Structural Variant Detection Tools Struggle with Whole Exome Sequencing (WES) Data

Lokesh Pugalenthi*, Rahul Nanduri*, Raymond Hong*, Rohit K. Prasad†, Dhivya Arasappan†, Jeanne Kowalski-Muegge‡

* Undergrad Research Assistant, College of Natural Sciences, University of Texas at Austin; †Faculty Collaborator, College of Natural Sciences, University of Texas at Austin;
‡ Faculty Collaborator, Live**STRONG** Cancer Institute, Dell Medical School, University of Texas at Austin

Detecting Large Variants in Multiple Myeloma can personalize treatment

- Whole Exome Sequencing (WES) provides a snapshot of the sample's exonic regions (exome).
- By comparing this to a typical exome, we can identify structural variants (SVs) in the sample.
- These SVs may play a role in the development of diseases such as multiple myeloma (MM).

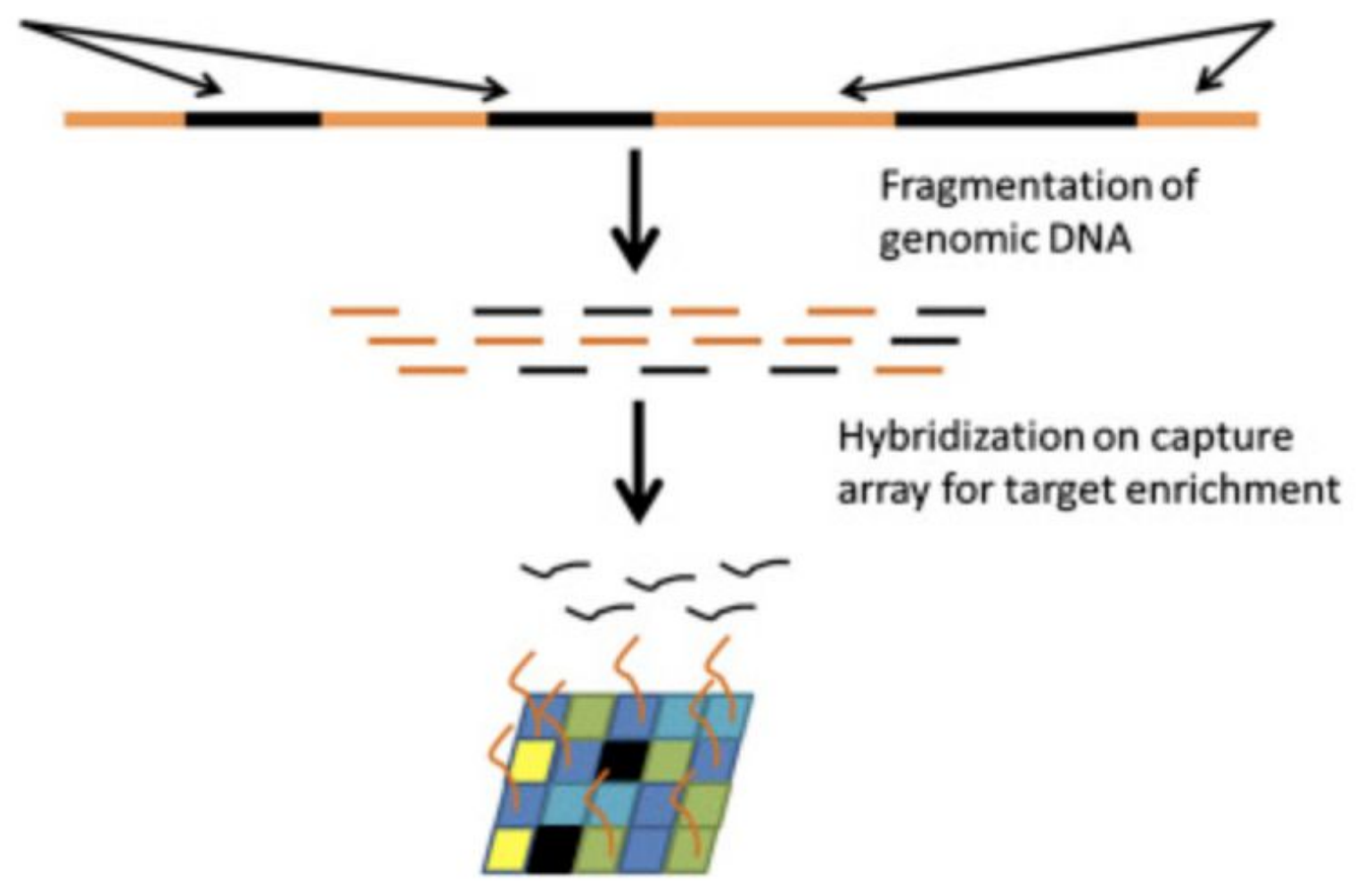
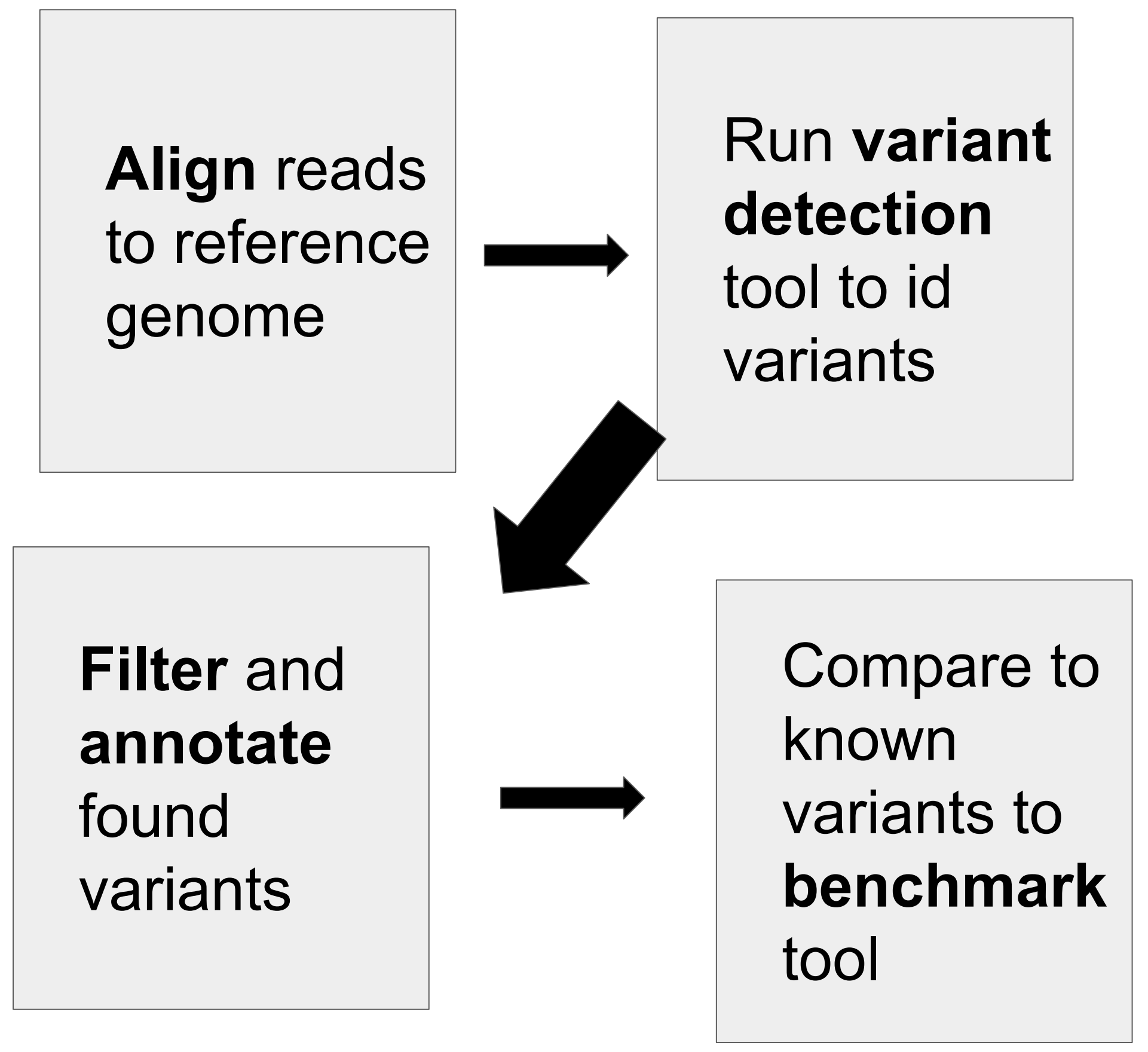

