Integrating single nucleotide variants (SNVs), copy number alterations (CNAs), and structural variants (SVs) into single-cell clonal lineage inference

Nishat Anjum Bristy[1], Xuecong Fu[2], Russell Schwartz[1,2]
[1]Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, Pittsburgh PA USA
[2]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh PA USA

Motivation and Background

Tumors develop through somatic mutations, where distinct mutations appear in specific subpopulation of cells (subclones) within the tumor. Understanding the **progression** of these mutations play a crucial role in identifying clinically relevant aspects of tumors, such as identifying markers of progression, metastatic growth and therapeutic responses [1]. These mutations include **single nucleotide variants (SNV), copy number aberrations (CNA) and structural variants (SVs)**. Since tumors progress through all three types of variants, different tumor subclones may be driven by different type of variants or by multiple variant types acting in combination. However, methods for reconstructing clonal lineage trees ("tumor phylogenetics") have traditionally focused on SNVs and to a lesser extent on CNAs. For **single-cell DNA sequences**, there were previously no methods for integrating all of these variant types into a single tree model [4]. Moreover, while robust methods have been developed for bulk sequencing (TUSV, TUSV-ext [2,3]), single cell DNA sequencing data presents its own challenges due to its distinct error profile and data artifacts, including allelic dropouts and high false positive and negative rates [4]. Here, we introduce a methods for inferring clonal lineage trees using SNV, CNA and SV derived from single-cell DNA sequencing data. The method offers promise fro better leveraging single-cell sequencing to provide a **personalized** and **comprehensive understanding** of how a tumor progresses in a patient's body.

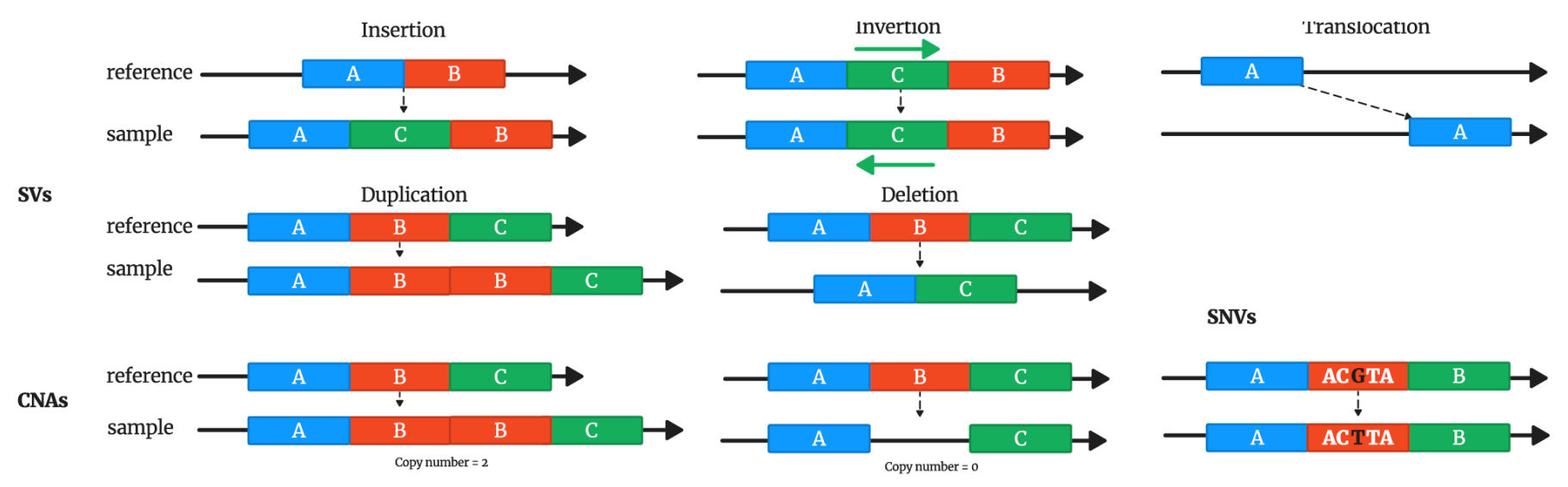

Figure 1: Common types of mutations observed in tumor evolution, including SNVs, CNAs, and SVs.

