Long-range Somatic Structural Variation Calling from Matched Tumor-Normal Co-assembly Graphs

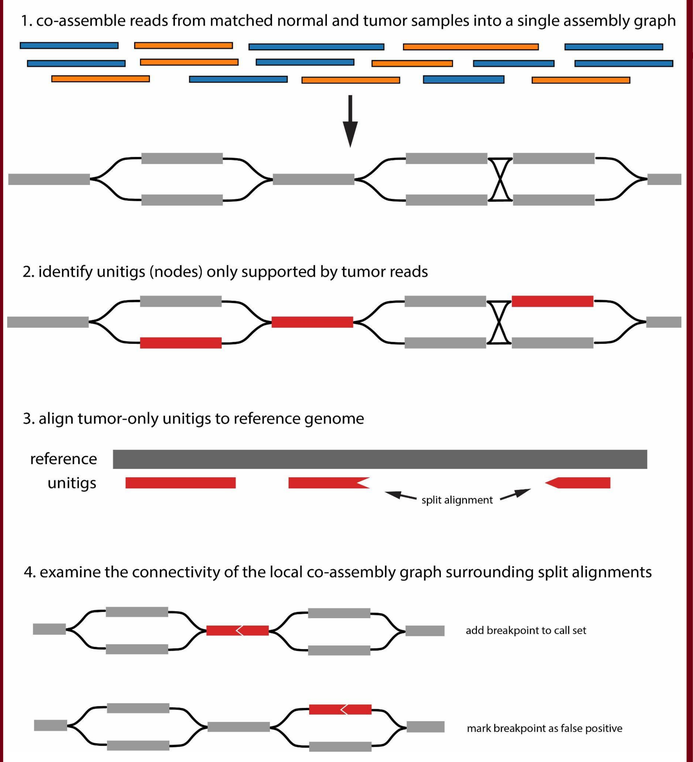
Megan K. Le[1], Qian Qin[2], Heng Li[3,4]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology; [2]Broad Institute of MIT and Harvard;
[3]Department of Biomedical Informatics, Harvard Medical School; [4]Department of Data Science, Dana-Farber Cancer Institute

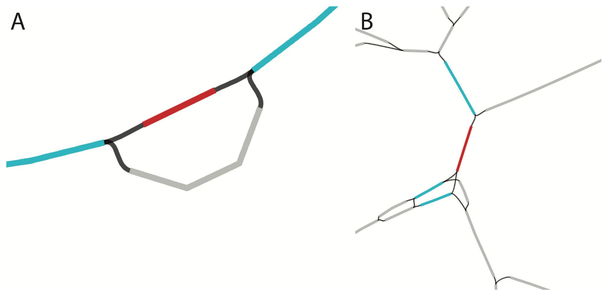
Motivations

- Large-scale somatic structural variations (SVs; rearrangements of segments of the genome at least 50 bp long) have been shown to play an important role in cancer development[1-4]
- However, existing **somatic SV callers still struggle with achieving high accuracy**, particularly when evaluated on precision
- **Co-assembly-based approaches**—in which reads from multiple samples are combined to create a single joint assembly—**have not yet been used for somatic SV calling** despite being successful for other applications (such as SV calling in microbiomes and copy number variation detection)[5-7]
- In this work, we developed colorSV, a method that identifies long-range SVs by examining the local structure of joint tumor-normal assembly graphs

Methodology



Methodology (cont.)



- The intuition behind our method is that **false positive** candidate breakpoints tend to be characterized by a **bubble-like topology** (panel A), where a parallel path of non-tumor-only unitigs also connect the candidate's neighbors
- **True somatic breakpoints** tend to connect two otherwise **locally disconnected subgraphs** (panel B), which correspond to distant portions of the genome (e.g., separate chromosomes)

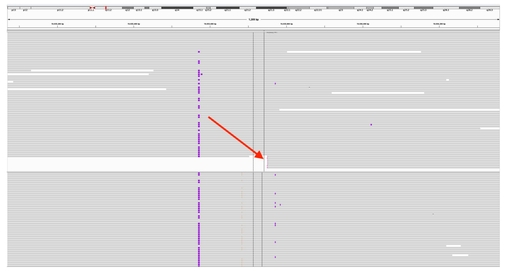
Results

Evaluation on the COLO829 Cell Line

Method	Sensitivity	Precision
colorSV	11 / 12	11 / 11
Sniffles2[8]	3 / 12	3 / 4
nanomonsv[9]	3 / 12	3 / 4
Severus[10]	10 / 12	10 / 11
SAVANA[11]	6 / 12	6 /7

