga Bugaeva, Pilvi Maliniemi, Wenche S. Prestvik, Eeva Leivo, Nicolas Kluger, Alexander Salava, Sanna Virtanen, Kirsi Jäntti, Olli Saksela, Kaisa Lehti, and et al. Tumour suppressor neuron navigator 3 and matrix metalloproteinase 14 are co-expressed in most melanomas but downregulated in thick tumours. *Acta Dermato-Venereologica*, 103:adv00883, Mar. 2023.

[135] Laura Amicone, Francesca M Spagnoli, Gerald Späth, Silvia Giordano, Cristina Tommasini, Silvia Bernardini, Veronica De Luca, Carlo Della Rocca, Mary C Weiss, Paolo M Comoglio, et al. Transgenic expression in the liver of truncated met blocks apoptosis and permits immortalization of hepatocytes. *The EMBO journal*, 1997.

[136] Michael D Kinnaman, Simone Zaccaria, Alvin Makohon-Moore, Brian Arnold, Max F Levine, Gunes Gundem, Juan E Arango Ossa, Dominik Glodzik, M Irene Rodríguez-Sánchez, Nancy Bouvier, et al. Subclonal somatic copy-number alterations emerge and dominate in recurrent osteosarcoma. *Cancer Research*, 83(22):3796–3812, 2023.

[137] Lavinia Spain, Alexander Coulton, Irene Lobon, Andrew Rowan, Desiree Schnidrig, Scott TC Shepherd, Benjamin Shum, Fiona Byrne, Maria Goicoechea, Elisa Piperni, et al. Late-stage metastatic melanoma emerges through a diversity of evolutionary pathways. *Cancer discovery*, 13(6):1364–1385, 2023.

[138] Yilei Fu, Winston Timp, and Fritz J. Sedlazeck. Computational analysis of dna methylation from long-read sequencing. *Nature Reviews Genetics*, 2025.

[139] Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Alexandra Bignell, Sanjay Boddu, Paulo R Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Mehrnaz Charkhchi, Alexander Cockburn, Luca Da Rin Fiorretto, Claire Davidson, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Gurpreet S Ghattaoraya, Jose Gonzalez Martinez, Cristi Guijarro, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Mike Kay, Vinay Kaykala, Tuan Le, Diana Lemos, Diego Marques-Coelho, José Carlos Marugán, Gabriela Alejandra Merino, Louisse Paola Mirabueno, Aleena Mushtaq, Syed Nakib Hossain, Denye N Ogeh, Manoj Pandian Sakthivel, Anne Parker, Malcolm Perry, Ivana Piližota, Irina Prosovetskaia, José G Pérez-Silva, Ahamed Imran Abdul Salam, Nuno Saraiva-Agostinho, Helen Schuilenburg, Dan Sheppard, Swati Sinha, Botond Sipos, William Stark, Emily Steed, Ranjit Sukumaran, Dulika Sumathipala, Marie-Marthe Suner, Likhitha Surapaneni, Kyösti Sutinen, Michal Szpak, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Brandon Walts, Elizabeth Wass, Natalie Willhoft, Jamie Allen, Jorge Alvarez-Jarreta, Marc Chakiachvili, Bethany Flint, Stefano Giorgetti, Leanne Haggerty, Garth R Ilsley, Jane E Loveland, Benjamin Moore, Jonathan M Mudge, John Tate, David Thybert, Stephen J

Trevanion, Andrea Winterbottom, Adam Frankish, Sarah E Hunt, Magali Ruffier, Fiona Cunningham, Sarah Dyer, Robert D Finn, Kevin L Howe, Peter W Harrison, Andrew D Yates, and Paul Flicek. Ensembl 2023. *Nucleic Acids Research*, 51(D1):D933–D941, 11 2022.

[140] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G Izuogu, Julien Lagarde, Fergal J Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C P Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M Schmitt, Eloise Stapleton, Marie-Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S Choudhary, Mark Gerstein, Roderic Guigó, Tim J P Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L Tress, and Paul Flicek. Gencode reference annotation for the human and mouse genomes. *Nucleic Acids Res*, 47(D1):D766–D773, Jan 2019.

[141] René Dreos, Giovanna Ambrosini, Rouayda Cavin Périer, and Philipp Bucher. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic Acids Research*, 43(D1):D92–D96, 11 2014.

