TITLE: Accurate detection of structural variation is hard

Kyle Lesack[1,2], Grace M. Mariene[1,2], Erik C. Andersen[3], James D. Wasmuth[1,2,*]

1. Faculty of Veterinary Medicine, University of Calgary, Alberta, Canada

2. Host-Parasite Interactions Research Training Network, University of Calgary, Alberta, Canada

3. Department of Molecular Biosciences, Northwestern University, IL, USA

* Corresponding author: jwasmuth@ucalgary.ca

ABSTRACT

The accurate characterization of structural variation is crucial for our understanding of how large chromosomal alterations affect phenotypic differences and contribute to genome evolution. Whole-genome sequencing is a popular approach for identifying structural variants, but the accuracy of popular tools remains unclear due to the limitations of existing benchmarks. Moreover, the performance of these tools for predicting variants in non-human genomes is less certain, as most tools were developed and benchmarked using data from the human genome.

To address this problem, multiple short- and long-read tools were benchmarked using real and simulated *Caenorhabditis elegans* whole-genome sequence data. To evaluate the use of long-read data for the validation of short-read predictions, the agreement between predictions from a short-read ensemble learning method and long-read tools were compared. The results obtained indicate that the best performing tool is contingent on the type and size of the variant, as well as the sequencing depth of coverage. These results also highlight the need for reference datasets generated from real data that can be used as 'ground truth' in benchmarks.

INTRODUCTION

Large alterations in chromosome structure contribute substantially to the genetic diversity observed in natural populations and play a fundamental role in the evolution of novel genes (Kaessmann 2010). These changes that span large segments of the genome (*e.g.*, > 100 bp) are termed structural variants (SVs) and include deletions, tandem and interspersed duplications, insertions, and inversions. SVs may be neutral, deleterious, or adaptive (Hurles et al. 2008), and are known to facilitate speciation (Mérot et al. 2020). SVs drive genome evolution using several mechanisms. For example, large heterozygous inversions can suppress recombination, thereby protecting locally adapted alleles (Faria et al. 2019). Also, copy number variation (CNV) is an important factor in genome evolution that describes the gain or loss of genes. CNVs are associated with a wide range of phenotypic effects due to the modulation of gene expression, including differential drug responses between individuals (Santos et al. 2018), HIV susceptibility (Liu et al. 2010), autism spectrum disorders (Vicari et al. 2019), and schizophrenia (Marshall et al. 2017). Gene duplication and subsequent diversification is a source of novel genes and functional diversification (Dos Santos et al. 2016; Storz et al. 2013; Marques et al. 2008). Importantly, SVs have also been proposed as a source of "missing heritability" seen in genome-wide association studies (Manolio et al. 2009).

Several approaches have been developed for the detection of SVs from paired-end whole-genome sequencing (WGS) data (Kosugi et al. 2019; Ho et al. 2020). Paired-end approaches detect SVs when the orientation of a mapped read-pair is inconsistent with the reference genome or when the alignment produces an unexpected insert size. Split-read approaches detect SVs by identifying individual reads that span a given variant, resulting in at least two partial alignments. CNVs may be detected using read-depth differences caused by gene loss or gain. Hybrid approaches that combine multiple signals are employed by many tools and, recently, ensemble methods that leverage multiple separate SV callers have been developed (Becker et al. 2018; Zarate et al. 2021).

The emergence of international 'sequence everything' projects (Lewin et al. 2018; Blaxter 2022) and continual reduction in sequencing costs enable researchers to study the role that SV plays in the evolution of their favourite species. Although numerous tools are available, most benchmarks are limited to the human genome and a limited range of sequencing depths. Due to the difficulty in validating SVs, benchmarks often rely upon simulated data or incomplete sets of experimentally validated variants (Heller and Vingron 2019). Long-read support has also been used to validate SV calls (Layer et al. 2014), but the accuracy of this method is unknown. Because genome properties, such as repeat content or heterozygosity, can differ significantly between species, it is unclear how callers benchmarked on the human genome will perform for other species.

Here, the performance of SV calling in *Caenorhabditis elegans* was evaluated using real and simulated data. Multiple commonly used short- and long-read SV calling tools were benchmarked using mock genomes containing simulated variants. Real data were used to assess the degree of overlap between individual callers and to determine if the results from an ensemble short-read approach were comparable to those obtained by long-read callers. The results shown here demonstrate that SV prediction depends highly on the tool used and that optimal tool choice for each platform depends on the type and size of SV.

Results

Structural Variation Prediction Using Simulated Short-Read Data

To evaluate the performance of various short-read structural variant calling methods, mock genomes containing simulated deletions, duplications, and inversions were created, and used to generate simulated Illumina reads at 5X, 15X, 30X, and 60X sequencing depth of coverage. The caller performance varied by variant type and depth (Table 1).

		Deletions			**Duplications**			**Inversions**		
Caller	Depth	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
BreakDancer	5X	0.93	0.76	0.84	0.99	0.75	0.85	0.95	0.90	0.92
cnMOPS[2]	5X	0.83	0.32	0.46	0.61	0.74	0.67	NA	NA	NA
CNVnator[2]	5X	**1.0**	0.36	0.53	**1.00**	0.65	0.79	NA	NA	NA
DELLY	5X	**1.0**	**0.85**	**0.92**	**1.00**	**0.87**	**0.93**	0.96	**0.94**	**0.95**
Hydra	5X	0.93	0.66	0.78	0.36	0.26	0.30	0.98	0.28	0.43
Lumpy	5X	**1.0**	0.76	0.87	**1.00**	0.75	0.86	**1.00**	0.90	0.94
FusorSV[1]	5X	1.0	0.84	0.92	1.00	0.88	0.94	0.98	0.92	0.95
BreakDancer	15X	0.96	0.89	0.92	**1.00**	0.86	0.93	0.93	0.91	0.92
cnMOPS[2]	15X	0.86	0.62	0.72	0.35	0.79	0.48	NA	NA	NA
CNVnator[2]	15X	**1.00**	0.66	0.79	0.99	0.73	0.84	NA	NA	NA
DELLY	15X	**1.00**	**0.96**	**0.98**	**1.00**	**0.92**	**0.96**	0.94	**0.95**	0.95
Hydra	15X	0.89	0.71	0.79	0.56	0.28	0.37	0.40	0.02	0.04