Fragmentation of genomic DNA
Hybridization on capture array for target enrichment

Figure 1: The methodology to Whole Exome Sequencing; Genomic DNA is fragmented and only regions that hybridize to the capture array (exons) are isolated and sequenced, providing a view of the exome

- Detecting SVs can lead to more personalised, affordable cancer treatment by predicting antigens specific to the SVs present in the patient's cancer cells (neoantigen).
- As WES is much cheaper than Whole genome sequencing (WGS), effective SV detection on WES data would drastically reduce the costs of detecting a patient's novel cancer-causing SVs
- However, as most SV detection tools are designed for WGS data, it's unclear how well they work with WES data.
- To this end, we benchmarked SV detection tools on 71 MM WES cell lines.

General Workflow



Align reads to reference genome

Run **variant detection** tool to id variants

Filter and **annotate** found variants

Compare to known variants to **benchmark** tool

Tools' output vary in SV type & length

Structural variants are alterations to the genome typically spanning more than a few hundred base pairs. The 6 tools benchmarked detected different types of SVs of different lengths.

Figure 2: Proportion of SV types detected by each tool.
Legend: BND: Breakend. DEL: Deletion, INS: Insertion

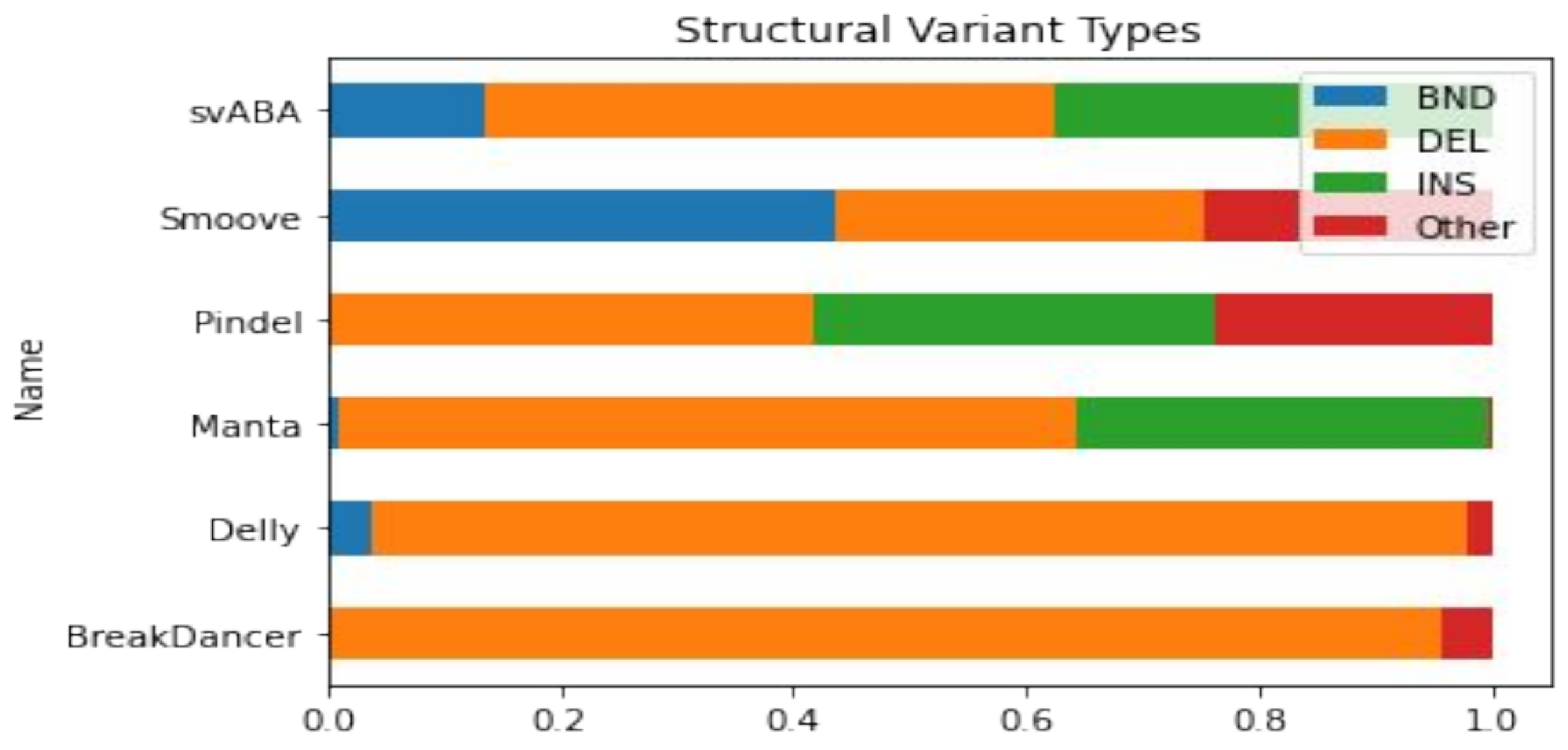
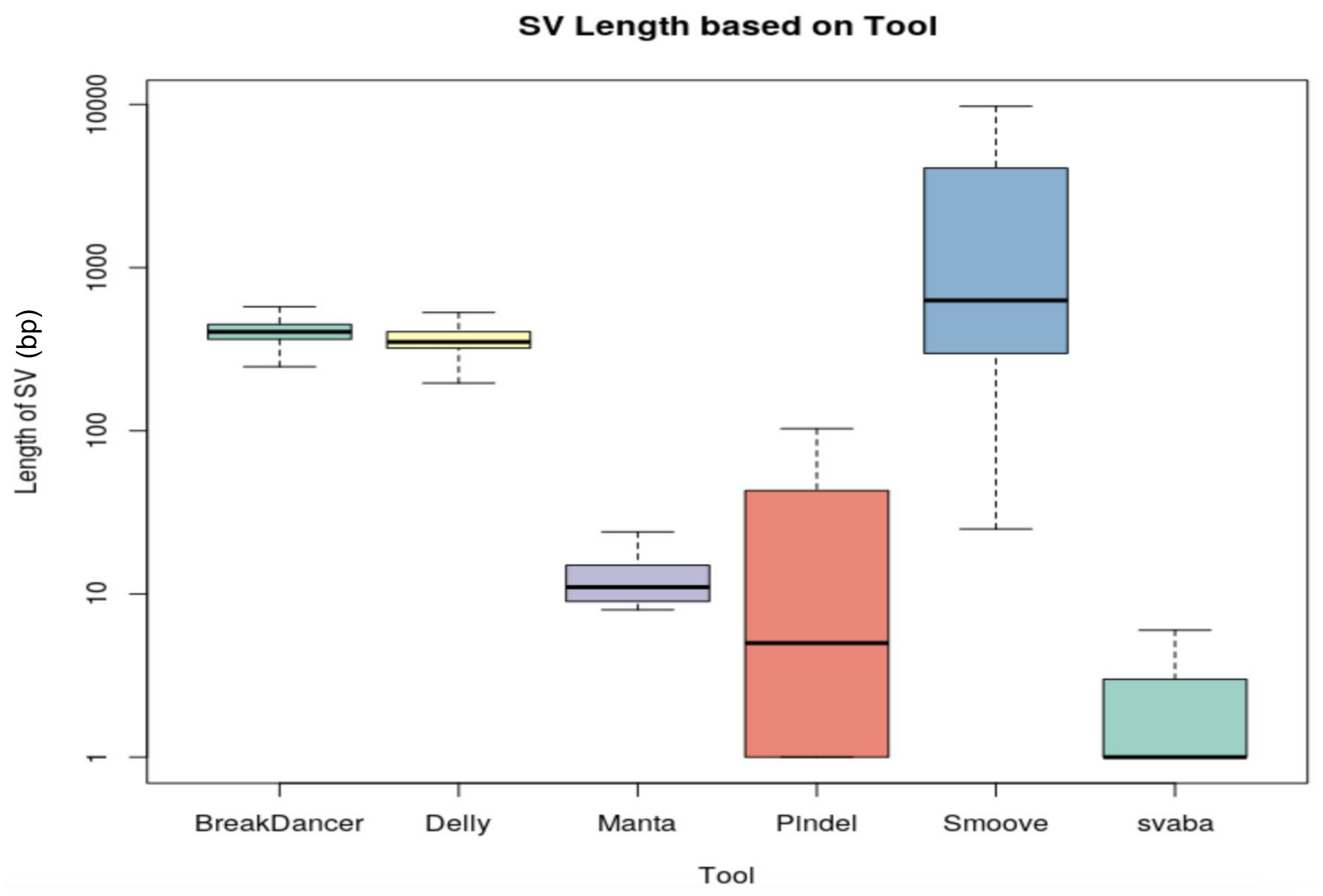