[142] Patrick Meylan, René Dreos, Giovanna Ambrosini, Romain Groux, and Philipp Bucher. EPD in 2020: enhanced data visualization and extension to ncRNA promoters. *Nucleic Acids Research*, 48(D1):D65–D69, 11 2019.

[143] Consortium ENCODE Project. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

[144] Yunhai Luo, Benjamin C Hitz, Idan Gabdank, Jason A Hilton, Meenakshi S Kagda, Bonita Lam, Zachary Myers, Paul Sud, Jennifer Jou, Khine Lin, Ulugbek K Baymuradov, Keenan Graham, Casey Litton, Stuart R Miyasato, J Seth Strattan, Otto Jolanki, Jin-Wook Lee, Forrest Y Tanaka, Philip Adenekan, Emma O'Neill, and J Michael Cherry. New developments on the encyclopedia of dna elements (encode) data portal. *Nucleic Acids Res*, 48(D1):D882–D889, Jan 2020.

[145] Benjamin C Hitz, Lee Jin-Wook, Otto Jolanki, Meenakshi S Kagda, Keenan Graham, Paul Sud, Idan Gabdank, J Seth Strattan, Cricket A Sloan, Timothy Dreszer, Laurence D Rowe, Nikhil R Podduturi, Venkat S Malladi, Esther T Chan, Jean M Davidson, Marcus Ho, Stuart Miyasato, Matt Simison, Forrest Tanaka, Yunhai Luo, Ian Whaling, Eurie L Hong, Brian T Lee, Richard Sandstrom, Eric Rynes, Jemma Nelson, Andrew Nishida, Alyssa Ingersoll, Michael Buckley, Mark Frerker, Daniel S Kim, Nathan Boley, Diane Trout, Alex Dobin, Sorena Rahmanian, Dana Wyman, Gabriela Balderrama-Gutierrez, Fairlie Reese, Neva C Durand, Olga Dudchenko, David Weisz, Suhas S P Rao, Alyssa Blackburn, Dimos Gkountaroulis, Mahdi Sadr, Moshe Olshansky, Yossi Eliaz, Dat Nguyen, Ivan Bochkov, Muhammad Saad Shamim, Ragini Mahajan, Erez Aiden, Tom Gingeras, Simon Heath, Martin Hirst, W James Kent, Anshul Kundaje, Ali Mortazavi, Barbara Wold, and J Michael Cherry. The encode uniform analysis pipelines. *bioRxiv*, Apr 2023.

[146] Federico Abascal, Reyes Acosta, Nicholas J. Addleman, Jessika Adrian, Veena Afzal, Rizi Ai, Bronwen Aken, Jennifer A. Akiyama, Omar Al Jammal, Henry Amrhein, Stacie M. Anderson, et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[147] Andrea Casazza, Veronica Finisguerra, Lorena Capparuccia, Andrea Camperi, Jakub M. Swiercz, Sabrina Rizzolio, Charlotte Rolny, Claus Christensen, Andrea Bertotti, Ivana Sarotto, Mauro Risio, Livio Trusolino, Jurgen Weitz, Martin Schneider, Massimilano Mazzone, Paolo M. Comoglio, and Luca Tamagnone. Sema3e–plexin d1 signaling drives human cancer cell invasiveness and metastatic spreading in mice. *The Journal of Clinical Investigation*, 120(8):2684–2698, 8 2010.

[148] Jiaxin Wen, Yanfeng Wang, Song Wang, Yuxin Liang, Xiaozhen Hu, Qiuxiang Ou, Hua Bao, Kuo Zhao, and Youyu Wang. Genetic and transcriptional insights into immune checkpoint blockade response and survival: lessons from melanoma and beyond. *Journal of Translational Medicine*, 23(1):467, 2025.

[149] M. Perego, M. Maurer, J X Wang, S. Shaffer, A C Müller, K. Parapatics, L. Li, D. Hristova, S. Shin, F. Keeney, S. Liu, X. Xu, A. Raj, J K Jensen, K L Bennett, S N Wagner, R. Somasundaram, and M. Herlyn. A slow-cycling subpopulation of melanoma cells with highly invasive properties. *Oncogene*, 37(3):302–312, 2018.

[150] Erbao Chen, Yuqian Mo, Jing Yi, Jie Liu, Ting Luo, Zheng Li, Zewei Lin, Yibing Hu, Zhilin Zou, and Jikui Liu. A novel hepatocellular carcinoma-specific mtorc1-related signature for anticipating prognosis and immunotherapy. *Aging (Albany NY)*, 15(16):7933–7955, Aug 2023.

