TT-Mars: Structural Variants Assessment Based on Haplotype-resolved Assemblies

Jianzhi Yang¹ and Mark Chaisson^{†1,*}

¹Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, 90007, USA.

*To whom correspondence should be addressed, mchaisso@usc.edu.

September 2021

Abstract

Variant benchmarking is a critical component of method development and evaluating the accuracy of studies of genetic variation. Currently, the best approach to evaluate the accuracy of a callset is the comparison against a well curated gold standard. In repetitive regions of the genome it may be difficult to establish what is the truth for a call, for example when different alignment scoring metrics provide equally supported but different variant calls in on the same data. Here we provide an alternative approach, TT-Mars, that takes advantage of the recent production of high-quality haplotype-resolved genome assemblies by evaluating variant calls based on how well their call reflects the content of the assembly, rather than comparing calls themselves. We used TT-Mars to assess callsets from different SV discovery methods on multiple human genome samples and demonstrated that it is capable at accurately classifying true positive and false positive SVs. On the HG002 personal genome, TT-Mars recapitulates 96.0%-99.6% of the validations made using the Genome in a Bottle gold standard callset evaluated by *truvari*, and evaluates an additional 121-10,966 variants across different callsets. Furthermore, with a group of high-quality assemblies, TT-Mars can evaluate performance of SV calling algorithms as a distribution rather than a point estimate. We also compare TT-Mars against the long-read based validation tool, *VaPoR*, and when assembly-based variant calls produced by *dipcall* are used as a gold standard. Compared with *VaPoR*, TT-Mars analyzes more calls on a long read callset by assessing more short variant calls (< 100 bases), while requiring smaller input. Compared with validation using *dipcall* variants, TT-Mars analyzes 1,497-2,229 more calls on long read callsets and has favorable results when candidate calls are fragmented into multiple calls in alignments. TT-Mars is available at <https://github.com/ChaissonLab/TT-Mars.git> with accompanying assembly data and corresponding liftover files.

1 Introduction

Large-scale sequencing studies enable association of genetic variation such as common structural variations (SVs) (Abel et al., 2020), or rare variation of any (consortium et al., 2015) with traits and disease. SVs include deletions, insertions, duplications, and rearrangements at least 50 bases that as a class have a considerable role in genetic diversity (Merker et al., 2018; Sanchis-Juan et al., 2018), developmental disorders (Y.-h. Jiang et al., 2013; Niemi et al., 2018), and cancer (Macintyre, Ylstra, & Brenton, 2016). Compared to variations such as SNVs, SVs have been historically more difficult to detect using high-throughput short-read data, particularly due to the vast diversity of sizes and breakpoint complexity that SVs span (Zhao et al., 2021). In recent years, single-molecule sequencing (SMS) has been used to generate high quality structural variant callsets (Sedlazeck et al., 2018; Huddleston et al., 2017; Stancu et al., 2017; Ebert et al., 2021) because long reads or their *de novo* assemblies span SVs, particularly in complex and repetitive regions (Mahmoud et al., 2019; S. Chen et al., 2019).

Despite recent performance gains in SMS instrument throughput (Eid et al., 2009; Clarke et al., 2009), short-read sequencing remains the primary technology for sequencing large populations (Collins et al., 2020; Abel et al., 2020). Many algorithms have been developed to detect SVs using different types of information from short-read sequencing (Layer, Chiang, Quinlan, & Hall, 2014; Rausch et al., 2012; X. Chen et al., 2016), but there are reports of both high rates of false positives (Mahmoud et al., 2019) and low recall rates (Sedlazeck et al., 2018). Additionally, detecting complex SVs is still challenging (Sedlazeck et al., 2018) and callers can produce different SV callsets on the same genome sample (Zook et al., 2020), which makes it more difficult to output a complete high-quality SV callset.

