Merging, Annotation, Validation, and Illustration (MAVIS) of Structural Variants from Long-Read Genome Sequencing

Jeremy Fan, Caralyn Reisle, Katherine Dixon, Kieran O'Neill, Steven Jones

Canada's Michael Smith Genome Sciences Centre, BC Cancer

Background

- Structural variants (SVs) are a class of DNA modifications characterized by large scale (>50 nucleotide) changes[1]
- SVs are known to have disproportionately large role relative to their abundance in the biology of rare diseases and cancer[1]
- SVs are classified as insertions, deletions, duplications, translocations, and inversions.
- Nanopore sequencing is long read (LR), amplification-free, pore-based sequencing technique[2]
- Whole genome DNA sequencing using Nanopore sequencing can reduce mapping ambiguity and resolves complex SVs of repetitive regions[2]
- Merging, Annotation, Validation, and Illustration (MAVIS) has been previously developed for short reads, but LR sequencing presents its unique challenges such as lower base calling quality

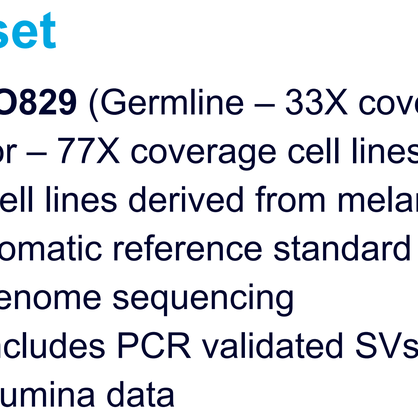
Objectives

1. Investigate and benchmark LR SV callers
2. Allow MAVIS to be compatible with LR input

Dataset

- **COLO829** (Germline – 33X coverage and Tumor – 77X coverage cell lines)
  - Cell lines derived from melanoma patient
  - Somatic reference standard for cancer genome sequencing
  - Includes PCR validated SVs from Illumina data

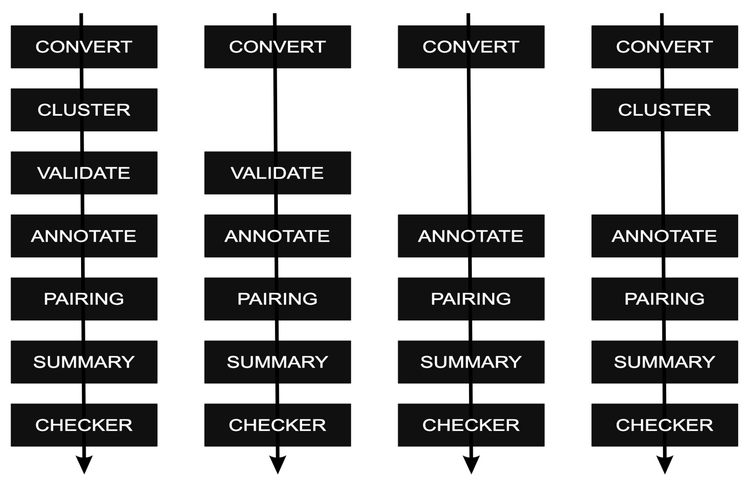
- **GM24385** (~35X coverage)
  - Genome in a Bottle gold standard containing 12,745 insertion and deletion
  - Data from son of an Ashkenazim Trio
  - Used as ground truth dataset



CIRCOS Figure of SV calls

Randomly selected 25000 SV observed in the COLO829 germline sample. Without post processing, determining relevant and clinically relevant SVs are a challenge.