Figure 3: Grouped box plot comparing the length (bp) of the structural variants detected by each tool. Note y-axis is log transformed.



All SV Detection tools tested have poor recall of experimentally determined SVs

Because these are established cell lines, we have 59 **experimentally determined** SVs that we expect to see in each cell-line.

We first computed recall of these SVs at the chromosomal level. SvABA and Delly had the best recall rates (78% and 28% respectively).

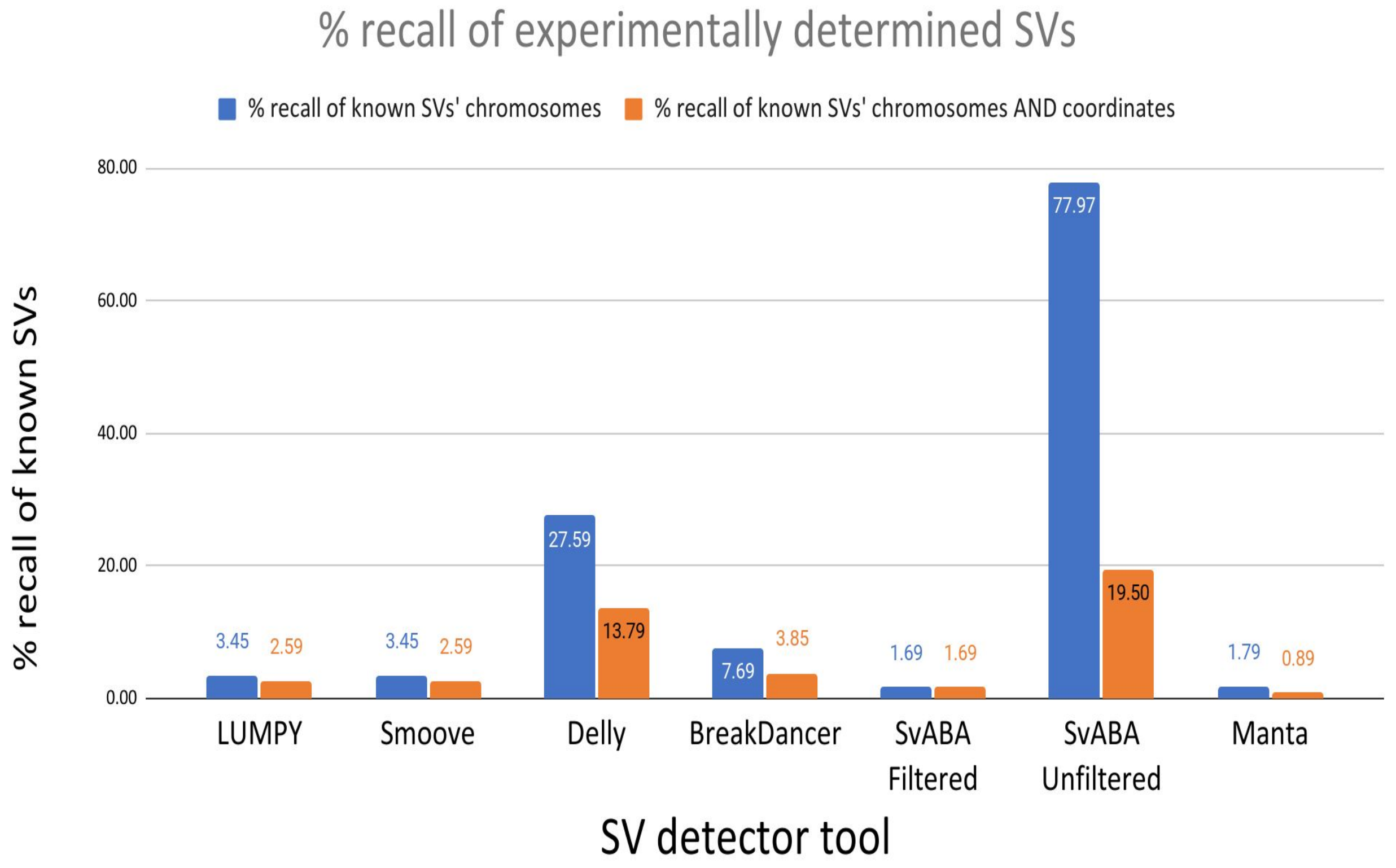
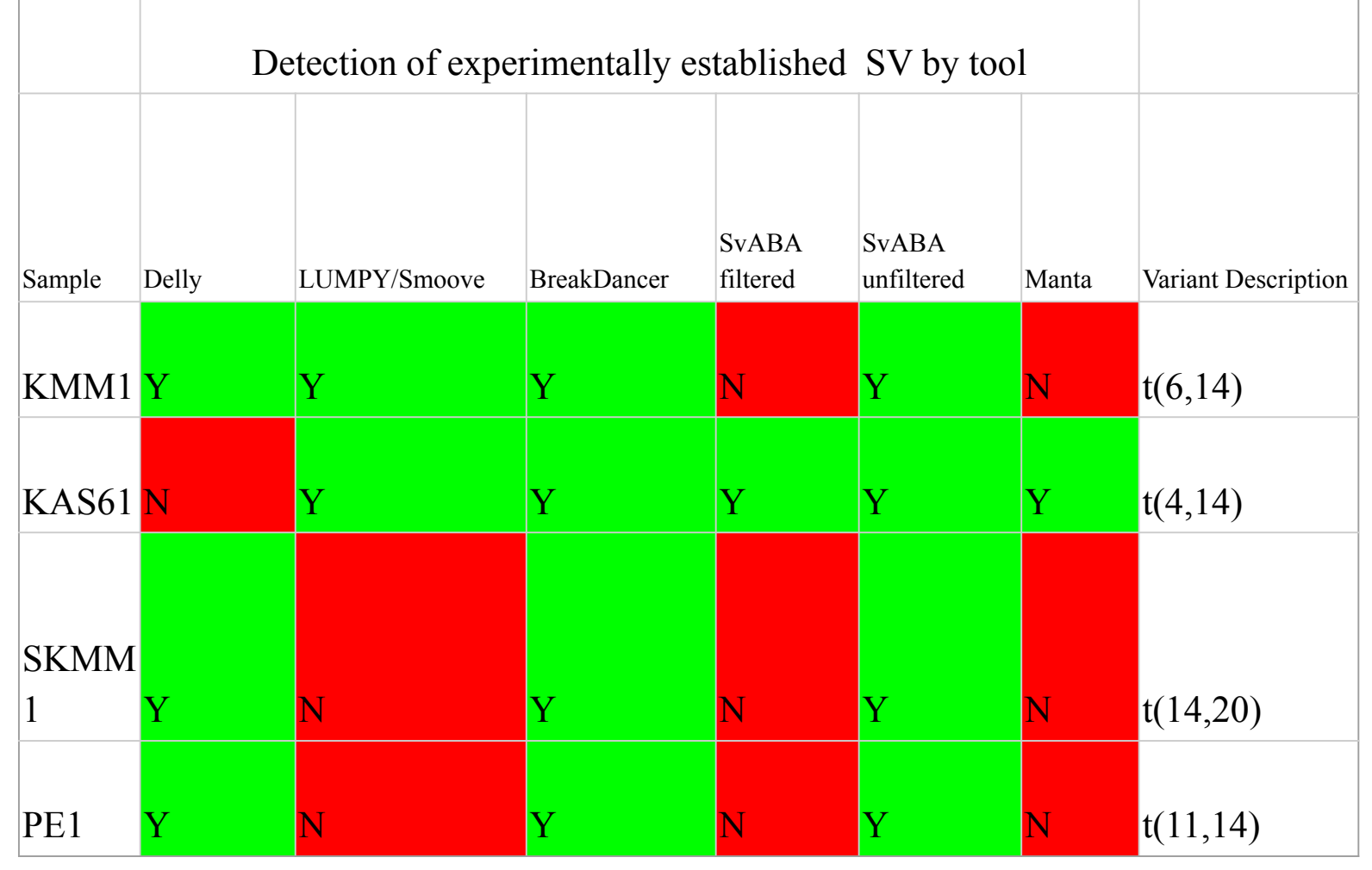