Lumpy	15X	**1.00**	0.93	0.96	**1.00**	0.87	0.93	**1.00**	0.93	**0.97**
FusorSV[1]	15X	1.00	0.94	0.97	1.00	0.94	0.97	0.99	0.93	0.96
BreakDancer	30X	0.95	0.88	0.92	**1.00**	0.86	0.93	0.93	0.91	0.92
cnMOPS[2]	30X	0.73	0.78	0.76	0.20	0.82	0.32	NA	NA	NA
CNVnator[2]	30X	0.99	0.79	0.88	0.99	0.8	0.88	NA	NA	NA
DELLY	30X	**1.00**	**0.97**	**0.98**	0.99	**0.93**	**0.96**	0.94	**0.96**	0.95
Hydra	30X	0.85	0.71	0.77	0.77	0.27	0.40	0.27	0.02	0.04
Lumpy	30X	0.99	0.93	0.96	**1.00**	0.90	0.95	**1.00**	0.94	**0.97**
FusorSV[1]	30X	1.00	0.94	0.97	1.00	0.93	0.97	0.98	0.92	0.96
BreakDancer	60X	0.94	0.88	0.91	**1.00**	0.86	0.93	0.93	0.92	0.92
cnMOPS[2]	60X	0.48	0.86	0.62	0.11	0.88	0.20	NA	NA	NA
CNVnator[2]	60X	0.81	0.88	0.84	0.95	0.82	0.88	NA	NA	NA
DELLY	60X	**0.98**	**0.97**	**0.98**	0.99	**0.93**	**0.96**	0.93	**0.95**	0.94
Hydra	60X	0.79	0.67	0.72	0.77	0.26	0.38	0.21	0.01	0.03
Lumpy	60X	0.41	0.94	0.57	**1.00**	0.91	0.95	**1.00**	0.94	**0.97**
FusorSV[1]	60X	1.00	0.95	0.97	1.00	0.93	0.96	0.99	0.94	0.97

1. FusorSV used a training model trained on simulated data for the other callers.
2. Tool doesn't predict inversions.

.

For deletion calls, DELLY had the highest F-measure scores at all sequencing depths, followed closely by BreakDancer. The sequencing depth had a stronger impact on the other variant callers, except for Hydra. The accuracies for cnMOPS, CNVnator, and Tigra all improved considerably above 5X, while increased false positives accounted for the decreased accuracy in Lumpy at 60X. The performance of FusorSV was similar the best performing tools for each variant type at all depths.

No single metric completely describes the performance of each variant caller. Because the precision, recall, and F1 scores were calculated based on the number of true positives, false positives, and false negatives, they fail to describe the performance of each caller at the base pair level. The Jaccard similarity score was used to describe the base pair overlap between the predicted variants and simulated variants. Although DELLY had the highest accuracy at all depths, its Jaccard value decreased from 0.99 at 30X depth to 0.74 at 60X depth (Figure 1). This decrease was caused by a 2.56Mbp false positive at the 60X depth. Differences between the

Jaccard and F1 scores can also be caused by size-dependent performance differences. At 5X depth, the CNVnator F1 and Jaccard scores were 0.53 and 0.91 respectively (Supplemental_Table_S1.xls). Because CNVnator performed well at predicting larger variants, a higher Jaccard score was obtained despite low accuracy for smaller deletion sizes. Conversely, the Hydra F1 scores ranged between 0.72 and 0.79, while its Jaccard scores ranged between 0.12 and 0.13. The large discrepancies between the F1 score and Jaccard scores resulted from good performance for smaller deletions but poor performance for larger sizes.

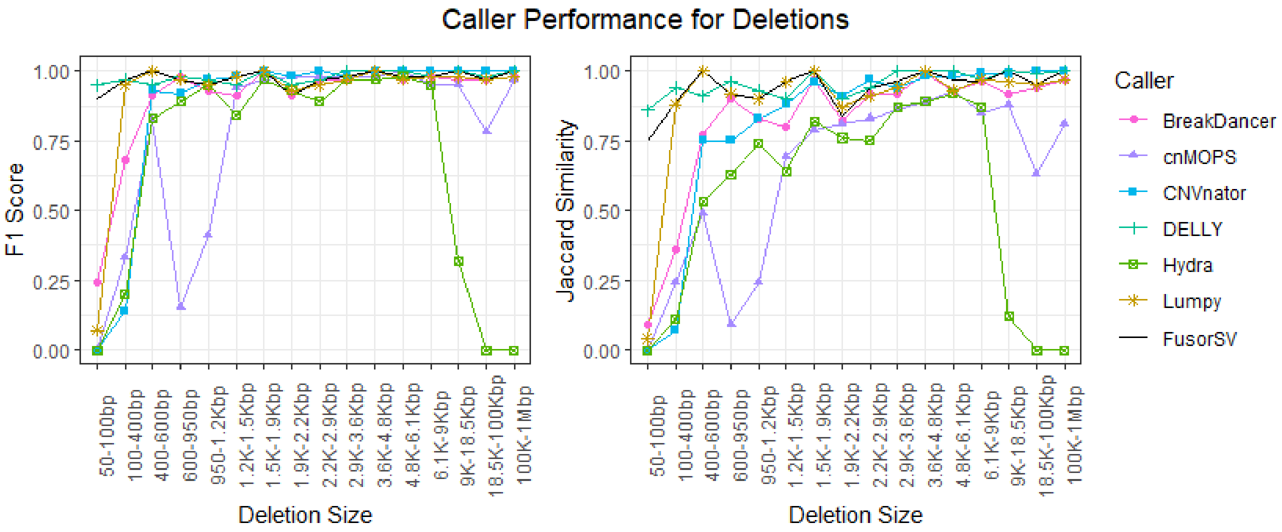

Figure 1 – Accuracy of predicted deletions from simulated short-read data. Results shown for 60X depth.