The limitations and challenges in SV detection, combined with an importance in clinical sequencing motivates the need for efficient and precise tools to evaluate SV callset accuracy. A common framework for this benchmark is to establish a ground truth as a consensus between calls from multiple sequencing technologies and SV discovery algorithms, and to compare new calls to the ground truth. One of the first approaches to generate a gold standard was the *svclassify* method that used machine learning to classify SVs as true positive (TP), false positive (FP) or unclear (Parikh et al., 2016). Recently the Genome in a Bottle Consortium (GIAB) made a high-quality benchmark set of large (≥ 50 bp) insertions and deletions using multiple SV callsets produced by a wide range of analysis methods (Zook et al., 2020). This benchmark set can be used to evaluate arbitrary SV callsets generated by different combinations of algorithms and sequencing data. The accompanying method, *truvari* (Zook et al., 2020), is used to compare SV calls based on agreement between the breakpoints in the test and benchmark calls, with the option to compare sequences of variants. To be considered a true positive, the test variant must be within a specified size and distance

of a ground truth call (Zook et al., 2020). This approach has been an invaluable standard for benchmarking method accuracy, however the notion of comparing callsets may be considered as a proxy comparison against the overarching goal of determining how well an algorithm estimates the content of a genome with a particular sequence input. In repetitive regions there may be multiple placements of breakpoints that have equal support for a variant, or similar placement of breakpoints that depend on parameters for scoring alignments (Chaisson et al., 2019). In part due to the breakpoint degeneracy, and difficulty in calling variants in repetitive regions, the benchmark callset excludes many variants in repetitive regions (Zook et al., 2020), although repetitive regions are enriched for SVs (Huddleston et al., 2017).

As *de novo* assembly quality has increased with improvements in algorithms and SMS technologies, assemblies have been used to generate SV callsets (Chin et al., 2020; English et al., 2015). SV calls based on haplotype-resolved assemblies are factored into the GIAB truth set, and haplotype-resolved assemblies have been used to create benchmark callsets for structurally divergent regions (Chin et al., 2020), however the difficulties in comparing calls in repetitive regions remains. Here we propose an alternative approach to validate variant calls by comparing the sequences implied by an SV call to assemblies rather than comparing variant calls. This alleviates the need for a curated gold-standard callset so that any genome with a haplotype-resolved assembly of sufficiently high quality may be included as a benchmark, and can help validate SVs in repetitive regions because explicit breakpoints are not compared. While this resource has historically not been available, recent advances in single molecule sequencing and assembly have enabled more routine generation of high-quality haplotype-resolved assemblies. The Human Genome Structural Variation Consortium (HGSVC) has generated 32 haplotype-resolved assemblies of human genomes at Phred quality scale over 40 with contig N50 values over 25 Mb (Ebert et al., 2021). The Human Pangenome Reference Consortium is using a combination of Pacific Biosciences HiFi sequencing reads and the hifiasm method (Cheng, Concepcion, Feng, Zhang, & Li, 2021) to generate haplotype-resolved assemblies with base quality approaching QV50 and assembly N50 over 40 Mb. We have implemented our approach as a method, TT-Mars (Structural Variants Assessment Based on Haplotype-resolved Assemblies), which assesses candidate SV calls by comparing the putative content of a genome given an SV call against a corresponding haplotype-resolved *de novo* assembly.

We demonstrate that within the regions annotated as high-confidence by GIAB, TT-Mars has consistent results with truvari analysis using the GIAB gold-standard callset. TT-Mars also has consistent results with two other benchmarking methods: VaPoR (Zhao, Weber, & Mills, 2017), which is a long-read validation tool, and dipcall (Li et al., 2018)+truvari using assembly-based variant calls as a gold standard. We demonstrate that compared to VaPoR, TT-Mars requires smaller input and shorter runtime to achieve comparable results, and validation of calls using TT-Mars is less dependent on alignment gap parameters compared to dipcall+truvari. Using 10 assemblies, we evaluate the distribution of call accuracy for three different short-read SV calling algorithms, LUMPY (Layer et al., 2014), Wham (Kronenberg et al., 2015), and DELLY (Rausch et al., 2012), as well as one long-read SV detection algorithm, pbsv (Wenger et al., 2019). The software is available at <https://github.com/ChaissonLab/TT-Mars.git>, which provides a utility to download assemblies and alignment maps.

2 Methods

2.1 Workflow

TT-Mars uses a haplotype-resolved long-read assembly, and a lift-over map to the human reference to assess each candidate SV call from a dataset with reads matched to the same genome as the assembly (Figure 1a-1c). Each SV is considered independently as an edit operation on the reference genome (e.g. deletion, duplication, etc.), and SV calls are evaluated by comparing both the reference modified by the edit operation and the original reference genome to the assembly (Figure 1d-1g).

