Improved variant calling and phylogeny inference from scATAC-seq

Megan Le[1,2], Ruitong Li[3,4], Daniel Schaffer[4,5,6]
[1]Department of Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139
[2]Dana-Farber Cancer Institute, Boston, MA 02215
[3]Bioinformatics and Integrative Genomics Program, Harvard Medical School, Boston, MA 02115
[4]The Broad Institute of MIT and Harvard, Cambridge, MA 02142
[5]Computational and Systems Biology Program, Massachusetts Institute of Technology
[6]Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology

Abstract

Cancer progression is driven by both genetic alterations and chromatin remodeling, yet little is known about the interplay between these two classes of events in shaping the clonal diversity of cancers. The use of scATAC-seq data for calling SNVs provides a promising alternative to scRNA-seq data, which does not sequence non-coding regions of the genome and may contain spurious mutations. Using the Monopogen framework, we performed germline mutation calling on data from single-cell ATACseq and DNAseq assays on the SNU601 gastric cancer cell line, with the scATAC-seq results containing 77.84% of the identified mutations from the scDNA-seq data with a 4.05% false positive rate. We then performed somatic mutation calling using LD refinement, finding 47.90% of scATAC-seq variants were also identified in the scDNA-seq results. We identified a high proportion of C>G mutations that were also mostly found in promoter regions, with the mutation data fitting to known mutation signatures related to chemotherapy treatment and mismatch repair. Though hierarchical clustering of the scATAC-seq SNV results was not effective in capturing known clonal structure, UMAP analysis reflected the highest-level clonal split through groupings of CNV clones 1 and 2 apart from clones 3, 4, 5, and 6. Additionally, pseudotime trajectory analysis was able to identify clone 3 as an intermediate clone before clones 4 through 6, which were grouped closer together in their UMAP embeddings. We analyzed correlations between the most variable peaks across clones and mutations with the best coverage but did not find any significant results. We also applied the CellPhy maximum likelihood estimate framework to our SNV results to perform phylogeny inference but were unable to reach convergence. Our work is a first step in exploring how scATAC-seq data can be used to improve mutation calling and to investigate cancerous cell population dynamics and chromatin accessibility changes during clonal evolution, which can ultimately lead to a better understanding of individual cancer evolution and prognosis.

Background

Since the 1970s, populations of cancer cells in tumors have been understood to start from a single cell and grow while subject to mutational and selective processes akin in miniature to those found on an organismal, evolutionary level [1]. In this model, individual cancer cells develop *de novo* mutations at a high rate, and cells with mutations that provide faster growth, better immune evasion, or other beneficial attributes to the cancer will have their mitotic descendents form a greater proportion of the cancer as it progresses [2]. Understanding the history and trajectory of cancer cells in this evolutionary context is of great importance to understanding the development, trajectory, and prognosis of the cancer as a whole, especially as the ‘fitness’ for which *de novo* somatic mutations in cancer cells are selected is closely tied to the malignancy of the cancer as a whole.

In the last decades, methods for studying the evolution and phylogenetic structure of tumors on the basis of sequence variation underwent their own complex evolutionary process. The resulting diverse population of methods rely on a variety of assays for several types of genomic or even epigenomic variation to fit one of several possible computational models [3]. However, most common methods tended to focus on either copy-number (structural) or single-nucleotide DNA variants and—after its widespread adoption—make use of next-generation sequencing as a direct source of variant information. While bulk sequencing assays (as well as some pre-sequencing assays) can directly capture heterogeneity present in a sample, they cannot identify how it is distributed. Consequently, these methods are applied to several samples taken from, *e.g.,* multiple pieces of a solid tumor or several metastases [3].

Traditional methods to infer tumor phylogeny or clonal history, which rely on multiple spatially or temporally heterogeneous samples, can only be performed on a solid tumor, and in particular are not applicable to cancer cell lines, whose 'evolutionary' history is therefore unknown. One alternative that can be applied to any set of cells is to use genetic variation to infer the phylogeny of cells in an apparently-homogenous sample, if such variation can be assayed. And, understanding how major cellular processes, such as gene regulation, vary along with clonal evolution and linking any such co-variation to any individual phenotype remains a major open problem.

It is apparent that single-cell sequencing assays represent a potent new source of raw data for the inference of phylogenetic structure within a single cancer sample [4,5]. One might first think to apply single-cell whole genome sequencing to the study of either copy number or single-nucleotide variants. However, single-cell DNA sequencing (scDNAseq) is especially vulnerable to the increased sparseness of reads from an individual cell [5]. This makes studying copy-number variation, which requires capturing with high confidence reads mapping to all copies of each variant region, and even single-nucleotide variants, very challenging in scDNAseq due to this low—and variable between cells—coverage. Additionally, from the perspective of one looking to repurpose an existing sequencing dataset, single-cell whole genome sequencing is relatively infrequently performed, as the genome in each cell is expected to generally be the same, so there is little gain from the added complexity of a single-cell assay and in fact loss (of a complete genome) due to the aforementioned read sparsity.