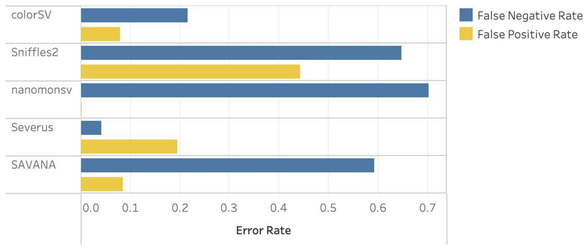
- We evaluated colorSV's ability to identify translocations against four other somatic SV callers using a reference call from Espejo Valle-Inclán *et al.*[12]
- **colorSV outperformed all other callers** in both metrics
  - colorSV identified a translocation that was not identified by any other caller
  - colorSV did not report a false positive that was reported by all other SV callers

Results (cont.)



- The other SV callers were unable to identify colorSV's novel translocation because they relied on alignments of ~20kb HiFi reads, and alignments at that position have a low mapping quality due to a large segmental duplication
- colorSV was able to extract the signal by aligning a ~60kb unitig to the region

Evaluation on the HCC1395 Cell Line



- To evaluate the ability of each tool to call translocations on the HCC1395 cell line, we combined the call sets of the other methods to generate reference sets
- colorSV had the **second highest approximated sensitivity**, being only outperformed by Severus
- colorSV also had the **second highest approximated precision,** being only outperformed by nanomonsv

	Sensitivity			Precision		
	Total Reference Set Size	Unique Calls in Reference Set	False Negative Rate	Total Call Set Size	Unique Calls in Call Set	False Discovery Rate
colorSV	60	13	0.2167	101	8	0.0792
Sniffles2	85	55	0.6471	79	35	0.443
nanomonsv	91	64	0.7033	28	0	0
Severus	47	2	0.0426	123	24	0.1951
SAVANA	81	48	0.5926	47	4	0.0851

Estimated false negative and false discovery rates for each evaluated SV caller. The reference sets used to calculate the false negative rates consisted of variants reported by at least two other methods. The reference sets used to calculate the false discovery rates were generated by taking the union of the other methods' call sets.

Discussion

- colorSV demonstrates **improved sensitivity and precision** over existing state-of-the-art methods for calling translocations on the COLO829 and HCC1395 cell lines
- By using an approach that leverages information from *de novo* co-assembly, colorSV is **less susceptible to errors that may arise as a result of germline SVs**
- The **use of unitigs rather than individual reads** for performing breakpoint identification **facilitate more accurate mapping** and subsequent SV detection
- colorSV is **limited by its reliance on current assembly tools** being able to generate accurate co-assembly graphs, meaning it is more likely to fail near complex regions or events
- This approach may be extended by using different criteria in the topology search to identify different types of structural variation
- The colorSV code and executable are available at github.com/mktle/colorSV

References

1. Zhang, Y. *et al.* Global impact of somatic structural variation on the DNA methylome of human cancers. *Genome Biol.* **20**, 209 (2019).
2. Hamdan, A. & Ewing, A. Unravelling the tumour genome: The evolutionary and clinical impacts of structural variants in tumourigenesis. *J. Pathol.* **257**, 479–493 (2022).
3. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
4. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event 364 during cancer development. *Cell* **144**, 27–40 (2011).
5. Nijkamp, J. F. *et al.* De novo detection of copy number variation by co-assembly. *Bioinformatics* **28**, 3195–3202 (2012).
6. Curry, K. D. *et al.* Reference-free Structural Variant Detection in Microbiomes via Long read Coassembly Graphs. 2024.01.25.577285 Preprint at https://doi.org/10.1101/2024.01.25.577285 (2024).
7. Xiao, C. *et al.* Personalized genome assembly for accurate cancer somatic mutation discovery using tumor-normal paired reference samples. *Genome Biol.* **23**, 237 (2022).
8. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* 1–10 (2024) doi:10.1038/s41587-023-02024-y.
9. Shiraishi, Y. *et al.* Precise characterization of somatic complex structural variations from tumor/control paired long-read sequencing data with nanomonsv. *Nucleic Acids Res.* **51**, e74 378 (2023).
10. Keskus, A. *et al.* Severus: accurate detection and characterization of somatic structural variation in tumor genomes using long reads. *MedRxiv Prepr. Serv. Health Sci.* 2024.03.22.24304756 (2024) doi:10.1101/2024.03.22.24304756.
11. Elrick, H. SAVANA. GitHub https://github.com/cortes-ciriano-lab/savana. Cortes-Ciriano lab (EMBL-EBI) (2023).
12. Espejo Valle-Inclan, J. *et al.* A multi-platform reference for somatic structural variation detection. *Cell Genomics* **2**, (2022).