[151] Mostafa Karimi Roshan, Arash Soltani, Anvar Soleimani, Kolsoum Rezaie Kahkhaie, Amir R Afshari, and Mohammad Soukhtanloo. Role of akt and mtor signaling pathways in the induction of epithelial-mesenchymal transition (emt) process. *Biochimie*, 165:229–234, 2019.

[152] Yochai Wolf, Osnat Bartok, Sushant Patkar, Gitit Bar Eli, Sapir Cohen, Kevin Litchfield, Ronen Levy, Alejandro Jiménez-Sánchez, Sophie Trabish, Joo Sang Lee, Hiren Karathia, Eilon Barnea, Chi-Ping Day, Einat Cinnamon, Ilan Stein, Adam Solomon, Lital Bitton, Eva Pérez-Guijarro, Tania Dubovik, Shai S. Shen-Orr, Martin L. Miller, Glenn Merlino, Yishai Levin, Eli Pikarsky, Lea Eisenbach, Arie Admon, Charles Swanton, Eytan Ruppin, and Yardena Samuels. Uvb-induced tumor heterogeneity diminishes immune response in melanoma. *Cell*, 179(1):219–235.e21, 2025/05/28 2019.

[153] Nicholas McGranahan and Charles Swanton. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell*, 168(4):613–628, 2017.

[154] Andrew McPherson, Andrew Roth, Emma Laks, Tehmina Masud, Ali Bashashati, Allen W Zhang, Gavin Ha, Justina Biele, Damian Yap, Adrian Wan, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature genetics*, 48(7):758–767, 2016.

[155] Lucy R Yates, Moritz Gerstung, Stian Knappskog, Christine Desmedt, Gunes Gundem, Peter Van Loo, Turid Aas, Ludmil B Alexandrov, Denis Larsimont, Helen Davies, et al. Subclonal

diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine*, 21(7):751–759, 2015.

[156] Tyler Funnell, Ciara H O'Flanagan, Marc J Williams, Andrew McPherson, Steven McKinney, Farhia Kabeer, Hakwoo Lee, Sohrab Salehi, Ignacio Vázquez-García, Hongyu Shi, et al. Single-cell genomic variation induced by mutational processes in cancer. *Nature*, 612(7938):106–115, 2022.

[157] Aino-Maija Leppä, Karen Grimes, Hyobin Jeong, Frank Y Huang, Alvaro Andrades, Alexander Waclawiczek, Tobias Boch, Anna Jauch, Simon Renders, Patrick Stelmach, et al. Single-cell multiomics analysis reveals dynamic clonal evolution and targetable phenotypes in acute myeloid leukemia with complex karyotype. *Nature genetics*, pages 1–14, 2024.

[158] Federico Gaiti, Ronan Chaligne, Hongcang Gu, Ryan M. Brand, Steven Kothen-Hill, Rafael C. Schulman, Kirill Grigorev, Davide Risso, Kyu-Tae Kim, Alessandro Pastore, Kevin Y. Huang, Alicia Alonso, Caroline Sheridan, Nathaniel D. Omans, Evan Biederstedt, Kendell Clement, Lili Wang, Joshua A. Felsenfeld, Erica B. Bhavsar, Martin J. Aryee, John N. Allan, Richard Furman, Andreas Gnirke, Catherine J. Wu, Alexander Meissner, and Dan A. Landau. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature*, 569(7757):576–580, 2019.

[159] Ronan Chaligne, Federico Gaiti, Dana Silverbush, Joshua S Schiffman, Hannah R Weisman, Lloyd Kluegel, Simon Gritsch, Sunil D Deochand, L Nicolas Gonzalez Castro, Alyssa R Richman, Johanna Klughammer, Tommaso Biancalani, Christoph Muus, Caroline Sheridan, Alicia Alonso, Franco Izzo, Jane Park, Orit Rozenblatt-Rosen, Aviv Regev, Mario L Suvà, and Dan A Landau. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat Genet*, 53(10):1469–1479, Oct 2021.