To speed up the reference/assembly comparison, only the local region surrounding the SV call are compared using an assembly-genome orthology map constructed from whole-genome alignments (Figure 1b). Each classification is made as a comparison between the relative scores of the alignment of the reference/modified reference and assembly at the SV locus, and/or the relative length the two, as detailed below.

2.2 Validation

The haplotype-resolved assemblies are pre-processed by TT-Mars to generate an orthology map between the assembly and the reference for regions that are annotated as assembled without error and have a clear 1-1 relationship. The contigs are aligned using *lra* (Ren & Chaisson, 2021) with the options `-CONTIG -p s`, from which an orthology map at fixed intervals (20 bp by default) is generated using *samLiftover* (<https://github.com/mchaisso/mcutils>) (Figure 1b). When two contigs have overlapping alignments on the reference, the alignment of the shorter contig is trimmed to exclude the overlap (Figure 1a). Next, a high-confidence filter is created by excluding centromeres, intervals where too few chromosomes are covered by assemblies (with number of X chromosomes supplied as a parameter), and low-quality assembled regions flagged by raw-read alignments (Supplementary Figure 1). SV calls that are flagged as analyzed if they are within the high-confidence filter.

The orthology map is accessed as a function `Lookup( $c, p$ )` that for a position p on chromosome c returns the assembly contig and position on the contig corresponding to the last position on the orthology map at or before p . To evaluate each SV call, we compare the local region surrounding the SV in either the original or modified reference to both assembled haplotypes (Figures 1b, 1c). For calls with a defined starting position s and ending position e , such as deletion calls (Figure 1d), this region may be defined as a region on the reference defined by these positions, expanded by a small number of bases on either side (w bases, $w=500$), and the corresponding intervals on the assembly defined as `Lookup( $c, s - w$ )` and `Lookup( $c, e + w$ )`.

Figure 1: TT-Mars Workflow. **a**, Assembly contigs are mapped to the reference, and the shorter of two overlapping contigs is trimmed to generate a unique mapping. **b**, the alignment is used to construct an orthology map at specific intervals (e.g. every 20 bases). **c**, For an SV, called at an interval $[s, e]$ on the reference, TT-Mars searches the orthology map for matches outside the interval that most closely reflects the length of the SV. In this example, the interval $[a, b]$ immediately flanking a deletion SV maps to an interval on the assembly that does not reflect the SV, but a wider search in the orthology map shows that $[c, b]$ spans a deletion in the assembly. **d-g** illustrate validation details by a deletion example. **d**, TT-Mars takes the candidate call with w flanking bases on both sides. The interval on the reference is compared with an interval on the assembly (**e**) before and after the SV operation. In **f** and **g**, the deletion operation removes the corresponding sequence on the reference, and is validated if the modified reference is more similar to the assembly than the original.

We assume that true-positive calls will produce a modified reference sequence that matches one or both haplotypes with high identity, as shown in the top figures in Figures 1e-1g, and that false-positive calls will produce a modified reference sequence that is different from both haplotypes as shown in the bottom figures in Figures 1e-1g. However, in highly repetitive regions, such as variable-number tandem-repeats, the alignment used to generate the orthology map may be different from that which is used to generate the SV call, and the interval defined directly from fixed offsets from the SV boundaries may not reflect this.

To account for potentially inconsistent breakpoints, we search for a combination of boundaries that maximises the congruence between the modified reference and assembly. Two criteria are used to quantify the discrepancy of the reference and contig intervals. First, relative length is defined as:

$$relative\ length = \frac{contig\ interval\ length - reference\ interval\ length}{sv\ length}. \quad (1)$$

Second, the contig interval is aligned to the reference interval before and after the SV operation. The relative score is calculated by

$$relative\ score = \frac{alignment\ score\ after - alignment\ score\ before}{|alignment\ score\ before|}, \quad (2)$$

where *alignment score after* is the alignment score of the contig interval and the reference interval after the SV operation and *alignment score before* is the alignment score before the SV operation. Figure 1c illustrates how inconsistent alignments in repetitive regions may cause incorrect inference of SV breakpoints in a lifted region.