- **SNVs, CNAs and SVs are markers of tumor evolution.**
- **By analyzing how these somatic mutations accumulate over the course of clonal evolution, recurring features of tumor growth can be identified.**
- **We make use of all of these mutation types to build single-cell tumor evolutionary histories to help understand tumor progression.**

Methods

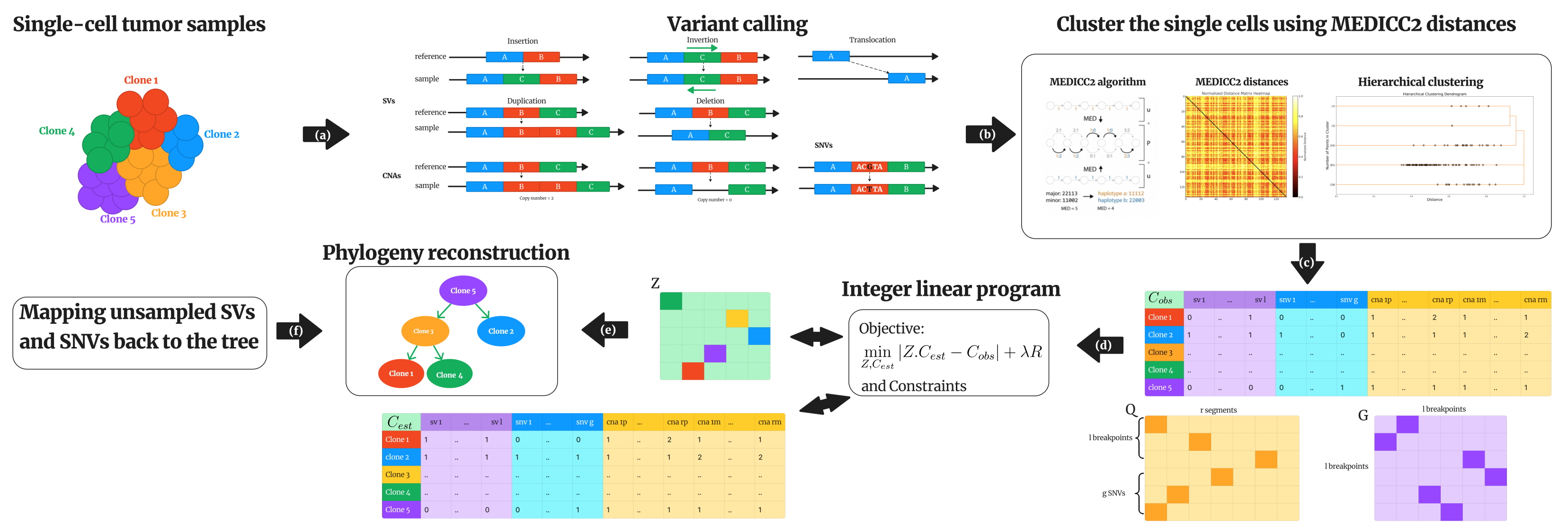

Figure 2: (a) Our method takes as input scDNA-seq variant calls for SNVs, CNAs and SVs. (b) Then, we cluster the single-cells using their MEDICC2 distances [5] into a user-defined number of clones based on a model of copy number evolution. (c) Next, we build a variant matrix C_{obs}; SNVs and SVs are represented as binary numbers, and CNAs are represented as positive integers. Since SVs are large scale genomic changes, they are represented as a pair of breakpoints (non-adjacent segments on the genome). We also construct SNVs and breakpoint-to-segment mapping matrix Q (each SNV and SV belongs to a corresponding segment in the genome), and breakpoint-to-breakpoint mapping matrix G (each SV is represented as a pair). (d) Then we reconstruct the evolutionary tree using a method based on integer linear programming (ILP), a widely used framework for solving hard optimization problems, that identifies a clonal lineage tree optimizing for the difference between observed and estimated variants and a minimum evolution tree reconstruction cost, R. (e) We build the evolutionary tree on a subsampled set of variants. (f) And at the end, we map the unsampled SNV and SV breakpoints back to the edges of the evolutionary tree.

