Assessing Short-Read Utility for SVs

Adam English[1,2], Vipin Menon[1], Rob Flickinger[2], Ginger A. Metcalf[1], Richard A. Gibbs[1], Fritz J Sedlezeck[1]

Human Genome Sequencing Center, Baylor College of Medicine, Houston TX Spiral Genetics, Seattle WA

Baylor College of Medicine

HGSC HUMAN GENOME SEQUENCING CENTER

Abstract

Structural Variation (SV) is a major source of genomic variation and the cause of multiple genomic diseases. SVs (insertions/deletions larger than 50bp) are difficult to identify within Illumina short-read data, prompting the use of long-read sequencing. Despite the overall success of long-read methods for detecting SVs, the data generation costs and sample requirements are limiting. Improved characterization of the performance of short-read sequencing for SV detection - including a detailed comparison with long-read methods, the impact of different genome references, and the source of sample data - can therefore enhance routine SV analyses.

We leveraged previously published long-read, haplotype-resolved Human genome assemblies to create high-confidence sets of SVs (hcSVs) from **36** individuals from 5 ethnicities, employing 3 different references (Hg19; GRCh38; CHM13). Using short-read samples of the same individuals, We tested the 'genotypability' of hcSVs with short-reads by leveraging two graph-reference based genotypers (BioGraph, Paragraph). We found that the average Human genome contains **26,672** SVs (10k deletions, 16k insertions) compared to GRCh38 and **24,525 SVs** (12k deletions, 12k insertions) compared to CHM13. We note an increase in the SV counts in samples of African ancestry (**~29k** GRCh38 SVs) compared to those of non-African ancestry (**~25k** GRCh38 SVs). The short-read methods are able to accurately genotype ~**80**% of hcSVs, highlighting that the underlying signals for SVs are conserved across sequencing technologies. However, we observe a bias against variants within Simple Repeats.

This work highlights that while short-read sequencing genotypes fewer SVs than long-reads, much of the signal elucidated by both approaches may still allow for a better understanding of SVs and their geno- and phenotypic implications when appropriate tools and analytical methods are employed.

Truvari when are two SVs the same?

github.com/spiralgenetics/truvari

spiralgenetics/truvari

Structural variant toolkit for VCFs

pypi package 3.0.0 | pylint 10 | FuncTests passing | coverage 90%

Python ☆ 111 Forks 30

```
█ = Deletion ^ = Insertion
```

```
--refdist REFDIST (500)
  Max reference location distance

    ACTGATCATGAT
     |--    --|

  Calls are within reference distance of 2
```

```
--pctovl PCTOVL (0.0)
  Min pct reciprocal overlap

    ACTGATCATGA
                     [2,7)
                     [4,8)

  variants have 0.6 reciprocial overlap
```

```
--pctsize PCTSIZE (0.7)
  Min pct allele size similarity

    ACTGATCATGA
                  -> 5bp
                  -> 4bp

  variants have 0.8 size similarity
```

```
--pctsim PCTSIM (0.7)
  Min percent allele sequence similarity

    A-CTG-ACTG
     ^   ^         haplotypes
     |   └ACTG -> CTGACTGA
     └CTGA     -> CTGACTGA

  haplotypes have 100% sequence similarity
```

Per-Sample Structural Variants

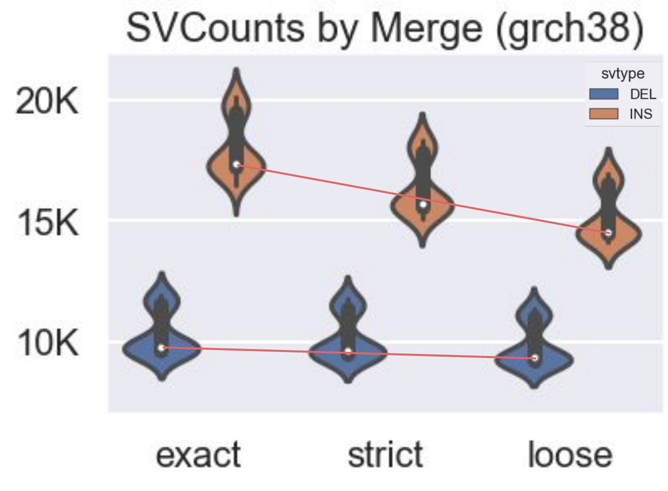
Name	pctsim	pctsize	refdist
exact	100	100	0
strict	0.95	0.95	500
loose	0.70	0.70	1000