Figure 4: Contrasting tool recall rates at the chromosomal (indicated by blue) and coordinate level (indicated by orange). SvABA's unfiltered output has best recall of experimentally determined SVs (19.5%)

However, recall of these SVs at the coordinate level was much lower among these tools. Strikingly, SvABA's unfiltered output recall rate was 58% lower!

Overlapping SV output

Only four experimentally determined SVs were detected by >= three tools



	Detection of experimentally established SV by tool						
Sample	Delly	LUMPY/Smoove	BreakDancer	SvABA filtered	SvABA unfiltered	Manta	Variant Description
KMM1	Y	Y	Y	N	Y	N	t(6,14)
KAS61	N	Y	Y	Y	Y	Y	t(4,14)
SKMM 1	Y	N	Y	N	Y	N	t(14,20)
PE1	Y	N	Y	N	Y	N	t(11,14)

Table 1: Experimentally determined SVs detected by at least 3 tools

As the tools' output had poor recall and little overlap, we can be most confident in SVs detected by more than one tool.

We characterised these cell-lines with 5783 SVs found by 3 tools w/ an error range of 100 bp..

	# of SVs detected by N tools				
Error range (+= bps)	2	3	4	5	6
0	5029	1070	73	3	0
10	9449	2573	294	5	0
100	20054	**5783**	862	20	0
1000	59280	17997	4880	456	0

Table 2: Number of SVs detected N tools with varying margins of allowed bp differences

Conclusions

- Different published cancer studies have used these tools to identify SVs in exome data
- Given their poor recall rate with established SVs, we question these tools' applicability to WES data.
- A reasonable approach may be considering only SVs detected by more than one tool.
- We identified 5783 SVs to characterize these cell lines.

Limitations

- Lack of coordinate info for known SVs
- Only evaluated tool with 59 known SVs even though each tool identifies 1000s of SVs
- No tool accounted for a specific type of SVs: Copy Number Variants (CNVs)

Future steps

- **Calculate structural variant burden** for each cell-line's output VCFs
- **Characterise CNVs** in each MM cell-lines.
- Use characterised SVs to **detect neoantigens.**
- **Visualize** characterised SVs.
- **Display characterisation** of MM cell-line with an interactive R shiny app.

Acknowledgements

- **LiveSTRONG Cancer Institute** who inspired and assisted us in our work.
- Peers in the **Big Data in Biology FRI** stream.
- **Texas Advanced Computing Center (TACC)** for large computing resources to parallelize our analysis.
- **Biomedical Research Computing Facility (BRCF)** for local computing and storage.
- **Keats lab at the Translational Genomics Research Institute** for the sequencing data.

References

- Yokoyama, Toshiyuki T., and Masahiro Kasahara. "Visualization Tools For Human Structural Variations Identified By Whole-Genome Sequencing". *Journal Of Human Genetics*, vol 65, no. 1, 2019, pp. 49-60. *Springer* Science And Business Media LLC, doi:10.1038/s10038-019-0687-0.
- "Exome Sequencing - An Overview | Sciencedirect Topics". Sciencedirect.Com, 2020, https://www.sciencedirect.com/topics/neuroscience/exome-sequencing.
- "What Are Whole Exome Sequencing And Whole Genome Sequencing?". Genetics Home Reference, 2020, https://ghr.nlm.nih.gov/primer/testing/sequencing.