For duplication calls, DELLY had the highest F-measure scores at all depths, followed closely by BreakDancer. The accuracies of CNVnator, Hydra, and Lumpy improved with increased depth between 5X and 30X, while only a slight decrease was observed for Hydra at 60X depth. The accuracy of Tigra was highest at 15X and decreased with further increases in depth. Increased depth decreased the cnMOPS accuracy due to higher false positive rates. The performance of FusorSV was similar the best performing tools for each variant type at all depths.

The performance of cnMOPS and Hydra were both dependent on the size of the predicted duplications. The rate of false positive duplications increased with higher sequencing depths for cnMOPS and was biased towards smaller duplication sizes (Figure 2). All predicted duplications from Hydra were below 10 Kbp, which led to a Jaccard score of 0.01 at all depths (Supplemental_Table_S2.xls). At 60X depth, DELLY predicted a 2.32 Mbp false positive duplication, which resulted in a decreased Jaccard score (0.83).

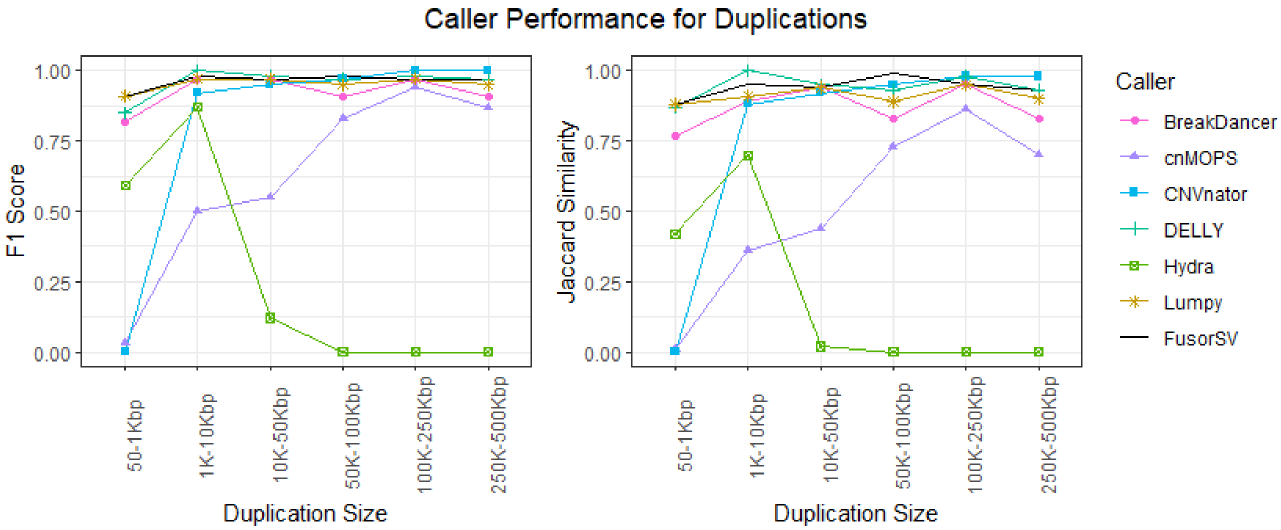

Figure 2 – Accuracy of predicted duplications from simulated short-read data. Results shown for 60X depth.

BreakDancer, DELLY, and Lumpy performed well for the prediction of inversions, as the accuracy of each tool was at least 0.92 at all depths. The accuracy for Hydra was considerably lower at 5X and decreased with increasing depth. The precision, recall, and F1 score for FusorSV was similar to the best performing tools for each variant type at all depths. However, lower Jaccard scores were observed in FusorSV (Supplemental_Table_S3.xls) due to several multiple megabase spanning false positives that were not predicted by other callers.

The performance of BreakDancer, Hydra, and DELLY were dependent on the size of the predicted inversions (Figure 3). Both BreakDancer and DELLY performed better for higher inversions sizes, while the performance of Hydra decreased with increasing size. A 4.27 Mbp false positive in DELLY at 15X, 30X, and 60X resulted in decreased Jaccard scores.

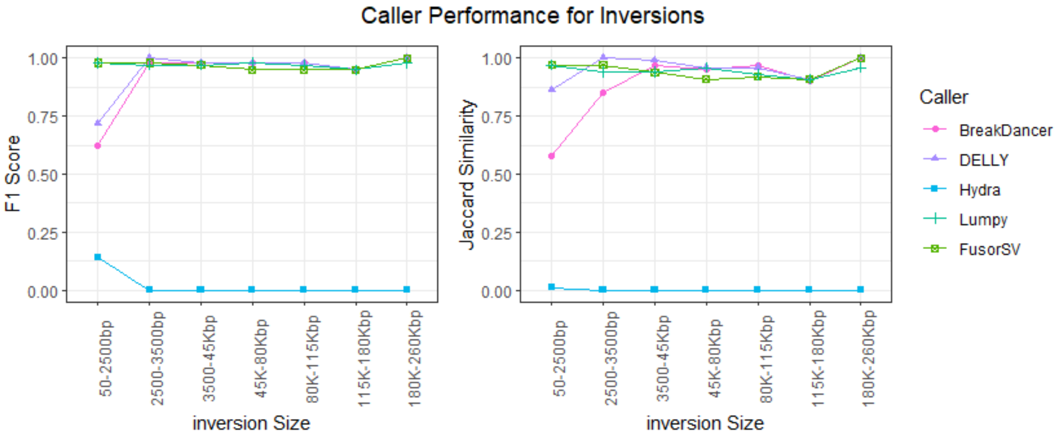

Figure 3 – Accuracy of predicted inversions from simulated short-read data. Results shown for 60X depth.

Prediction of Known Structural Variants in *C. elegans*

BC4586 is a *C. elegans* strain containing experimentally validated structural variants (Maroilley et al. 2021). Publicly available Illumina sequencing data allowed us to determine if the short-read SV callers in the SVE/FusorSV pipeline can resolve a 3910bp deletion (DEL-1; IV:9,853,675–9,857,585), a 552bp tandem duplication (DUP-1; IV:9,853,123– 9,853,675), and a 4812bp inversion (INV-1; IV:9,857,585–9,862,397). Only cnMOPs and CNVnator were able to resolve the deletion. Among the callers capable of predicting inversions, BreakDancer, DELLY, and Hydra predicted the inversion. Each caller predicted the tandem duplication. Both BreakDancer and DELLY predicted multiple overlapping duplications spanning DUP-1.

