Impact of Seed Selection on Subclonal Reconstruction Solutions

ANNA NEIMAN-GOLDEN*, PHILIPPA STEINBERG*, Lydia Y. Liu, Takafumi N. Yamaguchi, Yash Patel, Paul C. Boutros

Boutros Laboratory, Department of Human Genetics, David Geffen School of Medicine, UCLA

OVERVIEW

CANCER BACKGROUND

Many tumors start from a single cell that contains somatic driver mutations[1]. This ancestral cell replicates, and its descendants accumulate further mutations. Some cells are able to outcompete their neighbors and form distinct cell clusters with shared mutations (subclones)[1]. This results in intra-tumoral heterogeneity, which makes cancer prognosis and personalized treatment more difficult.

Subclonal reconstruction (SRC) quantifies intra-tumoral heterogeneity by looking at subclone attributes, such as the number and genotype of subclones, the cancer cell fraction (CCF), and aims to decipher how the subclones evolved through space and time[1].

KEY SRC TERMS

Single Nucleotide Variant (SNV): Mutation at a single base in the genome[2].

Copy Number Aberration (CNA): Deletion or amplification of large genomic regions[3].

Subclone: Population of cancer cells that is the descendant of an ancestral cell with a distinct set of mutations[2]. (SRC output: Unique clusters.)

Cancer Cell Fraction (CCF): Proportion of cancer cells with the mutation or belong to a subclone[2].

RESEARCH QUESTION

1. How much do SRC pipeline results vary based on tools used?

2. How much do SRC pipeline results vary due to the initializing seed?

DATASET

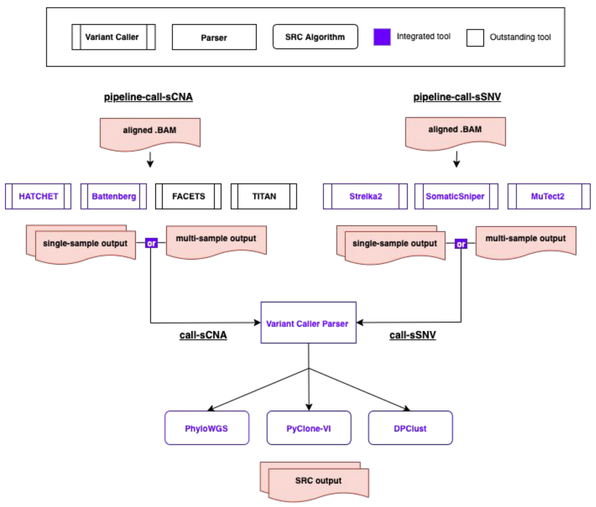
We received head and neck tumor samples from 14 patients with 1 primary and 2 lymph samples each.

SEED SELECTION

A seed is a starting value for random number generation in a probabilistic algorithm. A random seed ensures reproducibility of results and does not introduce noise that impacts the algorithm outcome, such as generating more statistically related numbers than by random chance. To minimize patterns or relatedness between seeds we generated 10 random seeds using a pseudo-random number generator (the python function *random.sample()*) initialized with an integer of random bytes generated through the command line.

DEVELOPING & RUNNING THE PIPELINE

EXPANDING THE SRC PIPELINE TO MORE TOOLS



The Boutros lab has developed subclonal reconstruction (SRC) pipelines that use somatic single nucleotide variant (sSNV) and somatic copy number aberration (sCNA) mutation caller output as input to SRC algorithms that quantify intra-tumoral heterogeneity. We have expanded the pipeline by creating parsers that extract the *mutation id*, *sample name*, *reference reads,* and *variant reads* from different sSNV-callers and tailored this variant data to the SRC algorithm input requirements. The parsers work for sSNV-caller multi-sample (all samples in one VCF) and single-sample (each sample in one VCF) mode. Additionally, the parser for the sCNA-callers extracts s*ample purity*, and distinguishes between clonal and subclonal CNA.

VARIABILITY ACROSS SRC PIPELINE TOOLS

We tested our data on three SRC pipeline combinations. We conducted both single-region (sr) (SRC on one sample) and multi-region (mr) (SRC on multiple samples).

