Benchmarking Multi-sample Subclonal Reconstruction Tools using TRACERx Data

JACOB VALENZUELA[1], Helena Winata[2], Yash Patel[3], Paul Boutros[3,4]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA 2 Graduate Programs in Biosciences, UCLA
3 Department of Human Genetics, David Geffen School of Medicine, UCLA 4 Institute for Urologic Oncology, David Geffen School of Medicine, UCLA

Introduction

Subclonal reconstruction (SRC) provides a framework for studying genetic diversity within tumors, highlighting subpopulations of cancer cells, key mutations, and understanding mutation timing and origins. Initially performed with single biopsy samples, this method offered limited insight into the full spectrum of tumor mutations. Multi-sample reconstruction significantly improves resolution, but most methods struggle to accurately and efficiently reconstruct evolutionary paths from such data. To overcome this, we developed Efficient Multi Sample Inference of Cancer Phylogeny (EMulSI-Phy), which uses a rule-based approach that leverages fundamental principles of cancer evolution. Through this study we aim to demonstrate its capabilities relative to established methods in the field.

Background

The TRACERx dataset, the basis for our benchmarking study, consists of comprehensive genomic and clinical data from 421 non-small cell lung cancer (NSCLC) patients. This dataset is particularly valuable because it includes multi-sample whole-exome sequencing data from various stages of tumor development, allowing us to capture the complexity of tumor evolution. By leveraging this data, we can rigorously evaluate EMulSI-Phy against other leading tools such as PyClone, PyClone-VI, and CONIPHER. Our comparison focuses on essential metrics like runtime, memory usage, and accuracy in reconstructing tumor evolutionary history. This benchmarking effort is crucial for establishing a standardized pipeline for SRC, ultimately improving the tools available for cancer research.