[160] Yuelin Liu, Xuan Cindy Li, Farid Rashidi Mehrabadi, Alejandro A. Schäffer, Drew Pratt, David R. Crawford, Salem Malikić, Erin K. Molloy, Vishaka Gopalan, Steve M. Mount, Eytan Ruppin, Kenneth D. Aldape, and S. Cenk Sahinalp. Single-cell methylation sequencing data reveal succinct metastatic migration histories and tumor progression models. *Genome Research*, 2023.

[161] Mengyang Chen, Ruijiang Fu, Yiqian Chen, Li Li, and Shou-Wen Wang. High-resolution, non-invasive single-cell lineage tracing in mice and humans based on dna methylation epimutations. *Nature Methods*, 22(3):488–498, 2025.

[162] Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, Eirikur Briem, Kun Zhang, Rafael A Irizarry, and Andrew P Feinberg. Increased methylation variation in epigenetic domains across cancer types. *Nature Genetics*, 43(8):768–775, 2011.

[163] Kevin C. Johnson, Kevin J. Anderson, Elise T. Courtois, Amit D. Gujar, Floris P. Barthel, Frederick S. Varn, Diane Luo, Martine Seignon, Eunhee Yi, Hoon Kim, Marcos R. H. Estecio, Dacheng Zhao, Ming Tang, Nicholas E. Navin, Rahul Maurya, Chew Yee Ngan, Niels Verburg, Philip C. de Witt Hamer, Ketan Bulsara, Michael L. Samuels, Sunit Das, Paul Robson, and Roel

G. W. Verhaak. Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response. *Nature Genetics*, 53(10):1456–1468, 2021.

[164] Fengju Chen, Yiqun Zhang, Wei Li, Fritz J. Sedlazeck, Lanlan Shen, and Chad J. Creighton. Global dna methylation differences involving germline structural variation impact gene expression in pediatric brain tumors. *Nature Communications*, 16(1):4713, 2025.

[165] Shuang G. Zhao, William S. Chen, Haolong Li, Adam Foye, Meng Zhang, Martin Sjöström, Rahul Aggarwal, Denise Playdle, Arnold Liao, Joshi J. Alumkal, Rajdeep Das, Jonathan Chou, Junjie T. Hua, Travis J. Barnard, Adina M. Bailey, Eric D. Chow, Marc D. Perry, Ha X. Dang, Rendong Yang, Ruhollah Moussavi-Baygi, Li Zhang, Mohammed Alshalalfa, S. Laura Chang, Kathleen E. Houlahan, Yu-Jia Shiah, Tomasz M. Beer, George Thomas, Kim N. Chi, Martin Gleave, Amina Zoubeidi, Robert E. Reiter, Matthew B. Rettig, Owen Witte, M. Yvonne Kim, Lawrence Fong, Daniel E. Spratt, Todd M. Morgan, Rohit Bose, Franklin W. Huang, Hui Li, Lisa Chesner, Tanushree Shenoy, Hani Goodarzi, Irfan A. Asangani, Shahneen Sandhu, Joshua M. Lang, Nupam P. Mahajan, Primo N. Lara, Christopher P. Evans, Phillip Febbo, Serafim Batzoglou, Karen E. Knudsen, Housheng H. He, Jiaoti Huang, Wilbert Zwart, Joseph F. Costello, Jianhua Luo, Scott A. Tomlins, Alexander W. Wyatt, Scott M. Dehm, Alan Ashworth, Luke A. Gilbert, Paul C. Boutros, Kyle Farh, Arul M. Chinnaiyan, Christopher A. Maher, Eric J. Small, David A. Quigley, and Felix Y. Feng. The dna methylation landscape of advanced prostate cancer. *Nature Genetics*, 52(8):778–789, 2020.

[166] Mikhail Kolmogorov, Kimberley J. Billingsley, Mira Mastoras, Melissa Meredith, Jean Monlong, Ryan Lorig-Roach, Mobin Asri, Pilar Alvarez Jerez, Laksh Malik, Ramita Dewan, Xylena Reed, Rylee M. Genner, Kensuke Daida, Sairam Behera, Kishwar Shafin, Trevor Pesout, Jeshuwin Prabakaran, Paolo Carnevali, Jianzhi Yang, Arang Rhie, Sonja W. Scholz, Bryan J. Traynor, Karen H. Miga, Miten Jain, Winston Timp, Adam M. Phillippy, Mark Chaisson, Fritz J. Sedlazeck, Cornelis Blauwendraat, and Benedict Paten. Scalable nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nature Methods*, 20(10):1483–1492, 2023.