Caller	DEL-1	INV-1	DUP-1
BreakDancer	No	Yes	Yes[2]
cnMOPs	Yes	N/A[1]	Yes
CNVnator	Yes	N/A[1]	Yes
DELLY	No	Yes	Yes[3]
Hydra	No	Yes	Yes
Lumpy	No	No	Yes
FusorSV	No	Yes	Yes

1. Neither cnMOPs nor CNVnator predict inversions.

2. BreakDancer predicted two inversions spanning the INV-1 genome coordinates
3. DELLY predicted three inversions spanning the INV-1 genome coordinates

Structural Variation Prediction Using Simulated Long-Read Data

Simulated PacBio DNA sequencing data was used to evaluate using long-read sequencing to validate SV calls generated from short-read sequencing platforms. The mock genomes containing simulated deletions, duplications, and inversions were used to generate simulated PacBio reads at 5X, 15X, 30X, 60X, and 142X depth of coverage.

The performance of each caller varied considerably by variant type and depth (Table 2). For deletions, SVIM had a considerably higher accuracy at 5X depth (F1-score = 0.848) compared to pbsv (F1-score = 0.529) and Sniffles (F1-score = 0.052). SVIM had the highest accuracy at 15X (F1-score = 0.945), followed by Sniffles (F1-score = 0.918), and pbsv (F1-score = 0.566). Sniffles had the highest accuracy at 30X depth (F1-score = 0.993), followed by SVIM (F1-score = 0.959) and pbsv (F1-score = 0.599). SVIM had the highest accuracy at 60X (F1-score = 0.956), followed by Sniffles (F1-score = 0.929) and pbsv (F1-score = 0.781). At 142X, SVIM had the highest accuracy (F1-score = 0.956), followed by pbsv (F1-score = 0.793) and Sniffles (F1-score = 0.423).

The accuracy of predicted duplications at 5X was higher for SVIM (F1-score = 0.452), compared to pbsv (F1-score = 0.356) and Sniffles (F1-score = 0.033). At 15X, 30X, and 60X depth, Sniffles had the highest accuracy (15X F1-score = 0.909; 30X F1-score = 0.983; 60X F1-score = 0.935) compared to pbsv (15X F1-score = 0.378; 30X F1-score = 0.462; 60X F1-score = 0.554) and SVIM (15X F1-score = 0.571; 30X F1-score = 0.605; 60X F1-score = 0.636). SVIM had the highest accuracy at 142X (F1-score = 0.659), followed by pbsv (F1-score = 0.531), and Sniffles (F1-score = 0.321). Both pbsv and SVIM had lower recall than precision, indicating that missed variant calls decreased the accuracy of these callers. Lower recall also decreased the accuracy of Sniffles at 5X, 15X, and 30X depth, while lower precision contributed more at 60X and 142X depth.

The accuracy of predicted inversions was higher in pbsv at 5X (F1-score = 0.686), 60X (F1-score = 0.745) and 142X depth (F1-score = 0.760). Sniffles had the highest accuracy at 15X (F1-score = 0.978) and 30X depth (F1-score = 0.923). Lower precision in SVIM at 5X (0.457) accounted for lower accuracy (F1-score = 0.493). At 15X, the SVIM precision and F1-scores increased to 0.836 and 0.742, respectively. The highest SVIM precision (0.918) and accuracy (F1-score = 0.763) was obtained at 30X. The SVIM precision decreased at 60X (0.833) and 142X depth (0.153), which resulted in lower accuracy (60X F1-score = 0.732; 142X F1-score = 0.248). At each depth, recall contributed more to decreased accuracy in pbsv. For Sniffles, lower recall contributed more to lower accuracy at 5X and 15X depth, while lower precision contributed more to lower accuracy at the higher depths.

Table 2 – Performance of long-read structural variant callers on simulated data.

		Deletions			**Duplications**			**Inversions**		
Caller	Depth	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
pbsv	5X	0.486	0.580	0.529	**1.000**	0.217	0.356	**1.000**	0.522	**0.686**
Sniffles	5X	**1.000**	0.027	0.052	**1.000**	0.017	0.033	**1.000**	0.232	0.376
SVIM	5X	0.857	**0.84**	**0.848**	0.792	**0.317**	**0.452**	0.457	**0.536**	0.493
pbsv	15X	0.410	**0.913**	0.566	**1.000**	0.233	0.378	**1.000**	0.594	0.745
Sniffles	15X	**0.985**	0.860	0.918	**1.000**	**0.833**	**0.909**	**1.000**	**0.957**	**0.978**
SVIM	15X	0.979	**0.913**	**0.945**	**1.000**	0.4	0.571	0.836	0.667	0.742
pbsv	30X	0.444	0.920	0.599	**1.000**	0.300	0.462	**1.00**	0.594	0.745
Sniffles	30X	**0.993**	**0.993**	**0.993**	**1.000**	**0.967**	**0.983**	0.892	**0.957**	**0.923**
SVIM	30X	**0.993**	0.927	0.959	**1.000**	0.433	0.605	0.918	0.652	0.763
pbsv	60X	0.675	0.927	0.781	**1.000**	0.383	0.554	**1.000**	0.594	**0.745**
Sniffles	60X	0.867	**1.000**	0.929	0.906	**0.967**	**0.935**	0.328	**0.971**	0.491
SVIM	60X	0.979	0.933	**0.956**	**1.000**	0.467	0.636	0.833	0.652	0.732
pbsv	142X	0.693	0.927	0.793	**1.000**	0.361	0.531	1.000	0.614	**0.760**
Sniffles	142X	0.269	**0.996**	0.423	0.192	**0.967**	0.321	0.087	0.952	0.160
SVIM	142X	**0.972**	0.940	**0.956**	0.968	0.500	**0.659**	0.153	0.652	0.248

Agreement in Predicted Structural Variants for Wild *C. elegans* strains

To evaluate the usefulness of long-read DNA sequencing data to validate structural variants predicted from short-read technologies, we obtained data for 14 *C. elegans* wild strains with both Illumina and PacBio sequencing data.

