Linking ovarian cancer genomic instability and phenotypic dynamics at single cell resolution

by

Hongyu Shi

A Dissertation

Presented to the Faculty of the Louis V. Gerstner, Jr.

Graduate School of Biomedical Sciences,

Memorial Sloan Kettering Cancer Center

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

New York, NY

Dec, 2023

_____	_____
Sohrab Shah, PhD Dissertation Mentor	Date
_____	_____
Nikolaus Schultz, PhD Dissertation Mentor	Date

Abstract

Somatic copy number alterations (CNAs) play an important role in driving aberrant gene expression in cancer cells. Tumors with a high level of chromosomal instability tend to have prevalent subclonal CNAs and heterogeneous cancer cell populations. However, the direct and indirect mechanisms that subclonal CNAs contribute to clone-specific gene expression remain poorly understood. Dosage effect is one of the direct mechanisms, which describes the positive correlation between gene copy number and expression. With the emergence of single cell datasets profiling genetic, epigenetic and transcriptomic aspects of cancer cells, it is possible to better characterize genotype-phenotype interplay in cancer cells. However, new computational methods are needed to integrate multi-modality single cell datasets and model the influence of subclonal CNAs on gene expression.

We developed TreeAlign, which computationally integrates independently sampled single-cell DNA and RNA sequencing data from the same cell population by assigning transcriptional profiles to genomic subclones. Through explicitly modeling of gene dosage effects from subclonal CNAs **(Chapter 2)** and incorporation of allele-specific information **(Chapter 3)**, TreeAlign achieved improved clone assignment accuracy compared to existing methods. By fitting the model recursively on scDNA-based phylogeny, TreeAlign also helps refine subclone definition based on transcriptional divergence. Using TreeAlign,

we investigated clone-specific transcriptional programs in ovarian cancer and explored the role of copy number dosage effects in driving subclonal phenotypes **(Chapter 4)**. Our approach sets the stage for dissecting the relative contribution of fixed genomic alterations and dynamic epigenetic processes on gene expression programs in cancer.

Acknowledgements

Completing this PhD dissertation has been a great journey filled with challenges and joy, and I owe my deepest gratitude to those who have supported me along the way.

First and foremost, I would like to express my sincere appreciation to my advisors, Dr. Sohrab Shah and Dr. Nikolaus Schultz, for their unwavering guidance, expertise, and endless patience. Your mentorship has been instrumental in shaping my research and academic growth. I would also like to thank my advisory committee members Dr. Samuel Bakhoum and Dr. Christina Leslie for their valuable feedback and advice.

I am also immensely grateful to all my labmates and colleagues at Shah lab and Schultz lab. Your insights, constructive conversations, and shared enthusiasm made every day in the lab a meaningful one. Special thanks to Dr. Marc Williams, Dr. Gryte Satas, Adam Weiner, Dr. Andrew McPherson for their contributions on the TreeAlign model. Special thanks to our collaborators Dr. Junyue Cao and Aileen Ugurbil for their work on scATAC-seq in the MSK-SPECTRUM project.

I would also like to thank GSK graduate school and all the friends I made here. This fantastic program and the warm community at GSK made the past 5 years unforgettable.

Last, I extend my heartfelt appreciation to my family members. Your unconditional love and patience have been my constant motivation.

Table of Contents

List of Figures

Chapter 1

Background

1.1 Clonal evolution and tumor heterogeneity

Cancer arises from clonal evolution, which is the iterative process of genetic alteration accumulation, clonal expansion and selection [1]. Extensive research efforts have been directed towards characterizing the evolutionary trajectories of cancer cells in both patients and model systems, elucidating the emergence and disappearance of subclones carrying specific genetic alterations [2–5]. These studies provided valuable insights on clonal evolution in disease progression. However, as selective forces act upon phenotypic traits rather than genotypic variations, in addition to profiling genomic changes, it is also important to profile the phenotypes of cancer cells to gain a comprehensive understanding of cancer evolution.

Different genomic instability mechanisms shape the diverse landscape of cancer genomes[6, 7]. For example, exogenous mutagens and impaired missense repair mechanisms can lead to higher prevalence of point mutations [8, 9] which can be characterized by different

mutational signatures [10]. Whereas chromosomal instability gives rise to frequent amplifications and deletions [11]. These genetic distinctions can lead to various cancer cell phenotypes and subsequent divergent selection pressures [12]. For example, chromosome unstable tumors with frequent CNAs tend to exhibit higher rates of whole genome doubling (WGD) [13] and compromised immune responses compared to chromosome stable tumors [14]. The interplay between somatic alterations and the tumor microenvironment generates a spectrum of cancer cell phenotypes, enhancing the heterogeneous nature of tumors and contributing to the complex landscape of cancer biology [15, 16].

Chromosomal instability can leads to frequent CNAs, which are known to contribute to transcriptomic diversity in cancer cells [17]. It is well established that CNAs of driver oncogenes and tumor suppressors are causal determinants that change the fitness of cancer cells [18], leading to clonal expansions, clone-clone variation [3] and tumor evolution. In addition to impacting specific genes, CNAs often span chromosome arms or whole chromosomes and therefore potentiate transcriptional impact across hundreds of genes with a single genomic event through copy number (CN) dosage effects. CN dosage effects are defined as the positive correlation between CN (or gene dosage) and the corresponding gene expression [19]. It was observed that patient-to-patient transcriptomic differences in ovarian cancer cells was predominantly influenced by CN dosage effects [20]. Recent reports on the extent of cell-to-cell variation of CNAs in tumors (including in well understood oncogenes) [21] also raise the critical question of how granular subpopulations are phenotypically impacted by subclonal CNAs. Importantly, phenotypic impact of subclonal CNAs can have both cell intrinsic effects and act as cell-extrinsic determinants of the tumor microenvironment [20], further illustrating the importance of dissecting how CNAs modulate phenotypic intra-tumor heterogeneity. In this dissertation, we focus on the impact of gene

dosage effects from CNAs as a mechanism phenotypic diversification.

1.2 Genetic and transcriptomic ITH

Intra-tumor heterogeneity (ITH) is an important feature of cancer and can be observed across various levels, from genetic alterations to epigenetic states and transcriptional profiles [6]. Findings from previous studies demonstrate that ITH is associated with clinical outcomes [22, 23] and therapeutic responses [24, 25]. The emergence of treatment resistance can stem from the expansion of pre-existing subclonal populations [26–29] or development of drug-resistant cell states [30, 31], underscoring the adaptive nature of cancer cells. In this context, unraveling ITH across multiple levels and dissecting the clonal evolution processes that generate such diversity become critical in order to better understand the disease as a whole.

Extensive studies have investigated genetic ITH. In high-grade serous ovarian cancers (HGSCs), primary tumors were found to be clonally diverse and metastatic tumors were formed by monoclonal or polyclonal seeding [32]. With the TCGA cohort, Andor et al. revealed that across 12 cancer types, approximately 86% of tumors exhibited a minimum of two distinct clones [33]. Notably, tumors characterized by increased genetic ITH and coexistence of multiple clones demonstrated more aggressive histological features and elevated risk of mortality.

Genetic ITH poses challenges on the design of target therapy, as tumors with higher genetic ITH are more likely to harbor subclones with pre-existing resistance to treatment [34]. It was also observed that even within the same tumor, multiple mechanisms of resistance can

be acquired during targeted therapy [2]. On the other hand, genetic ITH, along with the mutational processes that generate it, also provide opportunities for new therapeutic strategies. Colorectal cancers with high microsatellite instability (MSI-H) have escalated mutation rates within microsatellite regions due to defects in DNA mismatch repair pathways [35]. MSI-H patients show suboptimal responses to chemotherapy [36], but conversely tend to be more sensitive to PD-1 blockade [37]. Tumors marked by pronounced chromosomal instability not only display an increased propensity for metastasis but also concurrently exhibit elevated inflammatory signals through the sGAS-STRING pathway [38], thereby potentially facilitating the utilization of immunotherapies.

Transcriptomic ITH denotes the variability in gene expression among cancer cells within a tumor. Transcriptomic ITH can originate from genetic ITH. Subclones in the same tumor have distinct genetic alterations which may lead to dysregulation of different gene expression programs in cancer cells. It was shown that transcriptomic ITH correlates with subclonal copy number alterations (CNAs) in lung cancers [39]. In breast cancer cell lines, genetic changes were found to be associated with differential activation of transcriptional programs, influencing cell morphologies and proliferation [40]. Aside from genetic influences, other factors also contribute to transcriptomic ITH. Research on monoclonal tumor xenografts of colorectal cancers has highlighted the substantial contribution of in vivo multilineage differentiation to transcriptional diversity [41]. Another investigation in lung squamous cell carcinoma has shown transcriptomic ITH which impacts cancer-associated pathways and proliferative capacities only had a weak correlation with genetic ITH [42]. To better understand transcriptomic ITH, it is critical to dissect its origin in different cancer types.

Understanding the spectrum of ITH and the interplay between its various layers is of

paramount importance. The progression of single-cell sequencing technologies [43], particularly those enabling simultaneous measurement of multiple modalities within individual cells, potentially allow us to have a more comprehensive view of ITH.

1.3 Approaches to study ITH in the context of clonal evolution

Different methods have been used to investigate the genomic and phenotypic dynamics within the context of clonal evolution [43, 44]. The common strategy involves: 1. Profiling the clonal architecture of cancer cells. 2. Assessing the phenotypes of corresponding cancer cell populations.

Lineage tracing refers to the group of approaches that monitor the progeny of individual cells and decipher their lineage relationships. This could be achieved with naturally occurring genetic alterations or artificial tags introduced into cells. The latter approach is only applicable in model systems. For instance, Neftel et al. used lentiviral vectors to introduce distinct and heritable genetic markers into cells [45]. These markers were subsequently deciphered through single-cell RNA sequencing (scRNA-seq). Through this method, they tagged cancer cells from glioblastoma patients and xenografted them into mouse brains. By analyzing the tumors formed, they found that cells possessing identical barcodes displayed distinct transcriptional cell states, thus illuminating the plasticity of states that persists independently of genetic background. In another investigation by Quinn et al., cancer cells were genetically engineered with barcodes that could be dynamically edited by Cas9, facilitating lineage tracing [13]. Subsequent mutations introduced into the barcodes were

captured through scRNA-seq, thereby enabling the reconstruction of the phylogeny of cancer cells. Utilizing this technology, the researchers quantified metastatic dissemination rates for subclonal populations and elucidated associated metastatic phenotypes.

Phylogenetic relationships can also be deduced from naturally occurring genetic modifications. This approach is frequently employed to characterize clonal structures in samples derived from cancer patients. Utilizing paired RNA-seq and whole-exome sequencing data from the TRACERx project, Martínez-Ruiz et al. and Frankell et al. investigated the transcriptional profiles and evolutionary patterns of primary and metastatic lung cancers respectively [4, 39]. The integration of these matched datasets facilitated the identification of genomic attributes, such as the proportion of the genome affected by subclonal somatic CNAs, which were linked to transcriptomic ITH. Nevertheless, due to limitations inherent in bulk tumor sampling, establishing connections between genotypic modifications and phenotypic traits at the subclonal level remains exceedingly challenging. Single-cell sequencing technologies have emerged as pivotal tools that allow the resolution of genomic and phenotypic profiles at the granularity of individual cells. Funnell et al., for instance, applied single-cell DNA- and RNA-sequencing to analyze primary triple-negative breast cancer (TNBC) and HGSC cells [21]. Their study revealed clone-specific amplifications in oncogenes, accompanied by concomitant up-regulation in gene expression. Paired single-cell datasets are valuable resources for dissecting the interactions between genotypic variations and phenotypic attributes at subclonal resolution.

1.4 Challenges in linking genotypes and phenotypes

Establishing a robust connection between genetic compositions and cancer cell phenotypes is challenging. Conventional bulk sequencing methods have been widely employed to delineate somatic alterations and concomitant phenotypic modifications [46–49]. Although paired datasets with both DNA and RNA sequencing make it possible to correlate these two aspects, the resolution of these studies are still limited at the level of patient-derived tumor samples, and unable to provide a comprehensive view of ITH at the subclonal or single cell level.

The accurate inference of phylogenetic relationships from whole exome sequencing (WES) or whole genome sequencing (WGS) data is circumscribed by the identification of shared and distinctive somatic alterations detected within samples [50, 51]. For bulk RNA sequencing, it is complicated by the coexistence of normal cells in the tumor microenvironment. Many computational methods have been proposed to deconvolute cell type specific expression from bulk RNA-seq data [52–55]. However, the performance of these methods tend to be highly influenced by the choice of data preprocessing approaches and selection of cell type markers [56]. With bulk RNA-seq alone, it is hard to distinguish cancer cell intrinsic and extrinsic transcriptional signals.

Conducting single-cell RNA **(Fig. 1.1)** and DNA **(Fig. 1.2, 1.3)** sequencing separately offers the capacity to profile a large number of individual cells and thereby provides a more comprehensive depiction of cell populations in tumors. In recent years, an increasing number of studies have appeared, generating multimodal datasets with single-cell DNA and single-cell RNA profiles [21, 57–59]. Noteworthy among these is the work of Andor et al. [57], who profiled the genomes of 8824 cells and the gene expression patterns of more

than 28,000 cells from diverse gastric cancer cell lines, using 10x scDNA- and scRNA-seq. Furthermore, Parra et al. [59] conducted a study centered on chromothriptic medulloblastoma, generating shallow scDNA- and scRNA-seq profiles encompassing 757 and 22,500 cells across 7 samples, respectively. The measurements of both genetic alterations and gene expression at single cell level allow us to further dissect different aspects of ITH. However, a comprehensive understanding of the intricate interplay between genomic and phenotypic changes requires the development of computational frameworks for integrating these diverse data modalities.

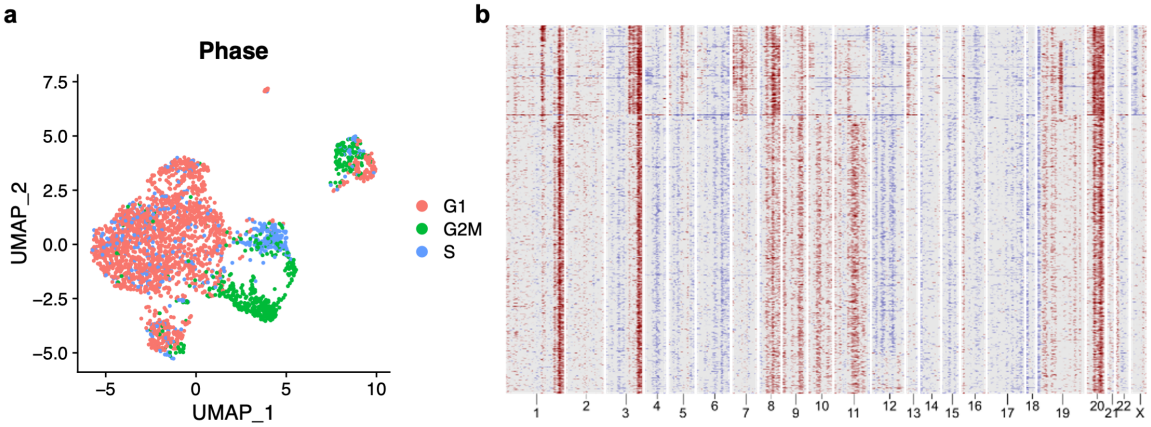

FIGURE 1.1: An example of scRNA data. **a,** UMAP embedding of cancer cells based on scRNA data from patient 045 colored by predicted cell cycle phase. **b,** InferCNV-corrected expression profiles for patient 045 [60]. InferCNV is a method for inferring CNAs from scRNA data.

1.5 High-grade serous ovarian cancers

High-grade serous ovarian cancer (HGSC) is the most lethal gynecological malignancy [61] and the archetype of cancer to study chromosomal instability. HGSCs are distinguished by frequent copy number alterations [17, 62–64] and extensive spread throughout

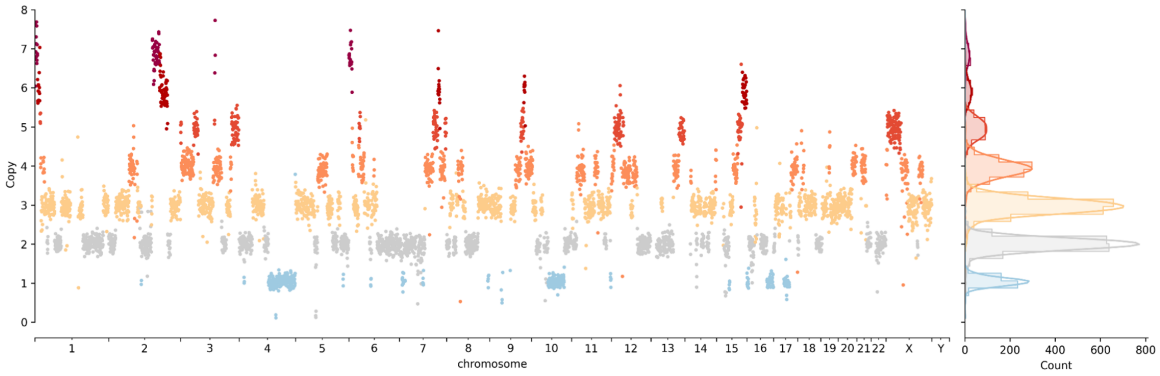

FIGURE 1.2: An example of a single cell copy number profile from scDNA data in patient 118.

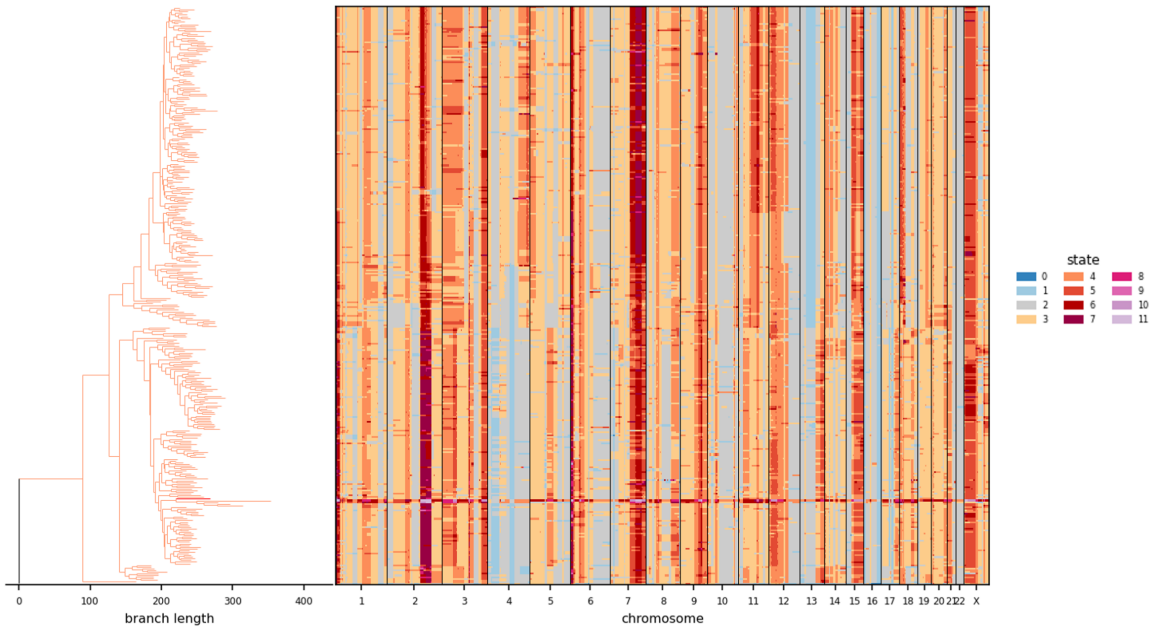

FIGURE 1.3: An example of a single cell phylogeny and CN heatmap built using scDNA data from patient 118.

the peritoneal cavity [32, 65, 66]. While recent advancements have introduced poly ADP-ribose polymerase (PARP) inhibitors targeting the prevalent homologous recombination deficiency (HRD) observed in HGSCs [67, 68], the disease remains largely untreatable in numerous instances, with a median survival period of 40.7 months [69].

Alterations in TP53 are almost universally present in HGSCs [70]. Inactivation of the homologous recombination repair pathway is also common and present in around 50% of cases through genetic alterations or epigenetic silencing of BRCA as well as other genes in this pathway [70, 71]. Point mutations in genes other than TP53 and BRCA1/2 are infrequent. Comparatively, copy number alterations are significantly more prevalent, impacting oncogenes such as CCNE1 and MYC, as well as tumor suppressors including RB1 and NF1 **(Fig. 1.4)**. In addition to specific gene alterations, HGSCs can also be stratified by varying mutational processes [72], including the homologous repair deficiency (HRD) subtypes and foldback inversion (FBI)-bearing subtype. Genomic instability, driven by these mutational processes, leads to an elevated degree of genetic ITH and shapes clonal evolution in HGSC.

1.6 Outline of this thesis

Our central hypothesis is that through explicitly modeling of (allelic) CN dosage effects, we can better integrate genomic and transcriptomic data and link cancer cell genotypes and phenotypes. Motivated by this hypothesis, the first aim of this thesis is to develop a new computational approach to integrate scDNA and scRNA datasets and infer CN dosage effects. In **Chapter 2**, we propose TreeAlign, a Bayesian probabilistic model to assign single cell expression profiles to a scDNA-based single cell phylogeny while inferring CN dosage effects. Our second aim is to further improve data integration by considering allele-specific

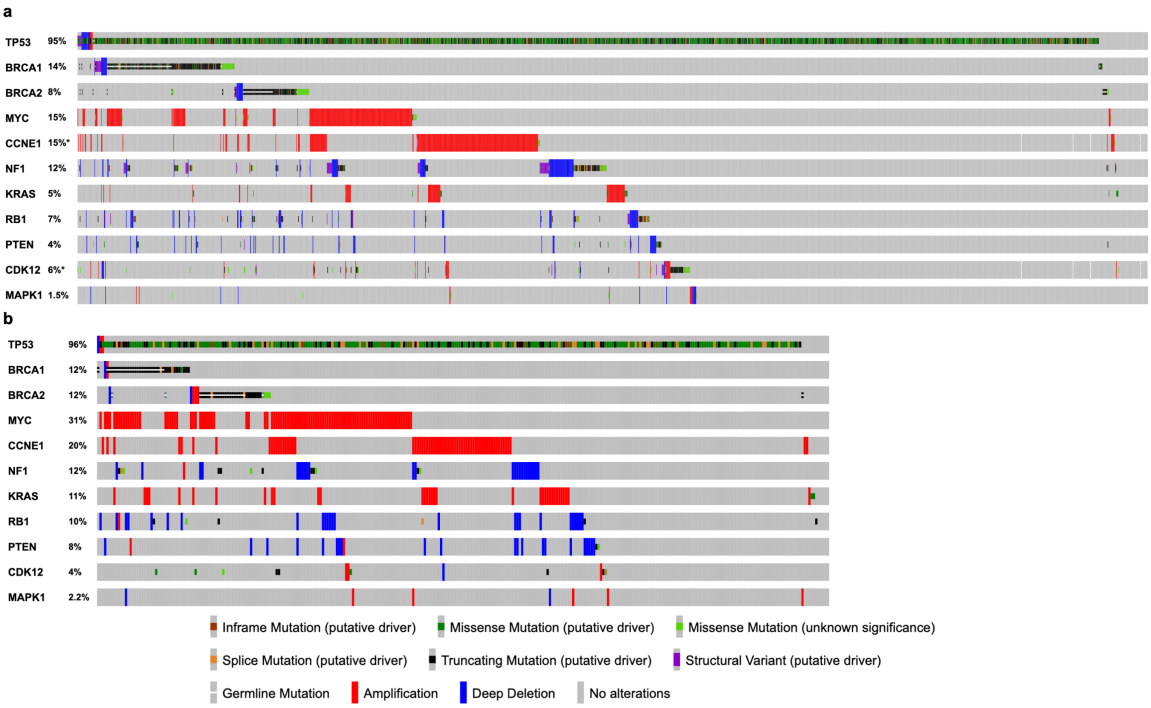

FIGURE 1.4: **a,** Genomic landscape of HGSCs in IMPACT. **b,** Genomic landscape of HGSCs in TCGA.

CN and gene expression. In **Chapter 3**, I will discuss extensions to TreeAlign which utilize allele-specific information to further improve the performance and allow us to characterize subclonal phenotypes of patient derived xenografts and cell lines derived from breast cancer and ovarian cancer patients. Finally, with the computational methods proposed, I will discuss clone-specific transcriptional heterogeneity in HGSCs and the underlying genetic and non-genetic origins of the heterogeneity **(Chapter 4)**. Interferon-related pathways were found to be frequently differently expressed between subclones highlighting their importance in driving subclonal phenotypic divergence. With TreeAlign, we were able to investigate clone-specific transcriptional phenotypes in the context of metastasis and whole genome duplication. Additionally, I will highlight the potential extension of the TreeAlign method for integrating additional data modalities such as scATAC. In this thesis, I mainly focus on single cell datasets from ovarian cancer. With the emergence of more multi-modal single cell datasets, I would expect that applications of the approaches described in this thesis can be expanded to a more diverse set of cancer types and experimental conditions.

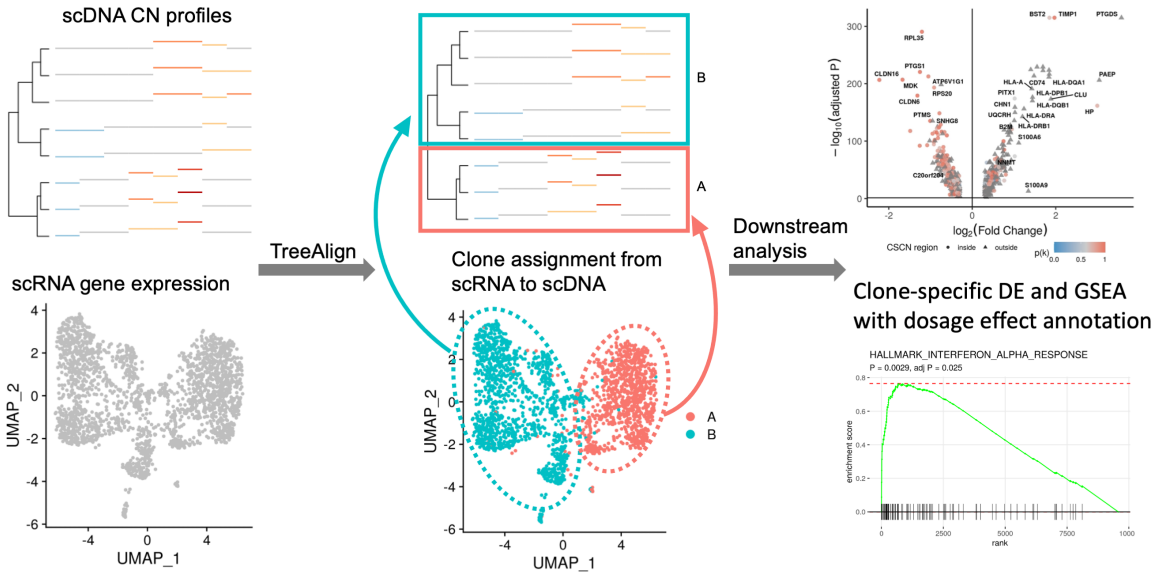

FIGURE 1.5: Graphical abstract of this thesis. We developed TreeAlign for scDNA and scRNA integration which allows for downstream clone-specific differential expression and gene set enrichment analysis with dosage effect annotations.

Chapter 2

TreeAlign for clone assignment and dosage effect inference

2.1 Introduction

Previous studies using bulk sequencing techniques have investigated the association between clonal CNAs and gene expression [46–49]. The expression level of a gene can be influenced by CN dosage effects reflected by the significant positive correlation between gene expression and the underlying copy number [19]. However, gene dosage effects are not deterministic and may be subject to compensatory mechanisms, rendering the impact of CNAs on expression as highly variable across the genome. Transcriptional adaptive mechanisms [73] including epigenetic modifications and downstream transcriptional regulation, can modulate CN dosage effects [74–76], further obscuring the direct impact of gene dosage. For example, the expression of certain immune response pathways often exhibit both CNA-dependent and CNA-independent expression [49].

Theoretically, measuring single cell RNA and DNA data should elucidate how genotypes

translate to phenotypes at single cell resolution. Technologies that sequence both RNA and DNA modalities co-registered in the same cell would be ideal for linking genomic alterations to transcriptional changes in tumor evolution. However, pioneering technologies [77, 78] have had limited throughput, lower quality and are still not mature enough for large-scale profiling of cancer cells. Sequencing single cell RNA or DNA independently allows more cells to be profiled and reveals a more comprehensive view of the cell populations, but requires computational integration of the two data modalities.

Several computational methods have been proposed for joint analysis of single cell DNA and RNA data. CloneAlign [79] is a probabilistic framework to assign transcriptional profiles to genomic subclones based on the assumption that the expression level of a gene is proportional to its underlying copy number. More recent methods SCATrEx [80] and CC-NMF [81] are also based on this assumption but use different methods to model the similarity between copy number profiles and gene expression patterns. However, these methods do not consider the possibility that transcriptional effects of copy number could be variable between genes and therefore lack the specificity to decipher genes that may be subject to dosage effects from those that are independent of CNAs. In addition, these methods require using predefined subclones from scDNA data or specify the number of subclones as input which may propagate errors of uninformative subclones or may miss more granular gene dosage effects. More importantly, the revelation of phenotypic plasticity as a driver of genetically independent transcription in cancer cells [82–84] motivates the need to disentangle genetic from epigenetic cell-to-cell mechanisms. No available methods directly model dosage effects of subclonal CNAs, which is critical to infer which genes are deterministically modulated by subclonal CNAs and which genes are independent of CNAs. Moreover, recent advances have illuminated the extent to which allele-specific CNAs can

mark clonal haplotypes both in DNA-based [21] and RNA-based [85] single cell analysis, illustrating both a methodological gap and analytical opportunity for integration.

In this study, we address the questions of how subclonal CNAs drive phenotypic divergence and evolution in cancer cells, and quantitatively encode dosage effects in this process. We present a new method, TreeAlign, to enumerate and define CNA-driven clone-specific phenotypes, and also a statistical framework to compare the transcriptional readouts of genomically defined clones. TreeAlign implements a Bayesian probabilistic model that maps gene expression profiles from scRNA to genomic subclones from scDNA which i) can refine subclone definition from single cell phylogenies through a recursive process suggested by transcriptional divergence, ii) explicitly models dosage effects of each gene. Through extensive simulation, we demonstrate that the TreeAlign outperforms alternative approaches in both terms of clone assignment and gene dosage effect prediction.

2.2 The total CN model for clone assignment and dosage effect inference

We developed TreeAlign, a probabilistic graphical model which maps scRNA sequenced cells to scDNA-derived subclones. TreeAlign employs a recursive algorithm for delineating subclones from phylogenies constructed using scDNA data, with guidance from gene expression information. The model jointly infers clone assignments and clone-specific CN dosage effects. The TreeAlign framework assumes a subset of genes with positively correlated expressions to their underlying copy numbers. For each gene, expression is modeled by k, where $k \in \{0,1\}$ is a Bernoulli variable such that the probability $p(k=1)$ represents the probability the gene has clone-specific CN dosage effects **(Fig. 2.1)**. This encoding

results in genes without dosage effects (low $p(k)$) to have little or no contribution to the clone assigning process.