Instead, widely-used single-cell sequencing assays include single-cell RNA sequencing (scRNAseq) and the single-cell assay for transposase-accessible chromatin with sequencing (scATAC-seq), which assay gene expression and gene regulatory activity, both processes that can vary greatly between cells [6]. In particular, single-cell, genome-wide profiling of accessible chromatin with scATACseq grants unprecedented scale and resolution to dissect phenotypic heterogeneity in the regulatory landscape. There have been emerging efforts in calling SNVs from sc-omics sequencing data (mainly scRNA-seq) in cancer samples. But they either suffer badly from poor accuracy or require specific sample composition such as diverse normal cell types in combination with tumor cells [7]. Further, compared to scRNA-seq, there are several apparent advantages to using scATAC-seq data as a source of single-cell variant data. First, scATAC-seq can be more reliably performed on formalin-fixed paraffin-embedded (FFPE) or otherwise aged samples. Additionally, ATACseq covers select non coding regions of the genome that are never sequenced in RNA-seq, and the coding regions of the genome are (presumably) under more constraint. Finally, RNA editing events may introduce spurious mutations in RNA-seq data in a non-clonal pattern. Therefore, mutation calling from profiles of accessible regions instead of mRNA abundance should be preferred for and will theoretically serve as a better lineage marker.

Yes, mutation calling from scATAC-seq is still a field left largely unexplored. Dou et al. cleverly used phasing of SNPs to improve the accuracy of nuclear single nucleotide variants (SNVs) [8], which are found to be consistent with lineage information (eg: differentiation during hematopoiesis). This method is called Monopogen. To investigate the possibility of using somatic SNVs as clonal markers to study tumor evolution at the single-cell level, we analyzed SNU601, a gastric tumor cell line with well-explored polyclonal structures [9]. Importantly, both scATAC-seq and scDNA-seq datasets are available for this cell line. We proceeded to 1) call SNVs with Monopogen on both scATAC-seq and scDNA-seq data to benchmark the accuracy of genotyping and analyze mutation signatures; 2) use well-defined CNV-based clones of SNU601 as the ground-truth for checking the clonal concordance of SNVs; and 3) attempt to construct a phylogeny tree with SNVs from the SNVs called scATAC-seq.

Results / Discussion

Somatic mutation calling and distribution

We began with datasets of single-cell sequencing reads from ATACseq [9] and DNAseq [10] assays on the SNU601 gastric cancer cell line, obtained in both cases by downloading the originally-deposited 10X .bam file from NCBI's SRA database. For both the scATAC-seq and scDNA-seq data, we first used the Monopogen preprocess module to remove reads with high alignment mismatches (using a threshold of 3 mismatches). We then performed germline SNV calling, using the 1000 Genomes Phase 3 project[1] for our external reference panel. We found that 77.84% of reported germline SNVs from the scDNA-seq data were also reported in the scATAC-seq data, and 4.05% of reported germline SNVs in the scATAC-seq data were not found in the scDNA-seq data.

We identified 11370 and 41063 somatic mutations from scATAC-seq and scDNA-seq from SNU601 after LD refinement. We found that 47.90% (n=5446) of detected somatic variants in scATAC were also detected as somatic variants in scDNA-seq. With a wider genome coverage in scDNA-seq, more somatic variants were found in scDNA-seq but also with markedly lower coverage per locus. In the scATACseq results, we find 1468 variants covered in more than 5% (n>111, either reference or alternative allele) of cells (**Fig. 1A**). But, in the scDNA-seq results, we only detect 155 variants in scDNA-seq with coverage in more than 5% (n>29) of cells (**Fig. 1B**).