Results

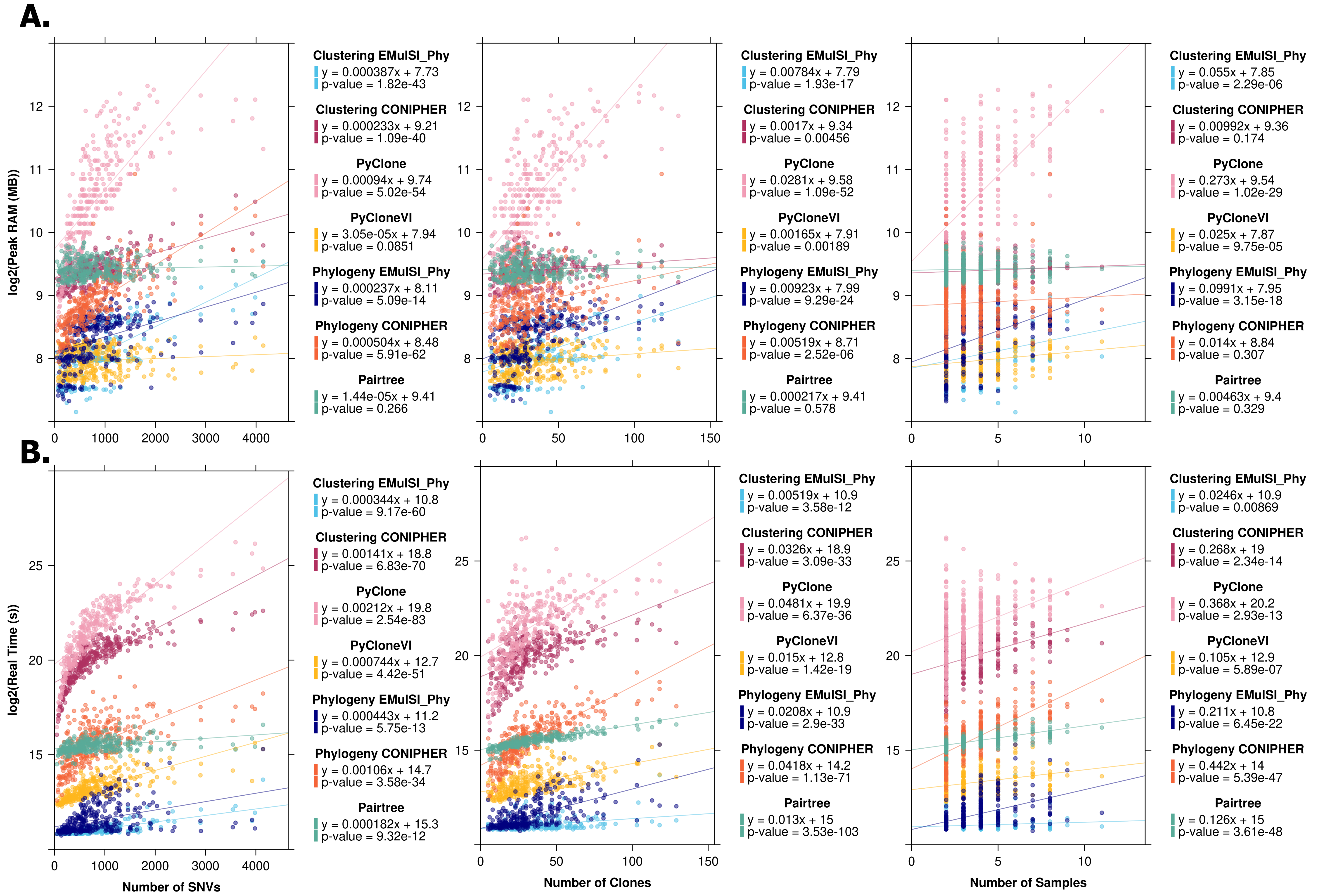

Figure 2. Comparison of computational resource requirements for EMulSI-Phy, CONIPHER, Pairtree, PyClone, and PyClone-VI. Each column shows scatter plots with linear regression, plotting metrics against covariates: Number of SNVs (left), Number of clones (middle), and Number of samples (right). The figure displays A) peak memory usage and B) runtime for each tool

Pairwise Wilcoxon Test Ranking (Highest to Lowest). *EMulSI-Phy has the shortest runtime for the clustering and phylogeny step, significantly (p-value < 2e…) outperforming the next fastest tool. The slope of the linear regression suggests that EMulSI-Phy's runtime increases more slowly with the number of clones and samples compared to all other tools. In terms of memory requirements, PyClone-VI and EMulSI-Phy perform equally well with no significant difference (p-value), both excelling in the clustering step, , while EMulSI-Phy excels significantly (p-value) in the phylogeny step.*