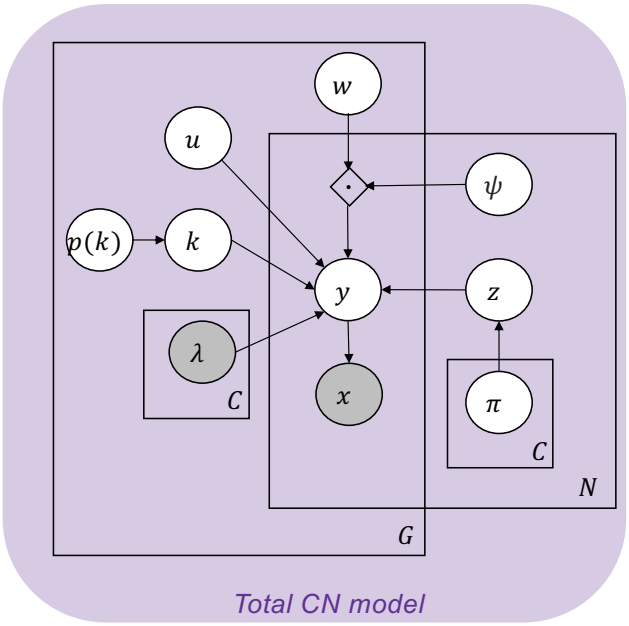

FIGURE 2.1: Graphical model of total CN TreeAlign.

To infer clone assignments and $p(k)$, TreeAlign requires three inputs: 1) a cell $\times$ gene matrix of raw read counts from scRNA-seq, 2) a cell $\times$ gene copy number matrix estimated from scDNA data and 3) A phylogenetic tree (or optionally, predetermined clone labels) from scDNA profiles. TreeAlign can either assign expression profiles to predefined clone labels, similar to CloneAlign [79] or can operate on a phylogenetic tree directly to assign cells to clades of the phylogeny **(Fig. 2.2)**. When using a phylogenetic tree, a Bayesian hierarchical model is recursively applied starting from the root of the tree, computing the probability that expression profiles in scRNA can be mapped to a subtree. The stopping condition of the recursion is satisfied when the genomic or phenotypic differences between

two subtrees become too small to allow confident assignment of expression profiles. See Methods for the complete explanation on model setup.

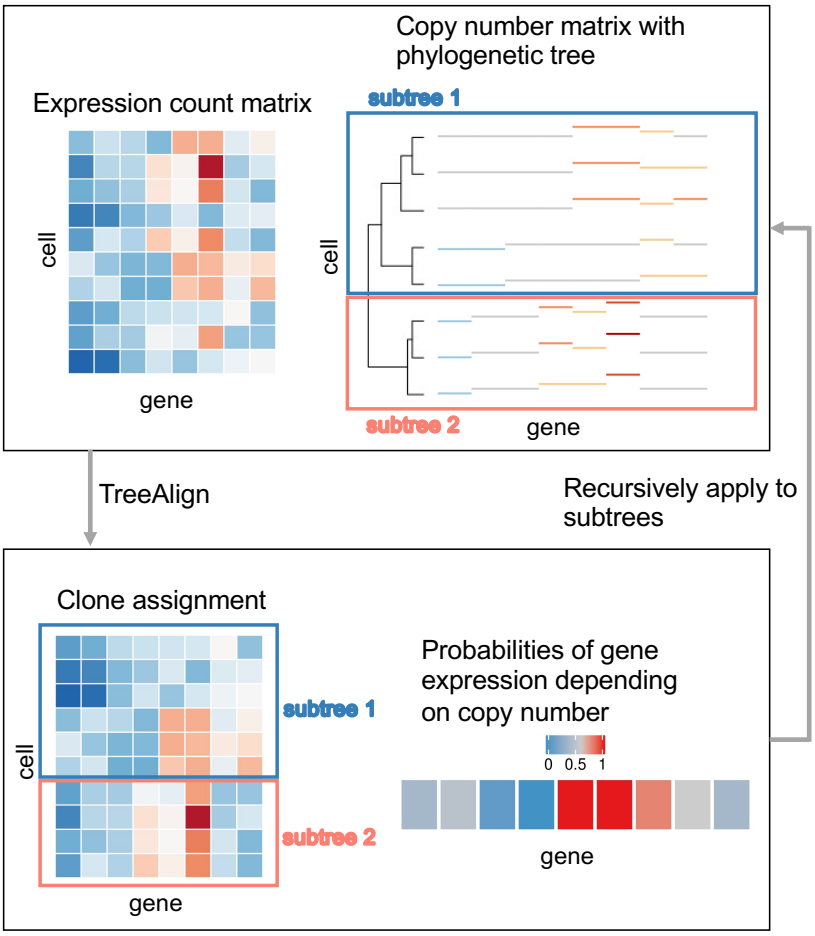

FIGURE 2.2: TreeAlign takes raw count data from scRNA-seq, the copy number matrix and the phylogenetic tree from scDNA-seq. By recursively assigning the expression profiles to phylogenetic subtrees, TreeAlign infers the clone-of-origin of cells identified in scRNA-seq and the dosage effects of subclonal CNAs.

2.3 Performance on simulated data

We first evaluated TreeAlign on synthetic datasets, quantifying the effect of three main parameters in the input data: number of cells (100 - 5000), number of genes (100 - 1000)

and proportions of genes with dosage effects (10%-100%). Simulations were performed using the generative model of CloneAlign [79]. We compared the performance of assigning expression profiles to ground truth predefined clones between TreeAlign, CloneAlign and InferCNV [60]. InferCNV was originally developed for inferring CNAs from gene expression data, but has also been repurposed for clone assignment in some studies [59]. InferCNV analysis in this context acts as a way of inferring clone assignment without the benefit of the scDNA data. Compared to CloneAlign and InferCNV, TreeAlign performed significantly better in terms of clone assignment accuracy especially in the regime where fewer genes exhibit dosage effects **(Fig. 2.3, B.2)**. For example, in the regime of 60% of genes with dosage effects (1000 cells, 500 genes), TreeAlign achieved mean clone assignment accuracy of 91.1%, compared to CloneAlign with 75.1% accuracy. The improvement in clone assignment accuracy was consistent across all cell and gene number simulation scenarios. We also tested performance with phylogenetic tree inputs to evaluate if TreeAlign could achieve similar results on tree input compared to pre-defined clone input. Similar to the "clone" regime, these simulations varied the proportion of genes with gene dosage effects in 10% increments. TreeAlign was able to assign expression profiles back to the corresponding clades of the phylogeny with similar accuracies compared to the clone input in regimes with $> 40\%$ genes with dosage effects **(Fig. 2.3, B.3)**. Together these evaluations reflect that the model effectively obviates a priori tree cutting without paying a penalty in accurate clone mapping.

We also evaluated the accuracy of predicting dosage effects for each gene in the input datasets. We compared the simulated and predicted (using $p(k)$ as an estimate) frequency of genes with CN dosage effects. For high expression genes, simulated and predicted frequencies were highly concordant **(Fig. 2.4)**. For datasets with $\geq 50\%$ of genes with dosage

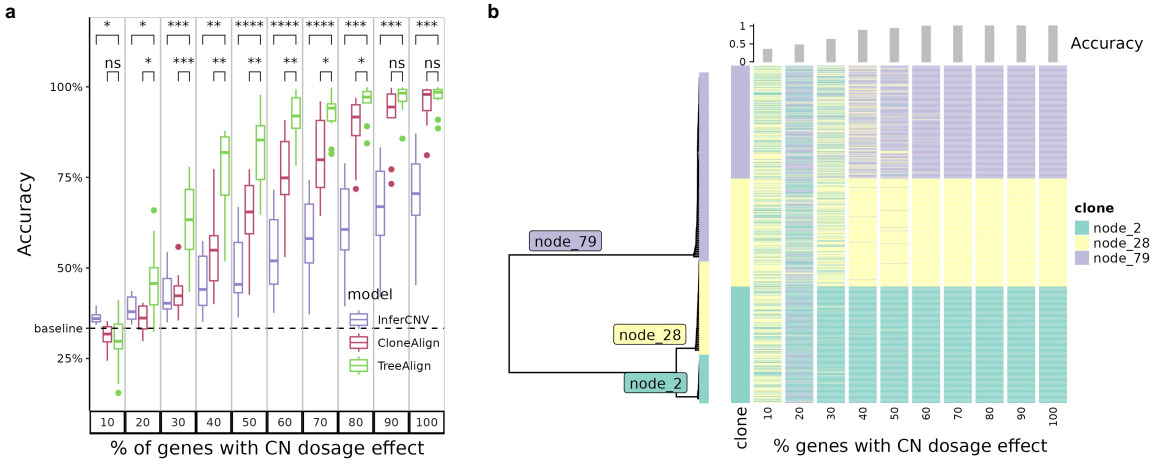

FIGURE 2.3: **a,** Clone assignment accuracy of TreeAlign, CloneAlign and InferCNV on simulated datasets (500 cells, 1000 genes, 3 clones) containing varying proportions of genes with CN dosage effects. $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. Brackets: Wilcoxon signed-rank test. **b,** Phylogenetic tree (left) of cells from patient 081 constructed using scDNA data. Heat map (right) of clone assignment by TreeAlign. Each column shows the assignment of simulated expression profiles to subtrees of the phylogeny. The bar chart above shows the overall accuracy of clone assignment.

effects, the mean area under the receiver-operator curve (AUC) was ≥ 0.99 for genes with relatively high expression level (genes in top 40% in terms of normalized expression levels) **(Fig. B.4)**. We compared $p(k)$ to a baseline estimation of CN dosage effects which is the per-gene Pearson correlation coefficient (R) of CN and expression after fitting CloneAlign. $p(k)$ from TreeAlign had an overall higher AUC compared to R from CloneAlign for predicting CN dosage effects. This establishes that $p(k)$ captures gene dosage effects and has the ability to distinguish genes with dosage effects from those without dosage effects.

2.4 Validation on real patient data

We next investigated TreeAlign's performance on real-world patient derived data. We first applied TreeAlign on single cell sequencing data from a HGSC patient (patient 022) [20].

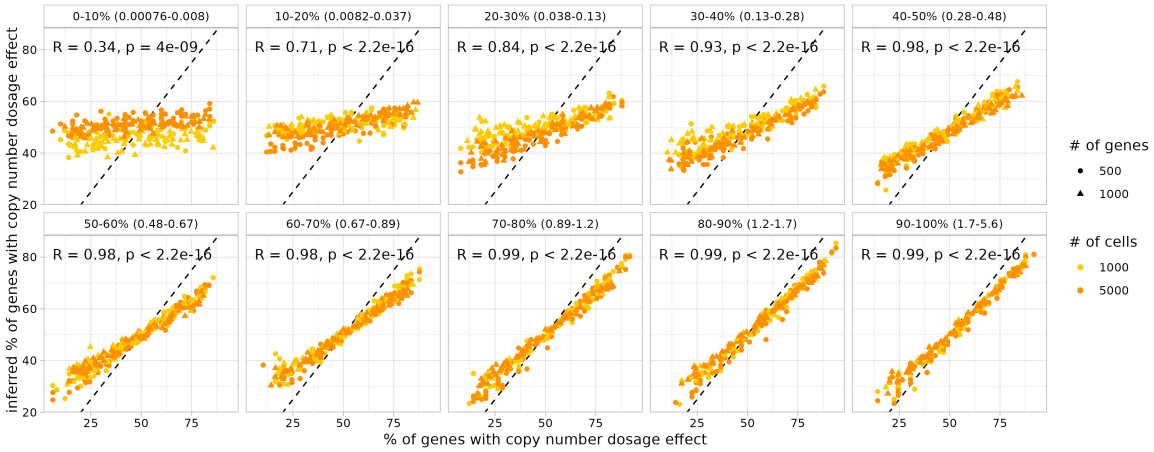

FIGURE 2.4: Scatter plots comparing inferred gene dosage effect frequencies and the simulated frequencies. Each panel groups genes with similar expression levels from low expression genes (0-10%, with normalized expression between 0.00076-0.008) to high expression genes (90-100%, with normalized expression between 1.7-5.6). Pearson correlation coefficients (R) and P values for the linear fit are shown.

Tumor samples were obtained from both left and right adnexa sites of the patient. scDNA ($n = 1050$ cells) and scRNA ($n = 4134$ cells) data were generated through Direct Library Preparation (DLP+) [86] and 10x genomics single-cell RNA-seq [87] respectively. 3579 (86.6%) ovarian cancer cells profiled by scRNA were assigned to 4 subclones identified by scDNA-seq **(Fig. 2.6)**. The expression profiles of clone C and D are overlapped on the UMAP embedding, while separated from the profiles of clone A and clone B, which coincides with the shorter phylogenetic distance between clone C and D **(Fig. 2.5)**. The separation of cells by assigned clones on the expression-based UMAP also suggests that the genetic subclones possess distinct transcriptional phenotypes.

We confirmed the clone assignment accuracy of TreeAlign by comparing the clonal frequencies estimated by RNA and DNA data **(Fig. 2.7)**. As both scRNA and scDNA data were generated by sampling from the same populations of cells, the clonal frequency estimated by the two methods should be consistent. Clonal frequencies in the left and

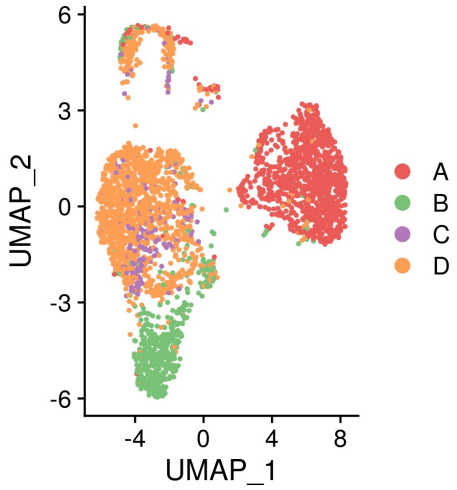

FIGURE 2.5: UMAP plot of scRNA-data from patient 022 colored by clone labels assigned by TreeAlign.

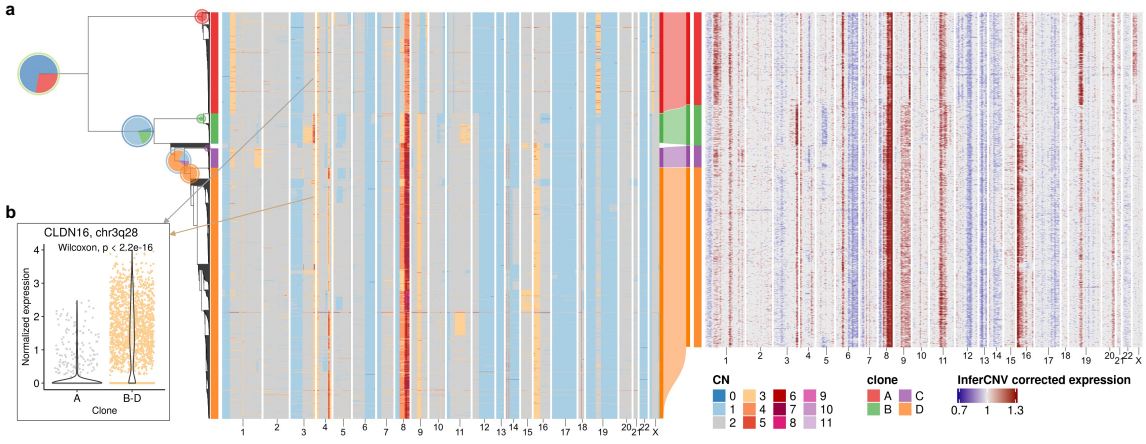

FIGURE 2.6: **a,** Single cell phylogenetic tree of patient 022 constructed with scDNA-data (left). Pie charts on the tree showing how TreeAlign assigns cell expression profiles to subtrees recursively. The pie charts are colored by the proportions of cell expression profiles assigned to downstream subtrees. The outer ring color of the pie charts denotes the current subtree. For example, the leftmost pie chart represents the proportions of cells assigned to the two main subtrees. The outer ring represents the root of the phylogeny. Red represents the subtree on the top or clone A. Blue represents the bottom subtree which contains clone B, C and D. Left heat map, total copy number from scDNA; right heat map, InferCNV corrected expression from scRNA; middle Sankey chart, clone assignments from RNA to DNA. **b,** Normalized expression of CLDN16 in clone A and clone B-D.

right adnexa sample from the two modalities were significantly correlated ($R = 0.99$, $P = 9 \times 10^{-7}$).

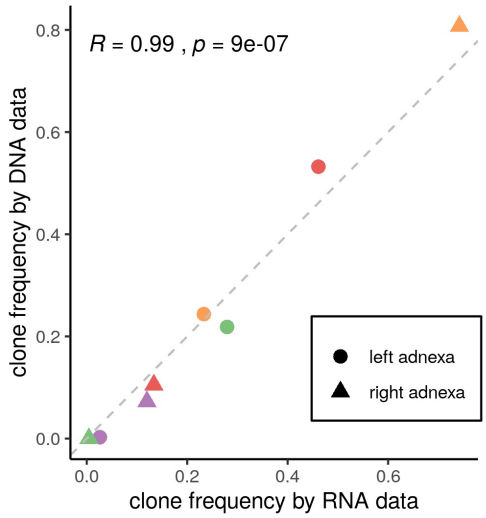

FIGURE 2.7: Correlation between clone frequencies of patient 022 estimated by scRNA-data (x axis) and scDNA-data (y axis).

In addition, CNAs inferred for scRNA cells using InferCNV [60] were concordant with the scDNA-based CNA of the clones to which scRNA cells were assigned **(Fig. 2.6)**. For example, notable clone specific copy number changes can be seen in both scDNA and scRNA on chromosome X in clone A. Clone B specific amplification on 3q, Clone C and Clone D specific amplification on 16p can also be observed in both scDNA and scRNA. By comparing the RNA-derived copy number profiles with scDNA data, we noticed that inferring copy number from RNA data is not always accurate. For example, the inferred profiles missed the focal amplification on chromosome 18.

We also held out genes from chromosome 9 and chromosome 12 and re-ran TreeAlign with the remaining genes. 98.8% cells were assigned consistently as compared to results using the full dataset. Clone level gene expression on chromosome 9 and 12 was consistent

with the corresponding copy numbers **(Fig. 2.8)**. These results demonstrated a proof of principle that TreeAlign can properly integrate scRNA and scDNA datasets and highlighted that scDNA-seq can provide valuable information on CNAs and tumor subclonal structures which would be difficult to detect with expression data only.

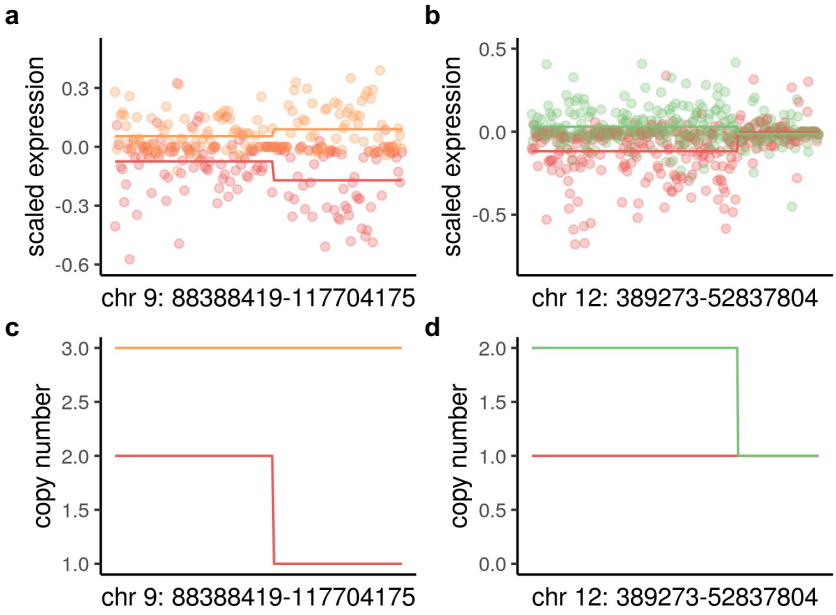

FIGURE 2.8: **a,** Scaled expression for regions on chromosome 9. **b,** Scaled expression for regions on chromosome 12. **c,** copy number profiles for regions on chromosome 9. **d,** copy number profiles for regions on chromosome 12 as a function of genes ordered by genomic location.

We then applied TreeAlign to data from an HGSC patient-derived xenograft (PDX) SA610X3XB03802 [21]. Compared to patient 022, clone-specific CN changes are less obvious in this sample **(Fig. 2.9)**. The two major subclones have CN difference on chromosome 2, 3, 4, 19 and X. TreeAlign was able to assign expression profiles to the two major clones. After reviewing the InferCNV output, we noticed that InferCNV was able to capture some of the CNAs such as the ones on chromosome 2, 19 and X but failed to recover the changes on chromosome 3 and 4.

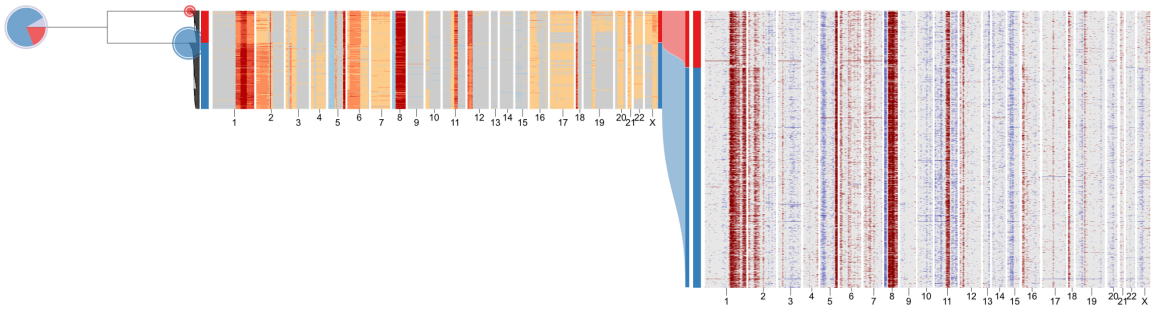

FIGURE 2.9: Total CN model of TreeAlign assigned expression profiles of SA610X3XB03802 to 2 subclones.

Finally, we applied TreeAlign to previously published data from a gastric cell line NCI-N87 generated by 10x genomics single-cell CNV and 10x scRNA assays [57]. TreeAlign assigned 3212 cells from scRNA to three clones identified in scDNA. The clonal frequencies estimated by both assays were closely aligned **(Fig. B.5)**. As for the patient 022 data, the scRNA cells showed subclonal copy number similar to the scDNA clones to which they were assigned, thus illustrating that TreeAlign also performs well with 10x scDNA data.

Chapter 3

Allele-specific TreeAlign and clone-specific CN dosage effects

3.1 Introduction

In the previous chapter, the total CN TreeAlign model was constructed by incorporating clone-specific CN dosage effects. In addition to altering gene expression, as CNAs usually affect one allele, this can lead to imbalanced copy numbers of maternal and paternal alleles, resulting in imbalanced gene expression levels from them **(Fig. 3.1)**. For example, genomic segments harboring loss of heterozygosity (LOH) deterministically leads to mono-allelic expression of genes in the segment while allelic imbalance owing to allele specific gains will skew the relative expression of specific alleles. Allele-specific CNAs have been extensively delineated with bulk DNA sequencing methods [88–90]. The investigation of allele-specific CNAs has revealed a more comprehensive view of the copy number landscape in cancers, including copy-neutral LOH and "parallel events" [21, 23], wherein different alleles acquire similar alterations at the same genomic locus in the same

tumor, implying the occurrence of convergent evolution.

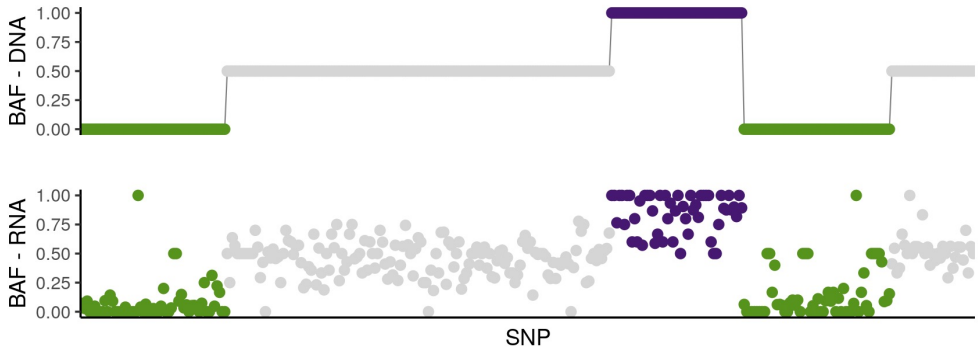

FIGURE 3.1: Allelic imbalance can be inferred from DNA data and RNA data. We assume a positive correlation between the two measurements to improve clone assignment. Y axis here represents B (minor) allele frequencies which is a metric for allelic imbalance.

The emergence of more single-cell sequencing datasets has engendered the development of novel computational tools, such as CHISEL [91], Alleloscope [92], and SIGNALS [21], designed to estimate allele-specific CN in individual cells. Single-cell datasets allow better characterization of cell-to-cell variability in allelic imbalance across cancer cell populations. For instance, SIGNALS facilitated the identification of losses of Chr 2q region on different alleles in a TP53-/- hTERT cell line, which was also found to be correlated with the down-regulation of genes situated within that genomic domain [21]. The presence of allele-specific CNAs, along with subsequent allele-specific expression (ASE), plays an important role in the trajectory of cancer evolution [93, 94]. For instance, 6p LOH has long been known as a prevalent event across diverse cancer types [95, 96], leading to HLA haplotype loss and decreased expression of HLA class I [97, 98]. This event potentially confers a selective advantage, as cancer cells undergoing HLA haplotype loss tend to evade immune surveillance more effectively [99].

The estimation of ASE can be derived from RNA-seq data through the analysis of read counts from heterozygous single nucleotide polymorphisms (SNPs). Various methods have been proposed to quantify ASE using bulk and single-cell RNA-seq data [100–103]. Given the inherent association between allele-specific CNAs and ASE, it is possible to leverage both aspects to integrate genomic and transcriptomic information. Several methods have been developed to infer allele-specific CNAs from single cell expression data and showed promising results [85, 104]. In this chapter, we will focus on building the allele-specific model of TreeAlign to take advantage of the allelic imbalance information. We showed that the incorporation of the allele-specific model enhances the clone-assignment performance of TreeAlign. By applying the integrated model that utilizes both total and allele-specific CN, to datasets from patient-derived xenografts (PDXs) and cell lines, we characterized clone-specific CN dosage effects and highlighted pathways that were differentially influenced by the cis-effects of CNAs.

3.2 The allele-specific model of TreeAlign

Allele-specific CNAs lead to allele-specific expression imbalance which is detectable in scRNA data [21, 91]. To exploit how allelic imbalance modulates allele specific expression, we extended TreeAlign to model both total CN and allelic imbalance **(Fig. 3.2, B.1)**. Given the B allele frequencies (BAFs) estimated from scDNA haplotype blocks using, for example, SIGNALS [21] and allele-specific expression at corresponding heterozygous SNPs in scRNA data, the allele-specific model contributes to clone assignment and infers the probability of the allele assignment $p(a = 1)$, $a \in \{0, 1\}$, which indicates whether the SNP is on allele B or not. The total copy and allele-specific components of the probabilistic

graphical model combine to form the "integrated model".

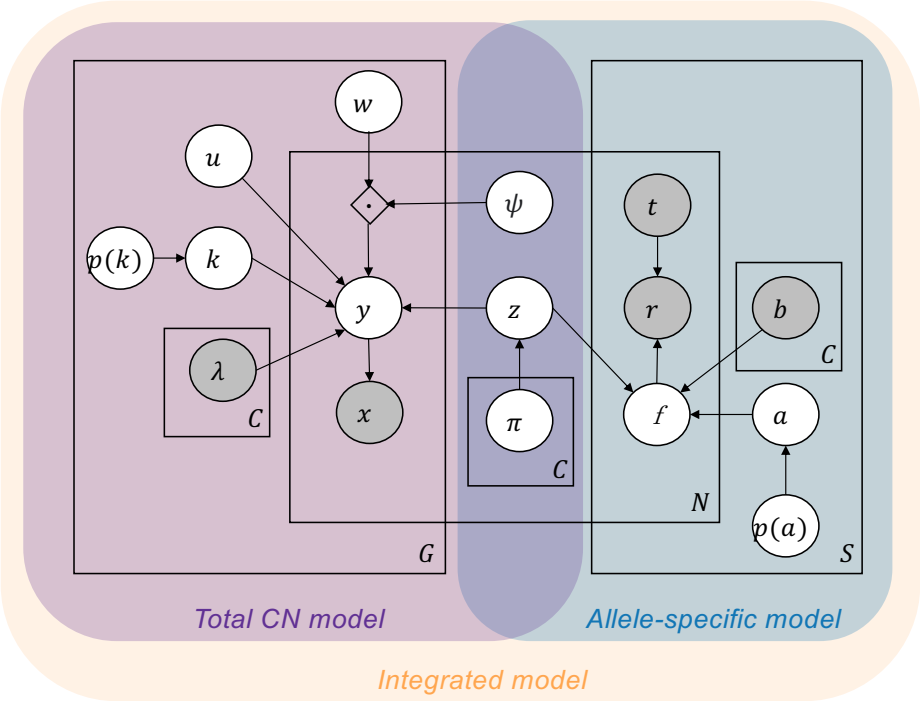

FIGURE 3.2: Graphical model of integrated TreeAlign.

The software for all models of TreeAlign (https://github.com/shahcompbio/TreeAlign) is implemented in Python using Pyro [105] and is publicly available. Our implementation allows users to run the total CN model, allele-specific model and integrated model by providing different inputs. See Methods for additional mathematical, inference and implementation details.

With synthetic dataset, we investigated how allele-specific information improves clone assignment. We simulated BAFs for varying numbers (0, 250, 500, 750 and 1000) of heterozygous SNPs with allelic-imbalance and simulated allele-specific expression from these SNPs using the generative model of allele-specific TreeAlign. We applied the integrated

model on these synthetic datasets which contained total CN and allelic information, and confirmed that clone assignment accuracy was improved when more SNPs were included **(Fig. 3.3)**.