[1] https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/

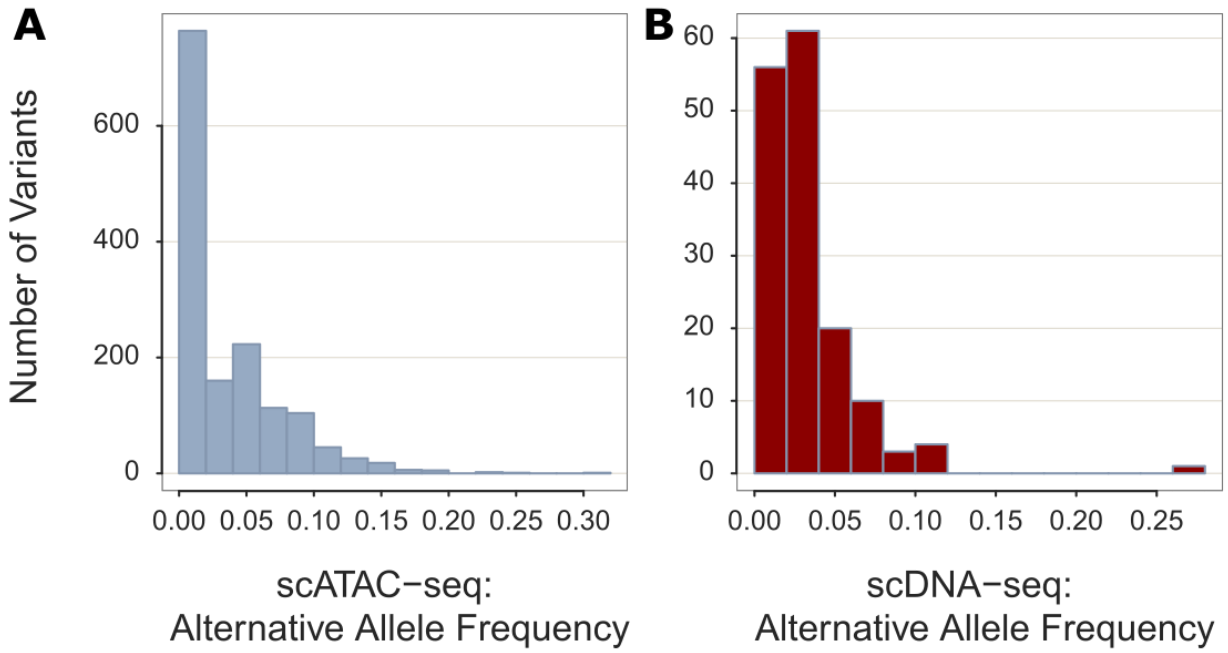

Figure 1: Histogram of alternative allele frequency among **(A)** scATAC-seq and **(B)** scDNA-seq variants with coverage in at least 5% of cells from the respective assays.

Out of the aforementioned somatic mutations called from the scATAC-seq with Monopogen, we first extracted the 96 mutation signature context (a point mutation with its left-right neighboring nucleotide; 4^3= 96 combinations) and visualized its distribution (**Fig. 2A**). A notably high percentage of mutations were found as C>G, especially in the context G_C and G_G. When fitting the mutation data to known mutation signatures, the top signatures we found were related to chemotherapy treatment (SBS86), mismatch repair (SBS15, SBS20, SBS21, and SBS26), and previously found with unknown origin (SBS17a, SBS17b). Since scATAC fragments are enriched in regulatory elements, we next sought to investigate potential strand bias (transcribed vs untranscribed) in mutations. In this analysis, we found that significantly higher rates of C>G mutation but lower rates of T>G mutations were found in transcribed strand vs untranscribed ones (**Fig. 2B,C**).

Lastly, we examined the distribution of mutations across the genome and important types of regulatory elements. We observed several areas of dropdown in a rainfall plot of mutations, which is suggestive of the existence of potential somatic mutation hotspots (**Fig. 2D**). Not surprisingly, most C>G mutations were found in promoter regions (**Fig. 2E**) .