Methods

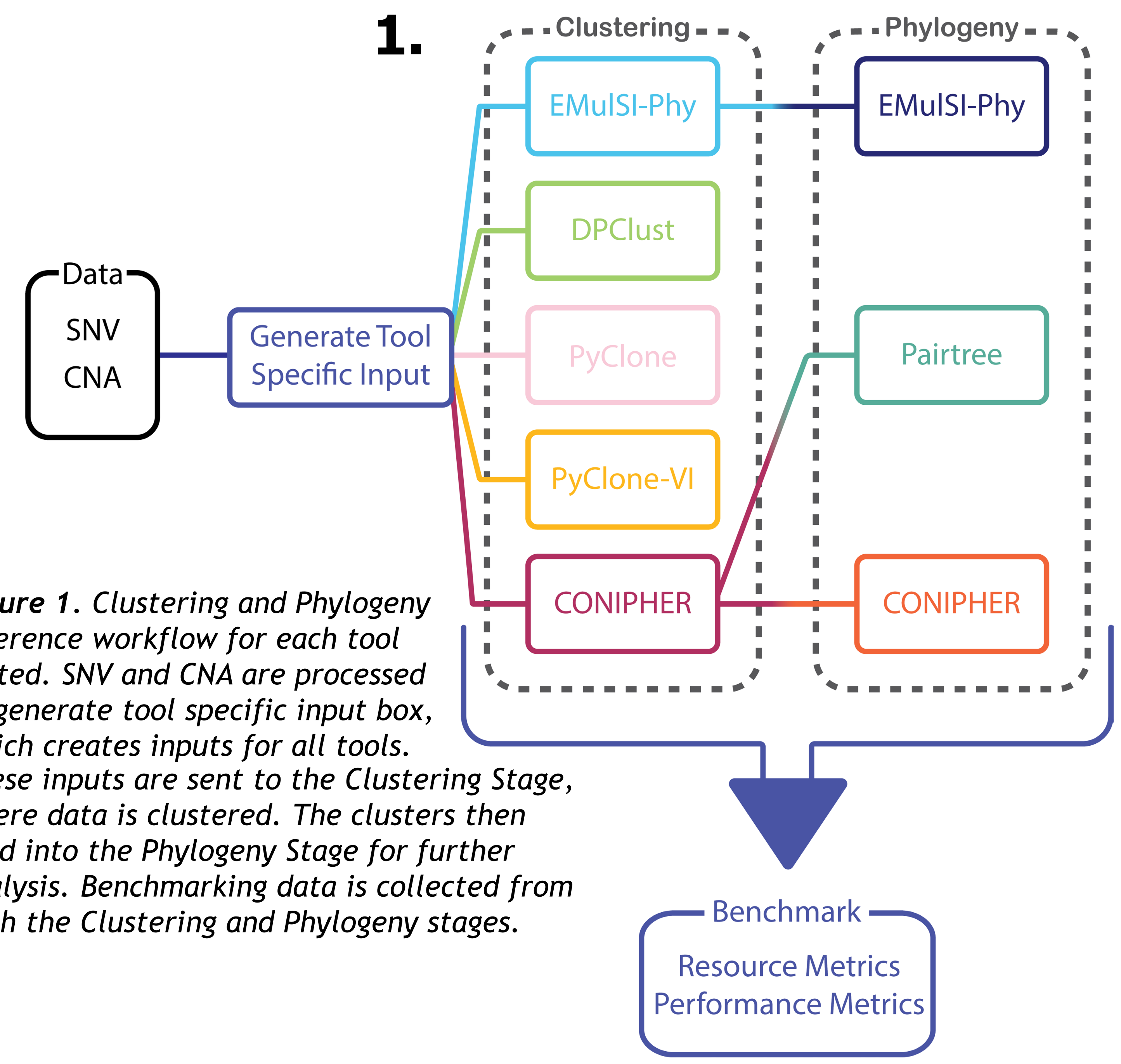

Figure 1. *Clustering and Phylogeny inference workflow for each tool tested. SNV and CNA are processed to generate tool specific input box, which creates inputs for all tools. These inputs are sent to the Clustering Stage, where data is clustered. The clusters then feed into the Phylogeny Stage for further analysis. Benchmarking data is collected from both the Clustering and Phylogeny stages.*

- Runtime and memory usage across tools are evaluated using pairwise Wilcoxon tests, with Bonferroni-adjusted p-values used to determine statistical significance.
- To evaluate performance, we will compare the clustering and phylogeny results to the published data using the Adjusted Rand Index (ARI) and Pearson's correlation, respectively. Additionally, to assess clustering independently of the ground truth, we will calculate the average silhouette width (ASW) and perform pairwise Wilcoxon tests on the ASW values.

Conclusion

In the benchmarking analysis, EMulSI-Phy performed well, ranking lowest in memory and runtime usage across most categories. Future work should focus on validating EMulSI-Phy's performance further by scoring it against the top-performing tools. This would strengthen the case for EMulSI-Phy as an efficient tool that maintains high SRC accuracy while minimizing resource consumption.

References

1. Al Bakir et al. The evolution of non-small cell lung cancer metastases in TRACERx. Nature 616, 534-542 (2023)
2. Roth et al. PyClone: statistical inference of clonal population structure in cancer. Nat Methods 11, 396–398 (2014
3. Gillis S, Roth A. PyClone-VI: scalable inference of clonal population structures using whole genome data. BMC Bioinformatics. 2020 Dec 10;21(1):571
4. Wintersinger et al. Reconstructing Complex Cancer Evolutionary Histories from Multiple Bulk DNA Samples Using Pairtree. Blood Cancer Discovery, 3(3):208-219, 2022