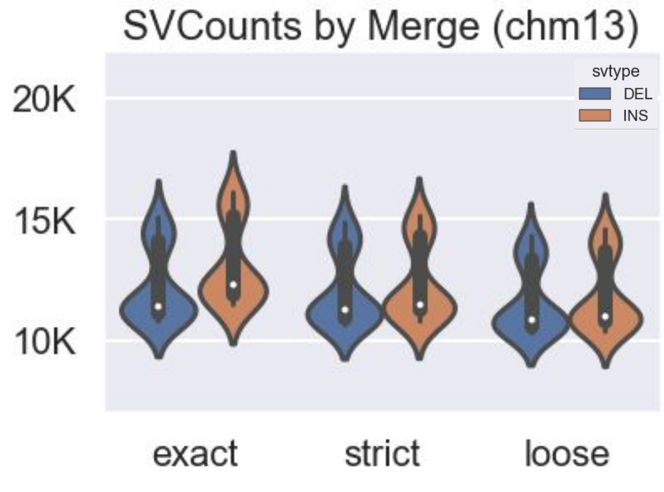
For each sample, variants called per-haplotype were compared to one another across three matching thresholds using haplotype-aware merging through `truvari collapse` (table above).

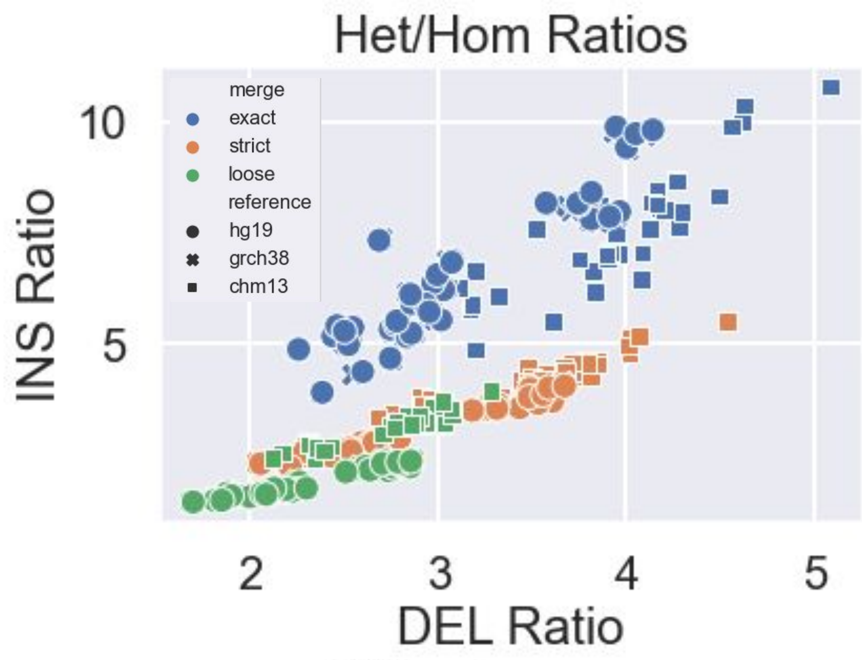
Not only do we see INS count changing dramatically by merging strategy, (grch38, top-left) but reference also having an effect because chm13 (bottom-left) has different counts and fewer calls collapsing as the merging strategy becomes more lenient.