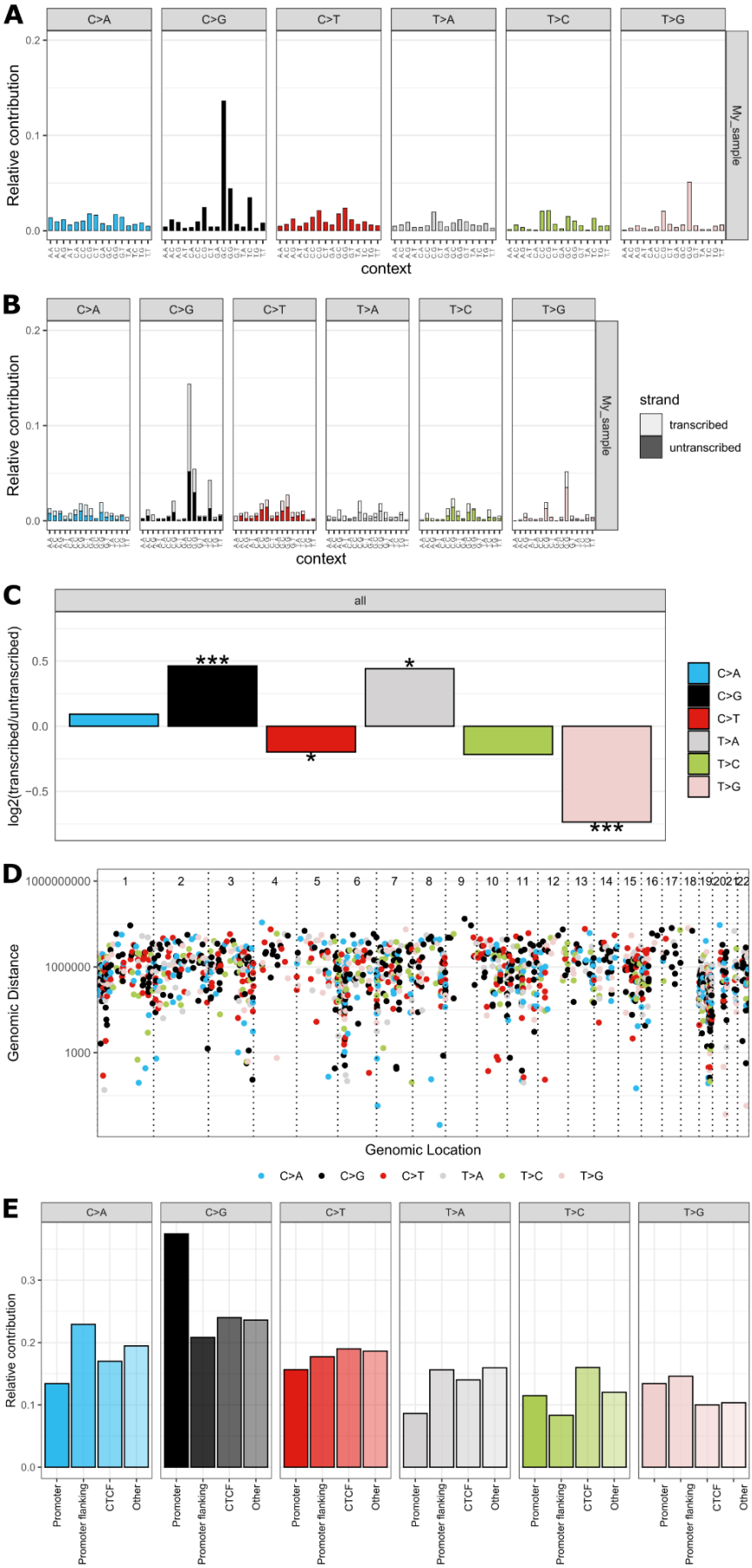

Figure 2: Characterization of somatic mutations called from the scATAC-seq data. **(A)** 96-block 3nt mutation signatures. **(B)** Bias towards transcription strands shown with the stacked bar plot and **(C)** significance testing on bottom. **(D)** Distribution of mutations across chromosome and genomic distance relative to left nearest neighbor mutation. **(E)** Distribution of mutations across the main classes of regulatory elements.

Mutation clustering and phylogeny inference

We next sought to assess the potential for performing SNV-based phylogeny inference by focusing on likely-informative CNVs called from the scATACseq. We started by selecting variants with allele frequency 0.1-0.9 and cells with at least 10% coverage of those mutations. This filtering step resulted in a cell by mutation matrix with dimension 857 by 104. A heatmap (**Fig. 3**) with hierarchical clustering (adjusted Jaccard distance) on cells and mutations did not immediately suggest a concordant relationship with the pre-defined clonal structure, surprisingly. With the assistance of population-level phasing implemented in Monopogen, the potential problem of sequencing artifacts and errors in individual cells should be largely alleviated. This is also confirmed by the relatively high consistency from mutation calling with two independent datasets (scDNA-seq and scATAC-seq). We hypothesize that the obscured SNV-clone concordance likely results from some combination of low capture efficiency with single-cell technology, allelic dropout, complicated copy number alterations, and inappropriate choice of distance metric for clustering. To assess whether any subtle phylogenetic signal is present, we attempted to instead infer a phylogenetic tree from forward to use the somatic variant matrix using a state-of-the-art mutation-based phylogeny-inference tool, CellPhy [11], which is based on the RAxML-NL tree inference program. Unfortunately, in all of our attempts the RAxML-NL optimizer failed—after hundreds or thousands of CPU hours—with the message that the optimizer "converged to a worse likelihood score." This suggests that there may be too much sparsity, or just not enough signal, in the mutation matrix to produce a complete phylogeny.