The predicted variants varied considerably between the callers (Table 3). The predicted deletions ranged from a median of 76 per strain in SVIM to 341 in cnMOPs. Conversely, cnMOPs only predicted a median of 5.5 duplications per strain compared to 129 in Sniffles. The median predicted inversions per strain ranged from 8 in Lumpy to 88 in BreakDancer.

Table 3 – Predicted deletions, duplications, and inversions in *C. elegans*

	Caller	**Deletions**		**Duplications**		**Inversions**	
		Median deletions per strain	Median genes spanned by deletions per strain	Median duplications per strain	Median genes spanned by duplications per strain	Median inversions per strain	Median genes spanned by inversions per strain
Long-read tools	Assemblytics	233.5	787	25	49.5	N/A[1]	N/A[1]
	MUM&Co	81	346	15.5	20	23.5	98.5
	PBSV	87	185	26	61.5	29	84
	Sniffles	171.5	385.5	129	393	63	206.5
	SVIM	76	119.5	71	112.5	21	61.5
Short-read tools	FusorSV	124.5	1363.5	128	1158	50	716.5
	BreakDancer	101	269.5	39.5	151.5	88	673
	cnMOPs	341	585.5	5.5	9.5	N/A[1]	N/A[1]
	CNVnator	269	2128.5	55	901	N/A[1]	N/A[1]
	DELLY	106	308	108	509.5	88	720
	Hydra	120.5	167.5	79.5	117	28.5	420

	Lumpy	107	340	107	551	8	36.5

1. Does not support inversions

Because the exact breakpoints for the same predicted variant may differ between callers, it is difficult to directly compare the agreement of calls generated by different tools. Therefore, predicted variants spanning protein-coding genes were used to compare caller agreement. Overlapping predictions between the long-read callers and FusorSV were compared to evaluate using long-read sequencing data for the validation of variants predicted from short-read data.

Among the total set of predicted deletions spanning genes, 555 were shared among all long-read callers. Of these deletions, 387 were predicted by FusorSV and 77% of the genes spanned by deletions predicted by FusorSV were not shared by at least one long-read caller (Figure 4; Supplemental_Table_S4.xls). Within the set of genes overlapping deletions predicted by SVIM, 97% were shared by at least one other caller. This overlap was considerably lower than the other long-read tools (Assemblytics = 71%, MUM&Co = 39%, pbsv = 31%, sniffles = 43%, and SVIM = 3%). The percentage of unique genes spanned by deletions also varied among the short-read callers (Supplemental_Table_S5.xls) and ranged from 3% in Lumpy to 48% in CNVnator.

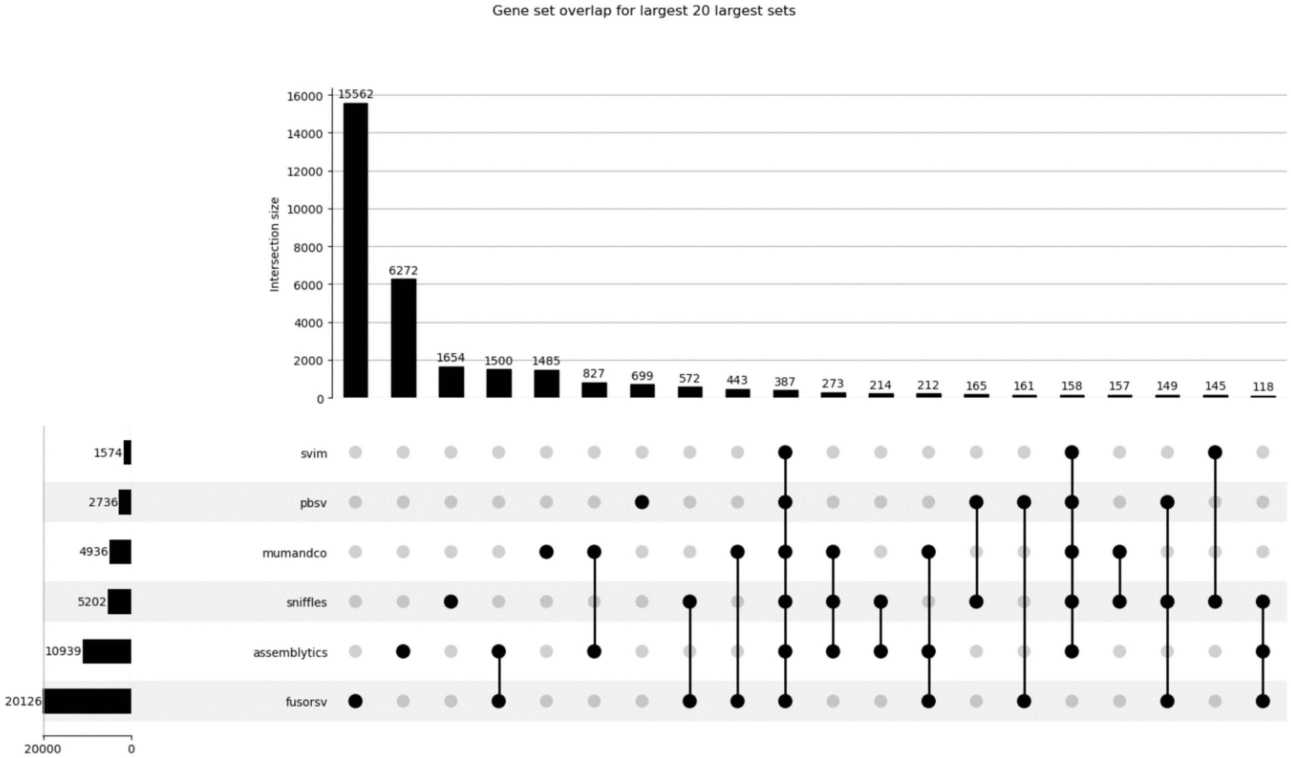

Figure 4 – Overlap of predicted deletions from long-read tools and FusorSV. The plotted results were limited to the 20 largest sets.