[167] Tobias Rausch, Rene Snajder, Adrien Leger, Milena Simovic, Mădălina Giurgiu, Laura Villacorta, Anton G. Henssen, Stefan Fröhling, Oliver Stegle, Ewan Birney, Marc Jan Bonder, Aurelie Ernst, and Jan O. Korbel. Long-read sequencing of diagnosis and post-therapy medulloblastoma reveals complex rearrangement patterns and epigenetic signatures. *Cell Genomics*, 3(4):100281, 2023.

[168] Ji-Yeon Kim, Kyunghee Park, Woong-Yang Park, Jin Seok Ahn, Young-Hyuck Im, Jeong Eon Lee, Seok Won Kim, Seok Jin Nam, Jonghan Yu, and Yeon Hee Park. Prognostic value of structural variants in early breast cancer patients. *NPJ Breast Cancer*, 10(1):64, 2024.

[169] Katharina Filipski, Michael Scherer, Kim N Zeiner, Andreas Bucher, Johannes Kleemann, Philipp Jurmeister, Tabea I Hartung, Markus Meissner, Karl H Plate, Tim R Fenton, et al. Dna methylation-based prediction of response to immune checkpoint inhibition in metastatic melanoma. *Journal for immunotherapy of cancer*, 9(7):e002226, 2021.

[170] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature Protocols*, 9(1):171–181, 2014.

[171] Matthew A Myers, Simone Zaccaria, and Benjamin J Raphael. Identifying tumor clones in sparse single-cell mutation data. *Bioinformatics*, 36(Supplement$_1$) : $i186 - -i193, 072020$.

[172] Andrew McPherson, Ignacio Vázquez-García, Matthew A. Myers, Duaa H. Al-Rawi, Matthew Zatzman, Adam C. Weiner, Samuel Freeman, Neeman Mohibullah, Gryte Satas, Marc J. Williams, Nicholas Ceglia, Danguolė Norkūnaitė, Allen W. Zhang, Jun Li, Jamie L. P. Lim, Michelle Wu, Seongmin Choi, Eliyahu Havasov, Diljot Grewal, Hongyu Shi, Minsoo Kim, Roland F. Schwarz, Tom Kaufmann, Khanh Ngoc Dinh, Florian Uhlitz, Julie Tran, Yushi Wu, Ruchi Patel, Satish Ramakrishnan, DooA Kim, Justin Clarke, Hunter Green, Emily Ali, Melody DiBona, Nancy Varice, Ritika Kundra, Vance Broach, Ginger J. Gardner, Kara Long Roche, Yukio Sonoda, Oliver Zivanovic, Sarah H. Kim, Rachel N. Grisham, Ying L. Liu, Agnes Viale, Nicole Rusk, Yulia Lakhman, Lora H. Ellenson, Simon Tavaré, Samuel Aparicio, Dennis S. Chi, Carol Aghajanian, Nadeem R. Abu-Rustum, Claire F. Friedman, Dmitriy Zamarin, Britta Weigelt, Samuel F. Bakhoum, and Sohrab P. Shah. Ongoing genome doubling shapes evolvability and immunity in ovarian cancer. *Nature*, 644(8078):1078–1087, 2025.

[173] Lucie C. Gaspard-Boulinc, Luca Gortana, Thomas Walter, Emmanuel Barillot, and Florence M. G. Cavalli. Cell-type deconvolution methods for spatial transcriptomics. *Nature Reviews Genetics*, 2025.

[174] Beibei Ru, Jinlin Huang, Yu Zhang, Kenneth Aldape, and Peng Jiang. Estimation of cell lineages in tumors from spatial transcriptomics data. *Nature Communications*, 14(1):568, 2023.

[175] Sarah Davidson, Mirjana Efremova, Angela Riedel, Bidesh Mahata, Jhuma Pramanik, Jani Huuhtanen, Gozde Kar, Roser Vento-Tormo, Tzachi Hagai, Xi Chen, Muzlifah A. Haniffa, Jacqueline D. Shields, and Sarah A. Teichmann. Single-cell rna sequencing reveals a dynamic stromal niche that supports tumor growth. *Cell Reports*, 31(7):107628, 2020.

[176] Miranda V. Hunter, Reuben Moncada, Joshua M. Weiss, Itai Yanai, and Richard M. White. Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. *Nature Communications*, 12(1):6278, 2021.