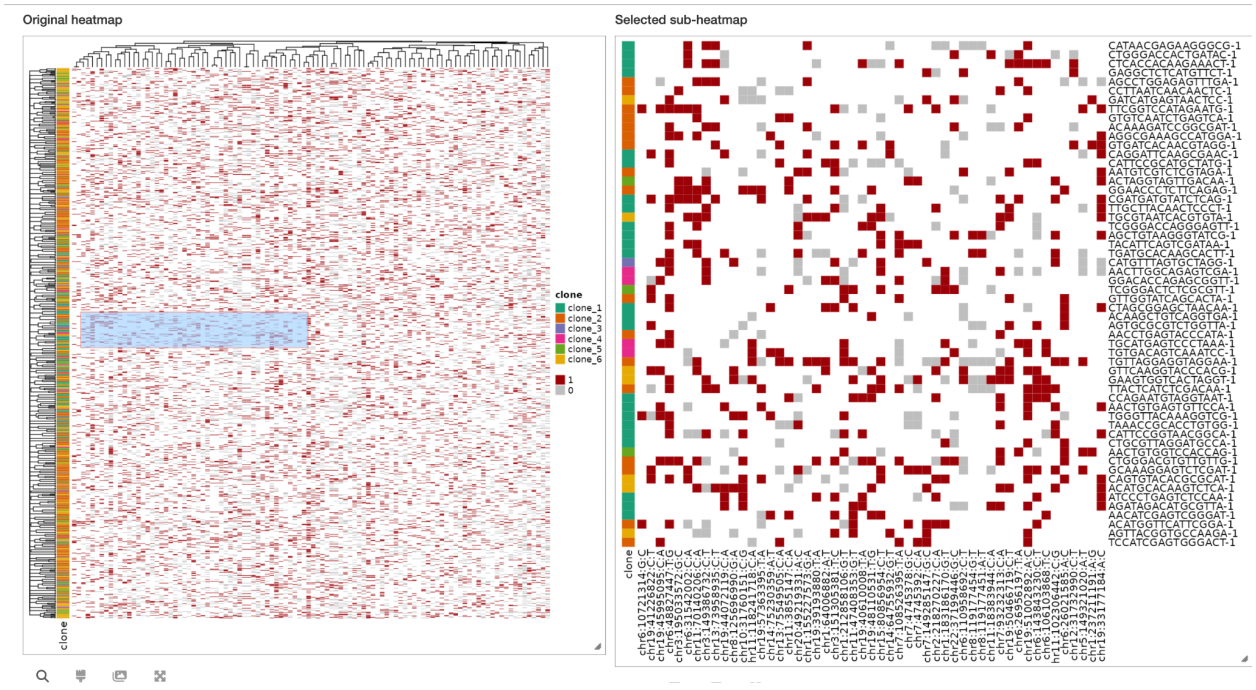

Figure 3: Heatmap of SNV alleles (columns) and cells (rows). Cells are labeled at left by their CNV-based clonal identity. Red denotes the alternate allele, gray denotes the reference allele, and white denotes no coverage of a particular SNV in a particular cell.

Peak calling, cell clustering, and concordance with mutations

Finally, we sought to integrate information about the regulatory landscape of each cell—obtainable by using the scATACseq dataset for its intended purpose—with the variant calling. We created a scATAC fragment file from the provided bam file using the sinto package. Furthermore, as suggested in recent work on using consensus peak sets for more meaningful and consistent interpretation of open chromatin region from scATAC-seq data [12], we generated the peak-by-cell matrix using consensus peaks with cPeaks (1.5×10^6 peaks) rather than calling *de novo* peaks. This sparse matrix was then input to the Signac package (version 1.12.0) [13] to create a Seurat object using CreateSeuratObject, retaining only peaks found in 10% out of the ~2500 cells with a CNV clone annotation (~100k, 6% of all consensus peaks). Second, we normalized the cell-by-peak matrix by the TF-IDF (term frequency–inverse document frequency) method using RunTFIDF and ran a singular value decomposition (SVD) on the TF-IDF normalized matrix with RunSVD. The combined steps of TF-IDF followed by SVD are known as latent semantic indexing (LSI) [14]; we refer to the reduced dimensions as LSIs later in text and figures. We retained the 2nd to 50th LSIs and identified clusters using SNN (shared nearest neighbor) graph clustering with FindClusters with a resolution of 0.8, with the SLM (Smart Local Moving) algorithm for modularity optimization. We dropped the 1st LSI because it was highly correlated with the sequencing depth in each cell.

We created a 2D embedding of each cell from the reduced dimension space (LSIs) using UMAP (**Fig. 4A-B**). By coloring cells based on 1) scATAC clusters determined by reduced dimensions LSIs (**Fig. 4A**) and 2) CNV clones (**Fig. 4B**), we can clearly see from the inconsistency between scATAC clusters and CNV clones that genetic clones do not always show distinct chromatin accessibility patterns. However, we do see that the highest-level clonal split, between CNV clones {C1,C2} and {C3,C4,C5,C6} [9], is somewhat reflected, indicating very limited inheritance of accessibility between cells.

In the absence of existing knowledge on phylogenetic history of tumor samples, an alternate non-phylogenetic way to retrospectively investigate the dynamics among cellular subpopulation in a snapshot sc-omic sequencing dataset is through pseudotime trajectory analysis. We utilized monocle3 to infer pseudo time based on the UMAP embedding created from the cell by peak matrix and a small subset of Clone 1 as root cells (N=11). The pseudotime trend is to some extent consistent with clonal labeling, which suggested a potential underlying continuum of chromatin accessibility change during clone evolution (**Fig. 4C-D**). The scATAC cell by peak matrix with Tn5 insertion count data is featured by sparsity and low count number, therefore it is usually directly approximated with binarization. Similar attributes are also found in cell by mutation count data, we therefore reasoned to try applying all pre-processing, dimension reduction and embedding aforementioned with cell by mutation as inputs and also performed pseudotime inference with the same subset of root cells. However, by coloring cells with the derived pseudotime from mutation profile to the same UMAP, more inconsistency with the CNV clones are observed, suggesting that mutation status input may be ill-suited for the dimension reduction operations commonly adopted in high-dimensional single-cell field for signal aggregation.