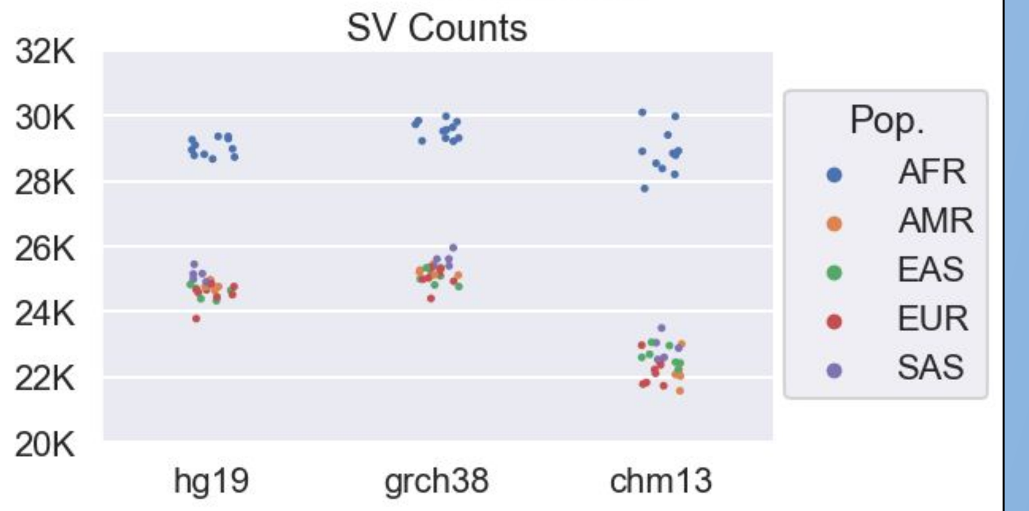
The Het/Hom ratio of DEL and INS (top-right) shows that more lenient thresholds allow HETs to find matches and become HOMs. However, the loose ratios (green) begin clustering by reference (shape), suggesting that calls may be over-collapsing

Given that no assembly process is perfect and alignment ambiguities still arise, we observe exact merge under-collapsing and over-estimating counts and het/hom ratios while loose over-collapses. Our final SV counts from strict collapsing (bottom-right) shows clustering by population.



