SVX: Population-scale merging of structural variants with tandem repeat-aware refinement

T. Mokveld[1], J. A. Lake[1], W. J. Rowell[1], E. Dolzhenko[1], M. A. Eberle[1]; [1]PacBio, Menlo Park, CA,

Introduction

Studies of structural variants (SVs) in large cohorts remain challenging due to high data volumes and imprecise breakpoints in low sequence complexity regions, such as tandem repeats (TRs), which account for ~70% of all SVs[1] and can exhibit extreme allelic heterogeneity (**Figure 1**)[2].

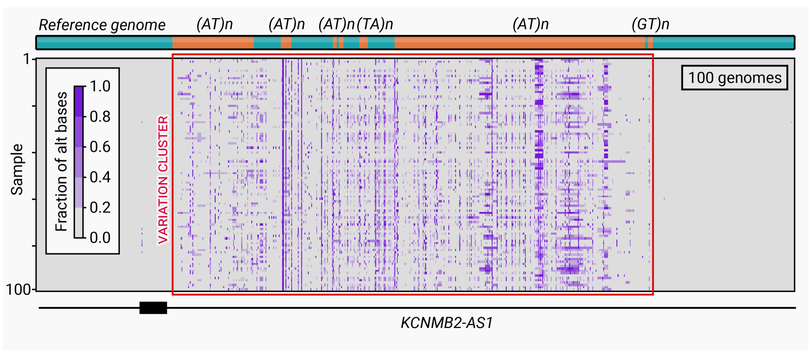

Figure 1. A highly variable TR region.
A 3 kb TR-rich region in the *KCNMB2* gene showing significant variation across 100 genomes. Each row denotes variation in a sample relative to the reference genome.

Repeat regions such as *KNMB2*, harbor numerous variant calls that are challenging to merge and interpret. Variability among SV-calling pipelines exacerbate these challenges, leading to distorted SV merging outcomes. Treating these loci as single units removes ambiguity and enables accurate allele resolution[3].

Overview of SVX

We developed SVX, a fast, memory-efficient SV merging tool implemented in Rust. Similarly to Jasmine[4], SVX uses SV proximity graphs to link SVs with similar breakpoints and lengths, then progressively further refines resultant SV clusters (**Figure 2**).

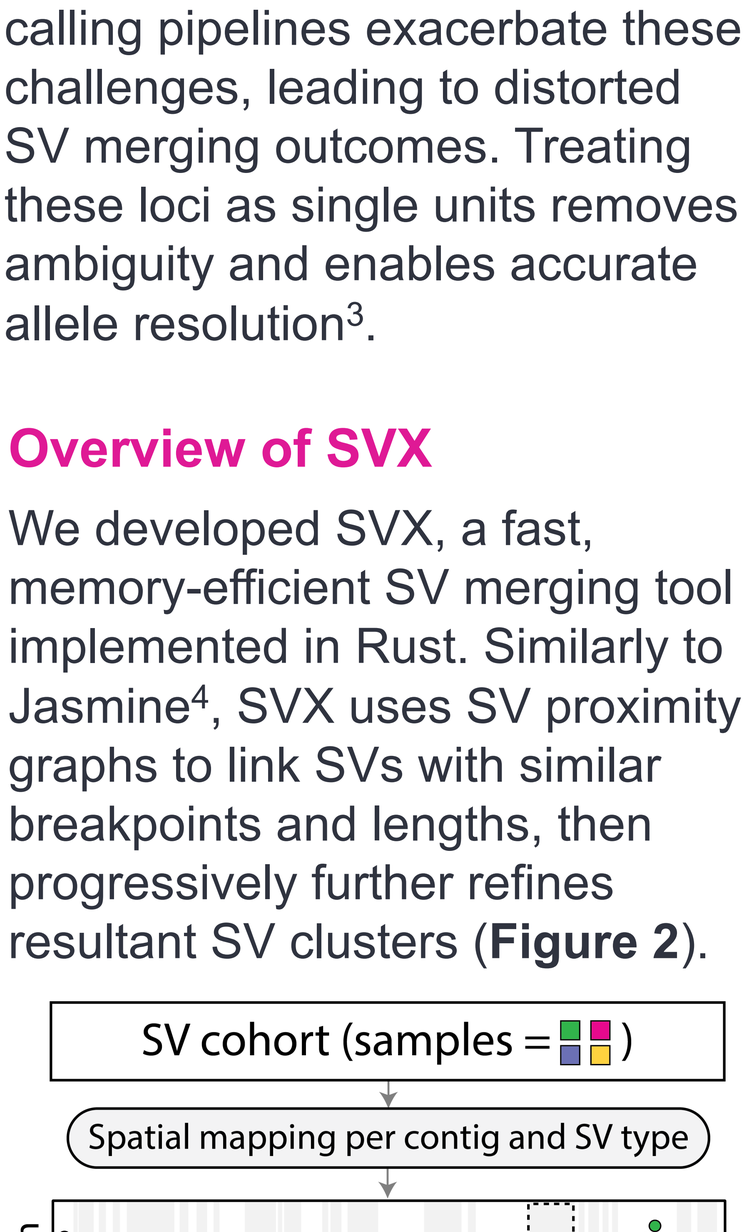

Figure 2. SVX overview. Overview of the proximity analysis, clustering, and refinement steps used by SVX.