[177] Bo Li and Colin N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[178] Kristiana Grigoriadis, Ariana Huebner, Abigail Bunkum, Emma Colliver, Alexander M. Frankell, Mark S. Hill, Kerstin Thol, Nicolai J. Birkbak, Charles Swanton, Simone Zaccaria, and Nicholas McGranahan. Conipher: a computational framework for scalable phylogenetic reconstruction with error correction. *Nature Protocols*, 19(1):159–183, 2024.

[179] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B. West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome Biology*, 16(1):91, 2015.

[180] Gryte Satas and Benjamin J Raphael. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics*, 33(14):i152–i160, 07 2017.

[181] Salem Malikic, Katharina Jahn, Jack Kuipers, S Cenk Sahinalp, and Niko Beerenwinkel. Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data. *Nat Commun*, 10(1):2750, Jun 2019.

[182] Salem Malikic, Farid Rashidi Mehrabadi, Simone Ciccolella, Md. Khaledur Rahman, Camir Ricketts, Ehsan Haghshenas, Daniel Seidman, Faraz Hach, Iman Hajirasouliha, and S. Cenk Sahinalp. PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research*, 29(11):1860–1877, October 2019.

[183] Yuelin Liu, Anton Goretsky, Ayse G. Keskus, Salem Malikic, Tanveer Ahmad, E. Michael Gertz, Farid Rashidi Mehrabadi, Michael Kelly, Maria Hernandez, Charlie Seibert, Juan Manuel Caravaca, Kayla Kline, Yongmei Zhao, Ying Wu, Biraj Shrestha, Bao Tran, Arindam Ghosh, Xiwen Cui, Antonella Sassano, Laksh Malik, Breeana Baker, Cornelis Blauwendraat, Kimberley J. Billingsley, Eva Perez-Guijarro, Glenn Merlino, Erin K. Molloy, S. Cenk Sahinalp, Chi-Ping Day, and Mikhail Kolmogorov. Long-read sequencing of single cell-derived melanoma subclones reveals divergent and parallel genomic and epigenomic evolutionary trajectories. *bioRxiv*, 2025.

[184] Cong Ma, Metin Balaban, Jingxian Liu, Siqi Chen, Michael J. Wilson, Christopher H. Sun, Li Ding, and Benjamin J. Raphael. Inferring allele-specific copy number aberrations and tumor phylogeography from spatially resolved transcriptomics. *Nature Methods*, 21(12):2239–2247, 2024.

[185] Andrew Erickson, Mengxiao He, Emelie Berglund, Maja Marklund, Reza Mirzazadeh, Niklas Schultz, Linda Kvastad, Alma Andersson, Ludvig Bergenstråhle, Joseph Bergenstråhle, Ludvig Larsson, Leire Alonso Galicia, Alia Shamikh, Elisa Basmaci, Teresita Díaz De Ståhl, Timothy Rajakumar, Dimitrios Doultsinos, Kim Thrane, Andrew L. Ji, Paul A. Khavari, Firaz Tarish, Anna Tanoglidi, Jonas Maaskola, Richard Colling, Tuomas Mirtti, Freddie C. Hamdy, Dan J. Woodcock, Thomas Helleday, Ian G. Mills, Alastair D. Lamb, and Joakim Lundeberg. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature*, 608(7922):360–367, 2022.

[186] Shadi Shafighi, Agnieszka Geras, Barbara Jurzysta, Alireza Sahaf Naeini, Igor Filipiuk, Alicja Raczkowska, Hosein Toosi, Łukasz Koperski, Kim Thrane, Camilla Engblom, Jeff E. Mold, Xinsong Chen, Johan Hartman, Dominika Nowis, Alessandra Carbone, Jens Lagergren, and Ewa Szczurek. Integrative spatial and genomic analysis of tumor heterogeneity with tumoroscope. *Nature Communications*, 15(1):9343, 2024.