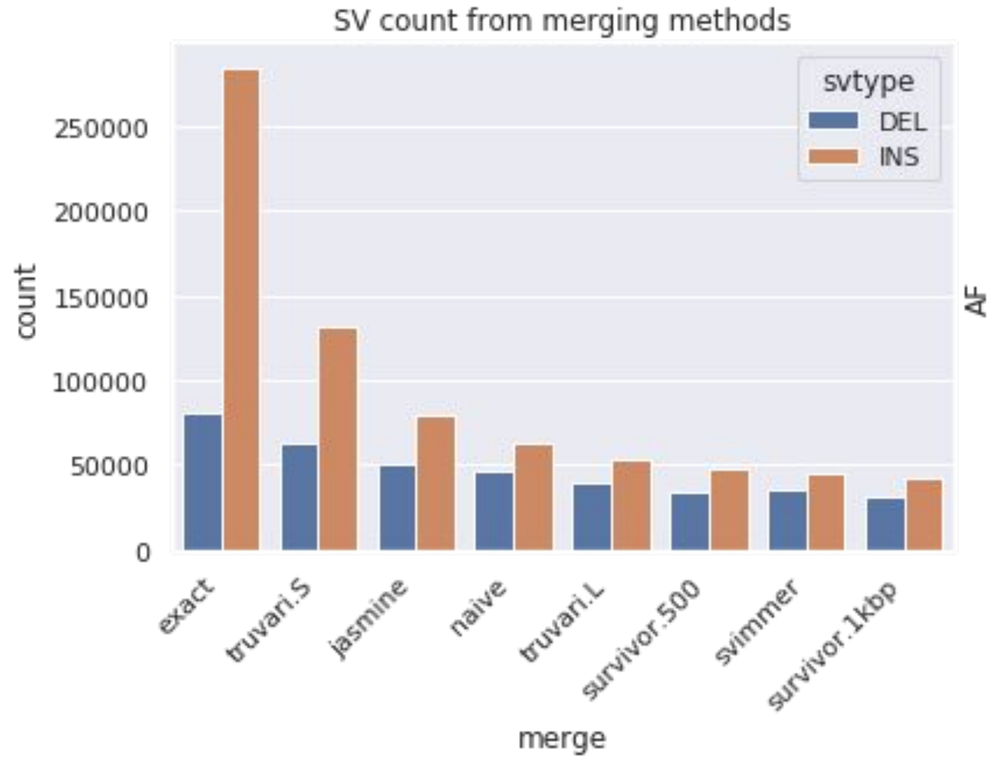


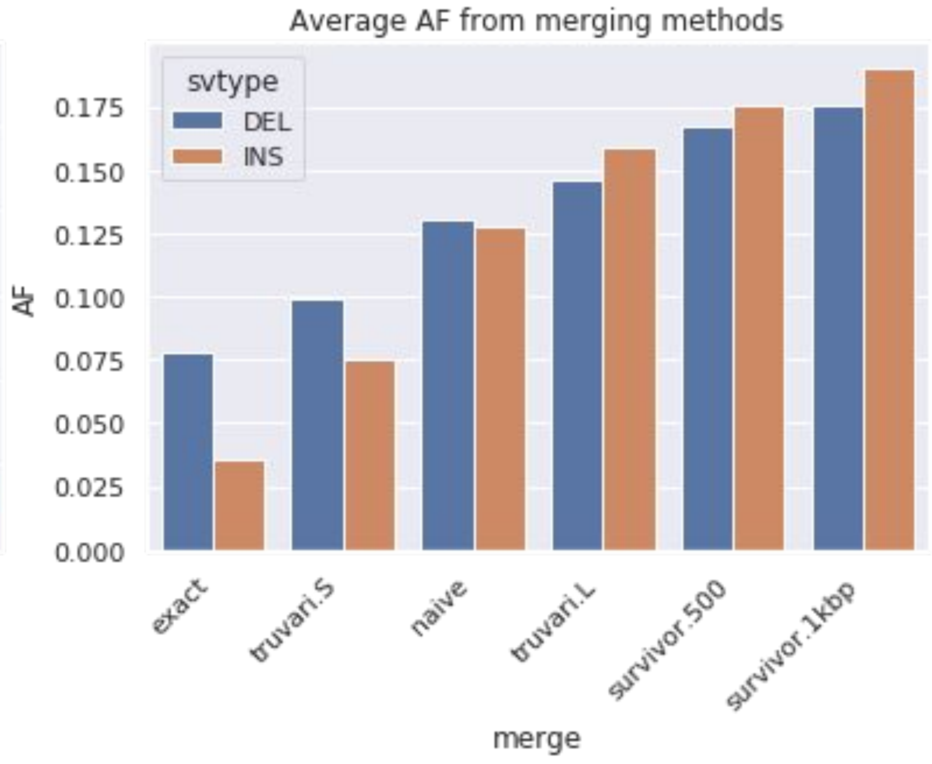




Comparing SV Merging Tools

A number of tools exist that merge variants between samples. For comparison, we merged the per-sample variation to create a pVCF using six tools: bcftools, truvari 'strict', truvari 'loose', naieve 50% reciprocal overlap, survivor with 500bp and 1kbp reference distance, jasmine, svimmer. Final SV counts (bottom-left) vary between tools, with bcftools - which employs an 'exact' matching - under-collapsing, and most tools over-collapsing. The effect of collapsing has a direct impact on allele frequency with less collapsing making variants seem less frequent and vice-versa (bottom-right). Note that Jasmine and svimmer do not preserve per-sample genotypes, which prevents AF calculations. Truvari strict and Jasmine have the best balance. However, Jasmine collapses more insertions. An example of where Bcftools, Truvari and Jasmine make different decisions is described on the right.





chr1:10627-10786 has a 29bp 'TAR1' microsatellite repeat.
The reference has 5 copies GGCGCGCCGCGCCGGCGCAGGCGCAGAGA = - or _

```
> POS
NS AC LN CN haplotype
```

```
           Exact                              Truvari                             Jasmine
> 10627                           > 10627                           > 10627
2 2 58  7      [--][_____]        3 4 58  7      [--][_____]        4 ? 58  7      [--][_____]
2 2 86  8     [---][_____]        4 4 86  8     [---][_____]
1 1 116 9    [----][_____]        2 2 116 9    [----][_____]
1 1 145 10 [-----][_____]         1 1 145 10 [-----][_____]         2 ? 145 10 [-----][_____]

> 10735
1 1 86  8  [_____][-*-]

> 10777                                                             > 10777
1 1 57  7  [_____][--]                                              4 ? 57  7  [_____][--]
1 1 57  7  [_____][-*]

> 10781
1 1 85  8  [_____][---]
```

- In a TAR1 microsatellite repeat, Bcftools shows 8 Insertions over 4 positions in the region. This region holds a 29bp simple repeat with 5 copies in the reference. The insertions expand the number of copies in the region from +2 (7 total copies) to +5 (10 total copies).