SVX handles tandem repeats

Recognizing the abundance and confounding influence of repeat-driven artifacts, SVX integrates curated TR catalogs and logic for refining variant boundaries in TR regions (**Figure 3**).

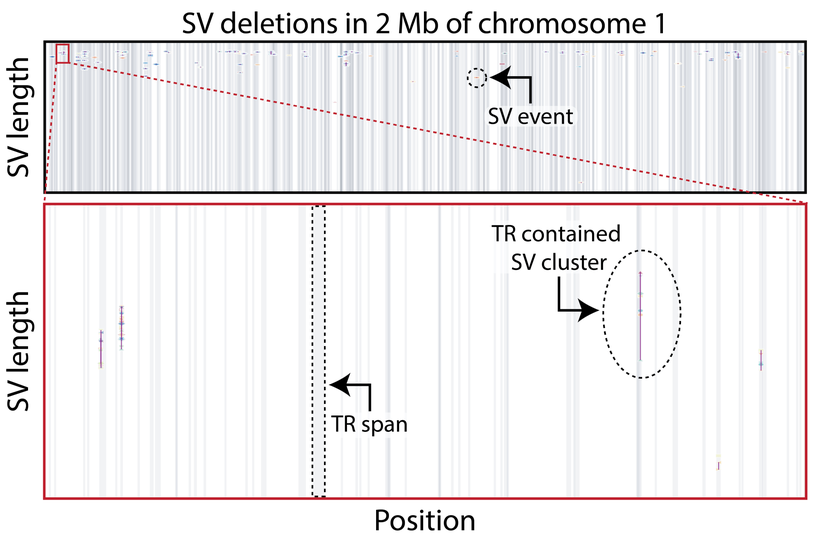

Figure 3. TR-annotated SV proximity graphs.
Before SVX merges SVs, individual SVs are first annotated using prior known TR definitions. During the merging process, SVs contained within the same TR region are clustered together.

SVX refines SV clusters

A greedy SV merging strategy can lead to variant under-merging or over-merging. SVX minimizes this by refining clusters after the initial greedy SV clustering (**Figure 4, 5**; Platinum Pedigree[5,6], n = 28).

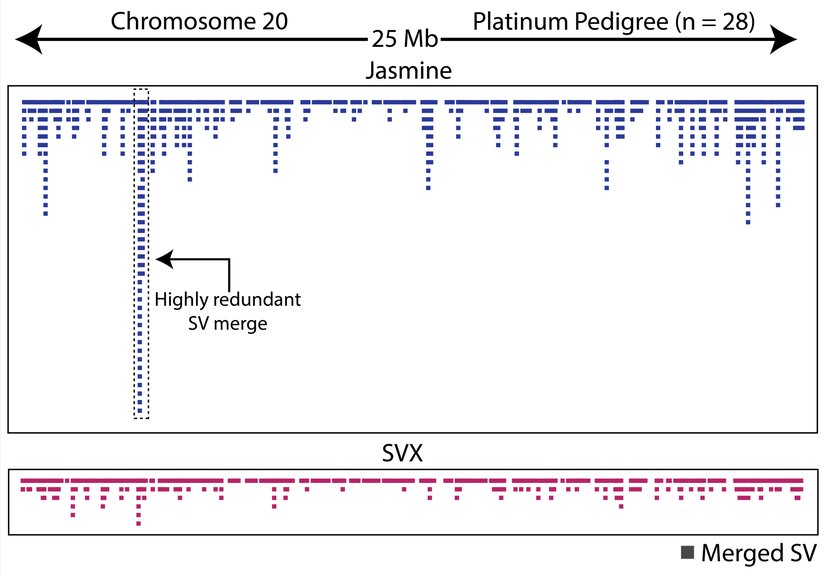

Figure 4. SV merging redundancy.
Positional mapping of merged SVs on chr20 in a 25 Mb window by Jasmine and SVX. Identical distance and similarity thresholds were used by both methods.