Validation on Simulated Datasets

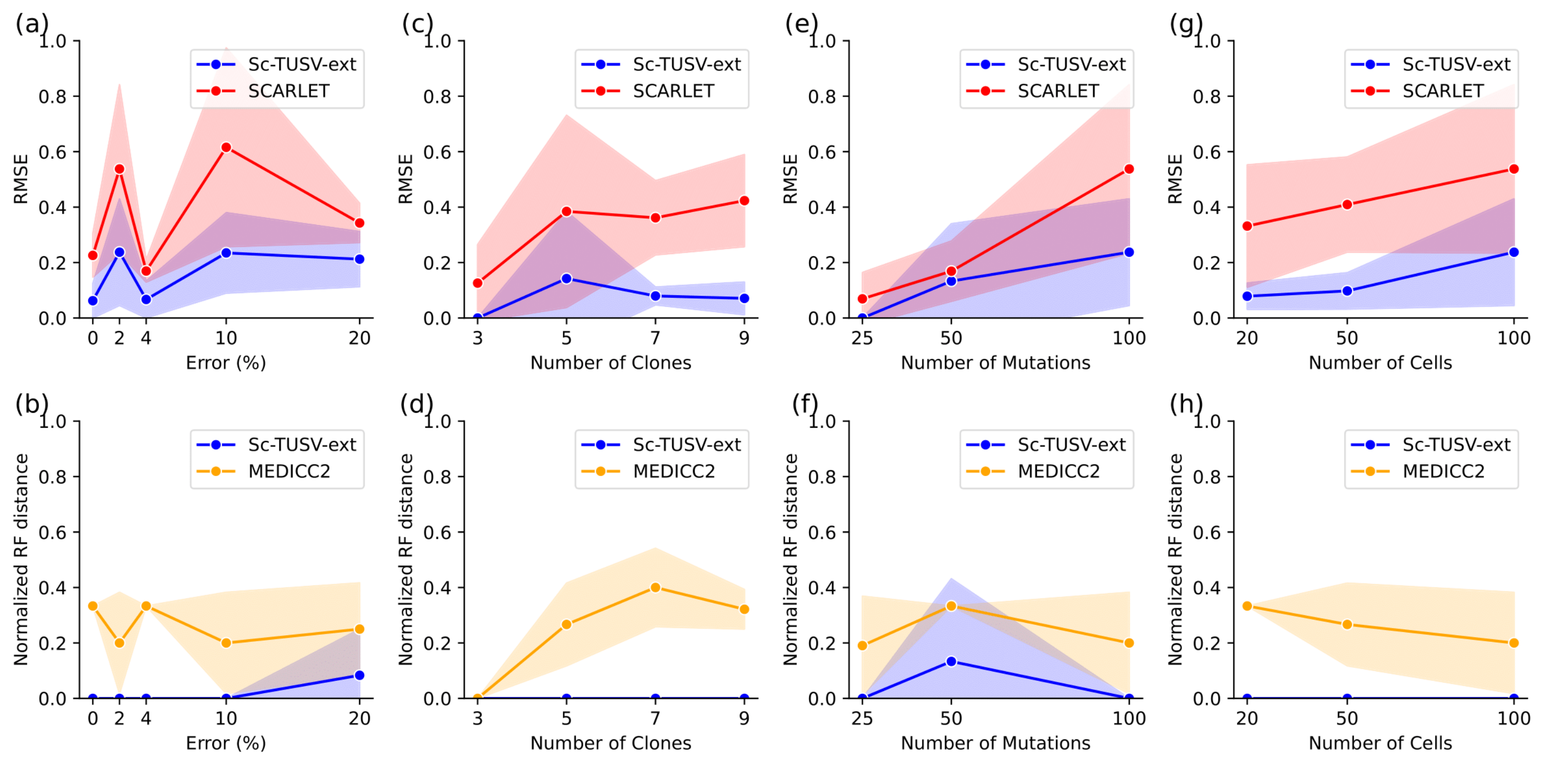

Figure 2: We validate the method using empirical datasets with different model conditions - (a) and (b): varying the level of errors in single-cell; (c) and (d): varying the number of tumor clones; (e) and (f) varying the number of SNV and SV mutations; (g) and (h): varying the number of single-cells.
(a), (c), (e) and (g) shows **root mean squared error (RMSE)** of the estimated SNVs with our method and SCARLET with different model conditions; (b), (d), (f) and (h) shows the **normalized RF distance** (a common method for comparing evolutionary trees) of the inferred phylogenies with our method and MEDICC2 for different model conditions. The solid lines represent the means of the simulations and shaded