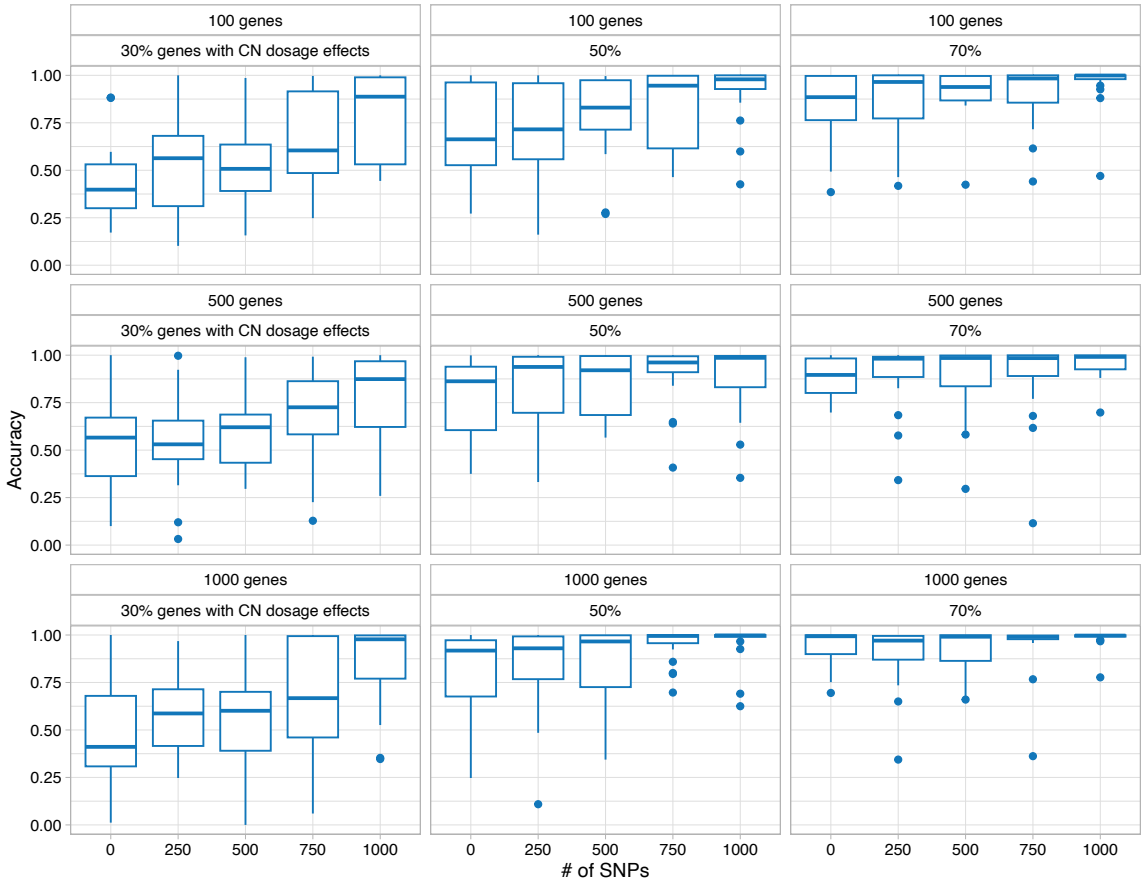

FIGURE 3.3: Accuracy of clone assignment for the integrated model of TreeAlign on simulated scRNA datasets as a function of varying numbers of heterozygous SNPs in input. Panels represent datasets with different numbers of genes and proportions of genes with CN dosage effects.

We also investigated the influence of inaccurate phylogeny input on TreeAlign performance. We randomly selected different proportions of CN profiles from scDNA and shuffled their cell labels in patient 022. With more cell labels being shuffled, the tree will become less accurate in reflecting the true phylogeny of the population. When less than

20% of cells were shuffled, TreeAlign was able to resolve the same number of subclones as with the original data **(Fig. 3.4)**. When more than 50% cells were shuffled, TreeAlign failed and assigned all expression profiles to the unassigned state. These results suggest that TreeAlign can tolerate inaccurate phylogeny input to some extent.

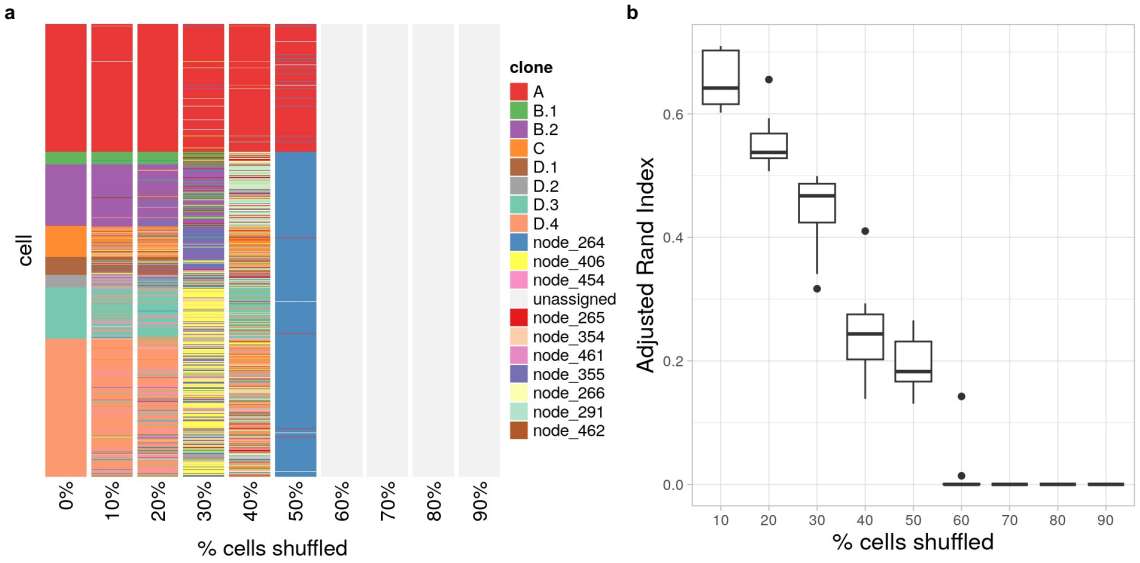

FIGURE 3.4: **a**, Heat map of clone assignment in patient 022. Columns represent input phylogenies with certain % of cell labels being randomly shuffled. **b**, Adjusted rand index of clone assignment using shuffled phylogenies in patient 022. Clone assignment results with the original phylogeny were used as ground truth for comparison.

3.3 Validation on real patient data

We next investigated the extent to which accurate clone assignment solely based on allele specific expression could be performed on real patient data. We inferred allele specific CN and BAF using scDNA data from patient 022 using SIGNALS [21]. The allele specific heat map **(Fig. 3.5)** revealed characteristic patterns of clonal LOH in whole chromosomes (e.g. chr 6,13, 14, 17) as well as subclonal losses (e.g. chr 9q in clone A and parallel

losses on chr 5 across multiple subclones). With the allele-specific model, we assigned cells from scRNA to clone A as identified by scDNA in patient 022. Clone assignments were consistent between the allele specific model and the total CN model with 87% cells concordant. The clone-specific frequencies of reads from B allele in scRNA accurately reflected scDNA BAF (**Fig. B.6**), with the exception of SNPs on chromosome X which showed allelic imbalance in scRNA but not in scDNA due to X-inactivation. The predicted allele assignments of SNPs from the allele-specific model were also consistent with haplotype phasing from scDNA reported by SIGNALS ($AUC = 0.84$) (Methods). These results suggest that allelic imbalance information can be effectively exploited for clonal mapping.

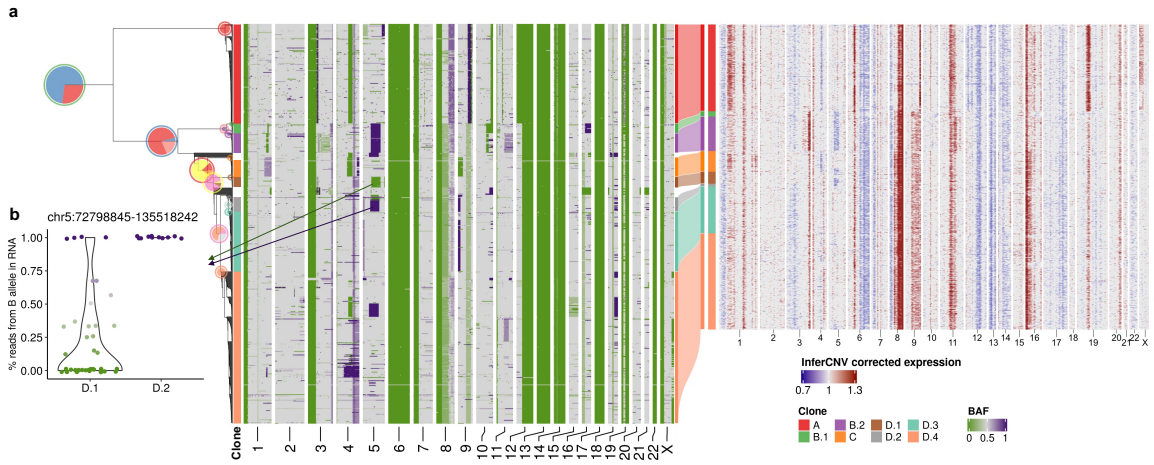

FIGURE 3.5: Integrated TreeAlign model assigns expression profiles to phylogeny of patient 022. Left heat map, single cell BAF profiles estimated from scDNA-data using SIGNALS, annotated with clone labels on the left side (BAF profiles without clone label represent cells ignored by TreeAlign) (Methods).

We then applied the integrated model utilizing both total CN and allele-specific information on data from patient 022. Relative to the total CN model, the integrated model mapped scRNA cells to smaller subclones **(Fig. 3.5)**. Specifically we note when considering allele specificity, Clone B was subdivided into two subclones (B.1 and B.2). Clone B.1 had an additional deletion at 16q leading to LOH, whereas Clone B.2 had an amplification at 11q

with increased BAF. Clone D was further divided into four subclones (D.1, D.2, D.3 and D.4). Clone D.1 and clone D.2 both had a deletion on chromosome 5, but the deletion events occurred on different alleles in the two subclones with different breakpoints, each of which was distinct from the 5q deletion on Clone B, indicative that parallel evolution is indeed reflected in transcription with the allele specific model.

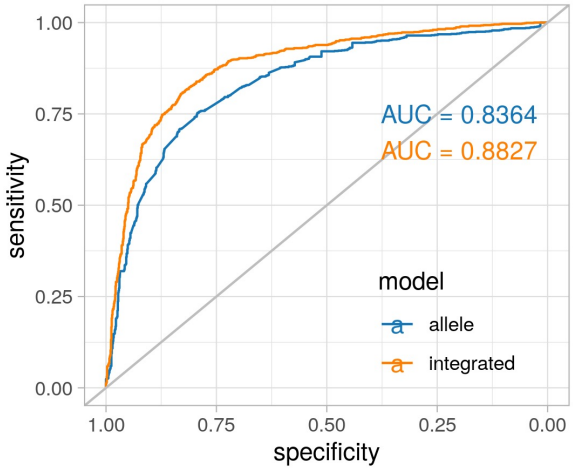

FIGURE 3.6: ROC curves for predicting $p(a = 1)$ with allele-specific TreeAlign and integrated TreeAlign.

We computed proportions of B allele reads at each heterozygous SNP for each of the subclones assigned from the scRNA data. Subclonal BAF estimated with scDNA data and proportions of reads from B allele from scRNA were significantly correlated ($0.25 < R < 0.53$ for each subclone, $P < 2.2 \times 10^{-22}$) **(Fig. 3.7, 3.8 and B.6)**, consistent with more accurate clone assignment. With integrated TreeAlign, we also achieved better performance for predicting allele assignment parameter a of SNPs compared to the allele-specific model **(Fig. 3.6)**. We note that recent identifications of parallel allele-specific alterations whereby maternal and paternal alleles are independently lost or gained in different cells [21, 23, 91] would further complicate clonal mapping, if allele specificity is not taken into account.

Here we show that mono-alleleic expression of maternal and paternal alleles is consistent with coincident maternal and paternal allelic loss in different clones **(Fig. 3.5)**. The allele-specific TreeAlign model correctly assigns cells at this level of granularity that would otherwise be missed.

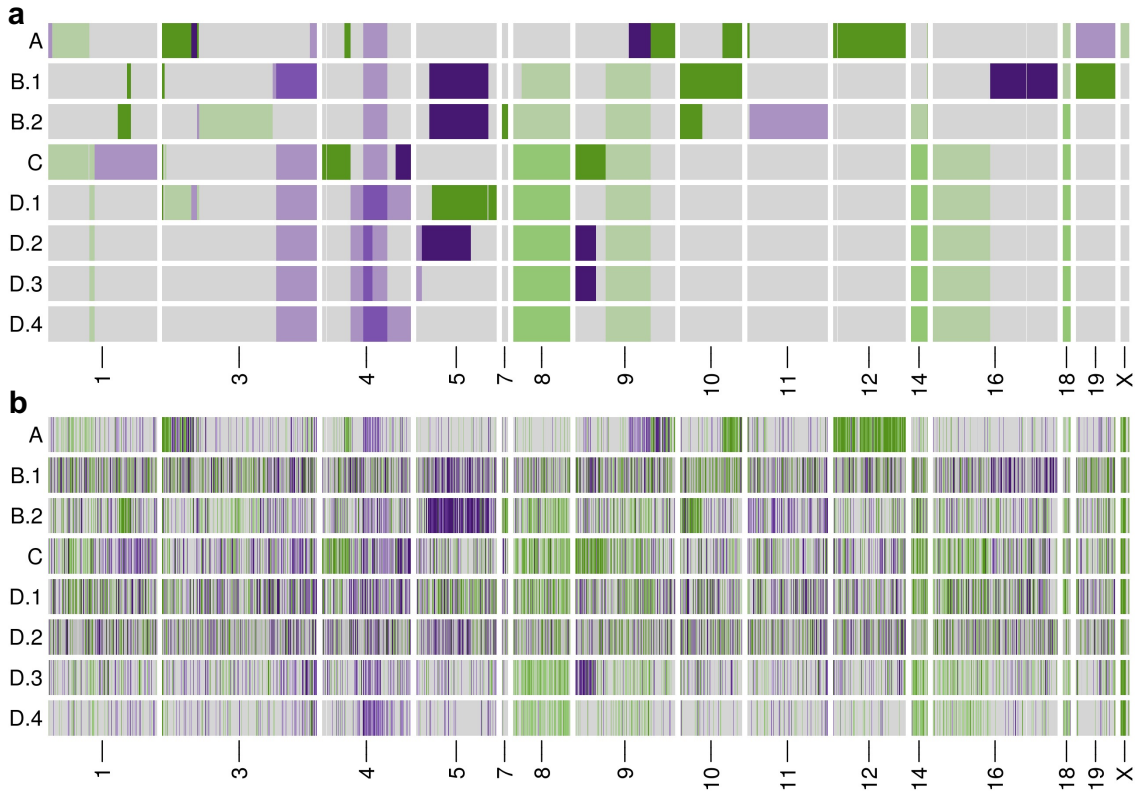

FIGURE 3.7: **a,** BAF of subclones with scDNA. **b,** Proportional of reads from B allele for subclones in scRNA.

We compared the performance of total CN, allele-specific and integrated TreeAlign using subsampled datasets of patient 022 and evaluating against results from the full dataset **(Fig. 3.9)**. The total CN and integrated model were robust to reduced numbers of cells. Compared to the total CN model, the integrated model performed significantly better when fewer genomic regions were included in the input suggesting it is more robust when there

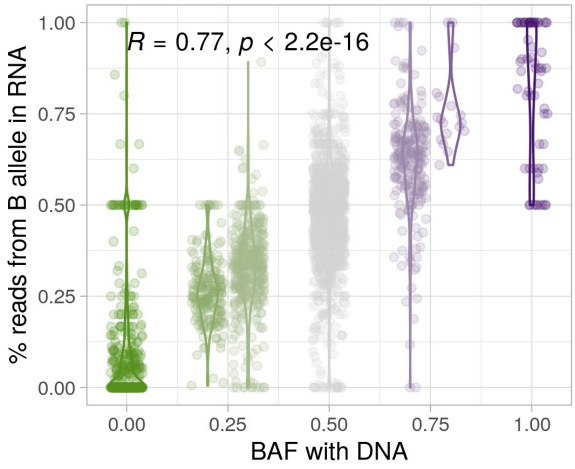

FIGURE 3.8: Correlation between % of reads from B allele in scRNA and BAF estimated with scDNA in patient 022. Annotations at the top indicate the Pearson correlation coefficient (R) and P value derived from a linear regression.

are fewer copy number differences between subclones. The allele-specific model without total CN is inferior, as expected.

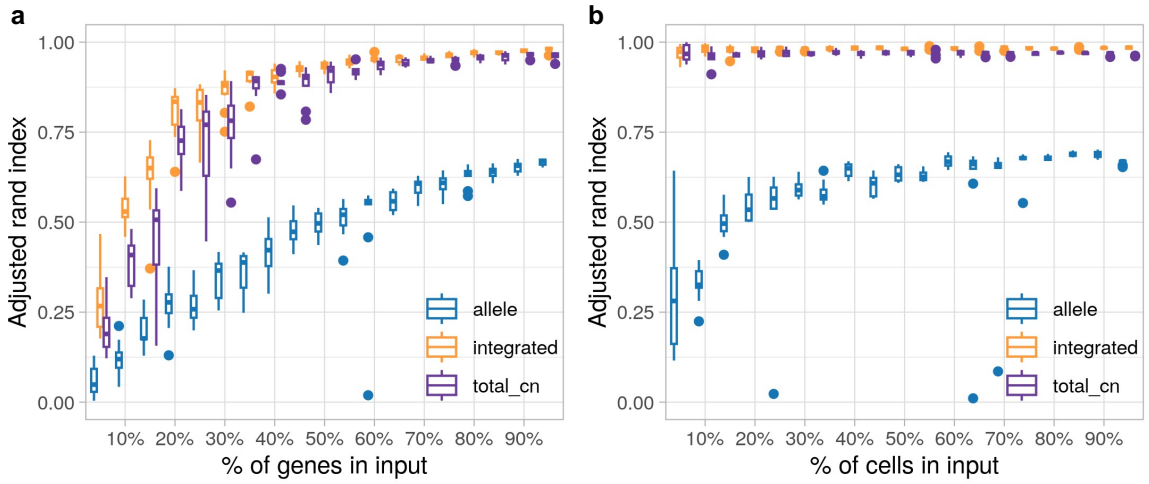

FIGURE 3.9: **a,** Robustness of clone assignment to gene subsampling in patient 022. Adjusted rand index was calculated by comparing clone assignments using subsampled datasets to the complete dataset. **b,** Robustness of clone assignment to cell subsampling in patient 022.

3.4 Inferring copy number dosage effects in human cancer data

We next compared the integrated model to the total CN model on a recently published cohort of cell lines and PDXs with scDNA and scRNA matched data from Funnell et al (referred to as "Signature cohort") [21]. We applied TreeAlign on data from PDXs of Triple Negative Breast Cancer (TNBC) ($n = 3$) and HGSC ($n = 6$). In addition we tested the model on one ovarian cancer control cell line and 6 184-hTERT cell lines engineered to induce genomic instability from a diploid background with CRISPR loss of function of TP53 combined with BRCA1 or BRCA2. Both integrated and total CN TreeAlign were run on matched DLP+ and 10x scRNA-seq data. The integrated model was fitted for 1-10 rounds **(Fig. B.8)** and the total CN model was fitted for 1-3 rounds **(Fig. B.7)** when we ran TreeAlign with the phylogeny input. The integrated model only failed to assign expression profiles to any subclones for cell line SA906a, due to a low number of genes ($n = 32$) with CN differences and heterozygous SNPs ($n = 7$) with BAF differences between subclones. In comparison, the total CN model failed in 8 cases due to lack of allelic information. As expected, the integrated model characterized more clones and achieved a lower number of cells not confidently assigned to a subclone **(Fig. 3.10)**. For cells that were assigned confidently by the integrated model but not the total CN model, their InferCNV-corrected expression showed higher correlation coefficients with the CN profiles of subclones assigned by the integrated model compared to random subclones **(Fig. B.10)**, implying better performance of the integrated model.

For high expression genes (mean normalized expression > 0.1) located in clone specific

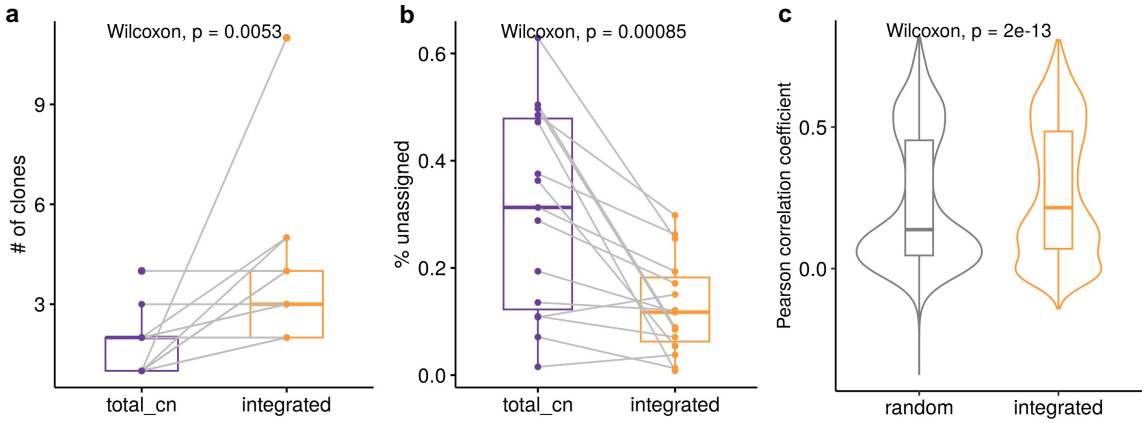

FIGURE 3.10: **a,** Number of clones characterized by total CN and integrated model (Wilcoxon signed-rank test). **b,** Frequencies of unassigned cells **(Methods)** from total CN and integrated model (Wilcoxon signed-rank test). **c,** Distribution of Pearson correlation coefficients (R) between scDNA-estimated total CN and InferCNV-corrected expression for cells assigned by the integrated model but unassigned by the total CN model. Left, correlation distribution calculated by comparing InferCNV profiles to CN profiles of a random subclone; Right, correlation distribution calculated by comparing InferCNV profiles to CN profiles of subclones assigned by integrated TreeAlign.

copy number (CSCN) regions, 76.7% (64.4% - 86.6% across cases) had $p(k) > 0.5$ suggesting their expression is dependent on copy number **(Fig. B.11, B.12)**. Taking together the simulation results and the fact that there are 13.4% - 35.6% genes with low CN dosage effects ($p(k) \leq 0.5$), we would expect benefits of incorporating k and $p(k)$ in TreeAlign as compared to CloneAlign. It was reported that cancer genes tend to have stronger CN-expression correlation compared to non-cancer genes in HGSCs [106]. We also observed concordant results that cancer genes annotated by COSMIC Cancer Gene Census [107] tend to have higher $p(k)$ compared to non-cancer genes suggesting stronger CN dosage effects in cancer genes **(Fig. B.12)**.

When we summarized $p(k)$ by genomic locations, we noticed that genes located at the same

CSCN region had more consistent $p(k)$. Notably, $p(k)$ of genes in a contiguous region exhibited significantly lower variation compared to randomly sampled genes across different regions **(Fig. 3.11)**. It should be noted that we only included CN events that span more than 10 genes in this analysis. Therefore, it is not known whether the conclusion still holds for more focal copy number events. In addition to broad regions of the genome, we note subclonal high-level amplifications affecting known oncogenes which have been identified previously [21]. Using TreeAlign, we also identified subclonal amplifications of oncogenes accompanied by consistent changes in gene expression. For example, in OV2295, subclonal upregulation of MYC expression coincides with the clone-specific MYC amplification with $p(k) > 0.8$ **(Fig. B.13)**. To investigate whether MYC pathway activation was also impacted by non-CNA driven effects, we performed pathway enrichment on genes with low $p(k)$ and found genes in the Hallmark MYC Target V1 gene set [108] significantly enriched in low $p(k)$ genes. Combined with HLAMP results, this suggests the pathway can be regulated by both CN dosage effects and other (potentially non-genomic) effects at the subclonal level **(Fig. B.14)**, further highlighting the importance of $p(k)$ for interpreting the mechanism of gene dysregulation.

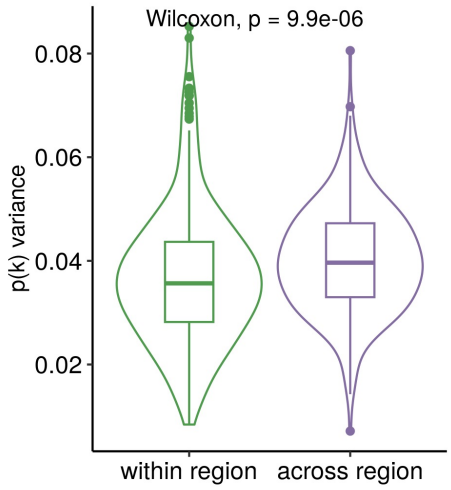

FIGURE 3.11: Variance of $p(k)$ sampled from the same genomic regions and across regions.

Chapter 4

Clone-specific phenotypes in HGSCs

4.1 Introduction

Investigations into the genetic ITH of HGSCs have relied on bulk and single cell DNA sequencing [21, 32, 62, 64], which unveiled dynamic changes in clonal composition based on tumor location and treatment. Clonal diversity is also closely linked to the surrounding immune microenvironment, with tumors situated in immune-rich environments demonstrating reduced diversity, implying the prevalence of purifying selection [66]. Leveraging scDNA-seq improves our ability to further disentangle genetic ITH in HGSC. By applying scDNA-seq to HGSC and TNBC PDXs, extensive cell-to-cell variations in amplitudes and breakpoints of CN events were identified and shown to be associated with the ongoing mutational processes [21].

Previous investigations have similarly delved into the transcriptomic ITH and characterized recurrent cancer cell states in HGSC through scRNA-seq [109, 110]. The integration of matched scDNA and scRNA datasets facilitated by TreeAlign can provide us a great opportunity to link the genetic and transcriptomic ITH.

We applied TreeAlign on data collected through MSK-SPECTRUM, which is an ongoing project at MSK aiming to discover spatio-temporal determinants of ovarian cancer evolution, treatment and response [20]. Multimodal data including genomic, pathologic, radiologic and clinical information were collected from HGSC patients. As a part of the project, scRNA-seq and scDNA-seq were conducted on tumor samples from multiple intraperitoneal sites such as left/right adnexa, omentum and ascites. The MSK-SPECTRUM cohort currently contains scRNA-seq and scDNA-seq data from more than 40 patients and is a valuable resource to investigate the transcriptional phenotypes of HGSCs.

This chapter is dedicated to the application of TreeAlign to HGSC samples from MSK-SPECTRUM and Signature cohort, with a primary focus on unraveling clone-specific transcriptional diversity and elucidating the phenotypes associated with metastasis and whole genome doubling.

4.2 Clone-specific transcriptional phenotypes

We sought to interpret clone-specific transcriptional phenotypes and phenotypic divergence during clonal evolution from TreeAlign mappings. For patient 022, differential expression and gene set enrichment analysis identified genes and pathways upregulated in each clone **(Fig. 4.1)**. In total, we found 1346 genes significantly upregulated (adjusted $P < 0.05$, MAST [111]) in at least one of the subclones in patient 022. 52.1% (701) of these genes were not located in CSCN regions, while 47.9% (645) genes were located within CSCN regions. For 90.7% (585/645) of genes in CSCN regions, $p(k)$ was > 0.5, reflecting probable gene dosage effects.

Immune related pathways such as IFN-α and IFN-γ response were differentially expressed,

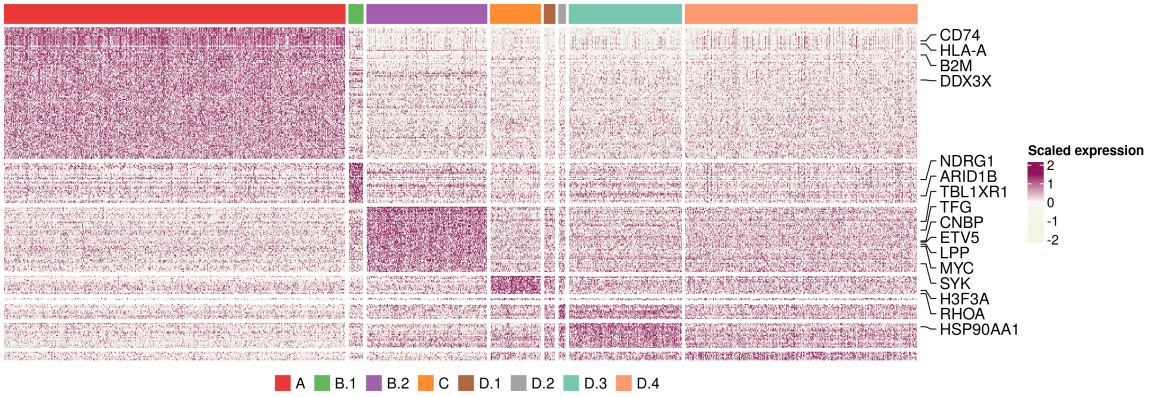

FIGURE 4.1: Scaled expression of upregulated genes in each subclone in patient 022, showing genes in rows and subclones in columns. Genes in the COSMIC Cancer Gene Census [107] are highlighted.

and with increased relative expression in clone A **(Fig. 4.2)**. Clone A contains cells from both right and left adnexa, thus dysregulation of these pathways cannot be simply explained by the microenvironment of clone A. Differential expression of immune related pathways was also found between more closely related subclones **(Fig. B.15)**. Compared to clone B.2, clone B.1 also has enriched expression in IFN-α and IFN-γ signaling pathways and downregulation in MYC targets V1 and G2M checkpoint gene sets. Clone D.4, compared to other clone D subclones, had down-regulated TNF-α signaling via NFκB. Seeking to explain the relative contribution of subclonal CNAs to differentially expressed pathways, we analyzed the proportion of differentially expressed genes found in subclonal CNAs for each pathway. Only 17.4% (4/23) of differentially expressed genes in the Allograft Rejection gene set are in CSCN regions compared to 61.5% (24/39) in the MYC Targets V1 gene set highlighting the distinct impact of subclonal CNA between pathways **(Fig. B.16)**.

We conducted a similar analysis on data from the Signature cohort. Differential expression analysis revealed varying proportions of DE genes located in CSCN regions ranging from

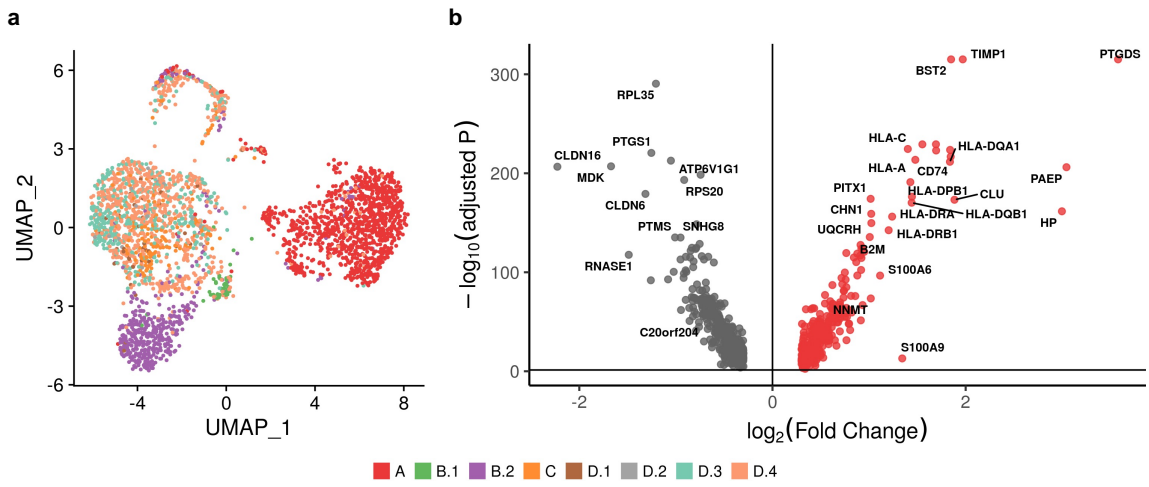

FIGURE 4.2: **a,** UMAP embedding of expression profiles from patient 022 colored by clone labels assigned by integrated TreeAlign model. **b,** Differentially expressed genes between clone A and other subclones (clone B-D) in patient 022.

1.3% to 63.9%, indicating that transcriptional heterogeneity due to cis-acting subclonal CNAs varied across tumors **(Fig. 4.3)**.

