Comprehensive Benchmarking of Structural Variant Callers

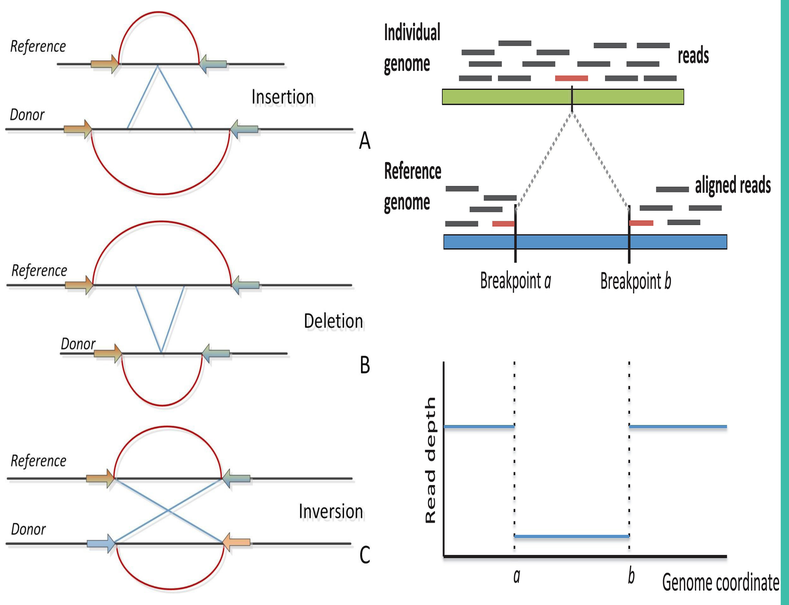
Varuni Sarwal[1], Ram Ayyala[1], Jacqueline Castellanos[1], Emily Wesel[1], Serghei Mangul[2,3],Eleazar Eskin[2,3]
[1] BIG Summer Program, Institute for Quantitative and Computational Biosciences,
[2] Department of Computer Sciences, UCLA
[3] Department of Human Genetics, UCLA

OBJECTIVES

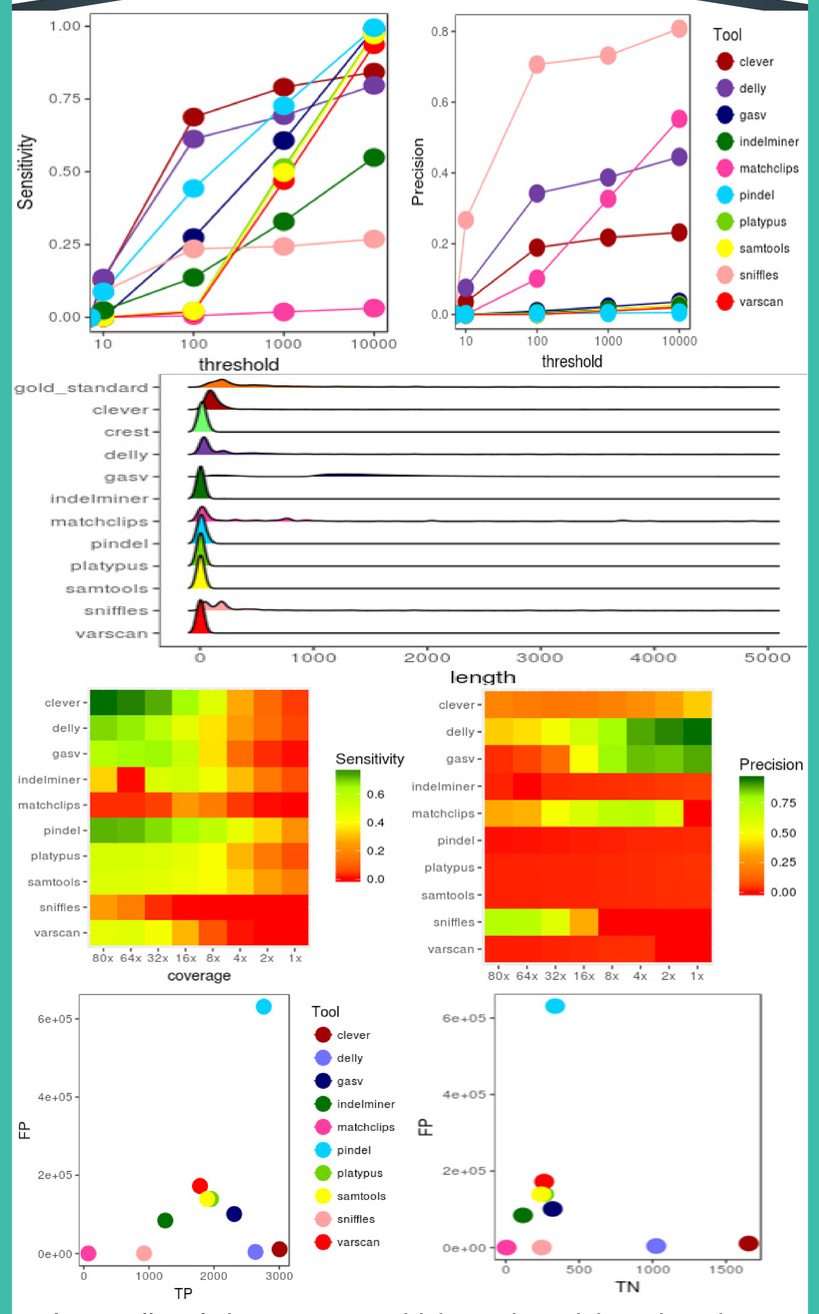
To comprehensively benchmark structural variant callers.

BACKGROUND

- Structural variants: Structural variation (SV) is generally defined as a region of DNA approximately 1 kb and larger in size and can include inversions, translocations, insertions and deletions, commonly referred to as copy number variants (CNVs)
- Deletions: A deletion is a type of structural variant where a region of DNA is absent from an individual but is present in the reference genome
- We chose deletions for our study because deletions are the easiest to detect



RESULTS



- Almost all tools have zero sensitivity and precision when the threshold is zero.
- Both sensitivity and precision increase with increasing threshold With increasing coverage, sensitivity generally increases and precision decreases.
- Breakdancer, lumpy and a pseudo tool, formed by combining delly, sniffles and clever, under different length ranges are the tools with the best balance of sensitivity and precision.
- Many tools overpredict deletions and have a high false positive rate, leading to a very high precision and a close to zero sensitivity.

METHODS

- We ran the tools on our raw data of 8 different strains of mouse chromosome 19, as well as nanopore sequenced human data.
- We compared the deletions detected by the tool and the deletions in a dataset consisting of PCR verified deletions, that we considered to be our gold standard,over a range of threshholds(0-10,000 bp) to calculate the sensitivity and precision of each tool.
- Because our data was high coverage, we subsampled our data over a wide range of coverages (80x-1x) to determine the effect of coverage on the performance of different tools.
- We also split the result of each tool based on the length of deletion predicted (0-50bp, 50-500bp, 500-1000bp, >1000bp) to study the behavior of each range.

FUTURE WORKS

- This study can be extended to study other structural variants like insertions, inversions and duplications to verify if the results are consistent.
- The same analysis can be preformed on other chromosomes of mice.

ACKNOWLEDGEMENTS

UCLA Computer Science
UCLA ZARLAB
UCLA QCBio