[187] Panagiotis Karras, Ignacio Bordeu, Joanna Pozniak, Ada Nowosad, Cecilia Pazzi, Nina Van Raemdonck, Ewout Landeloos, Yannick Van Herck, Dennis Pedri, Greet Bervoets, Samira Makhzami, Jia Hui Khoo, Benjamin Pavie, Jochen Lamote, Oskar Marin-Bejar, Michael Dewaele, Han Liang, Xingju Zhang, Yichao Hua, Jasper Wouters, Robin Browaeys, Gabriele Bergers, Yvan Saeys, Francesca Bosisio, Joost van den Oord, Diether Lambrechts, Anil K. Rustgi, Oliver Bechter, Cedric Blanpain, Benjamin D. Simons, Florian Rambow, and Jean-Christophe Marine. A cellular hierarchy in melanoma uncouples growth and metastasis. *Nature*, 610(7930):190–198, 2022.

[188] Tony Hui, Qi Cao, Joanna Wegrzyn-Woltosz, Kieran O'Neill, Colin A. Hammond, David J.H.F. Knapp, Emma Laks, Michelle Moksa, Samuel Aparicio, Connie J. Eaves, Aly Karsan, and Martin Hirst. High-resolution single-cell dna methylation measurements reveal epigenetically distinct hematopoietic stem cell subpopulations. *Stem Cell Reports*, 11(2):578–592, 2018.

[189] Ryan Poplin, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, Laura D. Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G. MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, November 2017.

[190] K. Wang, M. Li, and H. Hakonarson. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164, July 2010.

[191] Mohammed El-Kebir. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, 34(17):i671–i679, 2018.

[192] Simone Ciccolella, Camir Ricketts, Mauricio Soto Gomez, Murray Patterson, Dana Silverbush, Paola Bonizzoni, Iman Hajirasouliha, and Gianluca Della Vedova. Inferring cancer progression from single-cell sequencing while allowing mutation losses. *Bioinformatics*, 37(3):326–333, 2020.

[193] Palash Sashittal, Haochen Zhang, Christine A. Iacobuzio-Donahue, and Benjamin J. Raphael. ConDoR: tumor phylogeny inference with a copy-number constrained mutation loss model. *Genome Biology*, 24:272, 2023.

[194] 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 sigprofilerassignment. *Bioinformatics*, 39(12):btad756, 12 2023.

[195] Matúš Medo, Charlotte K. Y. Ng, and Michaela Medová. A comprehensive comparison of tools for fitting mutational signatures. *Nature Communications*, 15(1):9467, 2024.

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

[197] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[198] The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 03 2023.

[199] Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology*, 16(5):284–287, 2012. PMID: 22455463.

[200] Daniel Zingg, Julien Debbache, Simon M Schaefer, Eylul Tuncer, Sandra C Frommel, Phil Cheng, Natalia Arenas-Ramirez, Jessica Haeusel, Yudong Zhang, Mario Bonalli, et al. The epigenetic modifier ezh2 controls melanoma growth and metastasis through silencing of distinct tumour suppressors. *Nature communications*, 6(1):6051, 2015.

[201] Victoria K Hill, Jung-Sik Kim, and Todd Waldman. Cohesin mutations in human cancer. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1866(1):1–11, 2016.

[202] Yuchen Guo, Yi Chen, Lianghui Zhang, Ling Ma, Keyu Jiang, Gang Yao, and Lingjun Zhu. Tert promoter mutations and telomerase in melanoma. *Journal of Oncology*, 2022(1):6300329, 2022.

[203] Thierry Soussi. The p53 tumor suppressor gene: from molecular biology to clinical investigation. *Annals of the new York Academy of Sciences*, 910(1):121–139, 2000.

[204] Oxford Nanopore Technologies Ltd. modbam2bed. `https://github.com/epi2me-labs/modbam2bed`, 2023.

[205] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, 2014.

[206] Andrew R. Gilpin. Table for conversion of kendall's tau to spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement*, 53(1):87–92, 1993.

[207] Eric Talevich, A. Hunter Shain, Thomas Botton, and Boris C. Bastian. Cnvkit: Genome-wide copy number detection and visualization from targeted dna sequencing. *PLOS Computational Biology*, 12(4):1–18, 04 2016.

[208] T Tickle, I Tirosh, C Georgescu, M Brown, and B Hass. infercnv of the trinity ctat project.

[209] Nicole R Infarinato, Katherine S Stewart, Yihao Yang, Nicholas C Gomez, H Amalia Pasolli, Lynette Hidalgo, Lisa Polak, Thomas S Carroll, and Elaine Fuchs. Bmp signaling: at the gate between activated melanocyte stem cells and differentiation. *Genes Dev*, 34(23-24):1713–1734, Dec 2020.