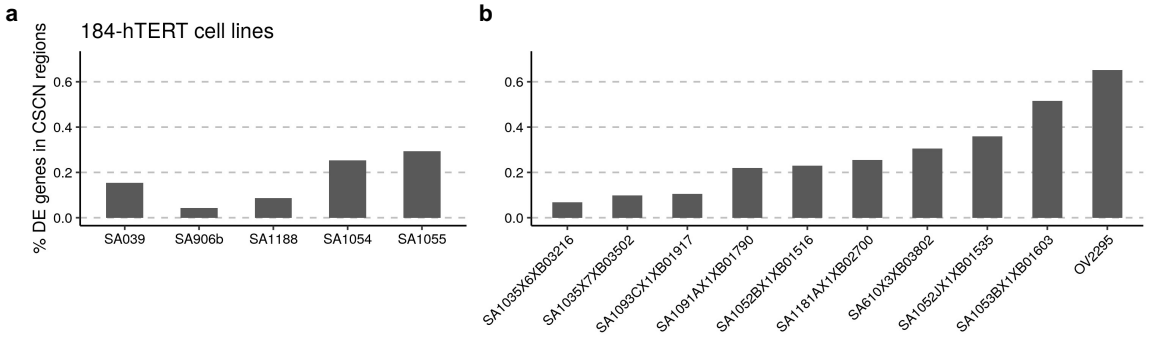

FIGURE 4.3: a-b, Proportions of subclonal differentially expressed genes located in CSCN regions for **(a)** 184-hTERT cell lines, **(b)** an HGSC control cell line and PDXs.

In addition to pathways such as KRAS signaling which are known to be important in these tumors, using GSEA, we found that IFN-α and IFN-γ response pathways also show variable expression between subclones of TNBC and HGSC PDXs frequently **(Fig. 4.5)**. We applied TreeAlign on additional patients from SPECTRUM with matched scRNA and

scDNA data and summarized frequently dysregulated pathways between clones **(Fig. 4.6)**. Immune related pathways such as IFN-α and IFN-γ response again show highly variable expression between clones. For example, in patient 105, clone E has upregulated expression of genes in IFN responses and antigen presentation **(Fig. 4.4)**. IFN signaling has important immune modulatory effects, and has been previously linked to immune evasion and resistance to immunotherapy [112]. The recurrent differential expression of immune related pathways between subclones suggests their importance in clonal divergence in these cancers of genomic instability.

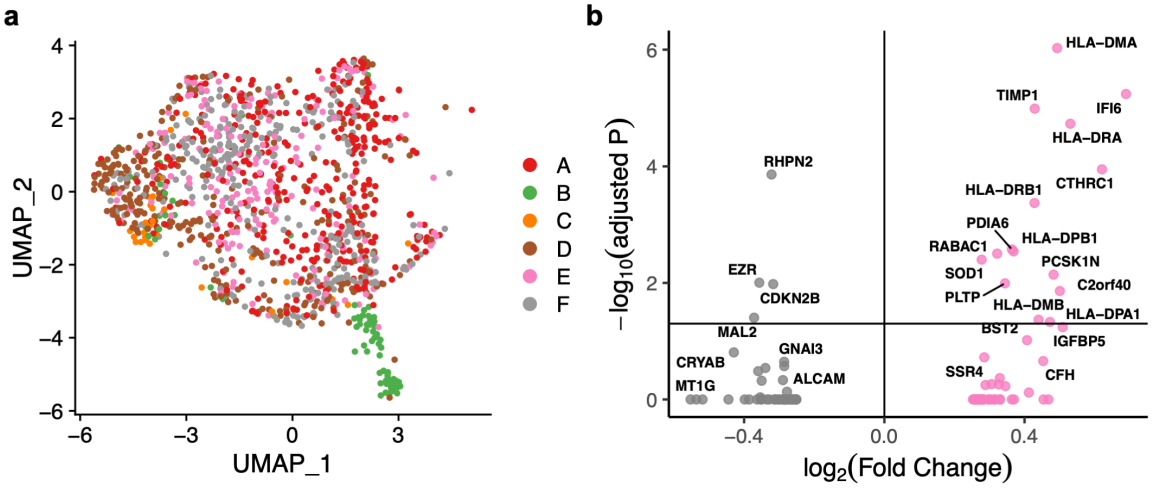

FIGURE 4.4: **a,** UMAP embedding of cancer cell expression profiles from patient 105 colored by clone assignment from TreeAlign. **b,** Differentially expressed genes between clone E and other subclones in patient 105.

To investigate transcriptional diversity within and across subclonal populations, we calculated Pearson correlation coefficients and Euclidean distance between cells using the top 20 principal components of the gene expression matrices. In addition to TreeAlign, we also used InferCNV to assign cells from scRNA to genomic clones. We found that cells sampled from the same TreeAlign clone or InferCNV clone tend to have higher correlation and

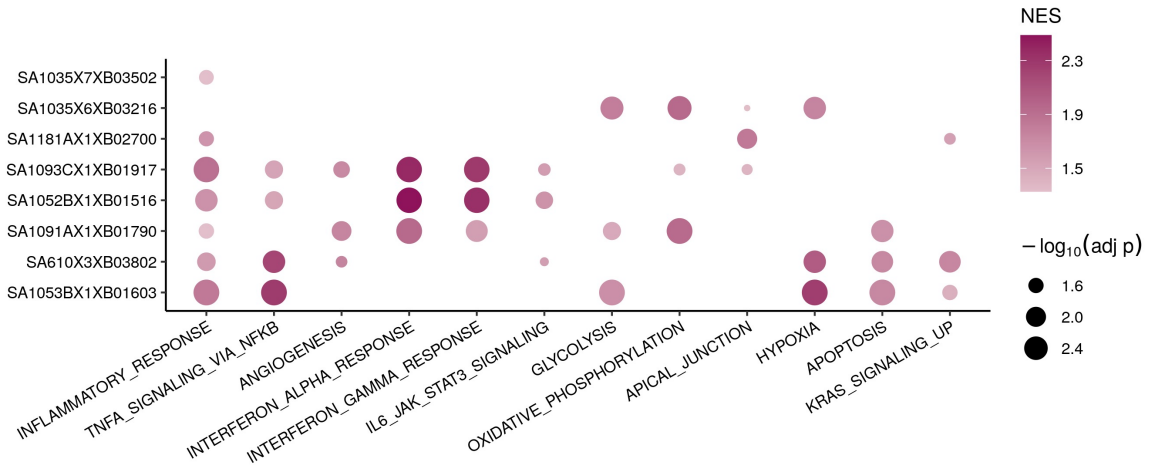

FIGURE 4.5: Pathways with clone-specific expression patterns in TNBC and HGSC PDXs.

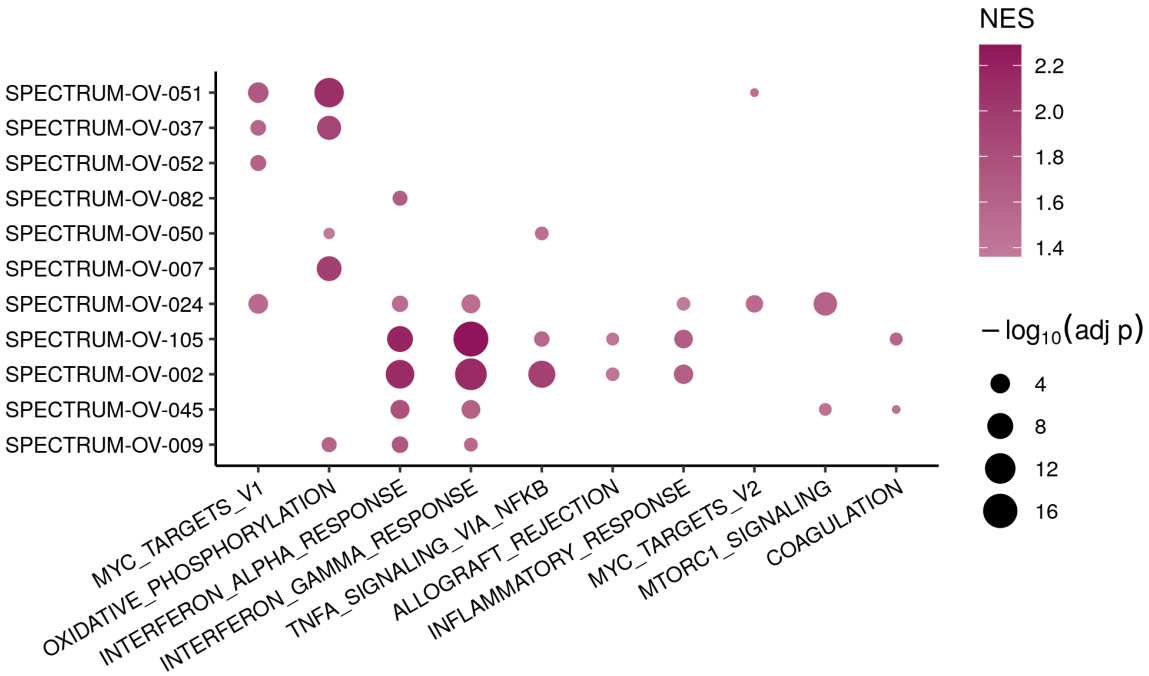

FIGURE 4.6: Pathways with clone-specific expression patterns in additional SPECTRUM patients.

lower distance **(Fig. 4.7)**, suggesting lower transcriptional diversity within the subclonal populations.

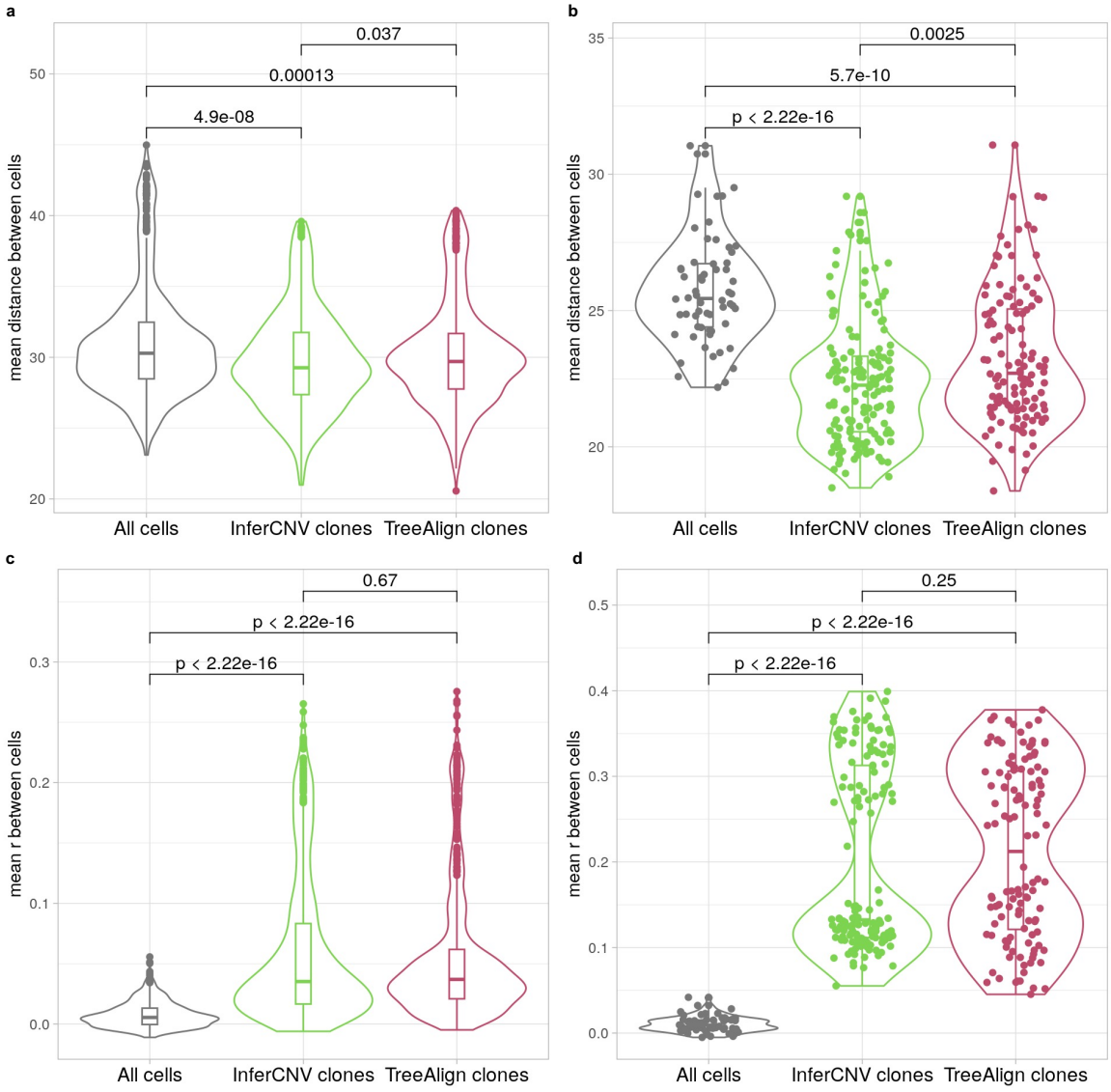

FIGURE 4.7: **a-b,** mean Euclidean distance between cells in scRNA-data sampled across or within subclones for **(a)** HSGC PDXs and cell lines and **(b)** patient 022. **c-d,** mean Pearson correlation coefficient between cells in scRNA-data sampled across or within subclones for **(c)** HSGC PDXs and cell lines and **(d)** patient 022.

4.3 Cis-effects of CNAs in HGSC metastasis

We also tried to delineate the transcriptional variances existing between primary and metastatic HGSCs and to elucidate whether these transcriptional discrepancies can be directly attributed to the cis-effects of CNAs. We focused on four patients (specifically, patients 009, 037, 081, and 083) from the SPECTRUM dataset **(Fig. B.20, B.21, 4.8 and B.22)**, where both scDNA and scRNA data were available, spanning the primary anatomical sites of left and right adnexa, in addition to metastatic locations such as the infracolic omentum. We employed DE and GSEA to compare the cancer cells from the primary and metastatic sites for each patient. Parameter $p(k)$, as inferred by TreeAlign, enabled us to discriminate whether a gene's expression relied upon clone-specific CN dosage effects.

Across all four patients, most genes displaying significant differential expression between primary and metastatic sites were not situated within CSCN regions. This suggests that the observed differences in their expression levels were not primarily attributed to cis-effects of clone-specific CNAs. We further categorized genes located within CSCN regions into two distinct groups based on their $p(k)$ values: genes with high $p(k)$ ($p(k) \geq 0.5$) and genes with low $p(k)$ ($p(k) < 0.5$). Subsequently, GSEA was applied independently to high $p(k)$ genes, low $p(k)$ genes, and genes located outside CSCN regions to identify pathways that exhibited enrichment within these gene sets.

In the case of patient 009, immune-related pathways such as interferon gamma signaling and PD-1 signaling were found to be upregulated in cancer cells at the metastatic site for genes located outside CSCN regions. This suggests that these pathways are not predominantly regulated by the cis-effects of CNAs. Similarly, for patient 081, interferon signaling

pathways exhibited upregulation in metastatic sites, but this upregulation was evident primarily in genes with low $p(k)$ or genes located outside CSCN regions, rather than in genes with high $p(k)$ **(Fig. 4.8)**. This observation indicates that gene regulatory mechanisms other than CN dosage effects assume a more significant role in modulating immune signaling at metastatic sites.

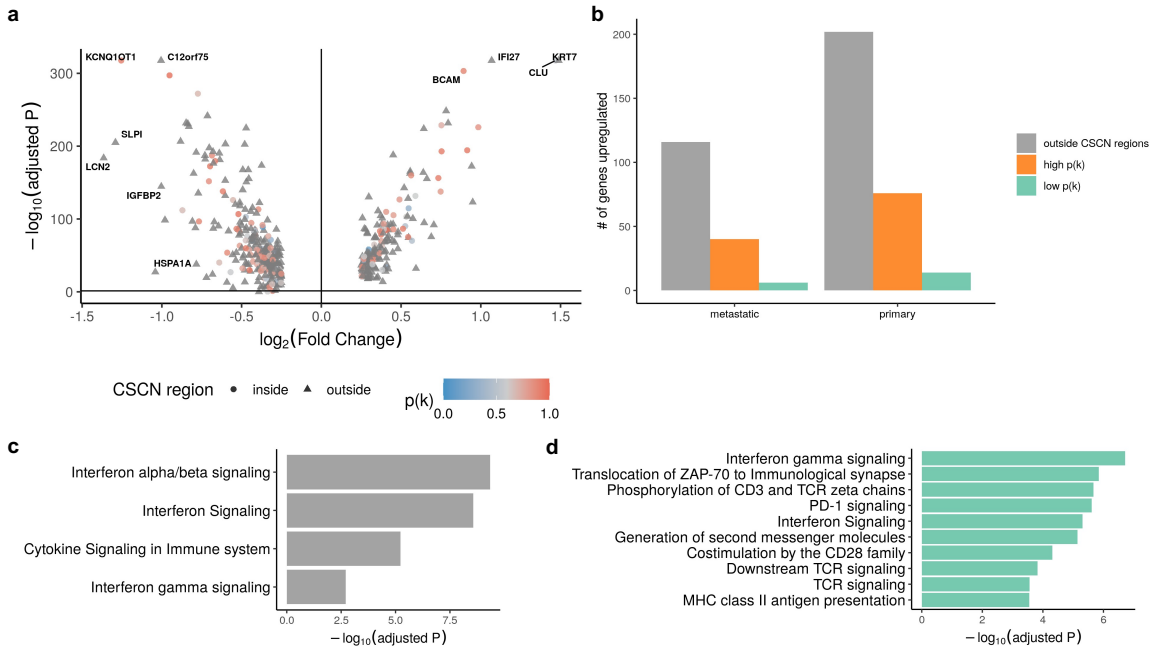

FIGURE 4.8: **a,** Differentially expressed genes between metastatic and primary sites in patient 081. **(b,)** Number of upregulated genes in metastatic and primary sites grouped by $p(k)$ level in patient 081. **c,** Upregulated gene sets among genes outside of CSCN regions in patient 081 metastatic site. **(d,)** Upregulated gene sets among low $p(k)$ genes in patient 081 metastatic site.

4.4 Clone-specific changes in WGD tumors

Whole genome doubling (WGD) is a frequently occurring event in cancer, detectable in approximately 30% of tumors [113–115]. This genomic alteration manifests early in the trajectory of tumorigenesis, allowing cancer cells to tolerate detrimental somatic mutations

within LOH regions [12]. Consequently, WGD is particularly favored in tumors marked by elevated rates of somatic CNAs [115]. WGD also exhibits an association with immune evasion. Tumors displaying WGD tend to exhibit reduced number of tumor-infiltrating leukocytes and diminished immune responses [14, 115]. Hypotheses have been formulated suggesting that WGD in tumors may lead to impaired antigen presentation or a reduced concentration of neoantigens, thereby contributing to the observed dampened immune responses [14]. In this section, we focused on exploring the transcriptional phenotypes linked to WGD and demonstrating the feasibility of distinguishing between closely related diploid and tetraploid subclonal populations through the utilization of TreeAlign.

Using scDNA data, we were able to estimate the prevalence of WGD in tumor samples from SPECTRUM. Notably, the distribution of WGD frequency exhibited a bimodal pattern within this cohort **(Fig. 4.9)**. Specifically, 25 samples had more than 80% of cells with WGD (referred to as "WGD samples"), while 38 samples exhibited fewer than 20% of cells with WGD (referred to as "nWGD samples"). Only three samples presented with WGD frequencies ranging between 20% and 80%. Moreover, it was observed that samples derived from the same patient often displayed consistent WGD status. Among the 19 patients for whom scDNA data was available at multiple anatomical sites, only patient 081 had both a WGD sample (from the infracolic omentum) and an nWGD sample (from the left adnexa), consistent with the notion that WGD is an early event in the progression of cancer.

Subsequently, employing DE and GSEA, we conducted a comparative analysis of cancer cells from WGD and nWGD samples with scRNA data **(Fig. 4.10)**. This analysis revealed that cancer cells originating from WGD samples exhibited a down-regulation in the expression of genes associated with interferon signaling pathways, alongside an up-regulation of

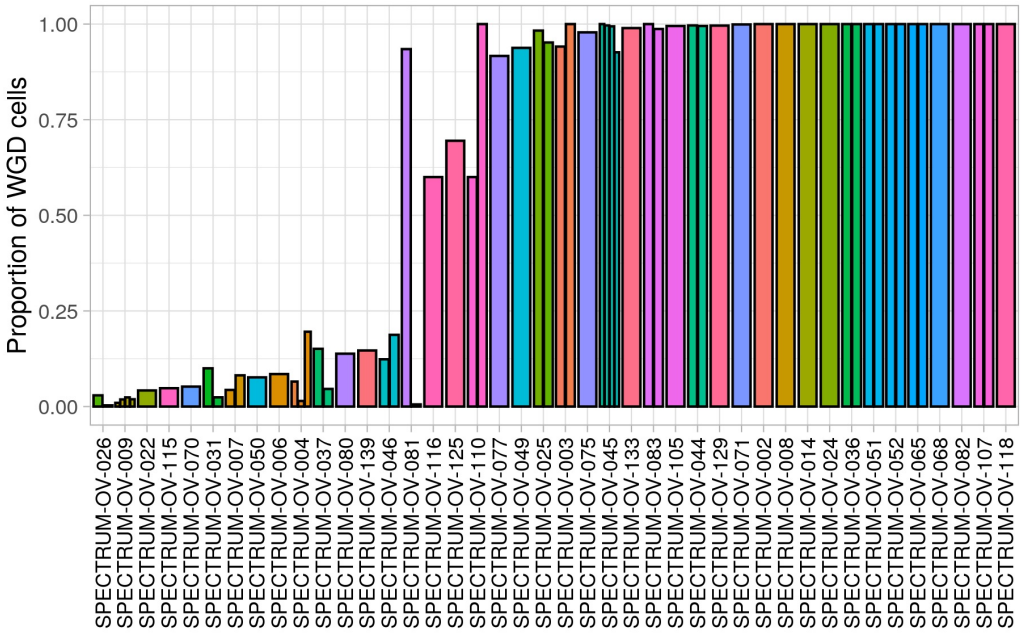

FIGURE 4.9: Frequencies of WGD cells in SPECTRUM samples

MYC targets V1, mitotic spindle and DNA repair pathways. These findings were consistent with prior analyses [115] and indicated that WGD in cancer cells may result in a diminished immune response and contribute to the down-regulation of interferon signaling.

Interestingly, in the case of patient 081 infracolic omentum, despite 91.76% (512 out of 558) of cancer cells in the tetraploid state, the remaining 8.16% cells are diploid, forming a distinct subclone as informed by scDNA data. We applied TreeAlign to assign cancer cell expression profiles from this sample to the two subclones with different ploidy. TreeAlign proficiently assigned 89.17% (1687 out of 1885) of scRNA cells to the tetraploid clone and 10.83% (145 out of 1885) to the diploid clone, in accordance with the scDNA-based estimations **(Fig. 4.11)**. Cells assigned to the tetraploid clone exhibited a higher total read count per cell and a greater proportion of reads originating from mitochondrial genes **(Fig. 4.12)**. Leveraging these cell assignments, we were able to undertake a comparative

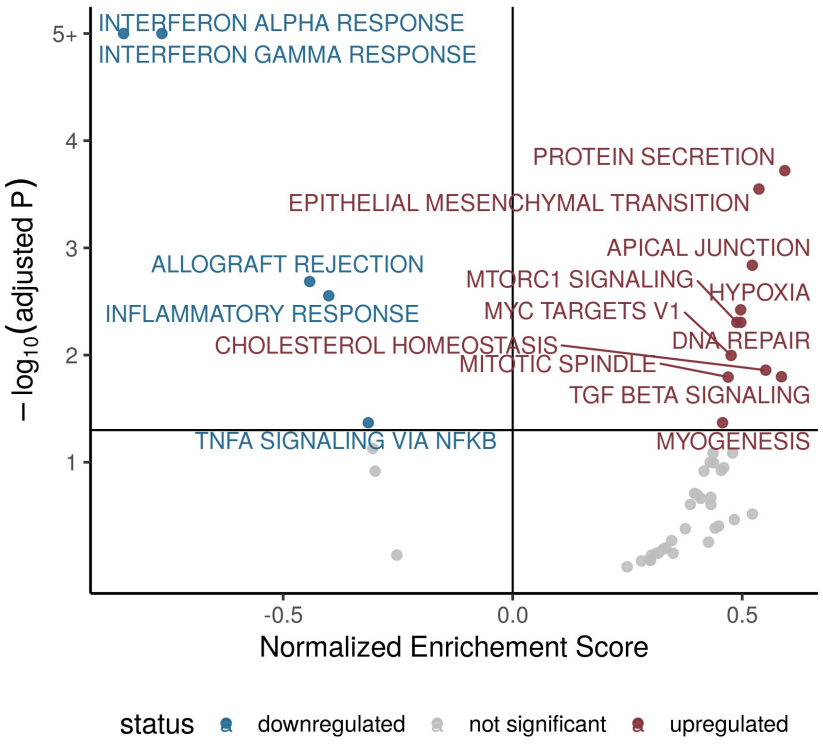

FIGURE 4.10: Volcano plot showing upregulated and downregulated pathways in WGD samples

analysis of gene expression between the closely related diploid and tetraploid subclones originating from the same anatomical site **(Fig. 4.13)**. In concordant with the broader sample-level comparisons, the gene set of MYC targets V1 was found to be up-regulated in the tetraploid clone. However, interferon gamma response pathways and related immune pathways displayed an up-regulation in the tetraploid cells, diverging from the sample-level results.

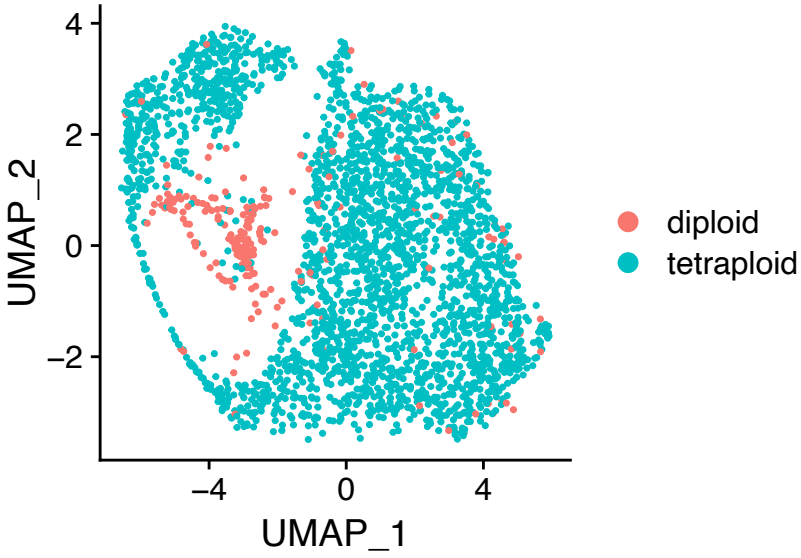

FIGURE 4.11: UMAP embedding of cancer cells from patient 081 infracolic omentum colored by ploidy state assigned by TreeAlign

Patient 081's unique presentation of mixed diploid and tetraploid cells at infracolic omentum, alongside the availability of matched scDNA and scRNA data, enabled TreeAlign to characterize WGD and nWGD clones originating from the same anatomical site, thus facilitating a comparative analysis of their transcriptional phenotypes in the same microenvironment. The presence of additional cases exhibiting similar characteristics could further harness the utility of TreeAlign in elucidating transcriptional changes associated with WGD.

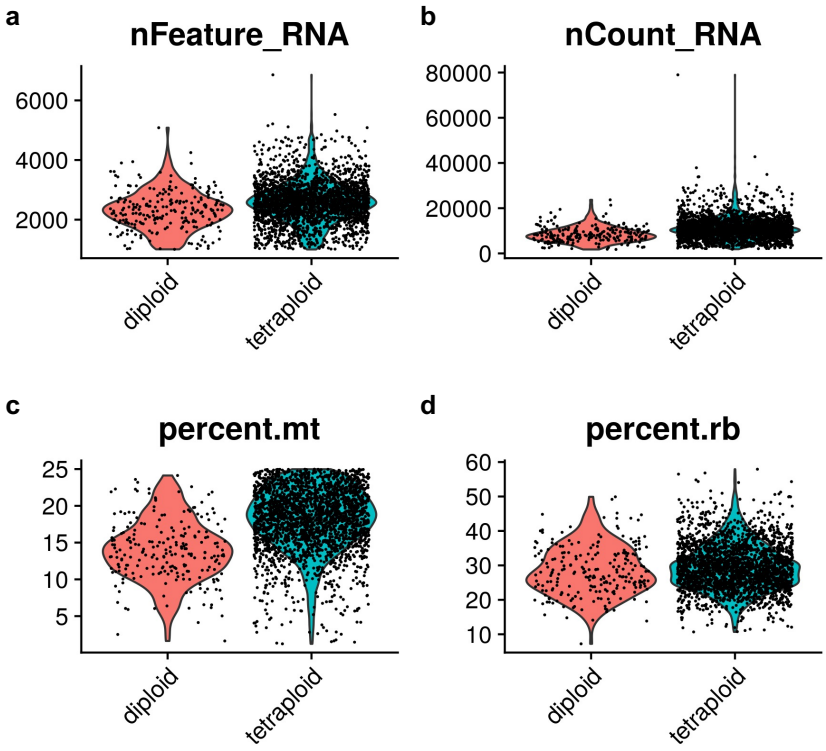

FIGURE 4.12: **a,** Number of genes detected per cell in patient 081 scRNA grouped by ploidy status. **b,** Number of reads per cell in patient 081 scRNA grouped by ploidy status. **c,** Percentage of reads from mitochondrial genes in patient 081 scRNA grouped by ploidy status. **d,** Percentage of reads from ribosomal genes in patient 081 scRNA grouped by ploidy status.

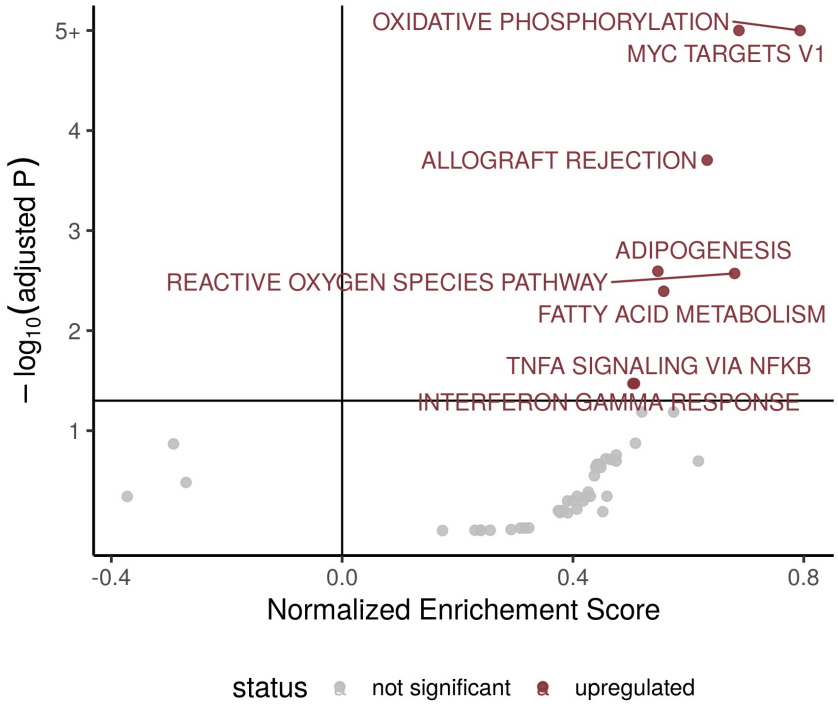

FIGURE 4.13: Volcano plot showing upregulated and downregulated pathways in WGD samples in WGD cells from patient 081 infracolic omentum

4.5 Potential extension of TreeAlign to other data modalities

Single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) is a powerful technique to probe the epigenetic landscape of individual cells by measuring chromatin accessibility. Beyond its primary usage in characterizing epigenetic signals, scATAC-seq data can also capture CN signals, as genomic regions experiencing gains or losses in copy number tend to have corresponding shifts in chromatin accessibility. Therefore, chromatin accessibility, as quantified by scATAC-seq, exhibits a positive correlation with the underlying CN status. To extract CN information from scATAC data, multiple methods have been developed [92, 116, 117], enabling the inference of both total and allele-specific CN profiles. However, due to relatively low coverage achieved at the single-cell level, scATAC-seq generally characterizes CN events with lower resolution compared to scDNA-seq[118].