- Truvari collapse preserves the left-aligned representations of all copies (7-10), merging the 3 other positions with their copy counterpart.

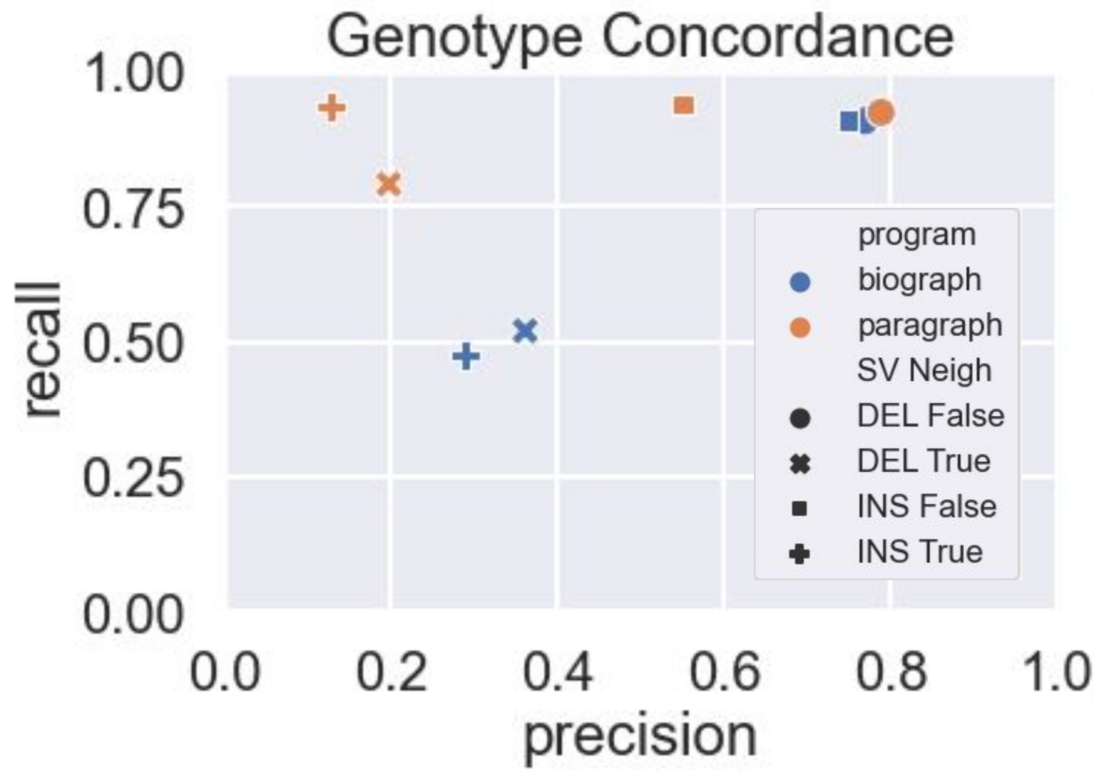
- Jasmine keeps 3 insertions over 2 positions with 7, 10, and 7 copies of the repeat.

- All other tools report a single insertion

Short-Read Genotyping

Using the pVCF, we ran two graph-based short-read genotypers: BioGraph and Paragraph. We then evaluated SV 'genotypability' of short-reads by checking if the genotypers can correctly determined the concordant genotype.

Precision of both genotypers drops significantly when SVs have a another call within 1Kbp (Neigh = True). While Paragraph has higher recall of events with Neighbors, this comes at the cost of low precision. The accuracy of the genotypes by Neigh is on the right.



Genotyping Balanced Accuracy

	Has Neighbor	
Program	**False**	**True**
BioGraph	90.4%	62.3%
Paragraph	88.4%	56.6%

With having a Neighbor being a strong indicator of if an SV is 'genotypable', we look at how frequently SVs with Neighbors are within Simple Repeats. Only 5% of SVs outside of Simple Repeats have Neighbors and 8% of SV within Simple Repeats lack Neighbors.

When looking at number of SVs in Simple Repeats by Reference/SVTYPE (below), we see an enrichment of SVs in SREPs over grch38.

SVs in Simple Repeats by Reference/SVTYPE