Among the total set of predicted duplications spanning genes, only 120 were shared among all long-read callers. Of these 97 were predicted by FusorSV (Figure 5; Supplemental_Table_S6.xls). 88% of the genes spanned by duplications predicted by FusorSV were not shared by at least one long-read caller. Within the set of genes overlapping duplications, MUM&Co contained the least unique predictions (4%) followed by SVIM (10%), Assemblytics (18%), pbsv (58%), and Sniffles (71%). The percentage of unique genes spanned by duplications also varied considerably among the short-read callers (Supplemental_Table_S7.xls) and ranged from 5% in BreakDancer to 77% in CNVnator.

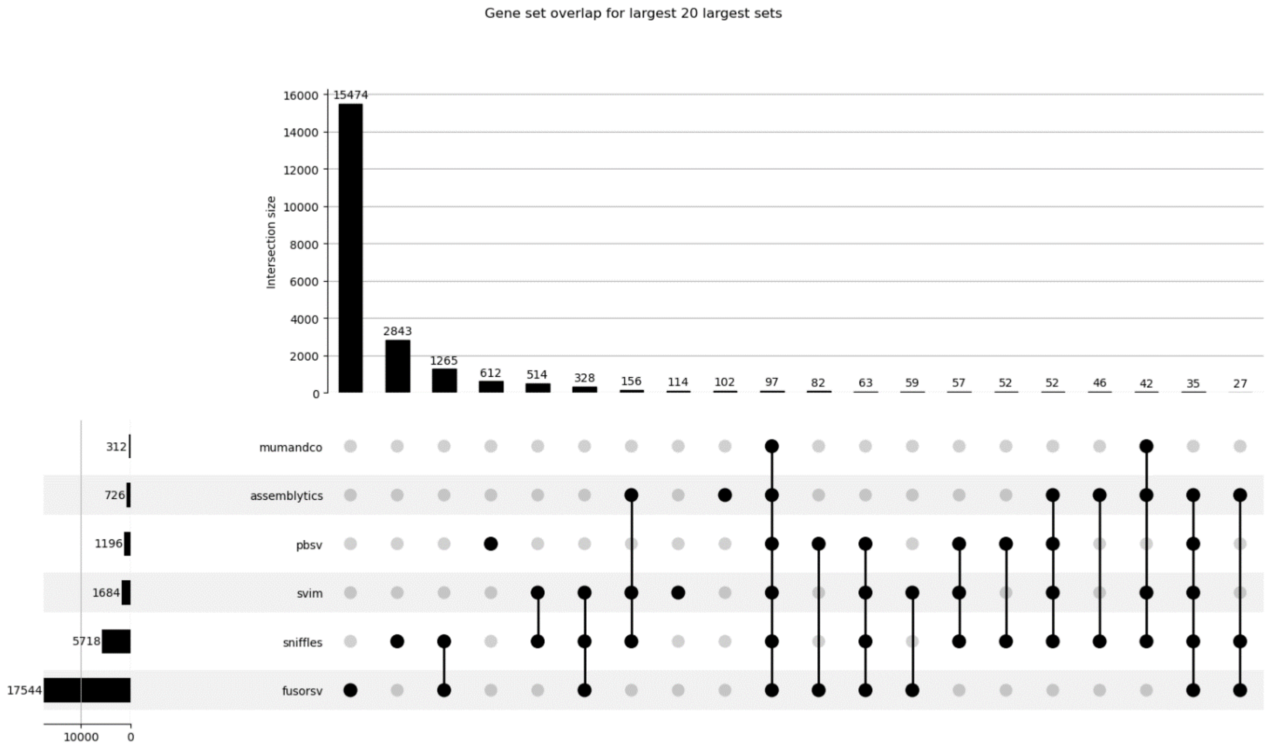

Figure 5 – Overlap of predicted duplications from long-read tools and FusorSV. The plotted results were limited to the 20 largest sets.

Among the total set of predicted inversions spanning genes, 224 were shared among all long-read callers. Of these inversions, 41 were predicted by FusorSV (Figure 6; Supplemental_Table_S8). Within the set of genes overlapping inversions, SVIM contained the least unique predictions (28%) followed by pbsv (49%), MUM&Co (63%), and Sniffles (74%). 90% of the genes spanned by inversions predicted by FusorSV were not shared by at least one long-read caller. The percentage of unique genes spanned by inversions also varied considerably among the short-read callers (Supplemental_Table_S9.xls) and ranged from 11% in Lumpy to 48% in Hydra.

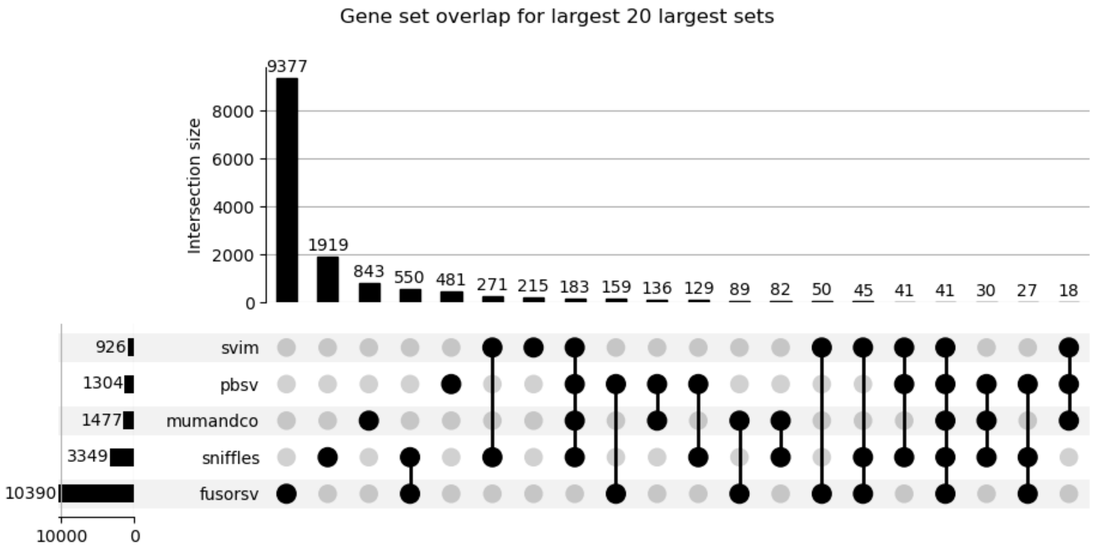

Figure 6 – Overlap of predicted inversions from long-read tools and FusorSV. The plotted results were limited to the 20 largest sets.