In this section, we explored the feasibility of integrating scDNA and scATAC, these distinct but complementary data modalities using the TreeAlign framework. Specifically, we generated scATAC data for two patients, patient 037 and patient 051 in SPECTRUM **(Fig. 4.14)**, in conjunction with their respective scDNA profiles **(Fig. 4.15)**. We merged scATAC profiles from the two patients. Our goal was to investigate whether TreeAlign could be employed to effectively integrate scDNA and scATAC data, thereby enabling us to assign scATAC profiles back to the corresponding copy number profiles of the originating patient.

In contrast to scRNA data, where we utilized cell $\times$ gene expression matrices as input, for scATAC, we generated cell $\times$ genomic region matrices of fragment counts, with each region spanning 500kb in length, as TreeAlign input. In parallel, for the copy number input,

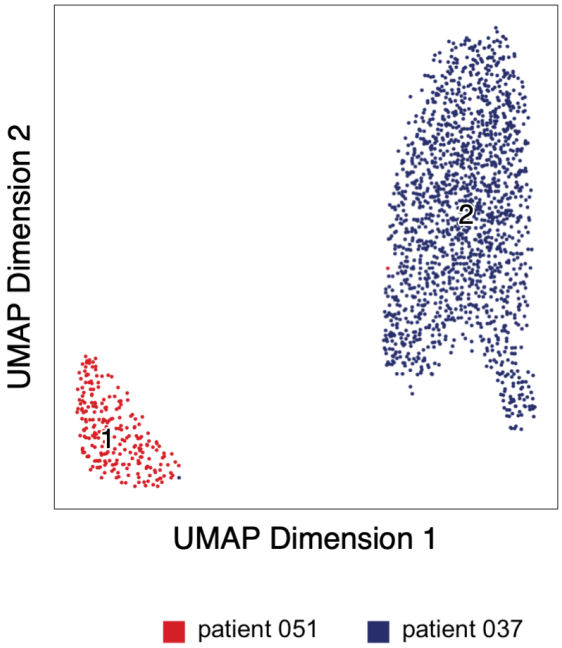

FIGURE 4.14: UMAP embedding of scATAC profiles of cancer cells from patient 037 and patient 051

we used cell $\times$ genomic region CN matrices rather than cell $\times$ gene matrices. TreeAlign successfully assigned scATAC profiles to the correct patient, achieving an accuracy of 96.81%. Although the divergence between CN profiles from different patients typically exceeds that between subclones within a patient, making scATAC profile assignment comparatively more straightforward in this scenario, the successful integration underscores the potential applicability of the TreeAlign framework to other data modalities.

In summary, our findings showcase the adaptability of TreeAlign for effectively merging scATAC and scDNA data, a development with the potential to enhance our understanding of the intricate interplay between genomic alterations and epigenetic states at the single-cell level. Furthermore, this study opens the door to the application of TreeAlign across a broader spectrum of data types, promising deeper insights into diverse aspects of cellular

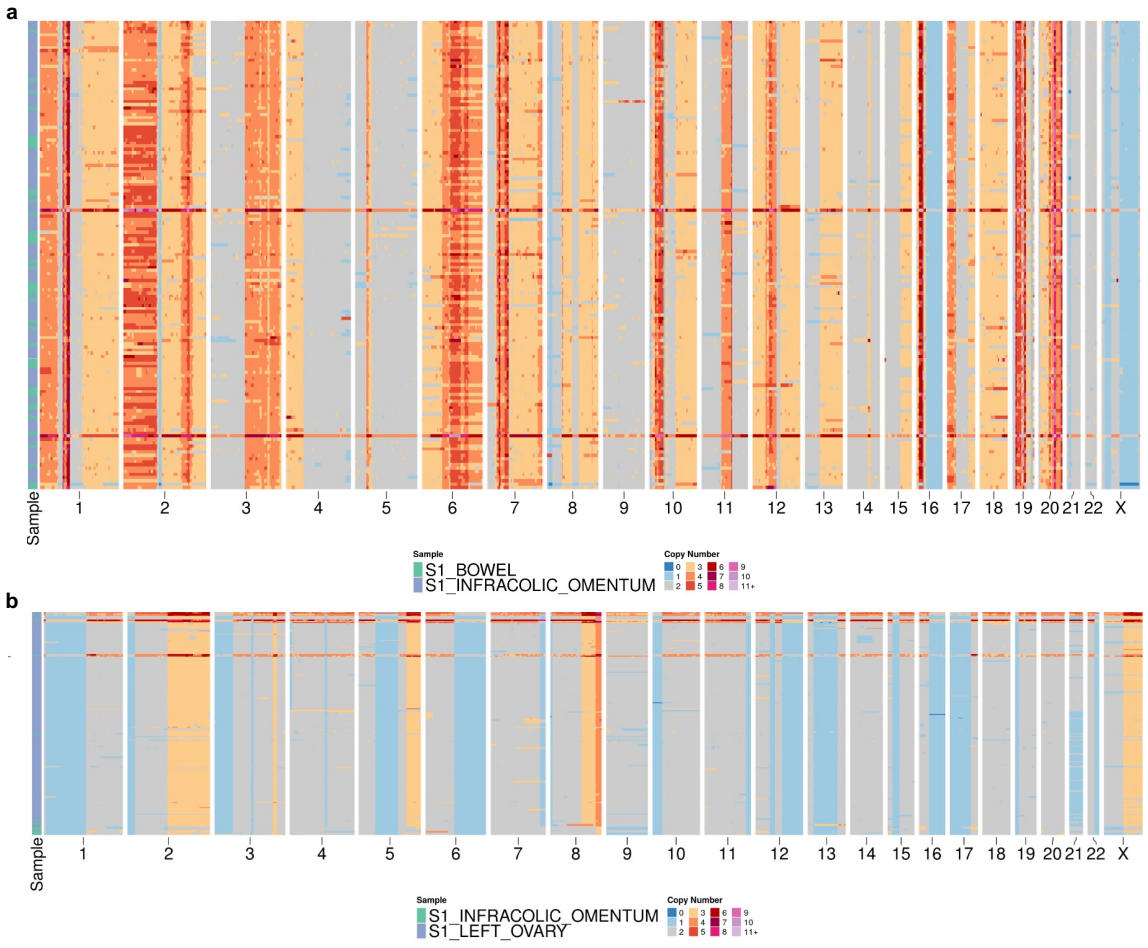

FIGURE 4.15: Heatmaps showing single cell CN profiles inferred from DLP+ data for patient 051 **(a)** and patient 037 **(b**

heterogeneity and its genetic underpinnings.

Chapter 5

Conclusion

TreeAlign establishes a probabilistic framework for integration of scRNA and scDNA data and inference of dosage effects of subclonal CNAs. TreeAlign achieves high accuracy of assigning single cell expression profiles to genetic subclones and was built to operate on phylogenetic trees directly, therefore informing phenotypically divergent subclones during the recursive clone assignment process. In addition to scRNA and scDNA integration, TreeAlign disentangles the cis dosage effects of subclonal CNAs which highlights highly regulated pathways in clonal evolution. The model also has improved flexibility allowing either total or allelic copy number or both to be used as input. With additional allele-specific information, TreeAlign has improved prediction accuracy and model robustness and is able to identify more refined clonal structures.

5.1 Limitations and future directions of TreeAlign

In terms of limitations, TreeAlign was designed to integrate matched scRNA and scDNA datasets. For partially matched datasets with different clonal compositions, TreeAlign

may have compromised performance. TreeAlign also assigns expression profiles based on clone-specific CNAs. For cancer types not driven by CN events, TreeAlign is not suitable due to lack of input features. The way TreeAlign encodes the relationship between gene expression and CN could also be further improved in the future. By default, TreeAlign truncates CNs > 10 to 10 and represents the CN-expression relationship with a linear function. Functions that are more biologically meaningful could be used to replace the current setup. Last, although we tried to address the issue of tree cutting by implementing an iterative process to allow TreeAlign define subclones informed by transcriptional profiles, the single cell phylogenetic tree itself is still purely based on scDNA data. Applying TreeAlign on trees constructed by different phylogenetic inference methods may result in different clones being characterized.

Based on TreeAlign, we may consider other model setup to address some of the current limitations. In CCNMF [81], which is another method for scDNA and scRNA integration, the authors use negative matrix factorization to co-cluster scDNA and scRNA profiles to define clones. I think it is possible to combine this approach and the dosage effect modeling in TreeAlign to allow defining clones or phylogenies based on both genomic and transcriptomic information. The underlying assumption is that there exists a set of clones with distinct CN profiles and these profiles can explain both scDNA and scRNA data. From the clone-specific CN profiles, we can model scDNA output using a hidden markov model and model scRNA output based on dosage effects similar to TreeAlign. The resulting model should be able to infer clone-specific CNs, CN dosage effects and assign cells from scDNA and scRNA to a set of clones. Additionally, I think it is possible to enforce phylogenetic constraints on the inferred clone-specific CN profiles, and through this way, we can also infer clone-based phylogenies. In this model setup, clones or phylogenies are defined by

both scDNA and scRNA which may be more informative for the integration task.

In this dissertation, we validated and compared TreeAlign to other methods using synthetic datasets. The synthetic datasets were produced from the generative model of CloneAlign. For Bayesian probabilistic modeling, it is a common approach to generate synthetic data from the generative model for validation. However, if the method outperforms competing methods on synthetic datasets, it does not mean that it will definitively perform better on real datasets since the assumption on the distributions of the real data could be inaccurate. Using synthetic datasets generated by an independent method unrelated to the models being evaluated may be a better approach for fairer comparisons. One potential solution here is using constrained autoencoders [119] to simulate CN-dependent scRNA data for validating TreeAlign and related methods.

We also expect potential extensions of TreeAlign for integration of other single cell data modalities such as single-cell epigenetic data. Current methods for integration of scRNA and scATAC data are primarily based on nearest neighbor graphs or other distance metrics to match similar cells across multimodal datasets. In Chapter 4.5, by mixing scATAC data from multiple patients, we demonstrated it is possible to assign scATAC profiles back to correct patients. The advantage of TreeAlign is that it estimates how well the expression of a gene matches with the given biological assumption, hence it is more interpretable and provides explanations for gene expression variations.

With development of technologies that allow sequencing multiple modalities (e.g. RNA and DNA) co-registered in the same cell, one question we may ask is whether TreeAlign would still be useful in the future. With such co-registered datasets, the clone assignment functionality of TreeAlign would no longer be needed, however, the ideas of inferring CN

dosage effects and modeling gene expression with genomic features are still useful for interpreting biological mechanisms that regulate cell phenotypes.

5.2 Understanding transcriptional regulations with TreeAlign

The emergence of more single cell multimodal datasets enable future studies to further reveal how genotypes translate to phenotypes and how ongoing mutational processes drive clonal diversification and evolution in cancer cells. In Chapter 4, we explored clone-specific gene expression profiles in HGSCs and characterized potential impact of metastasis and WGD on transcriptional phenotypes in cancer cells subclones. It remains an open question whether the CN-expression relation is consistent across tumors and whether application at scale can reveal phenotypic consequences of copy number alterations at subclonal resolution.

CN dosage effect is only one of mechanisms that genetic changes regulate gene transcription. Other genetic alterations such as point mutations and structural variations can also impact expression. Besides dosage effects, CN changes can also affect expression through other mechanisms. For instance, CN changes spanning promoter or enhancer regions can alter corresponding gene expression [120]. In addition to genetic changes, gene expression can also be affected by epigenetic changes including DNA methylation and histone modifications. For example, it was found that in ovarian cancers, the decreased copy number dosage effects of non-cancer genes coincided with increased DNA methylation level, suggesting methylation played a stronger role in regulating gene expression [106]. Finally, crosstalks between transcriptional pathways and interactions with the microenvironment can also affect expression regulations. Genes in immune related pathways are known to

have stronger transcriptional effects that offset dosage effects [49] and tend to be strongly regulated by immune cells in the tumor microenvironment. I think explicit probabilistic modeling of these processes, such as the dosage effect modeling in TreeAlign, will enhance our understanding of different mechanisms that influence cancer cell transcriptional phenotypes.

As TreeAlign also integrates allele-specific CN and expression, it would be interesting to investigate patterns of LOH and allele-specific expression on a subclone level as modulators of germline alterations and bi-allelic inactivation to better understand these events in the context of tumor heterogeneity and clonal evolution. With the emergence of more single-cell spatial transcriptomics data, it would also be interesting to explore clone-specific phenotypes and corresponding microenvironment together, and further dissect the origins of phenotypic divergence between clones. Previous research has established methods for inferring CNAs from spatial transcriptomics dataset without matching DNA data [121]. We expect that concepts introduced in TreeAlign will further facilitate the integration of single cell multimodal datasets and the interpretation of associations between modalities.

In conclusion, we anticipate that studying how copy number alterations impact gene expression programs in cancer applies broadly to different questions in cancer biology including etiology, tumor evolution, drug resistance and metastasis. In these settings, TreeAlign provides a flexible and scalable method for explaining gene expression with subclonal CNAs as a quantitative framework to arrive at mechanistic hypotheses from multimodal single cell data. Our approach provides a new tool to disentangle the relative contribution of fixed genomic alterations and other dynamic processes on gene expression programs in cancer.

Appendix A

Methods

A.1 TreeAlign total CN model

The TreeAlign model is a probabilistic graphical model as shown in **Fig. 2.1**. Here we describe the model in detail. Let X be a cell $\times$ gene expression matrix of raw counts from scRNA-seq for N cells and G genes, and x_{ng} be the scRNA read count for cell n and gene g. Let Λ be a gene $\times$ cell copy number matrix for G genes and C clones, and λ_{gc} be the copy number at gene g for clone c. To assign cells from the expression matrix to a clone in copy number matrix, we use a categorical variable z_n which indicates the clone to which a cell should be assigned. $z_n = c$ if cell n is assigned to clone c. z_n is drawn from a Categorical distribution with Direchlet prior.

$$z_{n=1...N} \sim Categorical(\pi) \tag{A.1}$$

$$\pi \sim Dir(\alpha) \tag{A.2}$$

To indicate whether the expression of a gene is dependent on the underlying CN, we introduced another indicator variable k_g. $k_g = 0$ if expression of gene g is not dependent on CN. $k_g = 1$ if expression of gene g is dependent on CN. k_g is a Bernoulli random variable with Beta prior.

$$k_{g=1...G} \sim Bernoulli(p(k_g)) \tag{A.3}$$

$$p(k_g) \sim Beta(\beta_1, \beta_2) \tag{A.4}$$

where we have $\beta_1 = 1, \beta_2 = 1$ as default.

Our assumption is that y_{ng}, the expected expression of gene g in cell n - will be proportional to the copy number of gene g in clone c to which cell n is assigned, if expression of gene g is dependent on copy number as indicated by k_g. Based on this assumption, our model is:

$$y_{ng} = E[x_{ng}|z_n = c] = l_n * \frac{[\mu_{g0} \times \lambda_{gc} \times k_g + \mu_{g1} \times (1 - k_g)] \times e^{\psi_n \cdot w_g^T}}{\sum_{g'=1}^{G} [\mu_{g'0} \times \lambda_{g'c} \times k_{g'} + \mu_{g'1} \times (1 - k_{g'})] \times e^{\psi_n \cdot w_{g'}^T}} \tag{A.5}$$

$$X_n = (x_{n1}, ..., x_{nG}) \tag{A.6}$$

$$Y_n = (y_{n1}, ..., y_{nG}) \tag{A.7}$$

$$X_{n=1...N} \sim Multinomial(l_n, Y_n) \tag{A.8}$$

where l_n is the total scRNA read count from cell n. Vector Y_n represents the expected read count for each gene in cell n. X_n is the actual read count from each gene in cell n we want to model. μ_{g0} is the per-copy expression of gene g if the expression is dependent on

copy number while μ_{g1} is the expression of gene g if its expression is independent of copy number. The intuition is that when $k_g = 1$, we expect the expression of g is proportional to its copy number; when $k_g = 0$, the expression of g is not dependent on the underlying copy number. We specified a softplus transformed Normal prior over the per-copy expression μ_{g0} and μ_{g1}.

$$\mu_{g0}, \mu_{g1} \sim \log(1 + e^{\mathcal{N}(\mu'_g, 10)}) \tag{A.9}$$

where we set μ'_g to the softplus inverse transformed mean read count of gene g across all cells.

The inner product $\psi_n \cdot w_g^T$ introduces noise into the model to avoid over-fitting. Their priors were set as described previously [79].

A.2 TreeAlign allele-specific model

To use allele specific copy number information for clone assignment, we set up a separate model, allele-specific TreeAlign which only takes in allele specific information. The input to allele-specific TreeAlign includes single cell level B allele frequencies at heterozygous SNPs estimated from scDNA data and read counts of reference allele and alternative allele of these SNPs from scRNA-data.

Let t_{ns} be the scRNA read count at a heterozygous SNP s in cell n, r_{ns} be the scRNA read count from the reference allele at heterozygous SNP s in cell n. Both t_{ns} and r_{ns} can be obtained by genotyping heterozygous SNPs of interests with tools such as cellsnp-lite [122].

With scDNA data, we can estimate b_{sc}, the BAF for SNP s and clone c using tools such as SIGNALS [21]. We assume that $f_{ns}(z_n = c)$, the expressed reference allele frequency at cell n and SNP s when cell n is assigned to clone c is controlled by the followings: 1). DNA BAF at that SNP of clone c and 2). whether the reference allele in scRNA data is B allele or not. We use a binary variable a_s to indicate whether the reference allele at SNP s should be assigned as B allele. a_s can be obtained using SIGNALS which can use information from scDNA to phase the SNPs in scRNA and assign alleles according. We can also treat a_s as a hidden variable and jointly infer it from the allele-specific model of TreeAlign. Comparing a_s inferred from TreeAlign to SIGNALS output allows us to estimate the performance of TreeAlign.

$$p(a_s) \sim Beta(\beta'_1, \beta'_2) \tag{A.10}$$

$$a_{s=1...S} \sim Bernoulli(p(a_s)) \tag{A.11}$$

$$f_{ns}(z_n = c) = a_s * b_{sc} + (1 - a_s) * (1 - b_{sc}) \tag{A.12}$$

$$r_{ns} \sim Binomial(t_{ns}, f_{ns}) \tag{A.13}$$

where we have $\beta'_1 = 1, \beta'_2 = 1$ as default.

The total CN model and allele-specific model share categorical variable z_n which indicates the clone assignment of cell n. Therefore, z_n can be inferred from the two models separately or combined depending on the input data provided. The integrated model is illustrated in **(Fig. 3.2)**. The prior distributions of all random variables are summarised in **(Fig. B.1)**.

A.3 Model implementation and inference

TreeAlign is implemented with Pyro [105] which is a universal probabilistic programming language written in Python and supported by PyTorch. Inference of TreeAlign is done by Pyro's Stochastic Variational Inference (SVI) functions automatically. Specifically, we use the *AutoDelta* function which implements the delta method variational inference [123]. The delta method variational inferences use a Taylor approximation around the maximum a posterior (MAP) to approximate the posterior. Optimization is performed using the Adam optimizer. By default, we set a learning rate of 0.1 and the convergence is determined when the relative change in ELBO is lower than 10^{-5} by default.

A.4 Incorporating phylogeny as input

In addition to the gene $\times$ clone copy number matrix, TreeAlign can also take the cell *times* gene copy number matrix from scDNA directly along with the phylogenetic tree constructed from this matrix as input. Starting from the root of the phylogeny, TreeAlign summarizes the copy number of gene g for each clade by taking the mode of copy number, and assigns cells from scRNA to clade-level CN profiles. This process is repeated recursively from the root of the phylogeny to smaller clades until: i) TreeAlign can no longer assign cells consistently in multiple runs (less than 70% cells have consistent assignments between runs by default), or ii) the number of genes located in CSCN regions becomes too small (100 genes in CSCN regions by default), or iii) Limited number of cells remain in scDNA or scRNA (100 by default). By default, TreeAlign also ignores subclades with less than 20 cells in scDNA. Some scRNA cells may remain unassigned to the scDNA phylogenetic tree. For a single cell, if the clone assignment probability < 0.8 or clone assignments

are not consistent in 70% of repeated runs, the cell will be denoted as unassigned. This feature is important to the model because there might be incomplete sampling of a given tumor, leading to a subclone only appearing in one of the two data modalities. Note, all parameters are fully configurable at run time by the user.

A.5 Benchmarking clone assignment and dosage effect prediction with simulations

Simulations were performed similarly as described previously. CloneAlign v.2.0 model was fit to the MSK-SPECTRUM patient 081 dataset to obtain the empirical estimations of model parameters. Then we simulated from CloneAlign considering the following scenarios: 1. Varying proportion (10%, 20%, 30%, ..., 90%) of genes with dosage effect. 2. Varying number of genes (100, 500 and 1000) in CSCN regions. 3. Varying number of cells (100, 1000 and 5000) in scRNA.

We compared TreeAlign to CloneAlign and InferCNV v.1.3.5 in terms of the performance of clone assignment. For CloneAlign, we summarized clone-level copy number by calculating the mode of copy number for each gene and ran CloneAlign with default parameters. For InferCNV, we used the recommended setting for 10x. 3,200 non-cancer cells were randomly sampled from the SPECTRUM dataset and used as the set of reference "normal" cells. To assign clones with InferCNV, we calculated Pearson correlation coefficient between InferCNV corrected gene expression profile (expr.infercnv.dat) and the clone-level copy number profiles from scDNA. Cells from scRNA-seq were assigned to the clone according to the highest correlation coefficient. Accuracy of clone assignment was computed

to compare the performance of the three methods. We also evaluated the TreeAlign's performance on predicting CN dosage effects.

To evaluate TreeAlign's performance on predicting CN dosage effects, we calculated the area under the curve (AUC) using $p(k)$ output by TreeAlign, and compared it to a baseline model. The baseline model for CN dosage effects was constructed by 1). assigning expression profiles to genomic clones using CloneAlign 2). calculating Pearson correlation coefficients (R) between normalized read count from scRNA and clone-specific CN from scDNA for each gene in input. The resulting R can be viewed as a metric for CN dosage effects. We calculated the baseline model AUC using R and compared it to TreeAlign model.

To demonstrate the performance of allele-specific TreeAlign, for the simulated datasets with 30% CN-dependent genes, we also simulated reference allele and total read counts for varying number of heterozygous SNPs (0, 250, 500, 75, 1000 and 1250) from the generative model of allele-specific TreeAlign. Adjusted rand index of clone assignments was calculated to evaluate the performance of the integrated TreeAlign model on simulated datasets with varying numbers of heterozygous SNPs.

To evaluate TreeAlign's performance inaccurate trees, we randomly shuffled labels for different proportions (10%, 20%,...,90%) of cells on the phylogeny of patient 22. TreeAlign was run with the shuffled phylogenies. Clone assignment results were compared to results obtained from the original phylogeny using adjusted rand index.

A.6 MSK SPECTRUM data

We obtained matched scRNA and scDNA from two HGSC patients (patient 022 and patient 081) from the MSK SPECTRUM cohort [20]. Samples were collected under Memorial Sloan Kettering Cancer Center's institutional IRB protocol 15-200 and 06-107. Single cell suspensions from surgically excised tissues were generated and flow sorted on CD45 to separate the immune component as previously described. CD45 negative fractions were then sequenced using the DLP+ platform as previously described [3, 21, 86].

A.7 Gastric cancer cell line data

Preprocessed scDNA data and scRNA count matrix of the gastric cancer cell line (NCI-N87) [57] were downloaded from SRA (PRJNA498809) and GEO (GSE142750). Copy number calling for scDNA were performed using the Cellranger-DNA pipeline using default parameters.

A.8 PDXs and additional cell line data

scRNA and scDNA from 6 HGSC PDX samples (SA1052BX1XB01516, SA1052JX1XB01535, SA1053BX1XB01603, SA1091AX1XB01790, SA1093CX1XB01917, SA1181AX1XB02700), 3 TNBC PDX samples (SA1035X6XB03216, SA1035X7XB03502, SA610X3XB03802), 1 ovarian cancer cell line (OV2295) and 6 hTERT-184 cell lines (SA039, SA1054, SA1055, SA1188, SA906a, SA906b) were obtained and processed as described previously [21].

A.9 scDNA data analysis

scDNA DLP+ data was processed as previously described [21, 86]. Cells with quality score > 0.75 and not in S-phase were retained for downstream analysis. Allele specific copy number was called using SIGNALS, which provides allele specific copy number of the from A|B in 500kb bins across the genome. A and B being the copy number of alleles A and B respectively with $total\ CN = A + B$. As the single cell data is sparse, only a subset of germline SNPs have coverage in each cell, therefore to produce the input required for TreeAlign (B-Allele frequencies per SNP per cell), we impute the BAF of each SNP assuming that a SNP will have the same BAF as the bin in which the SNP resides.

A.10 Clustering and phylogenetic inference

Clustering and phylogenetic inference of scDNA was performed using UMAP and HDBSCAN (parameters min_samples=20, min_cluster_size=30, cluster_selection_epsilon=0.2). For patient 022, we also constructed phylogenetic trees using Sitka38 as previously described.

A.11 Genotyping SNPs in scRNAseq cells

SNPs identified in scDNA-seq and matched bulk whole genome sequencing were genotyped in each single cell using cell-snplite with default parameters.

A.12 scRNA data analysis

scRNA data were processed as previously described7. Read alignment and barcode filtering were performed by CellRanger v.3.1.0. Cancer cell identification was performed with CellAssign. Principal-component analysis (PCA) was performed on the top 2000 highly variable features output by function FindVariableFeatures using Seurat v.4.2 [124]. UMAP embeddings and visualization were generated using the first 20 principal components. Unsupervised clustering was performed using FindNeighbors function followed by FindClusters function (resolution=0.2). To compare transcriptional heterogeneity across or within clones, we randomly sampled 100 expression profiles from the following groups: 1. all cancer cells in a patient/cell line/PDX 2. cancer cells in the same TreeAlign clone 3. cancer cells in the same InferCNV clone. Pearson correlation coefficients and Euclidean distance between the sampled expression profiles were calculated using the top 20 principal components.

A.13 Differential expression and gene set enrichment analysis

Differential expression analysis was performed using FindAllMarkers and FindMarkers function (test.use="MAST", latent.vars=c("nCount_RNA", "nFeature_RNA")) in Seurat v.4.2. Only G1 cells were used in differential expression analysis to avoid confounding of cycling cells. Cell cycle phase was annotated with CellCycleScoring function in Seurat.

We used the fgsea v.1.24.0 [125] package to conduct gene set enrichment analysis with

Hallmark gene sets (n=50) downloaded from MSigDB[126]. We set the following parameters for the gene set enrichment analysis: nperm=1000, minSize=15, maxSize=500.

A.14 Statistical analysis and visualization

Statistical tests and visualization were performed with R (v.4.2) package ggpubr (v.0.5.0) and ggplot2 (v.3.4).

A.15 Data Availability

Processed data containing input and output of TreeAlign have been deposited in Zenodo (https://doi.org/10.5281/zenodo.7517412). Raw scDNA data and scRNA count matrix of the gastric cancer cell line (NCI-N87) can be accessed from SRA (PRJNA498809) and GEO (GSE142750). Raw scDNA and scRNA data from Funnell et al. are available at https://ega-archive.org/studies/EGAS00001006343. Raw scRNA data for patient 022 and patient 081 are available at https://www.synapse.org/msk_spectrum.

A.16 Code Availability

The code is publicly accessible on a GitHub repository (https://github.com/shahcompbio/TreeAlign), which implements TreeAlign and describes how to generate simulated datasets.

Appendix B

Supplementary Figures

Variable	Distribution	Description
x_{ng}	Multinomial	Gene expression read count
y_{ng}	Deterministic f^n	Modeled expected expression
z_n	Categorical	Clone assignment indicator
π_c	Dirichlet	Prior probability of clone assignment
λ_{gc}		Copy number
μ_g	Softplus-Normal	Per-copy expression
k_g	Bernoulli	Copy number dependency indicator
$p(k)_g$	Beta	Prior probability of CN dependency
$\psi_n \cdot w_g^T$		Structured noise to avoid overfitting
t_{ns}		Total read count at SNPs in scRNA-seq
r_{ns}	Binomial	Reference allele count at SNPs in scRNA-seq
f_{ns}	Deterministic f^n	Reference allele frequency at SNPs in scRNA-seq
b_{sc}		B allele frequency at SNPs in scDNA-seq
a_s	Bernoulli	Allele assignment indicator
$p(a)_s$	Beta	Prior probability for allele assignment indicator

FIGURE B.1: Descriptions and prior distributions of random variables and data in TreeAlign model.

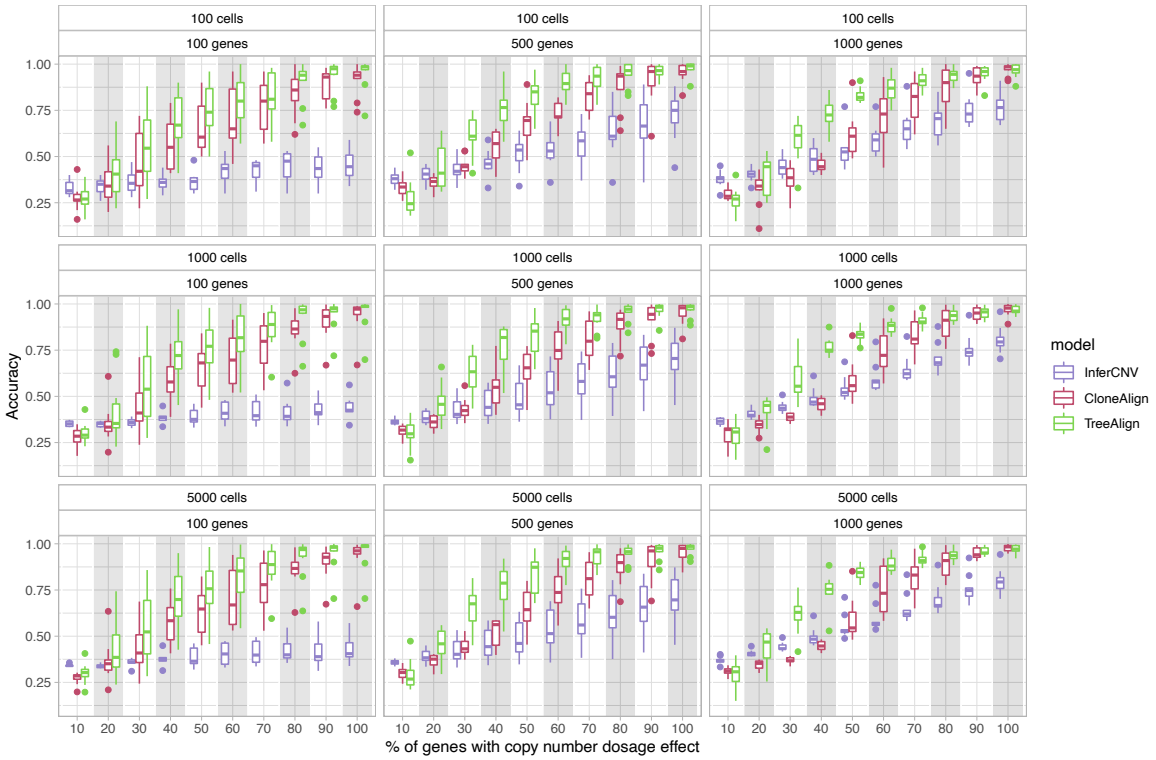

FIGURE B.2: Accuracy of clone assignment for TreeAlign, CloneAlign and InferCNV in simulated scRNA datasets as a function of varying proportions of genes with CN dosage effects. Panels represent datasets with different numbers of cells and genes.