area represents one standard deviation on either side of the mean. In each case, our method has less errors compared to SCARLET's estimated copy numbers and MEDICC2's evolutionary trees.

Results on Real Biological Datasets

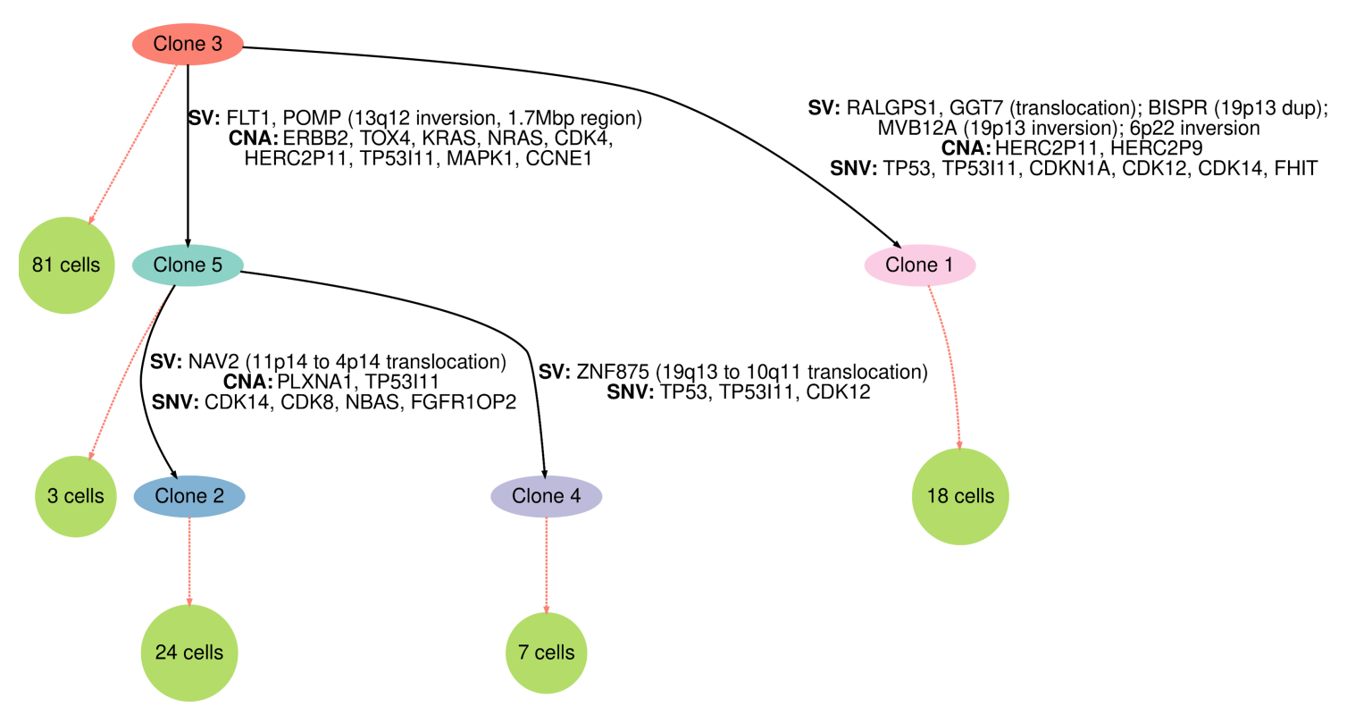
Clone 3
SV: FLT1, POMP (13q12 inversion, 1.7Mbp region)
CNA: ERBB2, TOX4, KRAS, NRAS, CDK4, HERC2P11, TP53I11, MAPK1, CCNE1
SV: RALGPS1, GGT7 (translocation); BISPR (19p13 dup); MVB12A (19p13 inversion); 6p22 inversion
CNA: HERC2P11, HERC2P9
SNV: TP53, TP53I11, CDKN1A, CDK12, CDK14, FHIT
81 cells
Clone 5
Clone 1
SV: NAV2 (11p14 to 4p14 translocation)
CNA: PLXNA1, TP53I11
SNV: CDK14, CDK8, NBAS, FGFR1OP2
SV: ZNF875 (19q13 to 10q11 translocation)
SNV: TP53, TP53I11, CDK12
3 cells
Clone 2
Clone 4
18 cells
24 cells
7 cells