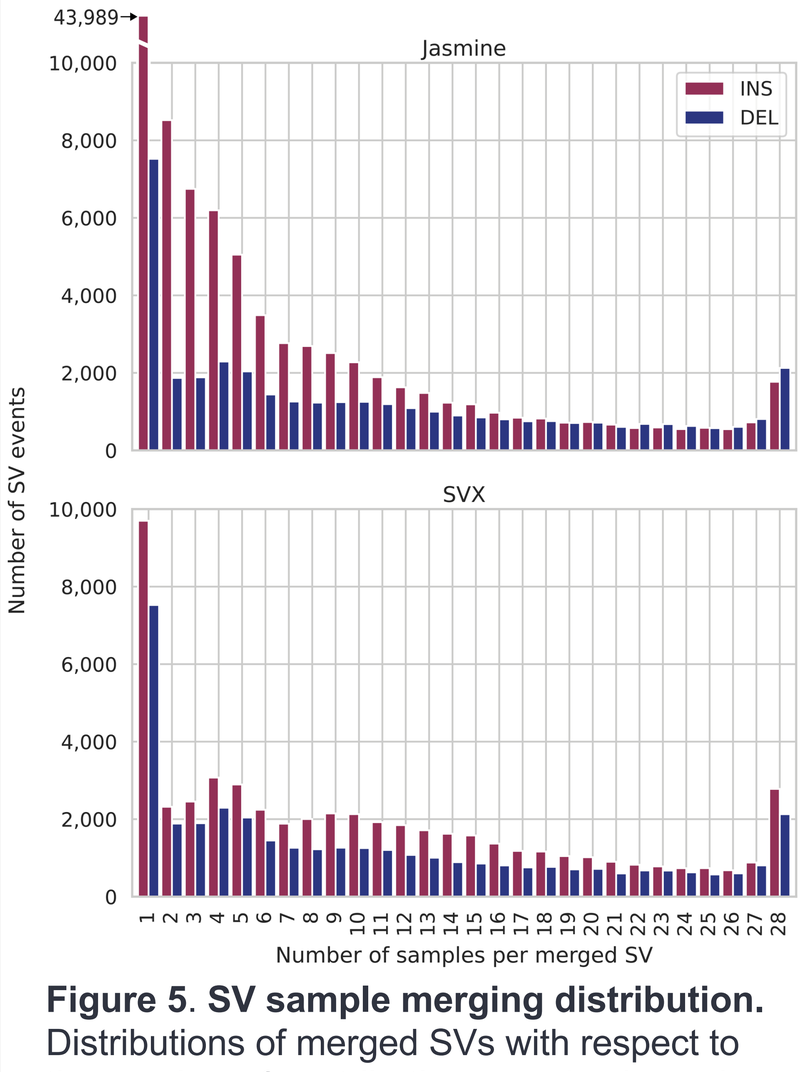

Figure 5. SV sample merging distribution.
Distributions of merged SVs with respect to the number of contributing samples in each SV, shown for Jasmine and SVX.

SVX is fast and efficient

Through minimization of both execution time and memory overhead, SVX makes iterative exploration of SV merging parameterization feasible across large-scale SV cohorts (**Table 1**).

Table 1. Benchmarking SVX.
Comparison of Jasmine and SVX on the HPRC (n = 100) and Platinum Pedigree (n = 28) cohorts. Jasmine did not complete SV merging within 24 hours for the HPRC cohort. BND SVs were excluded.

Cohort	Threads	Tool	Post merge SVs	Memory (GB)	Time	Speed up
Platinum Pedigree n = 28 SVs = 952,790	1	Jasmine	140,189	13.8	05:57:56	-
		SVX	92,367	**0.4**	**00:01:31**	**236**
	4	SVX	92,367	2.8	00:00:32	671
INS: 55.77% DEL: 38.91% BND: 4.68% DUP: 0.45% INV: 0.20%	8	SVX	92,367	3.1	00:00:23	933
	16	SVX	92,367	3.8	00:00:18	1193
	32	SVX	92,367	4.2	00:00:18	1193
HPRC n = 100 SVs = 3,525,391	1	Jasmine	-	>17.3	>24h	-
		SVX	291,040	**1.8**	**00:16:58**	-
	4	SVX	291,040	4.6	00:05:13	-
INS: 56.6% DEL: 39.57% BND: 3.34% DUP: 0.34% INV: 0.15%	8	SVX	291,040	8.1	00:03:54	-
	16	SVX	291,040	10.6	00:02:10	-
	32	SVX	291,040	11.9	00:01:55	-

Conclusion

- We developed SVX, an SV-merging tool intended for population-scale analysis, with an emphasis on ease of use.
- SVX is in early development. It currently *only* works with Sawfish, a recently developed SV caller[7]; we plan to add support for other SVs callers, copy number variants, and variant-specific logic beyond TRs.
- Documentation, a pre-release binary, and the source code are available on GitHub:

github.com/PacificBiosciences/svx

References

1. English, Adam C., et al. "Analysis and benchmarking of small and large genomic variants across tandem repeats."
2. Weisburd, Ben, et al. "Defining a tandem repeat catalog and variation clusters for genome-wide analyses and population databases."
3. Dolzhenko, Egor, et al. "Characterization and visualization of tandem repeats at genome scale."
4. Kirsche, Melanie, et al. "Jasmine and Iris: population-scale structural variant comparison and analysis."
5. Porubsky, David, et al. "Human *de novo* mutation rates from a four-generation pedigree reference."
6. Kronenberg, Zev, et al. "The Platinum Pedigree: A long-read benchmark for genetic variants."
7. Saunders, Christopher T., et al. "Sawfish: Improving long-read structural variant discovery and genotyping with local haplotype modeling."