TT-Mars validates separate classes of SVs by slightly different strategies. Deletion, sequence-resolved insertion calls are annotated as true positive if the relative length is close to 1 and the relative score is positive. Specifically, the true positive region of these two types is defined as (Figure 2a):

$$\begin{cases} -\alpha \times relative\ score + 1 - \beta \leq relative\ length \leq \alpha \times relative\ score + 1 + \beta, & 0 \leq relative\ score \leq \gamma \\ 1 - \delta \leq relative\ length \leq 1 + \delta, & relative\ score > \gamma \end{cases} \quad (3)$$

where α , β , γ and δ are empirical parameters.

Insertion SV calls that do not include the inserted sequence, such as many insertion calls made by short-read callers, are evaluated by relative length only, while inversion calls that do not change the length of the sequence are evaluated by relative score.

TT-Mars validates duplication calls by considering possibilities of both tandem duplication and interspersed duplication. A tandem duplication operation adds a copy of the duplicated sequence to the reference at the duplication locus in a tail to head orientation. The start and end coordinates of the duplication that are lifted over include both copies of the tandem duplication in true-positive calls. Similar to deletions and sequence-resolved insertions, it is validated by using both the relative length and the relative score. Interspersed duplications are evaluated by aligning the duplicated sequence to the assemblies using minimap2 (Li, 2018) allowing for multiple alignments. Interspersed duplications are validated if at least 90% of the duplicated sequence aligns to a sequence in the assembly annotated as an insertion relative to the reference.

3 Results

3.1 Validation of calls on GIAB HG002 assembly

On human genome sample HG002, we used TT-Mars to validate insertion, deletion, inversion and duplication calls produced by four methods: LUMPY, Wham, DELLY and pbsv (Table 1, Supplementary material). Translocation/break-end calls and calls on the Y-chromosome were ignored. The pbsv callsets, except the HG002 callset, were generated by the HGSVC (Ebert et al., 2021). Broadly, these calls are generated using the pbmm2 alignment method (using minimap2) and an SV detection algorithm designed for PacBio raw read data.

The computational requirements are modest. The runtime scales by the number of calls, most short read data sets can be evaluated within one hour using a single core, and pbsv (long-read) callsets within 9 hours, both using up to 28G of memory.

Figure 2: An example of classification of individual deletion calls by TT-Mars on HG002 made by DELLY **a** and LUMPY **b**. Each dot represents a candidate call. The classification boundaries determined empirically for human data are shown by the light green region. Panel **b** highlights different classifications made by TT-Mars and GIAB+truvari. **c**, The length distribution of SVs in the HG002 callsets by four SV detection algorithms. The distribution of calls that are analyzed or where TT-Mars does not provide an annotation (NA) are shown.

The analyzed rates are in the range of 92.2% to 96.7%. Among the analyzed calls, the TP rates range between 78.4% and 90.0%. Among the calls that are not analyzed (NA), a majority of calls (65.7%-78.4%) are in regions which are not covered by one or both assemblies and 14.8%-31.4% are in centromeres. Figure 2a shows an example of scatter plot of validation results from TT-Mars, where the TP and FP calls are separated as two clear clusters. Figure 2c gives the distribution of length of calls that are analyzed and missed by TT-Mars. The distribution conforms with the length distribution of the benchmark SV callset (Zook et al., 2020).

We compared the TT-Mars results to the GIAB HG002 benchmark callset validated using truvari (Table 2) set to search for calls within 1kb (e.g. `refdist`) and not using sequence identity comparison. For long-read callsets, truvari was used with the option that compares the sequences of SV for validation. Since the GIAB benchmark set only contains deletions and insertions, validation of other types of SVs are not included. Generally, the two validation methods have consistent TP and FP results. For SV that are evaluated by both methods (e.g. excluding the NA calls), 96.0%-99.6% of calls have the same classification results in the four callsets (Figure 2b). For the three short read callsets, there is a net increase of 121-1,476 calls analyzed by TT-Mars, and on the pbsv callset, TT-Mars analyzes 10,966 more calls. The regions annotated as confident for validating variant calls by TT-Mars and excluded by truvari include 141 Mbp. These

Table 1: TT-Mars results on callsets from four SV discovery algorithms on HG002.

Callers	TP	FP	NA	Total
Wham	1130	144	74	1348
LUMPY	2990	825	324	4139
DELLY	12784	1833	812	15429
pbsv	45625	5057	1741	52423

Table 2: Comparison of TT-Mars and GIAB benchmarks on HG002. The two methods have the same classification results on most analyzed calls, while TT-Mars analyzed more candidate calls on all the four callsets.