Figure 3a: Sc-TUSV-ext inferred tree for a high grade serous ovarian cancer patient, DG1134 (Funnel et al [6]), with **133 single-cells**. The edge labels represent subsets of mapped SNVs, CNAs and SVs. The number of cells mapped to the clones are shown with the dotted arrows.

Findings:

- High amplification of CCNE1 gene in clones 2, 4 and 5; consistent with the primary paper [4].
- One branch (clone 3 to clone 1) is mainly driven by TP53 SNV mutation.
- The other sub-tree is mainly driven by CNA type KRAS, NRAS, CCNE1 and SV type FLT1, POMP mutations, then further subdivided by a combination of mutations.

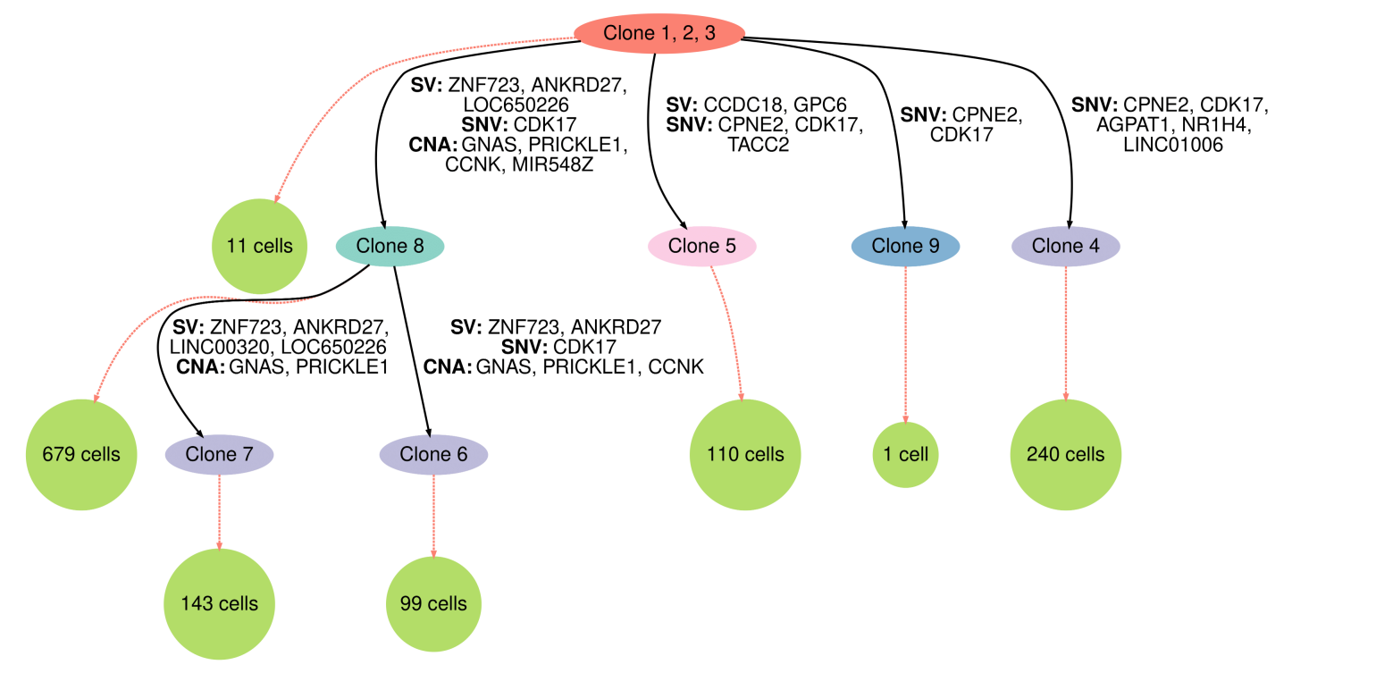

Figure 3b: Sc-TUSV-ext inferred tree for another high grade serous ovarian cancer patient, SA1049 (Funnel et al [6]), with **1283 single-cells**.

Findings:

- Almost all the SVs and CNAs belong to one part of the sub-tree (clone 1,2,3 to clone 8). While the other subtree is mostly SNV driven.
- This localization of events indicative of chromosome instability may be explained by a CCNK mutation in the first branch, as CCNK knockdown has been found to impair the DNA damage repair process.
- The recurring CDK17 mutations are likely an artifact of our method, for inferring an insufficient number of clones initially.

Discussion

- We introduced the first method to incorporate SNV, CNA and SV variants collectively in understanding tumor progression from single-cell data.
- Tests on simulated datasets demonstrate the robustness and accuracy of our method, and indicating how incorporating multiple data modalities can lead to improved accuracy relative to prior state-of-the-art methods.