- **<SNV-caller>-<CNA-caller>-<SRC-algoritm>(mode)**
- Strelka2-Battenberg-DPClust (sr) (14 patients)
- Strelka2-Battenberg-PyClone-VI (sr) (14 patients)
- Strelka2-Battenberg-PyClone-VI (ms) (7 patients)

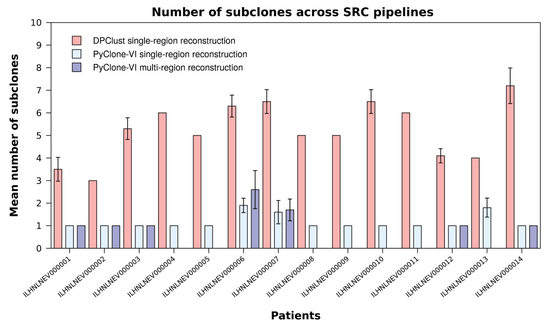


Figure 1: Variability Across SRC Algorithm

Previous research shows that the algorithm choice impacts the SRC results. We ran 3 SRC pipelines per patient across 10 random seeds. We extracted the number of subclones, CCF and number of SNVs per subclone. Subclone counts of 1 indicate all mutations belong to 1 cluster (monoclonal solutions). DPClust (sr) tends to call the highest number of clusters (polyclonal solutions) with error bars showing a standard deviation of 1. Running PyClone-VI (mr) (each patient with 1 primary and 2 lymph tumors) results in counting more subclones than PyClone-VI (sr).

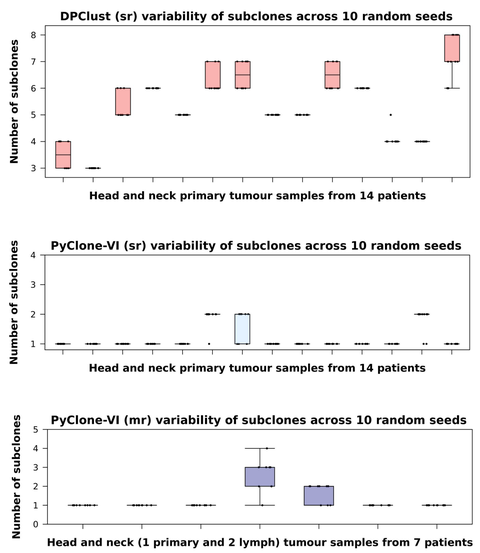


Figure 2a-c: Variability Across Seeds

Boxplots, showing median, Q1, Q3, whiskers as 1.5x the IQR and outliers of subclones per patient. For the PyClone-VI pipelines, we observe mostly the same number of clusters (1) across seeds per patient. We need to test further whether we see lower variability in the number of subclones specifically for monoclonal solutions. The DPClust algorithm only gave polyclonal solutions, which show greater variability in number of subclones across seeds for most patients.

RESULTS

SEED VARIABILITY

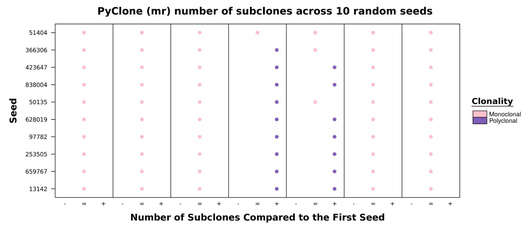


Figure 3: Relative Seed Variability

Compares the relative variability in number of subclones across seeds for the Strelka2-Battenberg-PyClone-VI multi-region pipeline. The reference seed (51404) classifies monoclonal solutions for all patients. For two patients, several seeds classify multiclonal solutions. There is variability, as all seeds differ in the number of subclones for at least one sample compared to the reference seed.

DISCUSSION

This project is the starting point for studying the impact of random seeds on subclonal reconstruction solutions. To strengthen our initial findings, we need to expand our analysis to more tool combinations, test more initializing seeds, apply our pipeline to larger tumor datasets, and quantify the impact on clinical settings.

Standardizing the pipelines across more SRC tools and quantifying variance due to seed selection will improve the reproducibility and accuracy of studying cancer evolution.

Further, seed selection is relevant to all fields that develop new statistical methods. This project is an appeal to other scientists to consider reporting their seed choices in their publications in all bioinformatics applications.

REFERENCES

[1] Salcedo, A. *Nat Biotechnol* **38,** 97–107 (2020).

[2] Tarabichi, M. *Nat Methods* **18,** 144–155 (2021).

[3] Zeira, R. *Bioinformatics*, **36**, i344–i352 (2020).

ACKNOWLEDGEMENTS

This project was funded by the National Institute of Health as part of the BIG summer research program. Special thanks to all members of the Boutros lab.

Contact pstein@berkeley.edu