TT-Mars																				
Wham						LUMPY				DELLY				pbsv						
	TP	FP	NA	Sum		TP	FP	NA	Sum		TP	FP	NA	Sum		TP	FP	NA	Sum	
GIAB	TP	882	3	21	906		2091	8	51	2150		2927	11	57	2995		8903	94	191	9188
	FP	1	13	0	14		22	218	3	243		4	412	6	422		286	165	16	467
	NA	118	24	20	162		582	215	99	896		1179	360	145	1684		9710	1463	721	11894
	Sum	1001	40	41	1082		2695	441	153	3289		4110	783	208	5101		18899	1722	928	21549

sequences predominantly overlap segmental duplications (99 Mbp) and highly repetitive sequences such as variable-number tandem repeats (4 Mbp).

Figure 3: **a**, SV metrics for 4 algorithms on 10 HGSCV genome samples. Benchmark results are given as a distribution by TT-Mars. **b**, The length distribution of results by dipcall+truvari, TT-Mars and VaPoR for pbsv HG00096 calls. The red solid and dashed lines indicate that dipcall+truvari and TT-Mars congruent calls have similar length distributions. The green and blue dashed lines show that VaPoR has more NA and disagreed results with TT-Mars for small SVs (size < 100 bases).

3.2 Performance on 10 HGSCV sample genomes

Gold-standard callsets such as those produced by the GIAB require intense effort to create. Because TT-Mars only needs haplotype-resolved assemblies, multiple samples may be evaluated in order to give benchmark results as a distribution. Recently, the HGSCV has generated 32 high-quality haplotype-resolved assemblies of human genomes (Ebert et al., 2021). We used TT-mars to evaluate three short-read and one long-read SV-discovery algorithms on ten samples sequenced and assembled by the HGSCV (Figure 3a). A second long-read SVs discovery method, cuteSV, had a low TP rate across every evaluation method, and is excluded from results. The number of analyzed and true positive sites assessed by TT-Mars are consistent across different samples, for both the short and long read callsets. The fraction of SV calls that are analyzed counts calls that are made by an algorithm in less repetitive regions of the genome that are amenable to long-read assembly, in particular excluding centromeres and high-identity long segmental duplications. Wham callsets have the highest analyzed rate (93.3% on average) among the three short reads algorithms. Pbsv callsets have the analyzed rate in the range of 93.0% and 95.2%. The TP rates on short-read callsets range between 63.9%-85.2%, while pbsv calls have TP rates from 71.4% to 94.3%.

The availability of assembled sequences allows TT-Mars to validate duplications, specifically checking the genomic organization is in tandem or interspersed. The average number of duplications called by Wham, LUMPY, DELLY and pbsv on the ten HGSCV genome samples are 1,463, 4,577, 761 and 1,458 respectively. Duplications that are tandemly organized may be validated as inserted elements at the location of the insertion because of the tail-to-head organization. Interspersed duplications are more challenging to validate because the target destination is unknown. Because the GIAB benchmark callset does not include duplication calls, we estimated the performance of TT-Mars using simulations. We ran 1,000 simulations of interspersed duplications with length uniformly taken from 100 to 100,000 on the HG002 assembly. Not all duplications

Table 3: TT-Mars validation results of simulated duplications.

	TP	FP	NA	Total
True DUP simulation	819	9	172	1000
False DUP simulation	0	882	118	1000

were analyzed by TT-mars because there was no requirements placed on the target site of duplications. In total, 82.8% duplications were analyzed, of which 98.9% were validated (Table 3), indicating that interspersed duplications from non-repetitive euchromatic regions of the genome may be validated by TT-mars. We also simulated 1000 false duplications, TT-Mars analyzed 88.2% of them of which all were annotated as FP (Table 3).

3.2.1 Comparison to VaPoR and dipcall

We compared our method to two other approaches for validating variant calls that use information from long reads or their assembly: VaPoR, which compares SV calls with individual long reads, and dipcall+truvari, where a callset from a *de novo* assembly is used as a gold standard for evaluating an SV callset. Both methods are compared with TT-Mars on callsets produced by the four SV discovery methods (Wham, LUMPY, DELLY and pbsv). Only insertion and deletion variants were in dipcall output, so that other types are not included in the comparison.