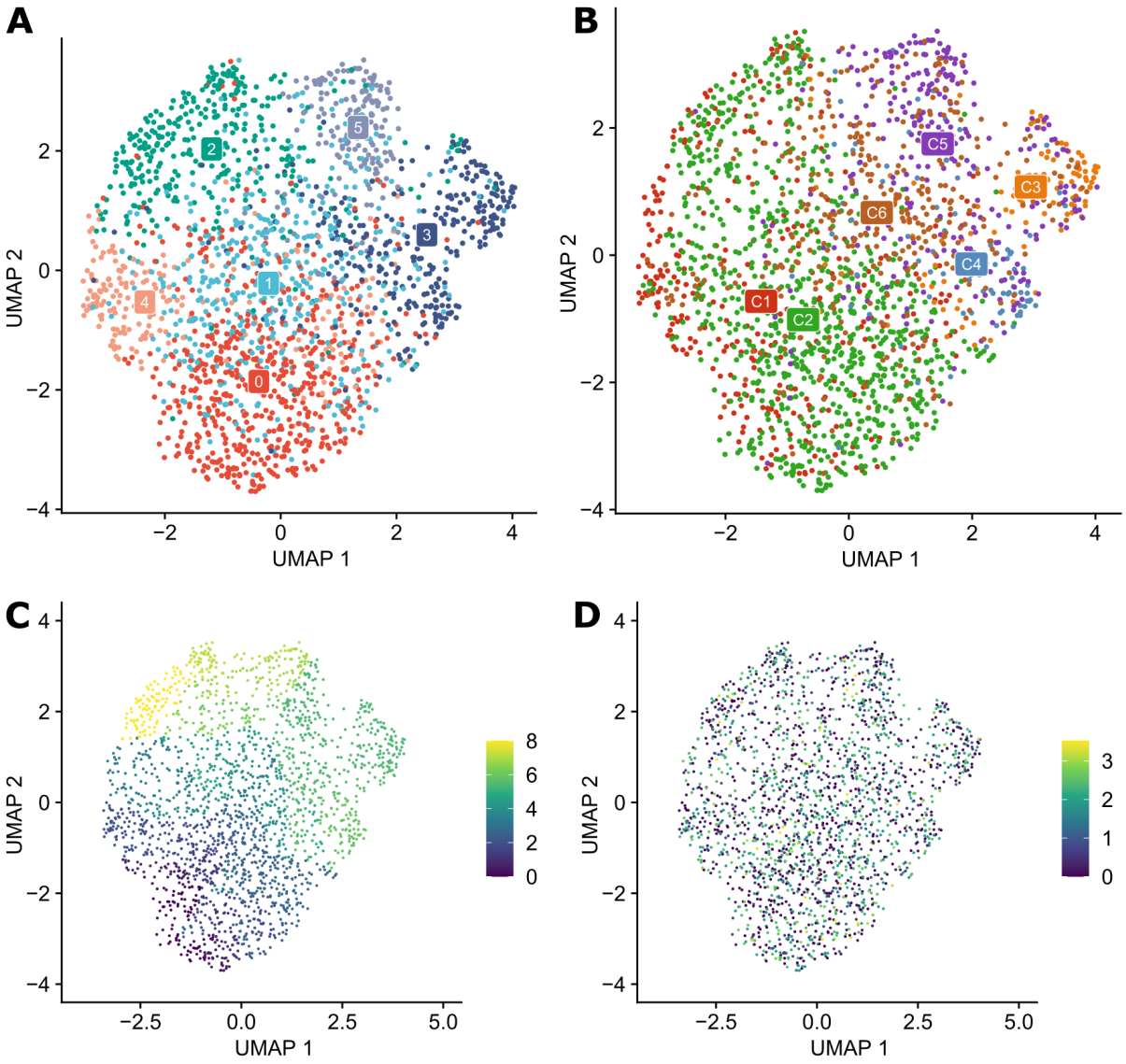

Figure 4: UMAP 2D embedding of cells colored by **(A)** clusters based on reduced dimensions LSIs **(B)** CNV clones replicated from Wu et al. **(C)** pseudotime derived from neighboring graphs based on cell by peak profile **(D)** pseudotime derived from neighboring graphs based on cell by mutation profile.

Lastly, we investigated potential mutation-peak linkage for a subset of most well covered mutations (N = 34) and their nearby peaks (50kb extended up- and downstream). Unfortunately, no clear separation in distributions of ATAC signals can be observed from the cells grouped by the mutation status (Ref: reference, Alt: alternative or ND:not detected). Two example mutations with marginal significance (Wilcoxon rank sum test) are shown below (**Fig. 5**). We further extended the analysis by differential chromatin accessibility across all peaks between cell populations grouped by mutation status. Unfortunately, none out of 34 mutations suggest differentially accessible peaks after adjustment of multiple testing.