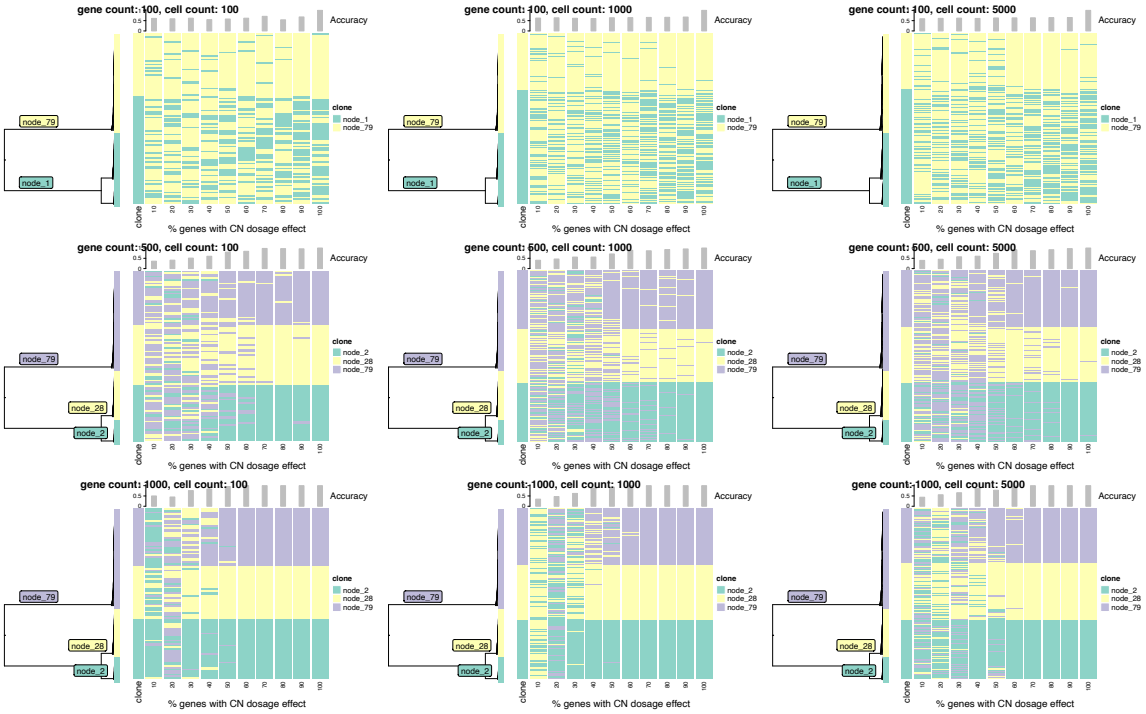

FIGURE B.3: Phylogenetic trees (left) constructed with scDNA-data from SPECTRUM-OV-081 along with Heat maps (right) showing clone assignment of simulated datasets by TreeAlign.

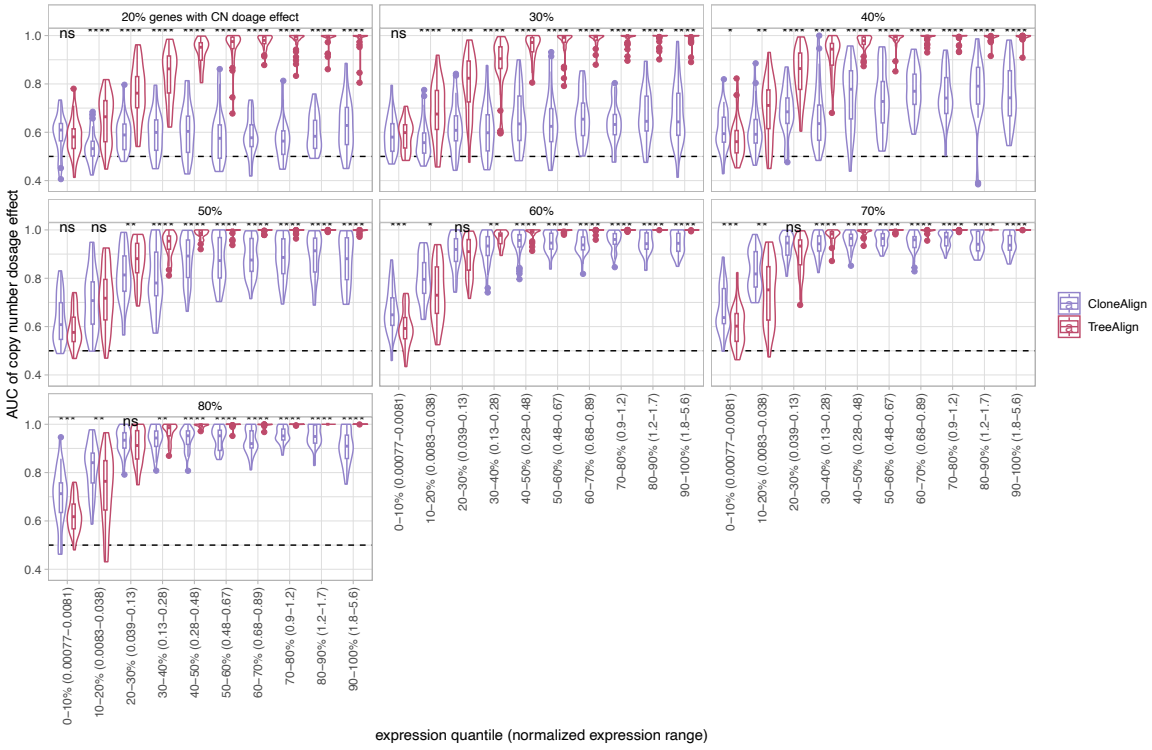

FIGURE B.4: *AUC* of CN dosage effect $p(k)$ predicted by CloneAlign and TreeAlign as a function of gene expression level. Genes were assigned to 10 bins based on expression level. Ranges of normalized expression for each bin were shown in brackets. Panels represent simulated datasets with varying gene dosage effect frequencies.

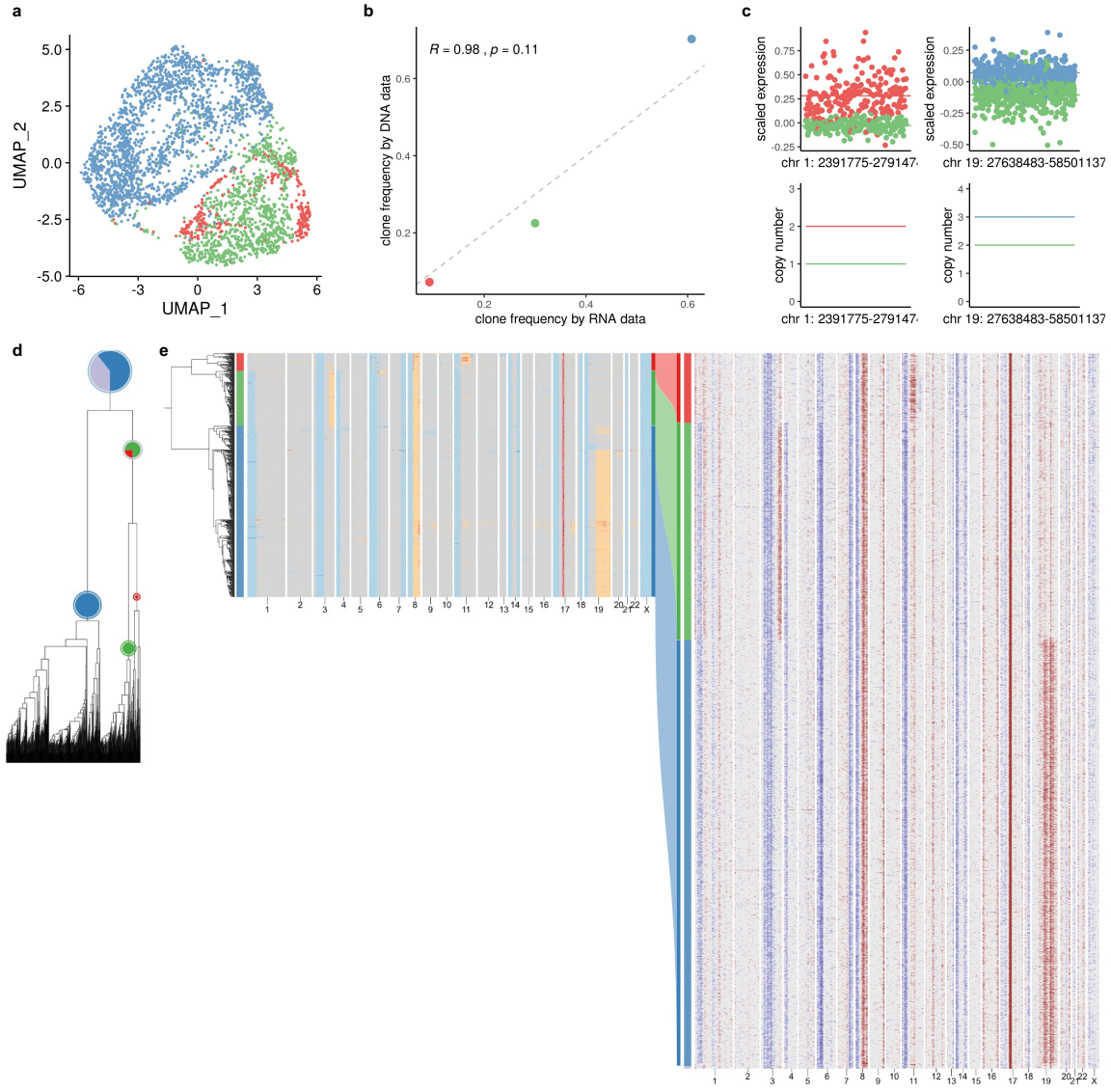

FIGURE B.5: **a,** UMAP plot of scRNA-data from gastric cell line NCI-N87 colored by clone labels assigned by total CN TreeAlign. **b,** Clone frequencies of NCI-N87 estimated by scRNA-data (x axis) and scDNA-data (y axis). **c,** Scaled expression and copy number profiles for regions on chromosome 1 and 19 as a function of genes ordered by genomic locations. d, Phylogenetic tree constructed with scDNA-data. **e,** Phylogenetic tree constructed with scDNA-data along with pie charts showing how TreeAlign assigns cell expression profiles to subtrees recursively. The pie charts are colored by the proportions of cell expression profiles assigned to downstream subtrees. The outer ring color of the pie charts indicates the current subtree. Heat maps of copy number profiles from scDNA (left) and InferCNV corrected expression profiles from scRNA (right). The Sankey chart in the middle shows clone assignment from expression profiles to copy number based clones by total CN TreeAlign.

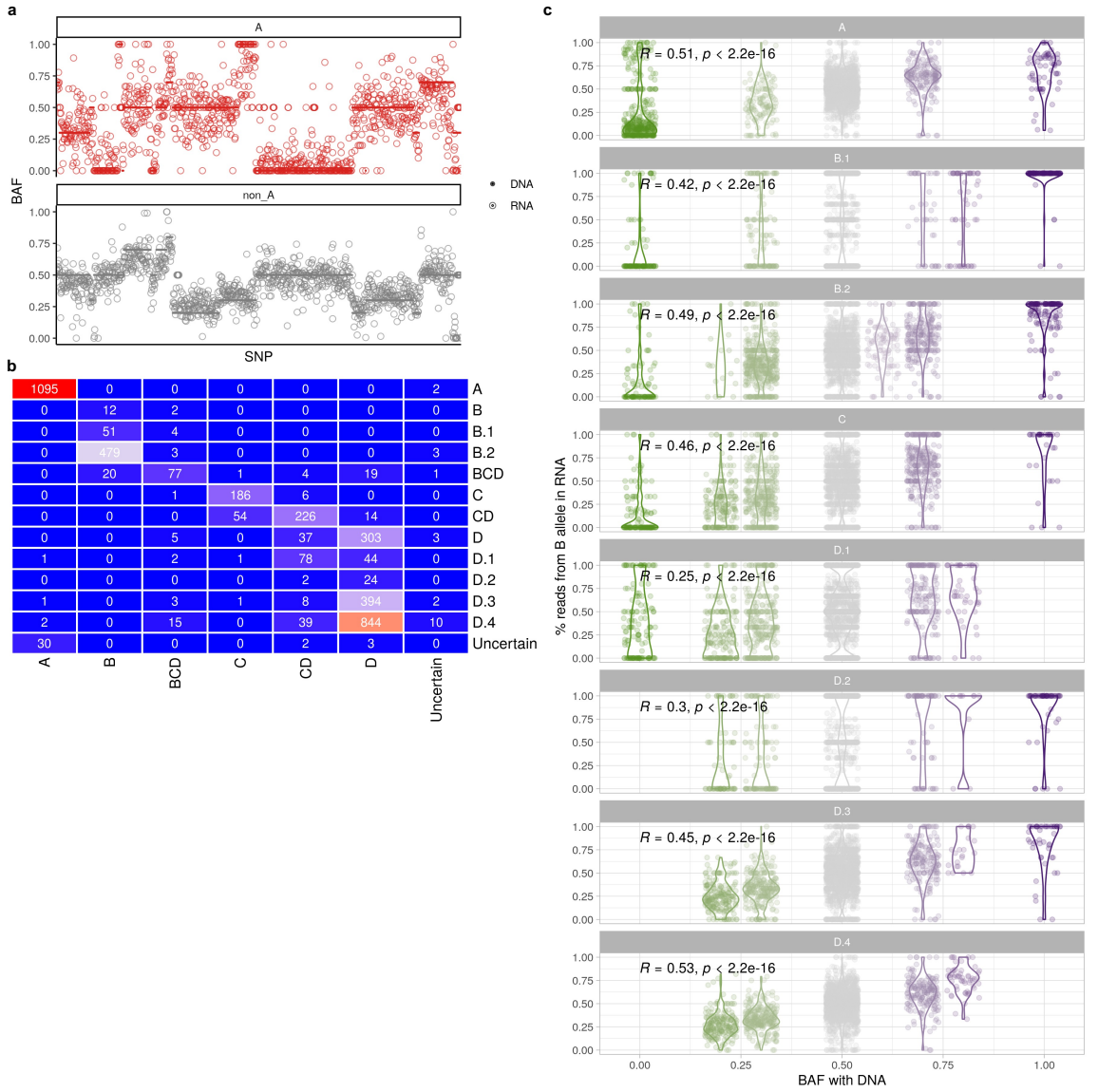

FIGURE B.6: **a,** BAF of heterozygous SNPs estimated from scRNA-data and scDNA-data for clone A and other clones (clone B - C) in patient 022 (ordered by gene location along chromosome). **b,** violin plot of BAF in SPECTRUM-OV-022 (Wilcoxon signed-rank test). **b,** Confusion matrix comparing clone assignment between total CN TreeAlign and integrated TreeAlign for patient 022. **c,** Correlation between proportions of reads from B allele in scRNA and BAF estimated from scDNA in patient 022 subclones (Wilcoxon signed-rank test).

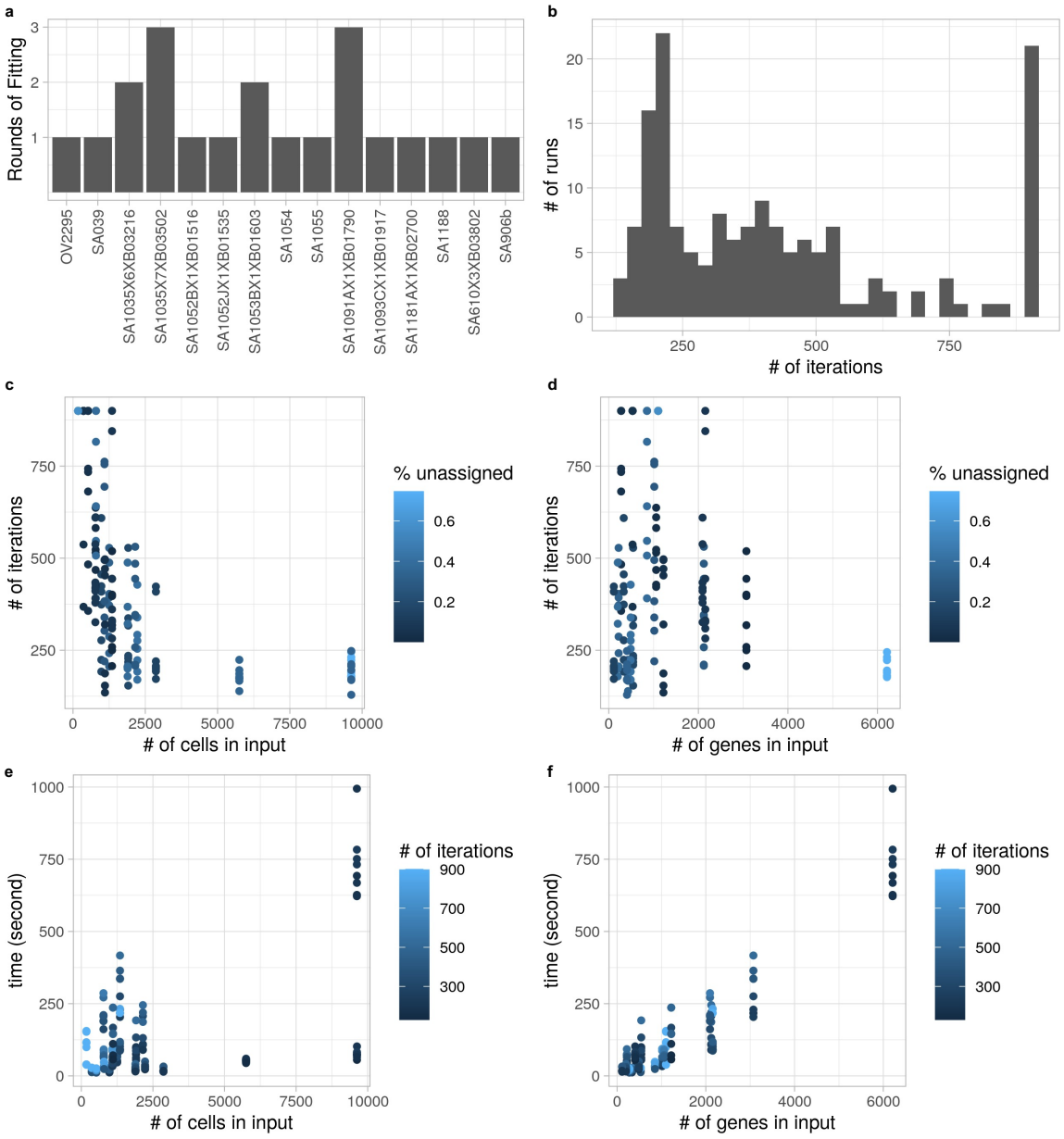

FIGURE B.7: **a,** times of fitting the total CN model with phylogeny input. **b,** Distribution of iterations for each inference run at convergence or before the maximum iteration of 900. **c,** Scatter plot showing the number of iterations and the number of cells in scRNA input for each run colored by frequencies of unassigned cells. **d,** Scatter plot showing the number of iterations and the number of genes in scRNA input for each run. **e,** Scatter plot showing the time to finish for each run as a function of the number of cells in scRNA input. **f,** Scatter plot showing the time to finish for each run as a function of the number of genes in scRNA input.

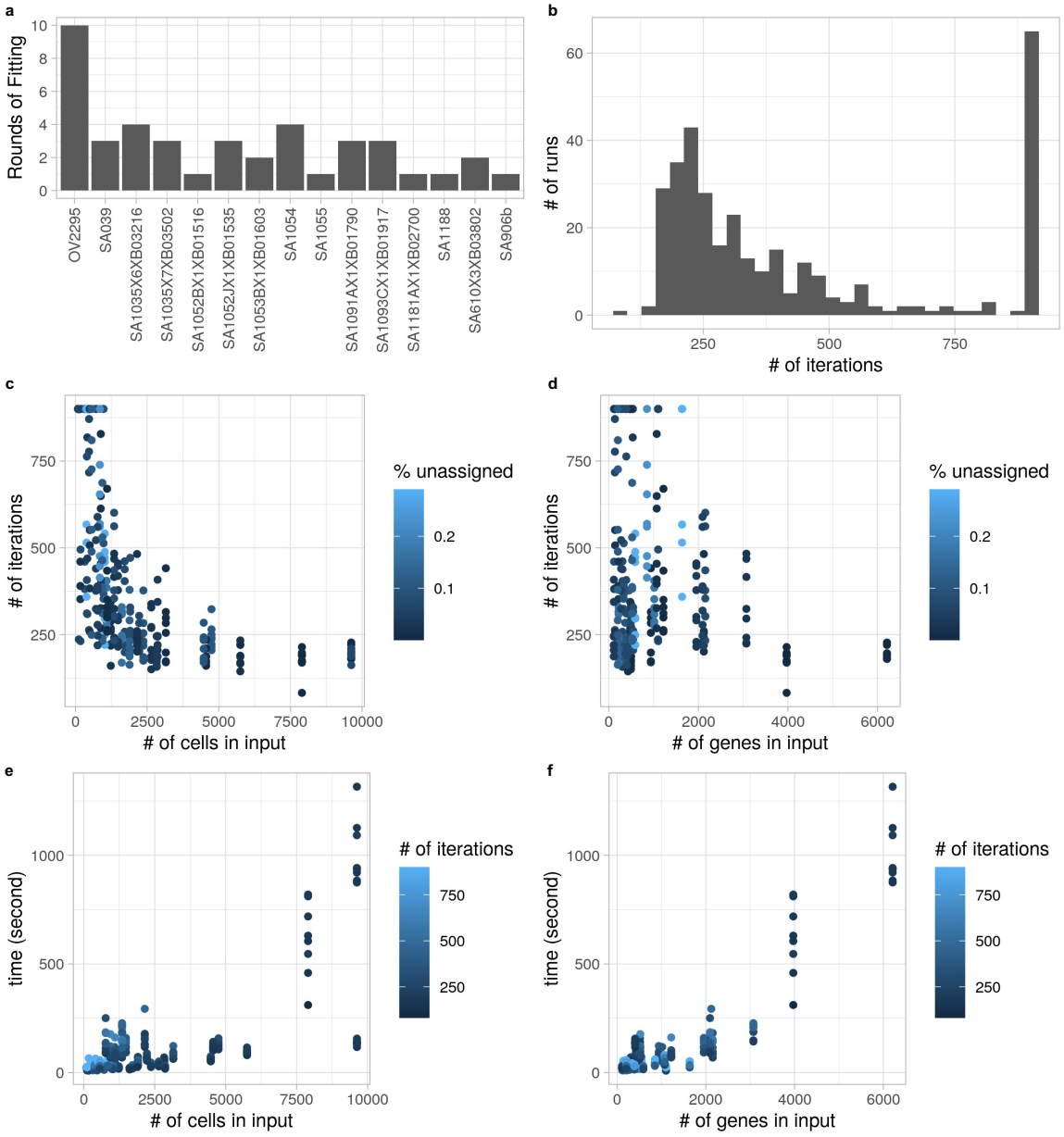

FIGURE B.8: **a,** rounds of fitting the integrated model with phylogeny input. **b,** Distribution of iterations for each inference run at convergence or before the maximum iteration of 900. **c,** Scatter plot showing the number of iterations and the number of cells in scRNA input for each run colored by frequencies of unassigned cells. **d,** Scatter plot showing the number of iterations and the number of genes in scRNA input for each run. **e,** Scatter plot showing the time to finish for each run as a function of the number of cells in scRNA input. **f,** Scatter plot showing the time to finish for each run as a function of the number of genes in scRNA input.

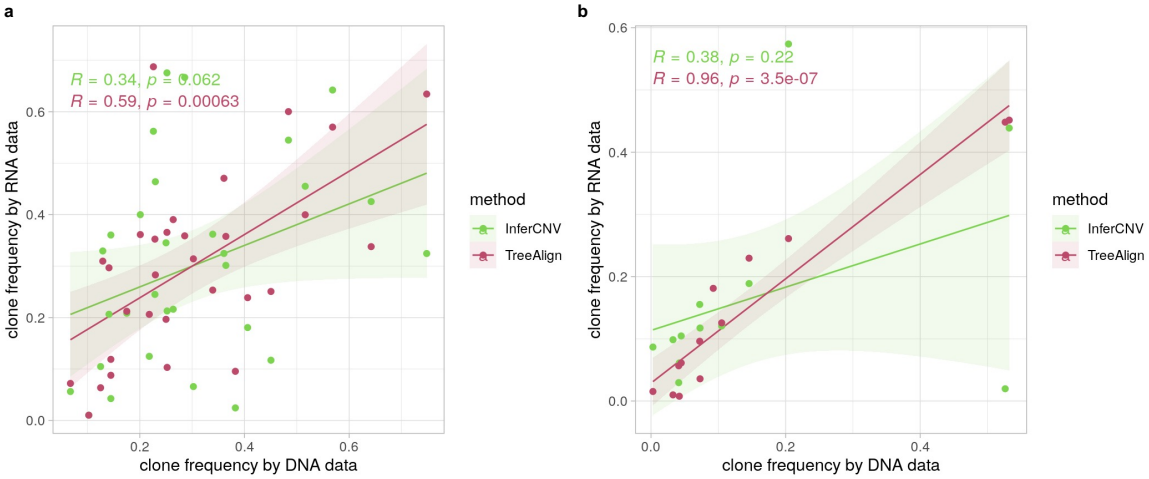

FIGURE B.9: **a-b,** Correlation between clone frequencies estimated by scRNA-data (x axis) and scDNA-data (y axis) by TreeAlign and InferCNV in **(a)** HSGC PDXs and cell lines and **(b)** patient 022.

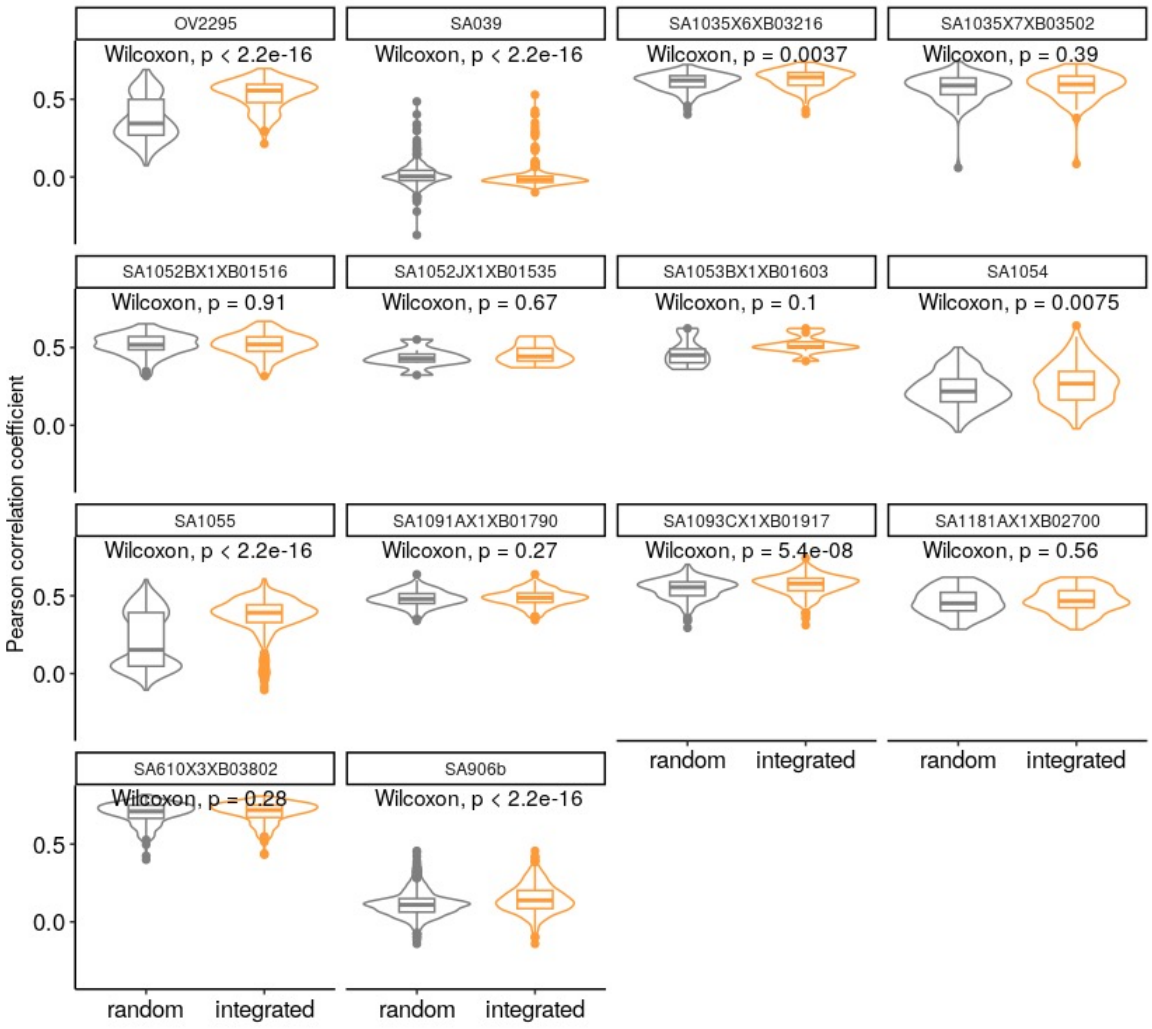

FIGURE B.10: Distribution of Pearson correlation coefficients (R) between scDNA estimated total copy number and InferCNV corrected expression for unassigned cells from total CN model. Left, correlation distribution calculated by comparing InferCNV profiles to CN profiles of a random subclone; Right, correlation distribution calculated by comparing InferCNV profiles to CN profiles of subclones assigned by integrated TreeAlign. Each panel represents results from a tumor sample/cell line.