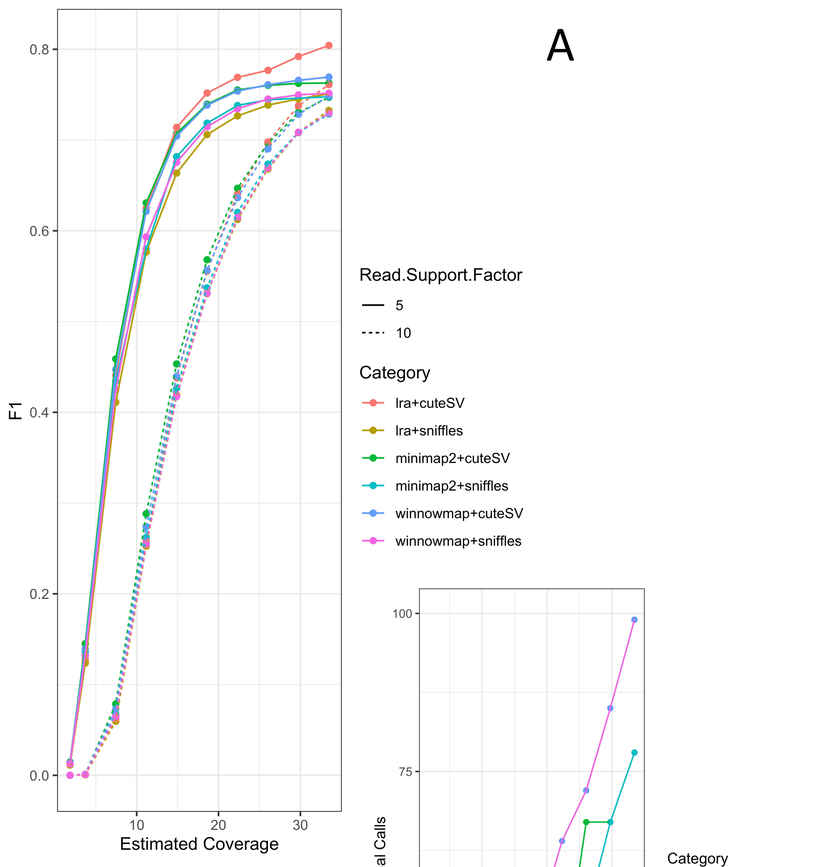
Methods



MAVIS Pipeline

MAVIS is split into a series of modules. Tool outputs are converted into a common MAVIS format, MAVIS clusters SVs based on proximity and type, validates the breakpoint using local breakpoint assembly, annotates with existing databases, pair transcriptional and genomic information together before ultimately.

Results



(A) F1 score of LR aligner and caller at 10 and 5 read support at different coverages (B) Number of translocations detected across different coverages in different LR aligner and callers

GM45385 dataset was serially downsampled by randomly selecting reads to replicate different levels of coverage.

Table 1. COLO829 LR calls were compared with previously PCR validated calls. Numbers of validated SVs with no support from manual interrogation of the alignment file.

Structural Variant Type	cuteSV called, (%)	Sniffles called, (%)	Number of validated SVs filtered out
Translocations	0 (0)	1 (50)	2
Tandem duplications	0 (0)	2 (50)	0
Deletions	18 (78)	23 (100)	2
Inversions	4 (80)	5 (100)	1
All Variants	22 (69)	31 (97)	1

Table 2. Adaptations made to MAVIS for LR, their corresponding rationale and the module they can be found within.

Adaptations	Rationale	Module
Integrate support of LR-SV caller input	LR-SV callers generate call patterns not seen from Illumina	Convert
Generate recommendation for blacklisted genomic regions	Filter out calls in known problematic regions in the genome due to sequencing artifacts	Cluster
Allow MAVIS file inputs for annotation of existing database	Allows for users to upload custom sets of annotations and of known SV calls	Summary

Conclusions

From benchmarking:

- Use minimap2 for aligner, LRA cannot identify translocations and custom flags are annoying to deal with
- A combination of Sniffles and cuteSV should be used

Future Directions

- Local breakpoint assembly should be implemented to verify breakpoint locations
- Illumina vs Nanopore MAVIS output comparisons should be done on known POG cases

References

1. Cretu Stancu, M., van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., de Ligt, J., Pregno, G., Giachino, D., Mandrile, G., Espejo Valle-Inclan, J., Korzelius, J., de Bruijn, E., Cuppen, E., Talkowski, M. E., Marschall, T., de Ridder, J., & Kloosterman, W. P. (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature communications, 8(1), 1326. https://doi.org/10.1038/s41467-017-01343-4
2. Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Partners