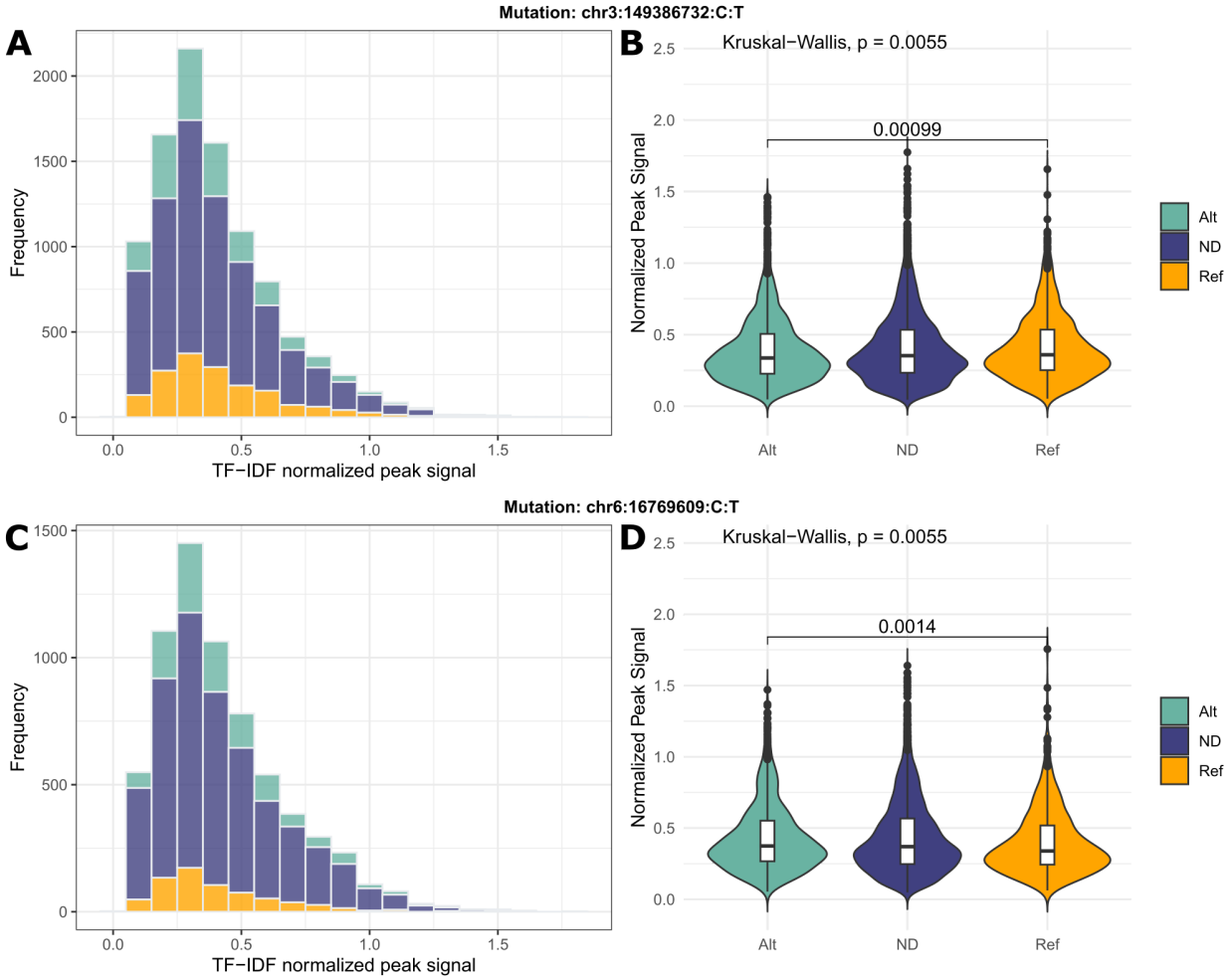

Figure 5: Distribution of weighted (TF-IDF) peak signals near (50kb up- and downstream) a zoomed-in somatic mutation, colored by cell mutation status. **(A)-(B)** Histogram and violin plus box plot for mutation chr6:16769609:C>T. **(C)-(D)** Histogram and violin plus box plot for mutation chr3:149386732:C>T. Group comparison between cell subpopulations carrying mutant (alternative) vs. reference allele were conducted with the Wilcoxon rank sum test.

Our code and scripts are available on GitHub at https://github.com/lijin0303/MLCB_project.

Future Goals

Despite the abject failure of CellPhy, there remains some possibility that a different phylogeny inference tool could extract a signal from the mutations calls. So, if we had more time, a next step would be to try a couple other methods for cell phylogeny inference, such as SCsnvcna [15] or CIMICE [16], to the scATAC-seq somatic variants. We would continue to focus on the ATAC-seq variants given the very poor coverage of the scDNA-seq. Unfortunately, the lack of concordance that we have observed with the known CNV-based phylogeny of this cell line suggests that the resulting SNV-based phylogenies may be, even in the best case, quite poor. We could also develop a more sophisticated method for constructing

phylogenies from the scATAC-seq data, as the hierarchical clustering does not seem to capture the high-level clonal structure. Additionally, though we recover the high-level clonal split structure in our UMAP analysis, we would like to do further work to improve resolution and capture the lower-level clonal splits by using additional filters such as expression levels, leveraging the additional sc-omics datasets available for the SNU601 cell line.