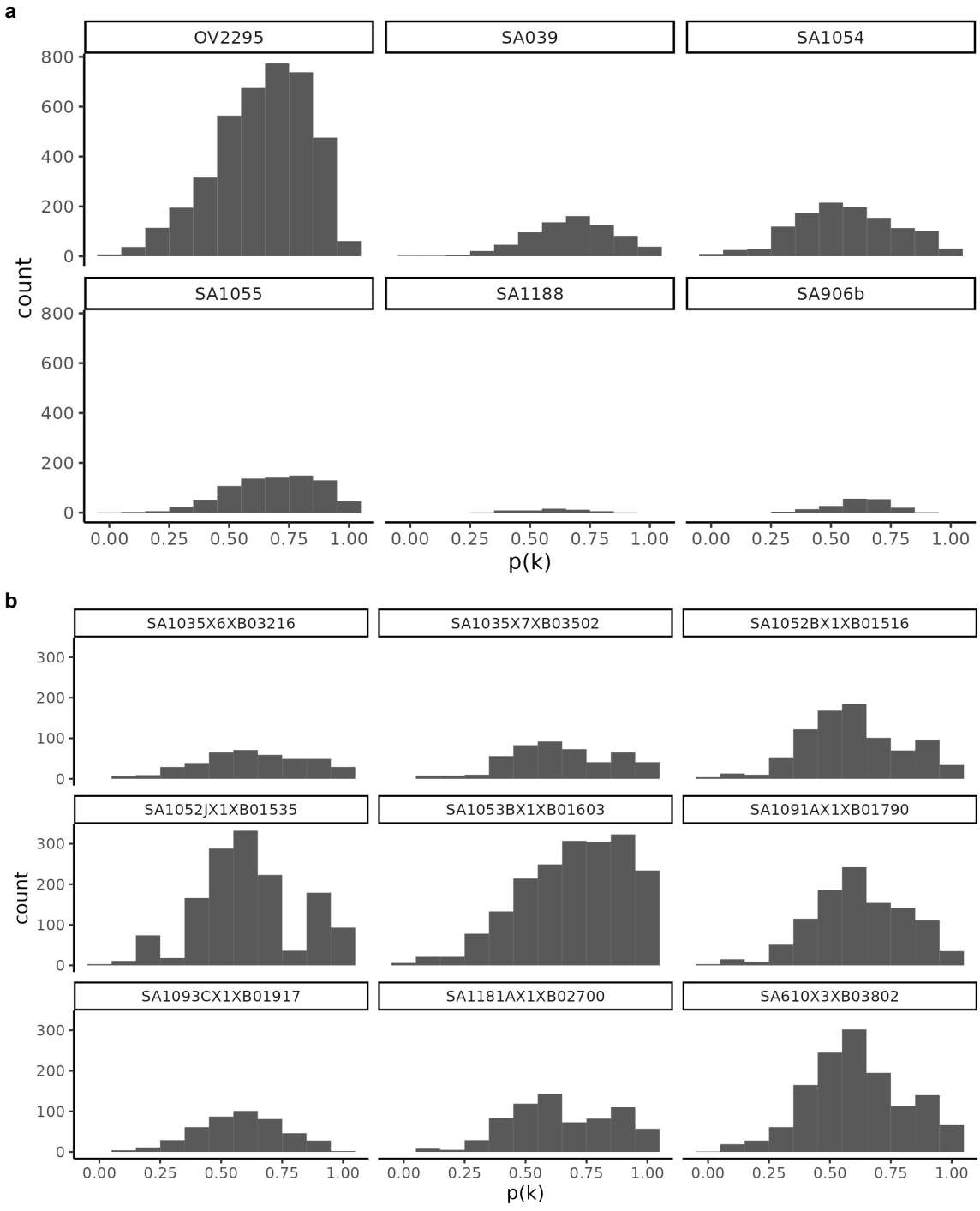

FIGURE B.11: **a,** Distribution of $p(k)$ in hTERT-184 and control cell lines. **b,** Distribution of $p(k)$ in PDXs.

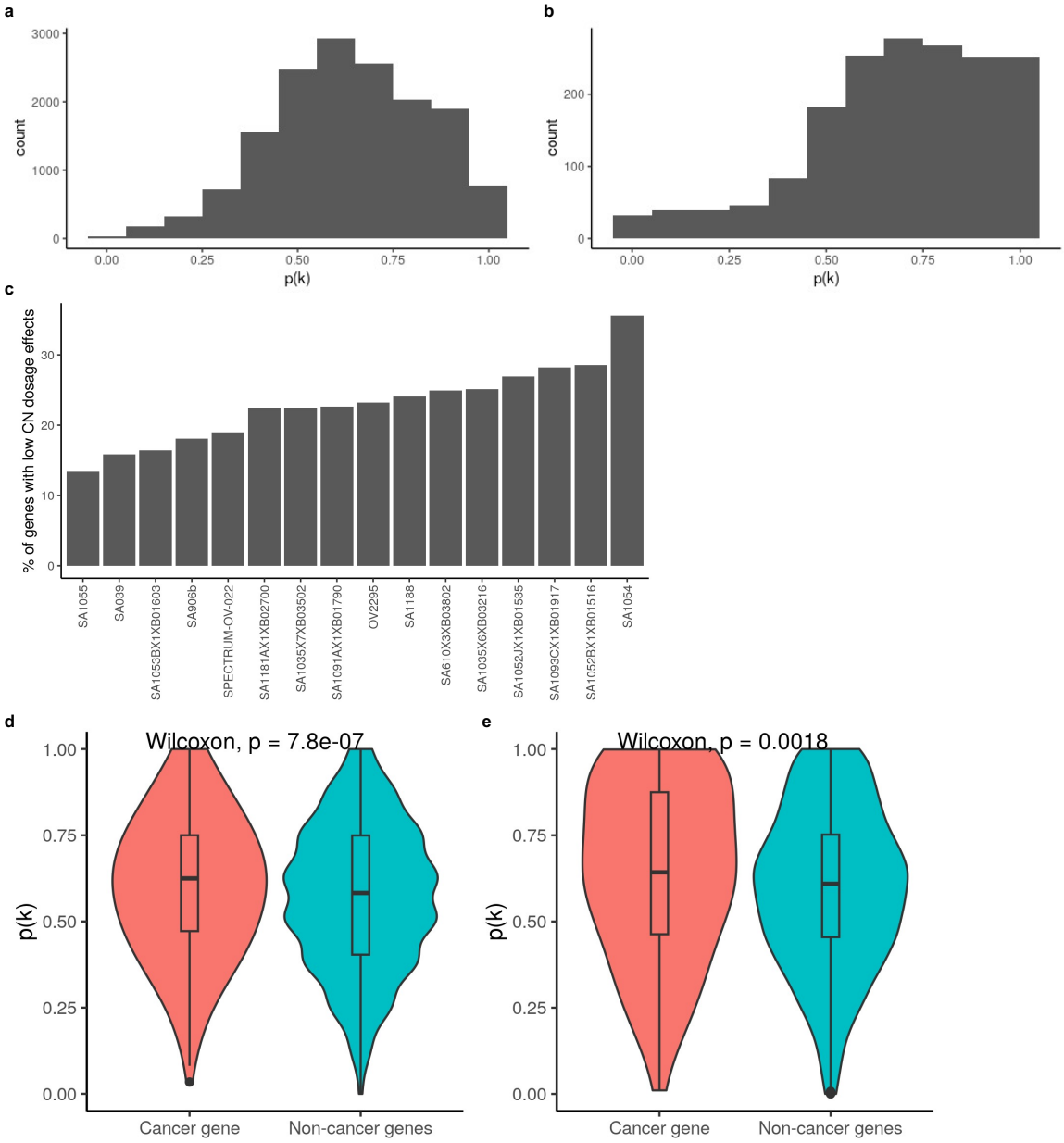

FIGURE B.12: **a,** Distribution of $p(k)$ in PDXs and cell lines. **b,** Distribution of $p(k)$ in patient 022. **c,** Proportions of genes with low CN dosage effects ($p(k) < 0.5$) in PDXs and cell lines. **d-e,** $p(k)$ for cancer genes and non-cancer genes in **(d)** PDXs and cell lines and **(e)** patient 022.

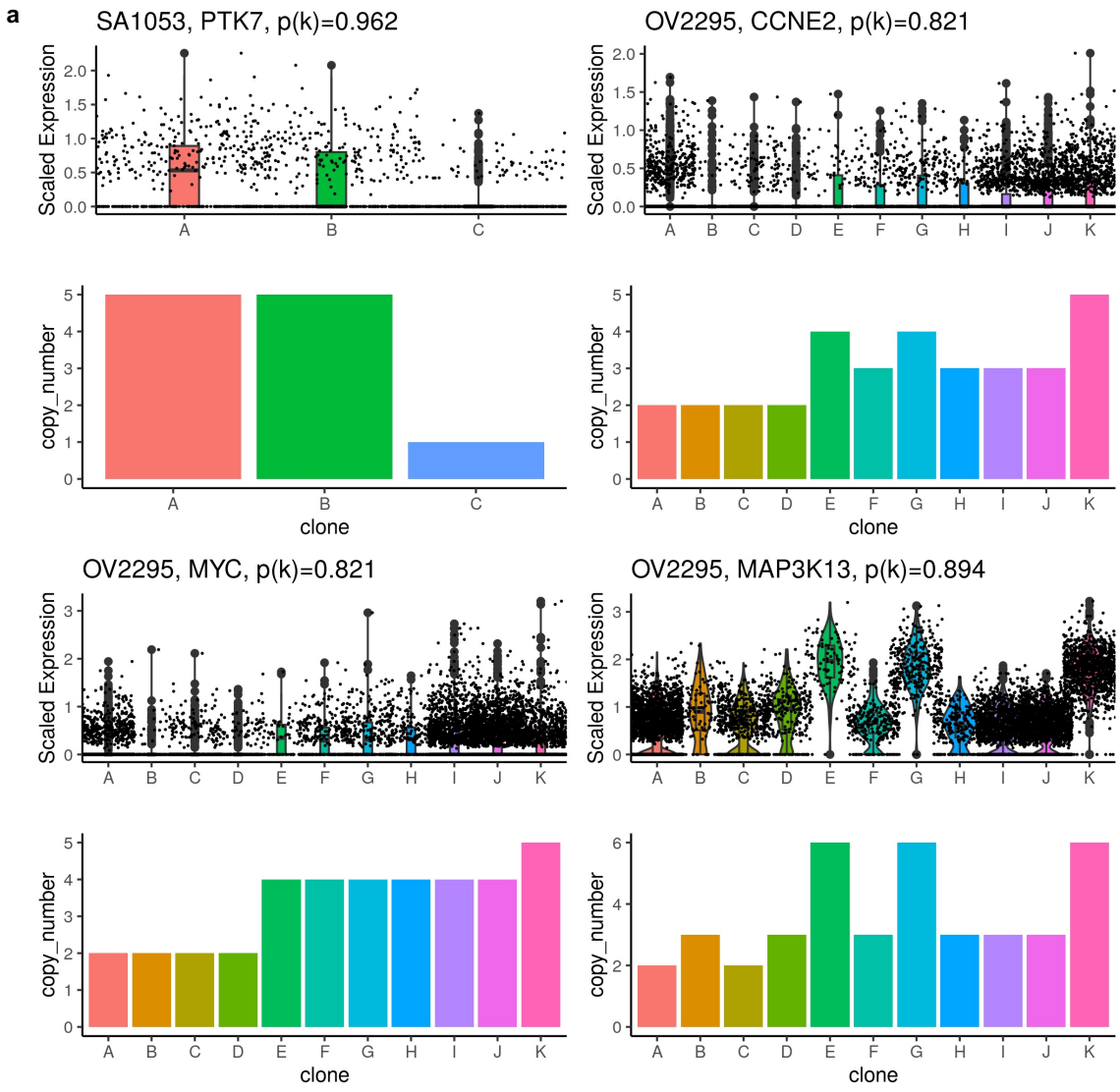

FIGURE B.13: **a,** Examples of genes with high level amplifications and high CN dosage effects.

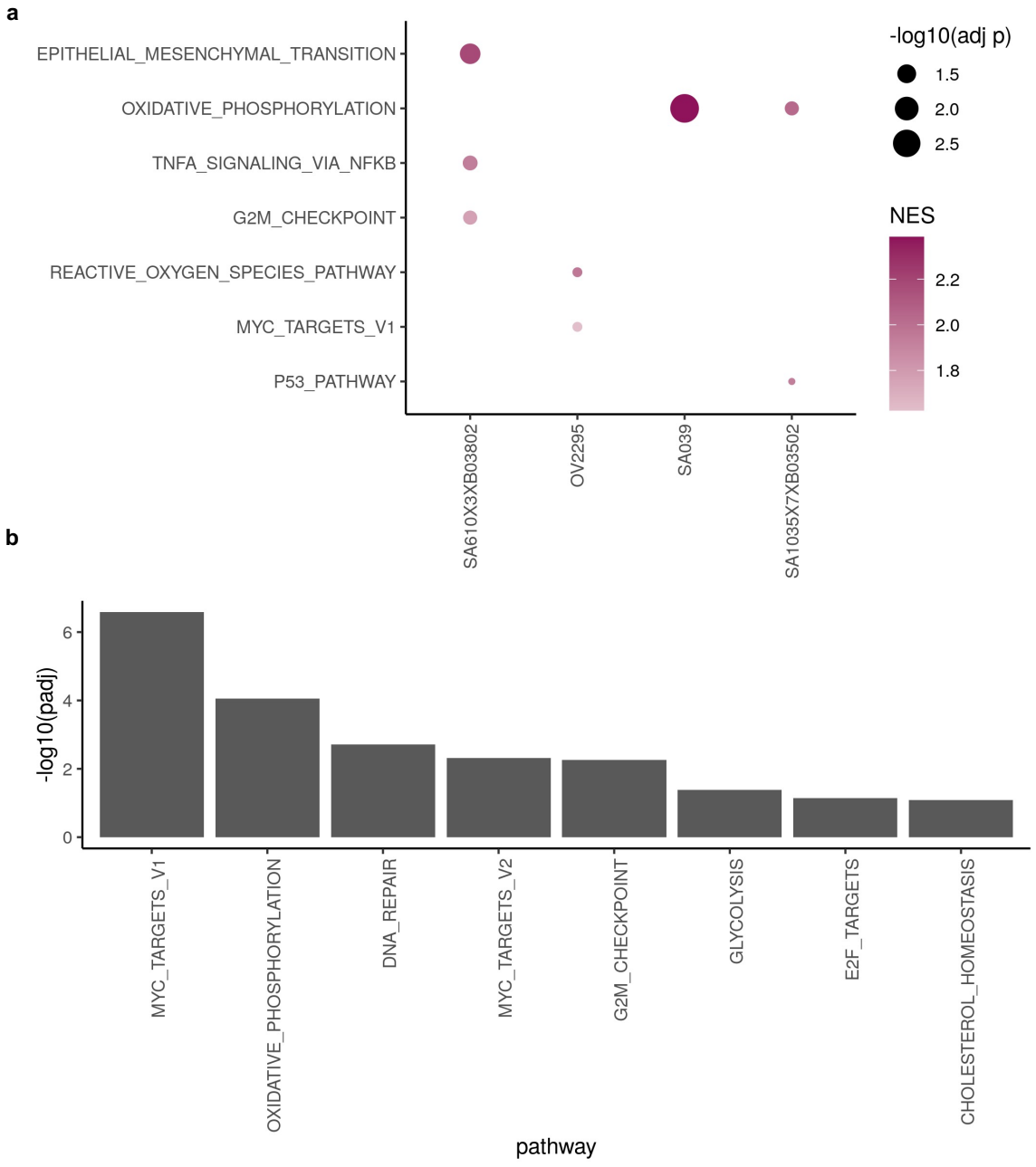

FIGURE B.14: **a,** Dot plot showing significantly enriched pathways in low p(k) genes. **b,** Significantly enriched pathways in low $p(k)$ genes from all PDXs and cell lines. $p(k)$ from all samples were combined before performing gene set enrichment analysis.

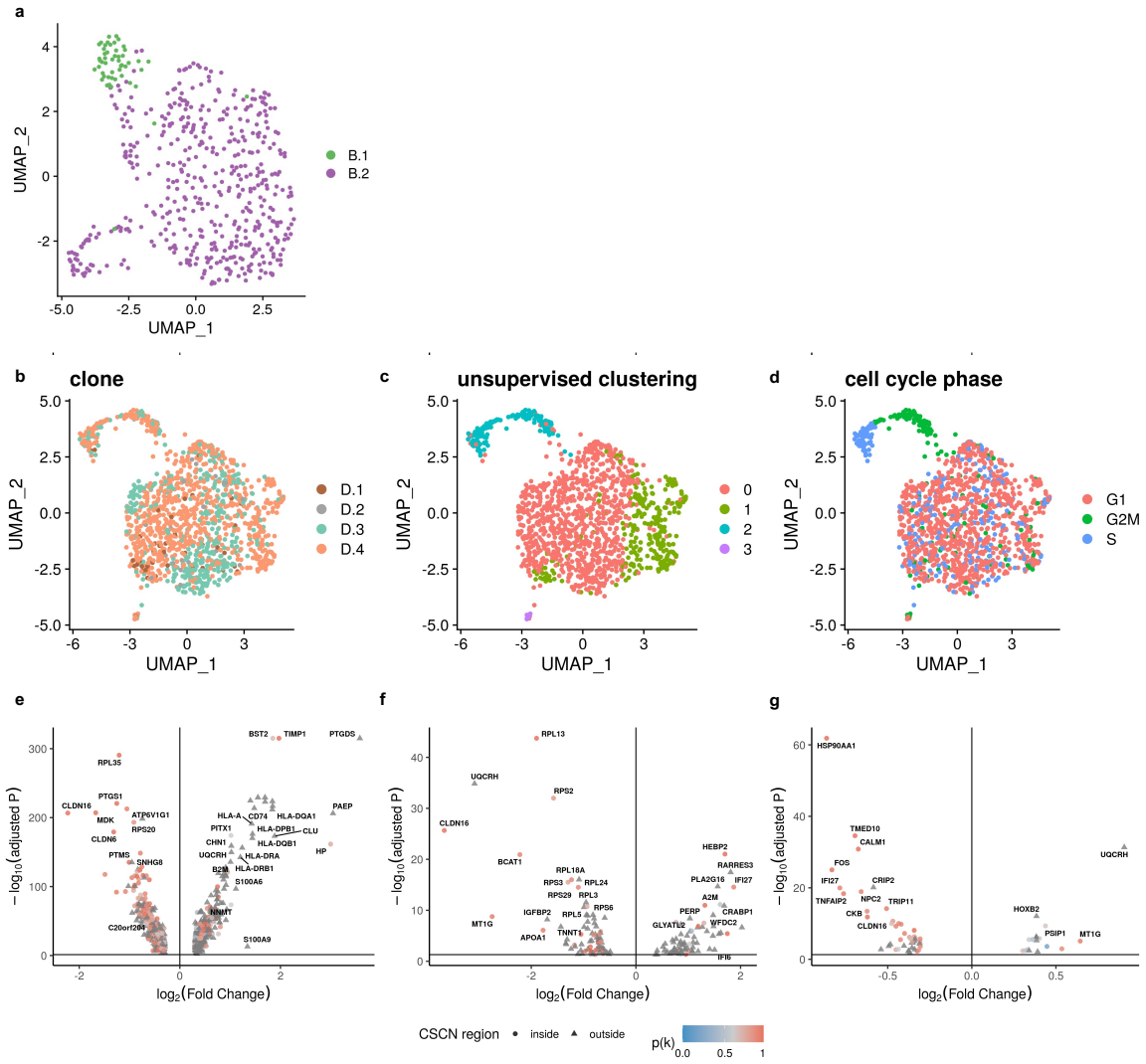

FIGURE B.15: **a,** UMAP plot of expression profiles of clone B.1 and B.2 in patient 022. **b,** UMAP plot of expression profiles of clone D.1, D.2, D.3 and D.4 in patient 022 colored by clone assignments. **c,** UMAP plot of expression profiles of clone D in patient 022 colored by Louvain unsupervised clustering. **d,** UMAP plot of expression profiles of clone D in patient 022 colored by cell cycle phase. **e,** Differentially expressed genes between clone A and clone B-D. **f,** Differentially expressed genes between cells in clone B.1 and B.2. **g,** Differentially expressed genes between cells in clone D.4 and D.1 - D.3.

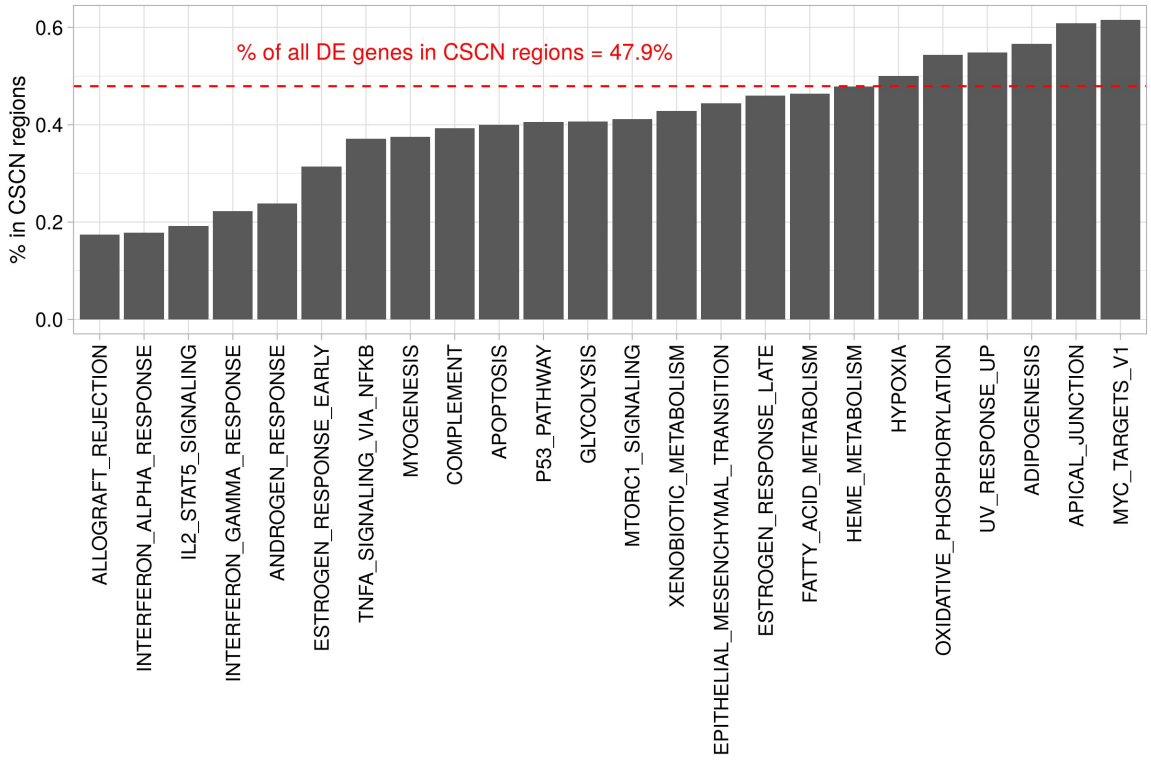

FIGURE B.16: Frequencies of DE genes in CSCN regions summarized by Hallmark pathways

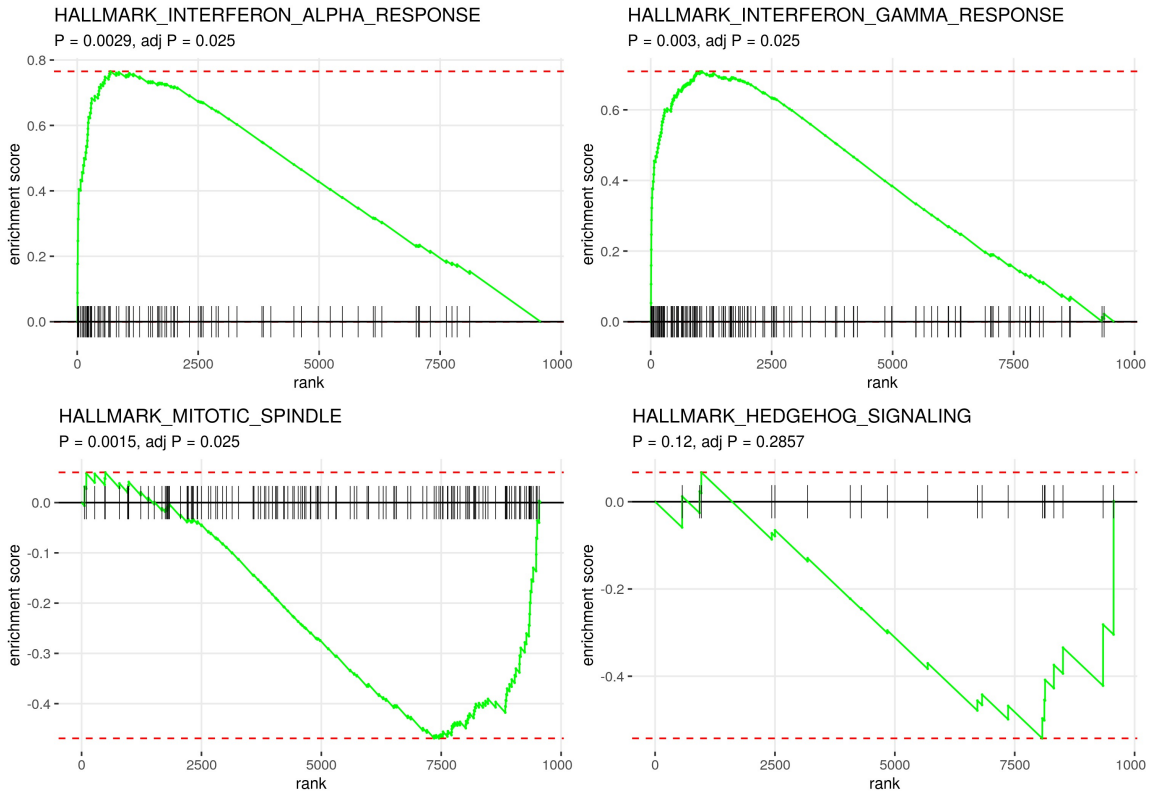

FIGURE B.17: Enriched and depleted pathways in clone A compared to other clones in patient 022

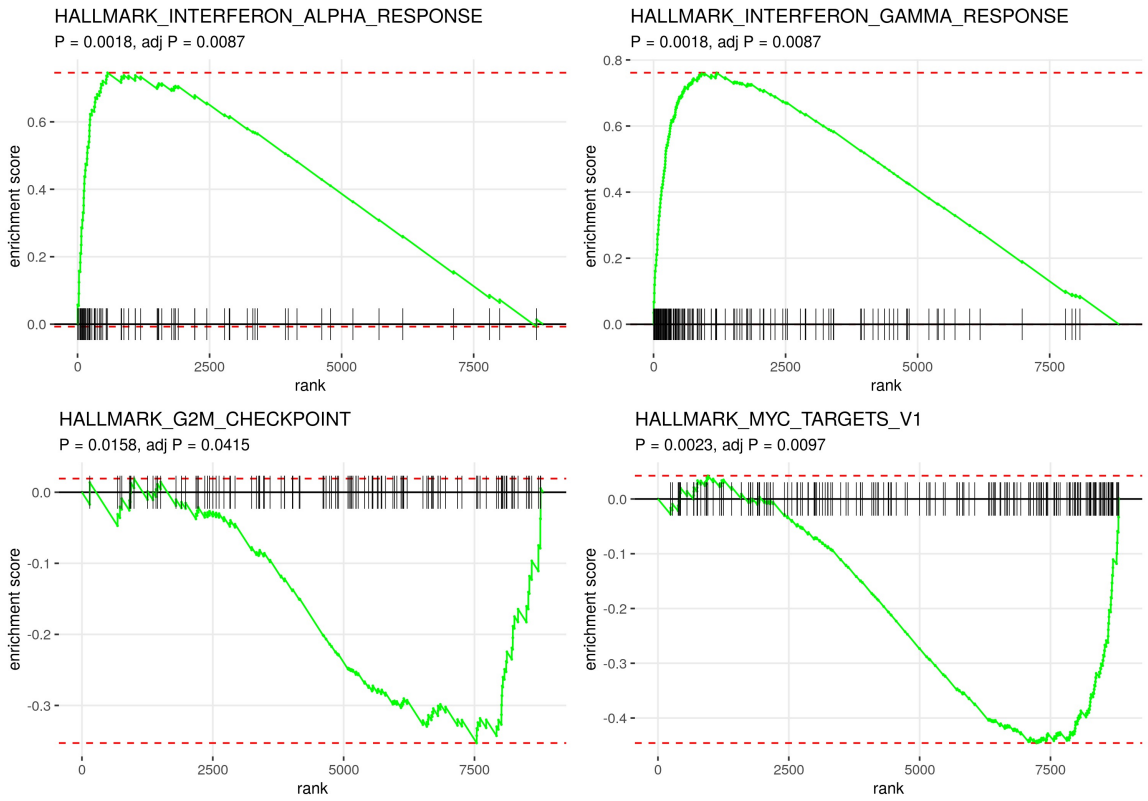

FIGURE B.18: Enriched and depleted pathways in clone B.1 compared to clone B.2 in patient 022

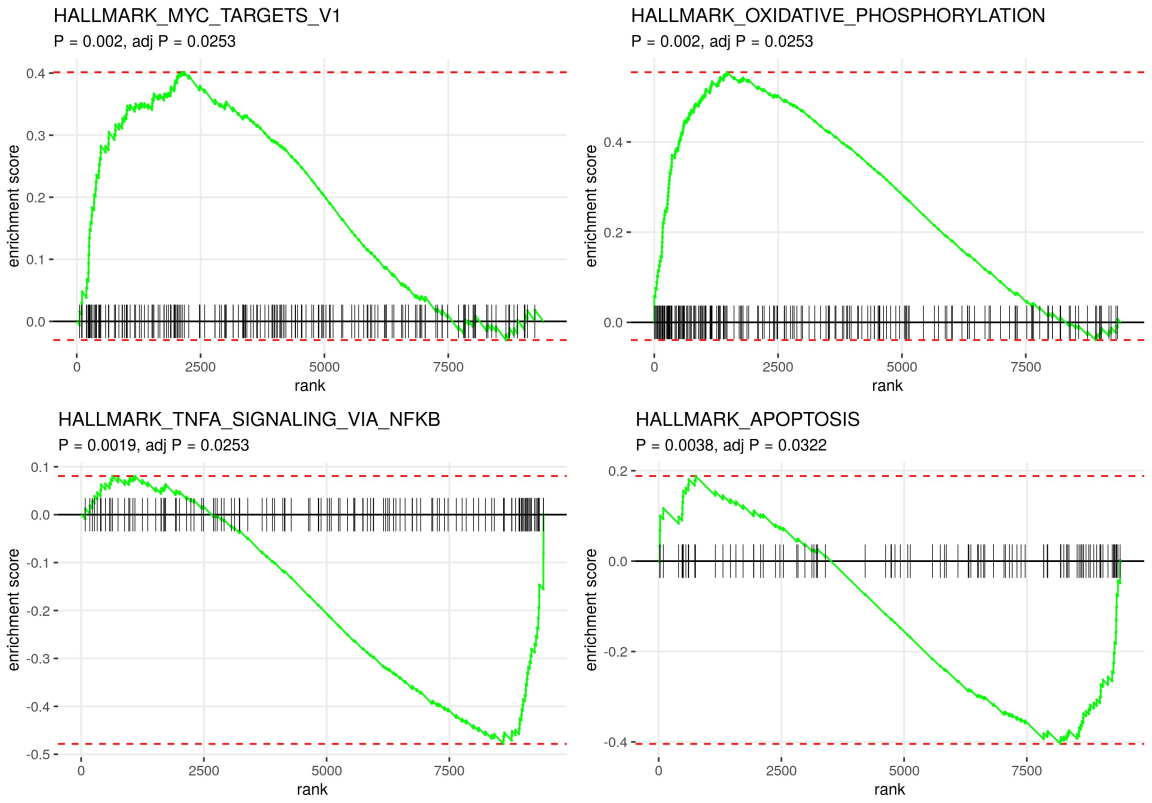

FIGURE B.19: Enriched and depleted pathways in clone D.4 compared to the rest of cells in clone D in patient 022