Discussion

Structural variant (SV) prediction is challenging for researchers studying non-human genomes. Most SV prediction tools were designed for the human genome, and benchmarks on other species are lacking (Cameron et al. 2019; Kosugi et al. 2019; Zarate et al. 2021; Becker et al. 2018). Without an adequate guide for tool selection, the accuracy of predicted SVs has a high degree of uncertainty. Here, we used simulated and real data to benchmark six short-read structural variant callers included in the SVE (Becker et al. 2018), a pipeline developed to be used with FusorSV, an ensemble learning method that leverages the strengths of each individual caller.

The results for the simulated short-read data suggest that deletions and duplications may be predicted with high confidence using BreakDancer and DELLY, and accurate inversion predictions may be obtained using BreakDancer, DELLY, and Lumpy. The FusorSV performance typically reflected that of the best performing individual tools, but occasionally predicted large megabase spanning false positives. WGS from the 1000

Genomes Project (1000GP) (Sudmant et al. 2015) may be used to benchmark SV prediction using real data. These data provide a high-confidence human truth set, as SVs were validated using a combination of short- and long-read WGS data, Moleculo synthetic long-read sequencing, microarray SV detection, and targeted long-read sequencing. Our results are in discordance with the values reported in the literature for benchmarks generated using human data from the 1000 Genomes Project (1000GP) (Becker et al. 2018). For example, the accuracy of deletion calls from 1000GP data were considerably lower for BreakDancer (F1-score = 0.47), DELLY (F1-score = 0.54), and FusorSV (F1-score = 0.62). Our results for duplications and inversions were substantially better than those reported for BreakDancer (duplication F1-score = 0.00, inversion F1-score = 0.08) DELLY (duplication F1-score = 0.01, inversion F1-score = 0.09), and FusorSV (duplication F1-score = 0.19, inversion F1-score = 0.45) based on 1000GP data. Differences in genome properties, such as repeat content, could account in part for the improvements seen here, but the usage of simulated data was likely a major factor.

For the simulated PacBio data, the performance of each caller varied considerably by variant type and depth. For deletions, SVIM had the highest accuracy at 5X (F1-score = 0.848), 15X (F1-score = 0.945), 60X (F1-score = 0.956), and 142X depth (F1-score = 0.956), while Sniffles had the highest accuracy at 30X depth (F1-score = 0.993). For duplications, SVIM had the highest accuracy at 5X (F1-score = 0.452), and Sniffles had the higher accuracy at 15X (F1-score = 0.909), 30X (F1-score = 0.983), and 60X depth (F1-score = 0.935). Again, SVIM had the highest accuracy at 142X depth (F1-score = 0.659). For inversion calls, a higher accuracy was obtained using pbsv at 5X (F1-score = 0.686), 60X (F1-score = 0.745) and 142X depth (F1-score = 0.760). Sniffles had the highest accuracy at 15X (F1-score = 0.978) and 30X depth (F1-score = 0.923).

It should be noted that SVIM generates a VCF file containing all candidate SV calls, including calls of low confidence. Each prediction includes a quality score between 0 and 100 that provides a confidence estimate. The authors recommend using a threshold between 10-15 for higher depth datasets (*e.g*., >40X) or a threshold that generates the expected number of predictions. Therefore, for datasets with lower sequencing depth, the analyst may be limited to selecting an arbitrary cut-off when the expected number of SVs is unknown. The thresholds we used for 5X (minimum QUAL = 2), 15X (minimum QUAL = 5), and 30X depth (minimum QUAL = 10) were proportional to the decrease in depth compared to the high depth specified by the SVIM authors. The performance for these cut-offs were similar to the optimum cut-off values that were calculated post-hoc(Supplemental_Table_ S10.xls, Supplemental_Table_S11.xls, Supplemental_Table_S12.xls), with the exception of inversions predicted at 142X depth, where a higher threshold is recommended.

Deletion benchmarks were previously described pbsv and Sniffles using data from the Database of Genomic Variants (MacDonald et al. 2014) and NCBI dbVar (MacDonald et al. 2014) projects. The precision and recall were quantified for different numbers of reads supporting the deletions. Precision values up to 0.91 and 0.81 were reported for pbsv and Sniffles, respectively. Lower recall values were observed for pbsv (up to 0.45) and Sniffles (up to 0.26). By contrast, we observed lower precision but higher recall using simulated data.

Illumina and PacBio sequencing data from 14 natural *C. elegans* strains were analyzed to determine the concordance between predicted SVs among both short- and long-read callers. Low agreement was observed among all predictions generated using either Illumina or PacBio data. Furthermore, many SV calls unique to a single caller were observed for predictions made using either short or long-reads. It is therefore difficult to ascertain the accuracy of structural variants described in the literature, as the considerably different results may be generated using different tools. Nonetheless, the simulated data suggests that short-read tools, such as DELLY, BreakDancer, and Lumpy are likely to provide more accurate SV calling across a range of depths compared to the other short-read tools that were included in these benchmarks. If training data are available, FusorSV may also be used, but large megabase spanning inversions should be interpreted with caution, as several large false positives were predicted by this tool. For SV prediction from long-reads, the simulated data suggested that neither pbsv, Sniffles, nor SVIM may be used with high-confidence for the prediction of duplications or inversions using lower depth data (e.g., 5X). Fewer generalizations can be made for higher depths. SVIM had the best performance for the prediction of deletions at 15X depth, while the performance of Sniffles was superior at 30X. Sniffles performed the best for the prediction of duplications and inversions at 15X and 30X depth. At 60X depth, SVIM had the highest accuracy for deletions, but Sniffles and pbsv demonstrated superior performance for duplications and inversions, respectively. At 142X depth, the SVIM accuracy was considerably higher than pbsv and Sniffles for deletions and duplications, while the accuracy of pbsv was considerably higher for inversions. If precision is less of a concern than recall, Sniffles may be the preferable choice for predicting duplications from long-read data.