Comparison with Original Proposal

We performed a number of analyses during project execution that were not outlined in the original proposal. First, we also performed SNV calling on DNA-seq data from the same cell line to give us a set of predictions that we could use as a comparison point for the scATAC-seq results. Second, since we saw low consistency between our initial hierarchical clustering of the detected SNVs with the known CNV clonal identities, we performed pseudotime trajectory analysis to further investigate cellular subpopulation dynamics. Third, in response to reviewer feedback, we analyzed peak-mutation correlations and the effects of cell-level coverage.

The key omission compared to our proposal is that, in light of the poor coverage of the single-nucleotide mutations, we were unable to infer a complete phylogeny. We therefore could not perform the proposed comparison to the CNV-based phylogeny. However, in light of the poor concordance between all individual SNVs and the CNV clonal identities, we can speculate that any such comparison would have been quite negative.

Commentary on Our Experience

The datasets were well-suited for the approach, as the wide range of available single-cell sequencing data for the SNU601 cell line provided multiple avenues with which we could examine variation and clonal structure. If we were to start the project over, one thing we could think about is possibly integrating these additional sc-omics datasets to make it easier to identify SNVs and perform phylogeny construction. One of the challenging aspects of our project came from the lack of initial concordance between detected SNVs and known clonal states, which made it too difficult to construct a clean phylogeny from the detected SNVs. Additionally, due to the sparsity of the data, current state-of-the-art phylogeny inference methods struggled to converge in a reasonable amount of CPU time (i.e., the methods timed out on the Engaging cluster compute nodes).

However, it was rewarding to explore the use of a novel SNV detection framework on scATAC-seq data, which previously had not been done before, and find some initial promising results in the SNV calling results and high-level clonal structure present in the UMAP analysis. We found that we were able to balance our time between execution and writing well by working on the report throughout the semester, as well as allocating substantial time in advance for the final compilation of results and writing.

Division of Labor

D.S. processed the scATAC-seq and DNA-seq data and worked on mutation-based phylogeny inference. M.L. performed germline SNV calling on the scATAC-seq and DNA-seq data and calculation of the cell-by-mutation coverage matrix from the BAM files. R.L. performed somatic mutation calling on the scATAC-seq and DNA-seq with LD refinement, characterization of somatic mutations, cell clustering

based on cell by mutation profile; peak calling and downstream analysis to compare peak-based clusters and CNV-based ones. M.L., R.L., and D.S. all contributed to interpreting results and writing the proposal and report.

References

1. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194: 23–28.

2. Bertram JS. The molecular biology of cancer. Mol Aspects Med. 2000;21: 167–223.

3. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. Nat Rev Genet. 2017;18: 213–229.

4. Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet. 2020;21: 410–427.

5. Evrony GD, Hinch AG, Luo C. Applications of Single-Cell DNA Sequencing. Annu Rev Genomics Hum Genet. 2021;22: 171–197.

6. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. 2015;25: 1491–1498.

7. Muyas F, Sauer CM, Valle-Inclán JE, Li R, Rahbari R, Mitchell TJ, et al. De novo detection of somatic mutations in high-throughput single-cell profiling data sets. Nat Biotechnol. 2023. doi:10.1038/s41587-023-01863-z

8. Dou J, Tan Y, Kock KH, Wang J, Cheng X, Tan LM, et al. Single-nucleotide variant calling in single-cell sequencing data with Monopogen. Nat Biotechnol. 2023. doi:10.1038/s41587-023-01873-x

9. Wu C-Y, Lau BT, Kim HS, Sathe A, Grimes SM, Ji HP, et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. Nat Biotechnol. 2021;39: 1259–1269.

10. Andor N, Lau BT, Catalanotti C, Sathe A, Kubit M, Chen J, et al. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. NAR Genom Bioinform. 2020;2: lqaa016.

11. Kozlov A, Alves JM, Stamatakis A, Posada D. CellPhy: accurate and fast probabilistic inference of single-cell phylogenies from scDNA-seq data. Genome Biol. 2022;23: 37.

12. Meng Q, Wu X, Li C, Li J, Xi X, Chen S, et al. The full set of potential open regions (PORs) in the human genome defined by consensus peaks of ATAC-seq data. bioRxiv. 2023. p. 2023.05.30.542889. doi:10.1101/2023.05.30.542889

13. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. Nat Methods. 2021;18: 1333–1341.

14. Srivatsan SR, McFaline-Figueroa JL, Ramani V, Saunders L, Cao J, Packer J, et al. Massively multiplex chemical transcriptomics at single-cell resolution. Science. 2020;367: 45–51.

15. Zhang L, Bass HW, Irianto J, Mallory X. Integrating SNVs and CNAs on a phylogenetic tree from

single-cell DNA sequencing data. Genome Res. 2023. doi:10.1101/gr.277249.122

16. Rossi N, Gigante N, Vitacolonna N, Piazza C. Inferring Markov Chains to Describe Convergent Tumor Evolution With CIMICE. IEEE/ACM Trans Comput Biol Bioinform. 2023;PP. doi:10.1109/TCBB.2023.3337258