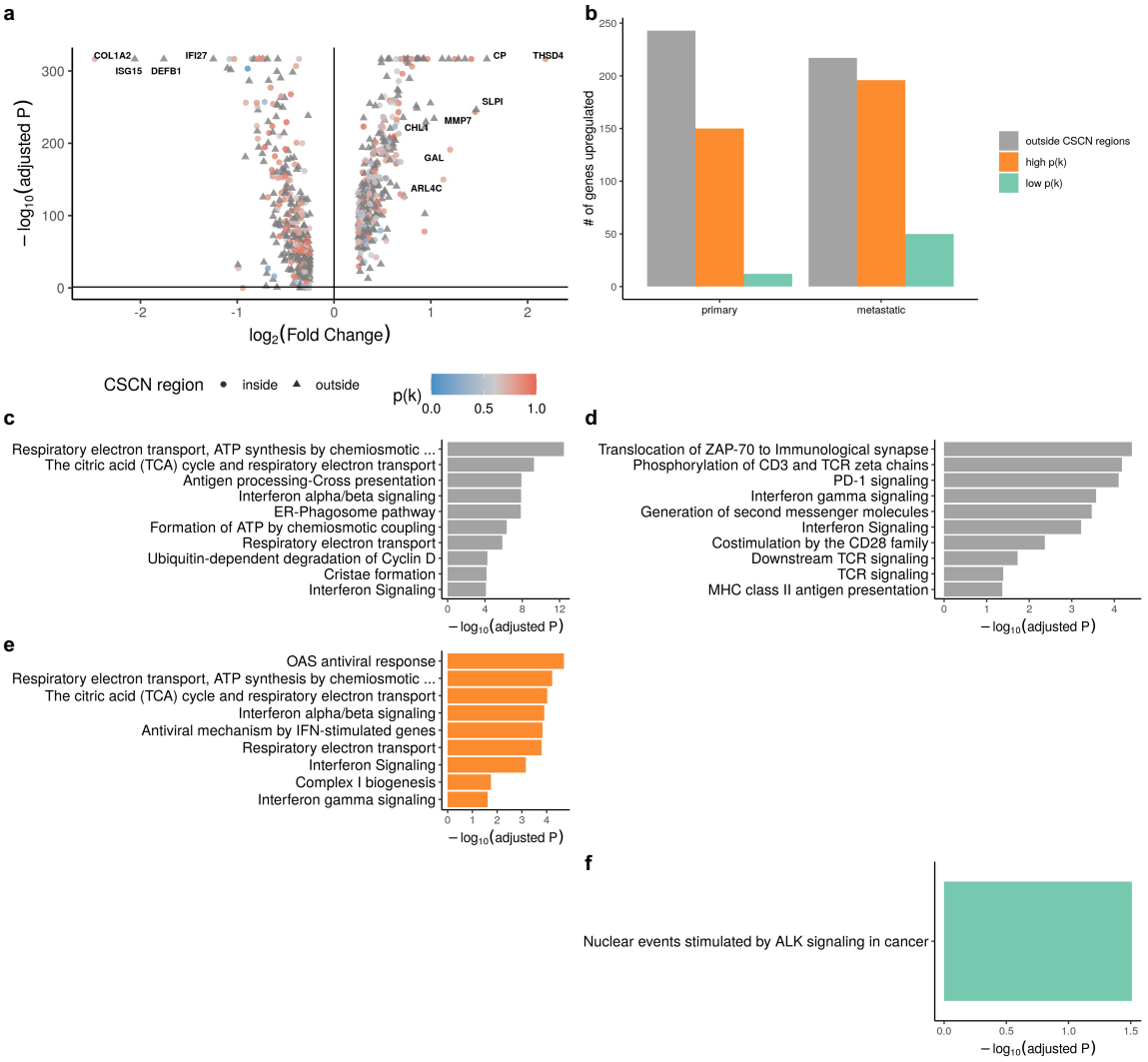

FIGURE B.20: **a,** Differentially expressed genes between metastatic and primary sites in patient 009. **b,** Number of upregulated genes in metastatic and primary sites grouped by $p(k)$ level in patient 009. **c-d,** Upregulated gene sets among genes outside of CSCN regions in patient 009 primary **(c)** and metastatic site **(d)**. **e,** Upregulated gene sets among high **p(k)** genes in patient 009 primary site. **f,** Upregulated gene sets among low $p(k)$ genes in patient 009 metastatic site.

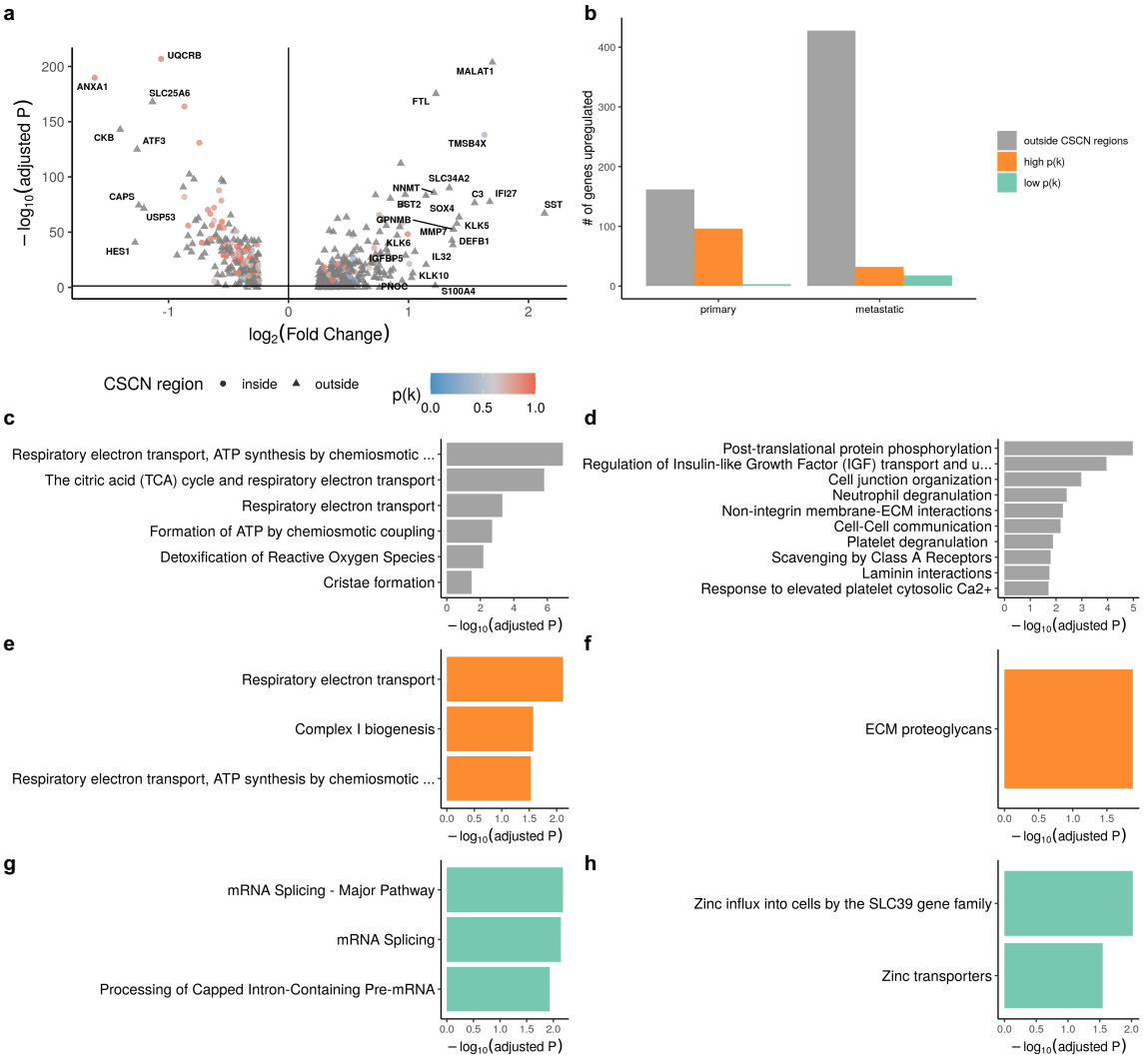

FIGURE B.21: **a,** Differentially expressed genes between metastatic and primary sites in patient 037. **b,** Number of upregulated genes in metastatic and primary sites grouped by $p(k)$ level in patient 037. **c-d,** Upregulated gene sets among genes outside of CSCN regions in patient 037 primary **(c)** and metastatic site **(d)**. **e-f,** Upregulated gene sets among high **p(k)** genes in patient 037 primary **(e)** and metastatic site **(f)**. **g-h,** Upregulated gene sets among low $p(k)$ genes in patient 037 primary **(g)** and metastatic site **(h)**.

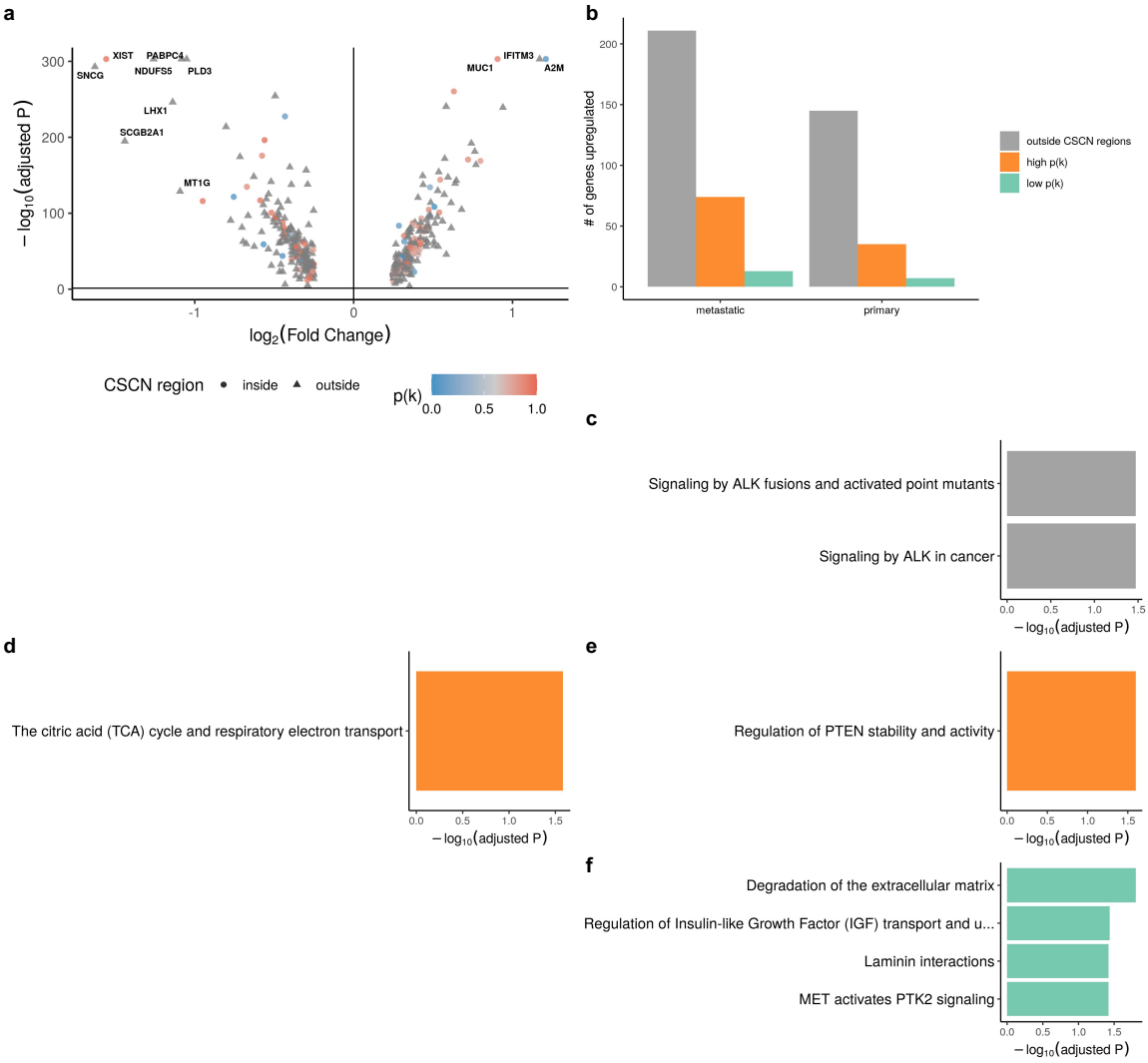

FIGURE B.22: **a,** Differentially expressed genes between metastatic and primary sites in patient 083. **b,** Number of upregulated genes in metastatic and primary sites grouped by $p(k)$ level in patient 083. **c,** Upregulated gene sets among genes outside of CSCN regions in patient 083 metastatic site. **d-e,** Upregulated gene sets among high **p(k)** genes in patient 083 primary **(d)** and metastatic site **(e)**. **f,** Upregulated gene sets among low $p(k)$ genes in patient 083 metastatic site.

Bibliography

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. en. *Nature* **481,** 306–313 (Jan. 2012).
2. Burrell, R. A. & Swanton, C. Tumour heterogeneity and the evolution of polyclonal drug resistance. en. *Mol. Oncol.* **8,** 1095–1111 (Sept. 2014).
3. Salehi, S. *et al.* Clonal fitness inferred from time-series modelling of single-cell cancer genomes. en. *Nature* **595,** 585–590 (June 2021).
4. Frankell, A. M. *et al.* The evolution of lung cancer and impact of subclonal selection in TRACERx. en. *Nature* **616,** 525–533 (Apr. 2023).
5. Al Bakir, M. *et al.* The evolution of non-small cell lung cancer metastases in TRACERx. en. *Nature* **616,** 534–542 (Apr. 2023).
6. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. en. *Nat. Rev. Cancer* **21,** 379–392 (June 2021).
7. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. en. *Nat. Rev. Clin. Oncol.* **15,** 81–94 (Feb. 2018).
8. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. en. *Cell* **150,** 1121–1134 (Sept. 2012).

9. Lee, V., Murphy, A., Le, D. T. & Diaz Jr, L. A. Mismatch Repair Deficiency and Response to Immune Checkpoint Blockade. en. *Oncologist* **21,** 1200–1211 (Oct. 2016).
10. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. en. *Nature* **578,** 94–101 (Feb. 2020).
11. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. en. *Nature* **606,** 984–991 (June 2022).
12. López, S. *et al.* Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. en. *Nat. Genet.* **52,** 283–293 (Mar. 2020).
13. Quinn, J. J. *et al.* Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. en. *Science* **371** (Feb. 2021).
14. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. en. *Science* **355** (Jan. 2017).
15. Vitale, I., Shema, E., Loi, S. & Galluzzi, L. Intratumoral heterogeneity in cancer progression and response to immunotherapy. en. *Nat. Med.* **27,** 212–224 (Feb. 2021).
16. Marusyk, A., Janiszewska, M. & Polyak, K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. en. *Cancer Cell* **37,** 471–484 (Apr. 2020).
17. Drews, R. M. *et al.* A pan-cancer compendium of chromosomal instability. en. *Nature* **606,** 976–983 (June 2022).
18. Tang, Y.-C. & Amon, A. Gene copy-number alterations: a cost-benefit analysis. en. *Cell* **152,** 394–405 (Jan. 2013).
19. Henrichsen, C. N. *et al.* Segmental copy number variation shapes tissue transcriptomes. en. *Nat. Genet.* **41,** 424–429 (Apr. 2009).

20. Vázquez-García, I. *et al.* Ovarian cancer mutational processes drive site-specific immune evasion. en. *Nature* **612,** 778–786 (Dec. 2022).
21. Funnell, T. *et al.* Single-cell genomic variation induced by mutational processes in cancer. en. *Nature* **612,** 106–115 (Dec. 2022).
22. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. en. *Nat. Genet.* **38,** 1043–1048 (Sept. 2006).
23. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. en. *N. Engl. J. Med.* **376,** 2109–2121 (June 2017).
24. Piotrowska, Z. *et al.* Heterogeneity Underlies the Emergence of EGFRT790 Wild-Type Clones Following Treatment of T790M-Positive Cancers with a Third-Generation EGFR Inhibitor. en. *Cancer Discov.* **5,** 713–722 (July 2015).
25. Chabon, J. J. *et al.* Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients. en. *Nat. Commun.* **7,** 11815 (June 2016).
26. Turke, A. B. *et al.* Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. en. *Cancer Cell* **17,** 77–88 (Jan. 2010).
27. Lim, Z.-F. & Ma, P. C. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. en. *J. Hematol. Oncol.* **12,** 134 (Dec. 2019).
28. Bozic, I. *et al.* Evolutionary dynamics of cancer in response to targeted combination therapy. en. *Elife* **2,** e00747 (June 2013).
29. Diaz Jr, L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. en. *Nature* **486,** 537–540 (June 2012).

30. Marine, J.-C., Dawson, S.-J. & Dawson, M. A. Non-genetic mechanisms of therapeutic resistance in cancer. en. *Nat. Rev. Cancer* **20,** 743–756 (Dec. 2020).

31. Hugo, W. *et al.* Non-genomic and Immune Evolution of Melanoma Acquiring MAPKi Resistance. en. *Cell* **162,** 1271–1285 (Sept. 2015).

32. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. en. *Nat. Genet.* **48,** 758–767 (July 2016).

33. Andor, N. *et al.* Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. en. *Nat. Med.* **22,** 105–113 (Jan. 2016).

34. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. en. *Cancer Cell* **27,** 15–26 (Jan. 2015).

35. Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. en. *Gastroenterology* **138,** 2073–2087.e3 (June 2010).

36. Jo, W.-S. & Carethers, J. M. Chemotherapeutic implications in microsatellite unstable colorectal cancer. en. *Cancer Biomark.* **2,** 51–60 (2006).

37. André, T. *et al.* Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. *N. Engl. J. Med.* **383,** 2207–2218 (Dec. 2020).

38. Bakhoum, S. F. & Cantley, L. C. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. en. *Cell* **174,** 1347–1360 (Sept. 2018).

39. Martínez-Ruiz, C. *et al.* Genomic-transcriptomic evolution in lung cancer and metastasis. en. *Nature* **616,** 543–552 (Apr. 2023).

40. Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. en. *Nature* **560,** 325–330 (Aug. 2018).

41. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. en. *Nat. Biotechnol.* **29,** 1120–1127 (Nov. 2011).

42. Sharma, A. *et al.* Non-Genetic Intra-Tumor Heterogeneity Is a Major Predictor of Phenotypic Heterogeneity and Ongoing Evolutionary Dynamics in Lung Tumors. en. *Cell Rep.* **29,** 2164–2174.e5 (Nov. 2019).

43. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. en. *Nat. Rev. Genet.* **22,** 3–18 (Jan. 2021).

44. Williams, M. J., Sottoriva, A. & Graham, T. A. Measuring Clonal Evolution in Cancer with Genomics. en. *Annu. Rev. Genomics Hum. Genet.* **20,** 309–329 (Aug. 2019).

45. Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. en. *Cell* **178,** 835–849.e21 (Aug. 2019).

46. Pollack, J. R. *et al.* Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. en. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 12963–12968 (Oct. 2002).

47. Jörnsten, R. *et al.* Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. en. *Mol. Syst. Biol.* **7,** 486 (Apr. 2011).

48. Ding, J. *et al.* Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. en. *Nat. Commun.* **6,** 8554 (Oct. 2015).

49. Bhattacharya, A. *et al.* Transcriptional effects of copy number alterations in a large set of human cancers. en. *Nat. Commun.* **11,** 715 (Feb. 2020).

50. Hong, W. S., Shpak, M. & Townsend, J. P. Inferring the Origin of Metastases from Cancer Phylogenies. en. *Cancer Res.* **75,** 4021–4025 (Oct. 2015).

51. Reiter, J. G. *et al.* Reconstructing metastatic seeding patterns of human cancers. en. *Nat. Commun.* **8,** 14114 (Jan. 2017).

52. Hao, Y., Yan, M., Heath, B. R., Lei, Y. L. & Xie, Y. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. en. *PLoS Comput. Biol.* **15,** e1006976 (May 2019).

53. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. en. *Nat. Methods* **12,** 453–457 (May 2015).

54. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. en. *Bioinformatics* **29,** 1083–1085 (Apr. 2013).

55. Racle, J. & Gfeller, D. EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data. en. *Methods Mol. Biol.* **2120,** 233–248 (2020).

56. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. en. *Nat. Commun.* **11,** 5650 (Nov. 2020).

57. Andor, N. *et al.* Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. en. *NAR Genom Bioinform* **2,** lqaa016 (June 2020).

58. Guo, L. *et al.* Single-Cell DNA Sequencing Reveals Punctuated and Gradual Clonal Evolution in Hepatocellular Carcinoma. en. *Gastroenterology* **162,** 238–252 (Jan. 2022).

59. Gonzalo Parra, R *et al. Single cell multi-omics analysis of chromothriptic medulloblastoma highlights genomic and transcriptomic consequences of genome instability* en. June 2021.

60. Tickle, T, Tirosh, I, Georgescu, C, Brown, M & Haas, B. inferCNV of the Trinity CTAT Project. *Broad Institute of MIT and Harvard.*

61. Torre, L. A. *et al.* Ovarian cancer statistics, 2018. en. *CA Cancer J. Clin.* **68,** 284–296 (July 2018).

62. Patch, A.-M. *et al.* Whole–genome characterization of chemoresistant ovarian cancer. en. *Nature* **521,** 489–494 (May 2015).

63. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. en. *Nature* **578,** 112–121 (Feb. 2020).

64. Wang, Y. K. *et al.* Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. en. *Nat. Genet.* **49,** 856–865 (June 2017).

65. Schwarz, R. F. *et al.* Spatial and temporal heterogeneity in high-grade serous ovarian cancer: a phylogenetic analysis. en. *PLoS Med.* **12,** e1001789 (Feb. 2015).

66. Zhang, A. W. *et al.* Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. en. *Cell* **173,** 1755–1769.e22 (June 2018).

67. Mirza, M. R. *et al.* The forefront of ovarian cancer therapy: update on PARP inhibitors. en. *Ann. Oncol.* **31,** 1148–1159 (Sept. 2020).

68. Liu, J. F., Konstantinopoulos, P. A. & Matulonis, U. A. PARP inhibitors in ovarian cancer: current status and future promise. en. *Gynecol. Oncol.* **133,** 362–369 (May 2014).

69. Gockley, A. *et al.* Outcomes of Women With High-Grade and Low-Grade Advanced-Stage Serous Epithelial Ovarian Cancer. en. *Obstet. Gynecol.* **129,** 439–447 (Mar. 2017).

70. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. en. *Nature* **474,** 609–615 (June 2011).

71. Konstantinopoulos, P. A., Ceccaldi, R., Shapiro, G. I. & D'Andrea, A. D. Homologous Recombination Deficiency: Exploiting the Fundamental Vulnerability of Ovarian Cancer. en. *Cancer Discov.* **5,** 1137–1154 (Nov. 2015).

72. Funnell, T. *et al.* Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. en. *PLoS Comput. Biol.* **15,** e1006799 (Feb. 2019).

73. Sztal, T. E. & Stainier, D. Y. R. Transcriptional adaptation: a mechanism underlying genetic robustness. en. *Development* **147** (Aug. 2020).

74. El-Brolosy, M. A. & Stainier, D. Y. R. *Genetic compensation: A phenomenon in search of mechanisms* 2017.

75. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. en. *Nat. Genet.* **47,** 115–125 (Feb. 2015).

76. Veitia, R. A., Bottani, S. & Birchler, J. A. Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation. en. *Trends Genet.* **29,** 385–393 (July 2013).

77. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. en. *Nat. Methods* **12,** 519–522 (June 2015).

78. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. en. *Nat. Biotechnol.* **33,** 285–289 (Mar. 2015).

79. Campbell, K. R. *et al.* clonealign: statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. en. *Genome Biol.* **20,** 54 (Mar. 2019).

80. Ferreira, P. F., Kuipers, J. & Beerenwinkel, N. *Mapping single-cell transcriptomes to copy number evolutionary trees* en. Nov. 2021.

81. Bai, X., Duren, Z., Wan, L. & Xia, L. C. *Joint Inference of Clonal Structure using Single-cell Genome and Transcriptome Sequencing Data* en. Oct. 2020.

82. Mu, P. *et al.* SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. en. *Science* **355,** 84–88 (Jan. 2017).

83. Chan, J. M. *et al.* Lineage plasticity in prostate cancer depends on JAK/STAT inflammatory signaling. en. *Science* **377,** 1180–1191 (Sept. 2022).

84. Johnson, K. C. *et al. Single-cell multimodal glioma analyses identify epigenetic regulators of cellular plasticity and environmental stress response* 2021.

85. Gao, T. *et al.* Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes. en. *Nat. Biotechnol.* (Sept. 2022).

86. Laks, E. *et al.* Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. en. *Cell* **179,** 1207–1221.e22 (Nov. 2019).

87. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. en. *Nat. Commun.* **8,** 14049 (Jan. 2017).

88. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. en. *Nucleic Acids Res.* **44,** e131 (Sept. 2016).

89. Chen, H., Bell, J. M., Zavala, N. A., Ji, H. P. & Zhang, N. R. Allele-specific copy number profiling by next-generation DNA sequencing. en. *Nucleic Acids Res.* **43,** e23 (Feb. 2015).

90. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. en. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 16910–16915 (Sept. 2010).

91. Zaccaria, S. & Raphael, B. J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. en. *Nat. Biotechnol.* **39,** 207–214 (Feb. 2021).

92. Wu, C.-Y. *et al.* Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. en. *Nat. Biotechnol.* **39,** 1259–1269 (Oct. 2021).

93. Robles-Espinoza, C. D., Mohammadi, P., Bonilla, X. & Gutierrez-Arcelus, M. Allele-specific expression: applications in cancer and technical considerations. en. *Curr. Opin. Genet. Dev.* **66,** 10–19 (Feb. 2021).

94. Valle, L. *et al.* Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. en. *Science* **321,** 1361–1365 (Sept. 2008).

95. Foulkes, W. D., Ragoussis, J, Stamp, G. W., Allan, G. J. & Trowsdale, J. Frequent loss of heterozygosity on chromosome 6 in human ovarian carcinoma. en. *Br. J. Cancer* **67,** 551–559 (Mar. 1993).

96. Feenstra, M *et al.* HLA class I expression and chromosomal deletions at 6p and 15q in head and neck squamous cell carcinomas. en. *Tissue Antigens* **54,** 235–245 (Sept. 1999).

97. Jiménez, P *et al.* Chromosome loss is the most frequent mechanism contributing to HLA haplotype loss in human tumors. en. *Int. J. Cancer* **83,** 91–97 (Sept. 1999).

98. Zhao, X. *et al.* Loss of heterozygosity at 6p21 and HLA class I expression in esophageal squamous cell carcinomas in China. en. *Asian Pac. J. Cancer Prev.* **12,** 2741–2745 (2011).

99. McGranahan, N. *et al.* Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. en. *Cell* **171,** 1259–1271.e11 (Nov. 2017).

100. Pirinen, M. *et al.* Assessing allele-specific expression across multiple tissues from RNA-seq read data. en. *Bioinformatics* **31,** 2497–2504 (Aug. 2015).

101. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. en. *Bioinformatics* **25,** 3207–3212 (Dec. 2009).

102. Edsgärd, D. *et al.* GeneiASE: Detection of condition-dependent and static allele-specific expression from RNA-seq data without haplotype information. en. *Sci. Rep.* **6,** 21134 (Feb. 2016).

103. Fan, J. *et al.* ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. en. *PLoS Genet.* **16,** e1008786 (May 2020).

104. Fan, J. *et al.* Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. en. *Genome Res.* **28,** 1217–1227 (Aug. 2018).

105. Bingham, E. *et al.* Pyro: Deep Universal Probabilistic Programming. *CoRR* **abs/1810.09538.** arXiv: 1810.09538. http://arxiv.org/abs/1810.09538 (2018).

106. Martins, F. C. *et al.* Clonal somatic copy number altered driver events inform drug sensitivity in high-grade serous ovarian cancer. en. *Nat. Commun.* **13,** 6360 (Oct. 2022).

107. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. en. *Nat. Rev. Cancer* **18,** 696–705 (Nov. 2018).

108. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. en. *Cell Syst* **1,** 417–425 (Dec. 2015).

109. Hu, Z. *et al.* The Repertoire of Serous Ovarian Cancer Non-genetic Heterogeneity Revealed by Single-Cell Sequencing of Normal Fallopian Tube Epithelial Cells. en. *Cancer Cell* **37,** 226–242.e7 (Feb. 2020).

110. Izar, B. *et al.* A single-cell landscape of high-grade serous ovarian cancer. en. *Nat. Med.* **26,** 1271–1279 (Aug. 2020).

111. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. en. *Genome Biol.* **16,** 278 (Dec. 2015).

112. Benci, J. L. *et al.* Tumor Interferon Signaling Regulates a Multigenic Resistance Program to Immune Checkpoint Blockade. en. *Cell* **167,** 1540–1554.e12 (Dec. 2016).

113. Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. en. *Nat. Genet.* **50,** 1189–1195 (Aug. 2018).

114. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. en. *Nat. Genet.* **45,** 1134–1140 (Oct. 2013).

115. Quinton, R. J. *et al.* Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. en. *Nature* **590,** 492–497 (Feb. 2021).

116. Nikolic, A. *et al.* Copy-scAT: Deconvoluting single-cell chromatin accessibility of genetic subclones in cancer. en. *Sci Adv* **7,** eabg6045 (Oct. 2021).

117. Ramakrishnan, A., Symeonidi, A., Hanel, P., Schubert, M. & Colomé-Tatché, M. *epiAneufinder: identifying copy number variations from single-cell ATAC-seq data* en. Apr. 2022.

118. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. en. *Nat. Biotechnol.* **37,** 925–936 (Aug. 2019).

119. Ayinde, B. O. & Zurada, J. M. Discovery through constraints: Imposing constraints on autoencoders for data representation and dictionary learning. *IEEE Syst. Man Cybern. Mag.* **3,** 13–24 (July 2017).

120. Gupta, S. *et al.* A Pan-Cancer Study of Somatic TERT Promoter Mutations and Amplification in 30,773 Tumors Profiled by Clinical Genomic Sequencing. en. *J. Mol. Diagn.* **23,** 253–263 (Feb. 2021).

121. Elyanow, R., Zeira, R., Land, M. & Raphael, B. J. STARCH: copy number and clone inference from spatial transcriptomics data. en. *Phys. Biol.* **18,** 035001 (Mar. 2021).

122. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. en. *Bioinformatics* (May 2021).

123. Wang, C. & Blei, D. M. Variational inference in nonconjugate models. *Journal of Machine Learning Research* (2013).

124. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell.* `https://doi.org/10.1016/j.cell.2021.04.048` (2021).

125. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv.* eprint: `https://www.biorxiv.org/content/early/2021/02/01/060012.full.pdf`. `https://www.biorxiv.org/content/early/2021/02/01/060012` (2021).

126. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (Oct. 2005).