Due to the runtime of VaPoR, we limited analysis of data from one sample with each SV discovery method (Table 4). On short read datasets, TT-Mars and VaPoR provide the same validation result on 80.5%-92.0% for calls analyzed by both methods, however there was a net increase of 302-1,863 calls that could be analyzed by TT-Mars. On the long read dataset, the two methods agree on 67.6% of calls analyzed by both methods and TT-Mars analyzes 39,187 more calls. Because of the efficiency of dipcall, we were able to compare all SV callsets on the ten sample genomes. The validation of short read callsets is similar, with the methods agreeing on validation results for 96.1%-99.8% of calls analyzed by both methods (Figure 4a). This is an expected result because of the relatively low repetitive nature of sequence where short-read algorithms detect SVs (Chaisson et al., 2019), and because both validation methods rely on the same assemblies. Similar to the VaPoR comparison, there was a net increase of 16-373 calls for which TT-Mars provides a classification compared to dipcall+truvari. On the ten pbsv callsets, TT-Mars and dipcall+truvari have agreeing classification for 83.8%-86.9% of the analyzed calls (Figure 4b), although TT-Mars analyzes 1,497-2,229 more insertions/deletions than dipcall+truvari.

Figure 4: Comparison of dipcall+truvari and TT-Mars on short read **a** and long read **b** callsets of 10 sample genomes. The combinations of TP and FP for dipcall+truvari/TT-Mars annotations are given with a horizontal scatter to distinguish points in individual categories.

The increased number of calls from long-read callsets provide granularity for which the validation approaches may be compared. We analyzed the classification results of TT-Mars, VaPoR, and dipcall+truvari on the pbsv callset on HG00096 (arbitrarily selected from the ten benchmark samples due to VaPoR runtime). There was no association between length (Figure 3b) nor genomic organization (Supplementary Figure 3) for whether TT-Mars and dipcall+truvari provided a validation, however the majority of the variants where VaPoR does not provide a call while TT-Mars does are less than 100 bases. To gauge the true results when two methods disagree, we randomly selected 50 (out of 2,434) calls with TT-Mars TP and dipcall+Truvai FP results for manual inspection using IGV (Robinson et al., 2011). There are various sources for the discrepancy of validation result (Supplementary 4.1). In 27 cases, a call made by pbsv is fragmented into multiple calls by dipcall (Figure 5), highlighting how the assembly-based approach can account for validation in repetitive DNA where breakpoints are uncertain. In 19 other cases there are correct variants near by and the validation depends on the genomic region considered by truvari. The remaining four cases do not show an SV in the alignments generated by minimap2 that are used for validation, but do show the SV in the LRA based alignments, indicating how call-based validation rather than assembly-based validation is reliant

Table 4: Comparison of TT-Mars and VaPoR. The two methods agree on most calls and can analyze a similar number of calls across short-read callsets. On the long read callset, TT-Mars evaluates 39,187 additional variants, the majority of which are under 100 bases.

		TT-Mars															
		Wham				LUMPY				DELLY				pbsv			
		TP	FP	NA	Sum	TP	FP	NA	Sum	TP	FP	NA	Sum	TP	FP	NA	Sum
VaPoR	TP	1503	98	76	1677	2614	288	282	3184	5211	405	480	6096	42268	8694	3531	54493
	FP	42	108	29	179	163	434	276	873	1009	627	581	2217	12243	1320	870	14433
	NA	267	149	80	496	345	515	1370	2230	2343	581	991	3915	37646	5942	2765	46353
	Sum	1812	355	185	2352	3122	1237	1928	6287	8563	1613	2052	12228	92157	15956	7166	115279
		HG00171				HG00096				HG03009				HG00096			

upon agreement of gap penalty between detection and validation algorithms. We also manually inspected the converse case 50 (out of 280) calls with TT-Mars FP and dipcall+truvai TP results. Of these, 47 calls have matched SVs in the assembly alignment used by dipcall (using minimap2) and 45 calls do not have matched SVs in the assembly used by TT-Mars (lra). These represent the cases the assembly has unexpected complicated alignment to the reference leading to spurious gaps (Supplementary 4.2).