- Application to real datasets demonstrate all variant type can collectively contribute to clonal evolution within a single cancer, as well as revealing potential differences in mutability phenotypes between sub-lineages and showing how the variant types may influence on another.

Future Direction and Limitations

Our method might be extended in a number of directions in future work:

- Incorporating complex structural variants like chromatin induced chromothripsis (chromosomal shattering).
- Automating identification of the number of clones from the dataset instead of using a user defined number.
- Inferring ancestral sub-clones.
- Developing better methods for handling large marker sets

Acknowledgements

We thank members of Schwartz Lab including Arjun Shrivastsa and Thomas Rachman, for their support and guidance. Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number R01HG010589. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Schwartz, R. and Schäffer (2017). The evolution of tumour phylogenetics: principles and practice.

[2] Fu, X. et al. (2022). Reconstructing tumor clonal lineage trees incorporating single-nucleotide variants, copy number alterations and structural variations.

[3] Eaton, J. et al. (2018). Deconvolution and phylogeny inference of structural variations in tumor genomic samples.

[4] Satas, G. et al. (2020). SCARLET: single-cell tumor phylogeny inference with copy-number constrained mutation losses.

[5] Kaufmann, T.L. et al. (2022). MEDICC2: whole-genome doubling aware copy-number phylogenies for cancer evolution.

[6] Funnell, T. et al. (2022). Single-cell genomic variation induced by mutational processes in cancer.