Figure 5: An example shows TT-Mars correctly validates an insertion but dipcall+truvai does not in a tandem repeat region, where the SV is split into two smaller insertions on the assembly alignment. The insertion with length 997 bases is on chromosome 4, coordinate 1046348. The first track is the pbsv callset, and the second track is the dipcall callset, followed by alignments used by TT-Mars (two tracks) and dipcall (two tracks). The bottom track is a tandem duplication region.

4 Discussion

Here we demonstrated an approach to validate SV using haplotype-resolved *de novo* assemblies where a call is validated by comparing the inferred content of a genome to the assembly, rather than comparing variant calls. This approach is implemented in the software TT-Mars, which is distributed along with accompanying data files required to benchmark variant calls on up to ten samples. The majority of calls that are evaluated by both the GIAB + truvai and TT-Mars (96.0%-99.6%) have the same classification, while evaluating 121-10,966 additional calls per dataset. Furthermore, we are able to evaluate 302-39,187 additional calls per callset compared to the long-read validation tool VaPoR, while not requiring users to download entire read data sets. We show that the assembly-content approach improves over benchmark data sets created from assembly alignment in repetitive regions of the genome because variants split into multiple calls from the assembly alignment are matched in the breakpoint searching procedure.

Duplication calls are a particularly challenging class of variant to validate. This has been done using comparison to array data (Sudmant et al., 2015), and more recently through semi-automated and automated visual inspection of calls (Belyeu et al., 2021). Ground truth SV callsets created from assemblies detect duplications as insertions relative to the reference and are annotated by the site of insertion in the genome, compared to duplication calls in short-read callsets that annotate the source interval on the genome that is duplicated. By checking for both tandem duplications, which can be confirmed as insertions in assembly callsets, as well as interspersed duplications using the assemblies, TT-Mars can help provide insight on the classes of variation detected by SV discovery algorithms.

Our analysis enables SV benchmarks to be reported as a spread over multiple genomes, rather than a point estimate on one curated data set, and provides insight on how extensible methods are on additional

genomes from a single curated callset. Fortunately, most methods had accuracy within 1%-2% across all genomes, indicating results are extensible across genomes and, for example, the results in current large-scale short-read based genomics studies likely hold valid. Currently, TT-Mars only validates deletion, insertion, inversion, and duplication calls. As additional methods emerge to detect more complex forms of variation (Lin et al., 2021), these models can be incorporated into the edit operations used by TT-Mars.

TT-Mars does not measure sensitivity because it does not fit into the paradigm of comparing inferred content, and requires a variant to be called. Studies that must evaluate sensitivity should consider variants detected by haplotype-resolved assemblies that are not within the vicinity of the validated calls, and should consider a class of variant that may have multiple representations when reporting results. Furthermore, because of the difference in scale of the number of variants discovered in short and long-read studies, it is important to consider the sensitivity of medically relevant variants compared to the total variant count.

5 Availability

The software is available at <https://github.com/ChaissonLab/TT-Mars.git>. The necessary lift-over files required to run TT-Mars as well as links to the original assemblies are provided through this repository.

References

- Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., Kanchi, K. L., ... others (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, 583(7814), 83–89.
- Alkan, C., Coe, B. P., & Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5), 363–376.
- Belyeu, J. R., Chowdhury, M., Brown, J., Pedersen, B. S., Cormier, M. J., Quinlan, A. R., & Layer, R. M. (2021). Samplot: a platform for structural variant visual validation and automated filtering. *Genome biology*, 22(1), 1–13.
- Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using dna microarrays. *Nature genetics*, 39(7), S16–S21.
- Chaisson, M. J., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., ... others (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature communications*, 10(1), 1–16.
- Chaisson, M. J., Wilson, R. K., & Eichler, E. E. (2015). Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, 16(11), 627–640.
- Chen, S., Krusche, P., Dolzhenko, E., Sherman, R. M., Petrovski, R., Schlesinger, F., ... others (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome biology*, 20(1), 1–13.
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., ... Saunders, C. T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32(8), 1220–1222.
- Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, 18(2), 170–175.
- Chin, C.-S., Wagner, J., Zeng, Q., Garrison, E., Garg, S., Fungtammasan, A., ... others (2020). A diploid assembly-based benchmark for variants in the major histocompatibility complex. *Nature communications*, 11(1), 1–9.
- Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore dna sequencing. *Nature nanotechnology*, 4(4), 265–270.
- Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., ... others (2020). A structural variation reference for medical and population genetics. *Nature*, 581(7809), 444–451.
- consortium, U., et al. (2015). The uk10k project identifies rare variants in health and disease. *Nature*, 526(7571), 82.
- Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., ... others (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., ... others (2009). Real-time dna sequencing from single polymerase molecules. *Science*, 323(5910), 133–138.
- English, A. C., Salerno, W. J., Hampton, O. A., Gonzaga-Jauregui, C., Ambreth, S., Ritter, D. I., ... others (2015). Assessing structural variation in a personal genome—towards a human reference diploid genome. *BMC genomics*, 16(1), 286.
- Huddleston, J., Chaisson, M. J., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., ... others (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome research*, 27(5), 677–685.

- Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., ... Wang, Y. (2020). Long-read-based human genomic structural variation detection with cutesv. *Genome biology*, 21(1), 1–24.
- Jiang, Y.-h., Yuen, R. K., Jin, X., Wang, M., Chen, N., Wu, X., ... others (2013). Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *The American Journal of Human Genetics*, 93(2), 249–263.
- Kronenberg, Z. N., Osborne, E. J., Cone, K. R., Kennedy, B. J., Domyan, E. T., Shapiro, M. D., ... Yandell, M. (2015). Wham: identifying structural variants of biological consequence. *PLoS computational biology*, 11(12), e1004572.
- Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). Lumpy: a probabilistic framework for structural variant discovery. *Genome biology*, 15(6), R84.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094–3100.
- Li, H., Bloom, J. M., Farjoun, Y., Fleharty, M., Gauthier, L., Neale, B., & MacArthur, D. (2018). A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nature methods*, 15(8), 595–597.
- Lin, J., Yang, X., Kosters, W., Xu, T., Jia, Y., Wang, S., ... Ye, K. (2021). Mako: A graph-based pattern growth approach to detect complex structural variants. *Genomics, Proteomics & Bioinformatics*. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1672022921001431> doi: <https://doi.org/10.1016/j.gpb.2021.03.007>
- Macintyre, G., Ylstra, B., & Brenton, J. D. (2016). Sequencing structural variants in cancer for precision therapeutics. *Trends in Genetics*, 32(9), 530–542.
- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., & Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it. *Genome biology*, 20(1), 246.
- Merker, J. D., Wenger, A. M., Sneddon, T., Grove, M., Zappala, Z., Fresard, L., ... others (2018). Long-read genome sequencing identifies causal structural variation in a mendelian disease. *Genetics in Medicine*, 20(1), 159–163.
- Niemi, M. E., Martin, H. C., Rice, D. L., Gallone, G., Gordon, S., Kelemen, M., ... others (2018). Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. *Nature*, 562(7726), 268–271.
- Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., ... others (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome biology*, 11(5), R52.
- Parikh, H., Mohiyuddin, M., Lam, H. Y., Iyer, H., Chen, D., Pratt, M., ... others (2016). svclassify: a method to establish benchmark structural variant calls. *BMC genomics*, 17(1), 64.
- Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), i333–i339.
- Ren, J., & Chaisson, M. J. (2021). Ira: A long read aligner for sequences and contigs. *PLOS Computational Biology*, 17(6), e1009078.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1), 24–26.
- Sanchis-Juan, A., Stephens, J., French, C. E., Gleadall, N., Mégy, K., Penkett, C., ... others (2018). Complex structural variants in mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome medicine*, 10(1), 1–10.
- Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature methods*, 15(6), 461–468.
- Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., ... others (2020). Nanopore sequencing and the shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature biotechnology*, 38(9), 1044–1053.
- Stancu, M. C., Van Roosmalen, M. J., Renkens, I., Nieboer, M. M., Middelkamp, S., De Ligt, J., ... others (2017). Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature communications*, 8(1), 1–13.
- Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., ... others (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571), 75–81.
- Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., ... others (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 37(10), 1155–1162.
- Zhao, X., Collins, R. L., Lee, W.-P., Weber, A. M., Jun, Y., Zhu, Q., ... others (2021). Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *The American Journal of Human Genetics*.
- Zhao, X., Weber, A. M., & Mills, R. E. (2017). A recurrence-based approach for validating structural variation using long-read sequencing technology. *GigaScience*, 6(8), gix061.
- Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., ... others (2020). A robust benchmark for detection of germline large deletions and insertions. *Nature biotechnology*, 1–9.