The concordance between long-read callers is pertinent if long-read data is to be used to validate candidate SVs called using short-read data. Although few predicted SVs were common to all long-read callers, a large majority of the predicted deletions and duplications from SVIM were supported by at least one other caller. Therefore, SVIM might provide a more conservative option for validating deletions and duplications predicted from short-reads. Although inversions predicted by SVIM had the highest support among other callers, over one quarter of these calls were unique to SVIM. Therefore, long-read data may be less reliable for validating inversions predicted from short-reads.

To assess the agreement between short- and long-read approaches, the agreement between FusorSV and the long-read tools was measured. For each variant type, over three quarters of the FusorSV predictions were not shared by any of the long-read tools. Because higher accuracy has been reported for long-read tools in the literature, it is likely that FusorSV generate many false positives. Low agreement was observed for many of the SVs predicted by the individual tools used to train FusorSV despite higher accuracy observed in the simulated data. This may reflect a limitation in using simulated data to train FusorSV, as simulated data may bias the FusorSV models towards callers that perform poorly on real data.

CONCLUSIONS

It is challenging to choose the appropriate tool for the prediction of structural variants from DNA sequencing data. Dozens of callers have been developed for calling structural variants using short-read data, but few independent benchmarks are available. Compounding this problem is the lack of benchmarks for non-human genomes. Here, multiple short-read and long-read callers were compared using both real and simulated *C. elegans* data. The results using simulated data showed that the performance of a given tool often varies considerably according to variant type and sequencing depth and that no single tool performed best for all situations. The predictions generated from real data showed low overlap among all callers and many predictions unique to individual tools.

Because variants predicted from short-reads depend highly on the tool used, the analyst may choose to validate these SVs using long-read data. However, the lack of a consensus among long-read callers suggests that using a long-read caller to generate a “truth-set” warrants caution. Nonetheless, most of the deletions and duplications predicted by SVIM and MUM&Co, respectively, were shared by at least one other caller. These tools may provide a more conservative approach for validating SV calls using long-reads. The availability of reference datasets for which the “ground truth” is known would provide valuable resources for improving our understanding of the best approaches for SV prediction in non-human organisms. Future benchmarking projects would benefit from publicly available data from strains with precise deletions of various lengths generated using CRISPR-Cas9 methods, as well as further lab strains with SVs validated manually using long-read technologies and PCR.

METHODS

STRUCTURAL VARIATION PREDICTION

Structural Variation Engine (SVE) and FusorSV (v0.1.3-beta) (Becker et al. 2018) were used to predict structural variants (deletions, duplications, and insertions) from real and simulated short-read sequencing data. SVE is an SV calling pipeline that produces VCF files compatible with FusorSV. FusorSV uses an ensemble learning approach to call structural variants using a fusion model trained using individual callers. The six structural variant callers included in SVE that support non-human genomes were evaluated here: BreakDancer (Fan et al. 2014), cnMOPS (Klambauer et al. 2012), CNVnator (Abyzov et al. 2011), DELLY (Rausch et al. 2012), Hydra (Lindberg et al. 2015), and Lumpy (Layer et al. 2014). The default SVE and FusorSV parameter settings were used.

Five tools were used to predict structural variants (deletions, duplications, and insertions) from real and simulated long-read sequencing data: Assemblytics (Nattestad and Schatz 2016) (v1.2.1), MUM&Co (Nattestad and Schatz 2016) (v2.4.2), pbsv (Pacific Biosciences) (v2.6.2), Sniffles (Sedlazeck et al. 2018) (v1.0.12a), and SVIM (Heller and Vingron 2019) (v2.0.0). For each tool, the recommended long-read aligner and default parameter settings were used. The genome assemblies and alignments required for Assemblytics and MUM&Co were created in Canu (Koren et al. 2017) (v2.2) and MUMMER (Marçais et al. 2018) (4.0.0rc1) respectively. The alignments used with pbsv were created using pbmm2 (v.1.7.0). The alignments used by Sniffles and SVIM were created using ngmlr (Sedlazeck et al. 2018) (v.0.2.7). Low confidence predictions below the SVIM quality score threshold were discarded using a different cut-off for each sequencing depth (5X = 2; 15X = 5; 30X = 10; 60X = 15, 142X = 15).

Custom Python (v3.7) and Bash scripts were used to select the final set of SV predictions to benchmark based on the following criteria: minimum size >= 100 bp, and vcf file FILTER flag = "PASS".

SIMULATED DATA

Svsim (v. 0.1.1) was used to create mock *C. elegans* genomes containing simulated structural variants (deletions, duplications, and inversions) based on the WormBase (Harris et al. 2004) (WBcel235; https://parasite.wormbase.org/Caenorhabditis_elegans_prjna13758/Info/Index/) reference assembly for the N2 strain. Because FusorSV uses the SV type and size as discriminating features to train the fusion model, training datasets with variable numbers of simulated deletions, duplications, and inversions were created ranging from 200 bp to 280 Kbp. 43 mock genomes were created for the small training dataset and included a total of 10 structural variants of each type per size bin used by FusorSV. The medium training dataset included 86 mock genomes with a total of 20 variants of each type per size bin, and the large training dataset included 129 mock genomes with 30 variants of each type per size bin. A testing dataset was created that included 129 mock genomes with 30 variants of each type per size bin.

Short-read DNA sequencing of the mock genomes was simulated with the randomreads.sh script included with BBTools (38.79). Paired-end reads of 100bp were simulated at 5X, 15X, 30X, and 60X depths of coverage using the Illumina error model with default settings. SimLoRD (v.1.0.4) was used to simulate PacBio sequencing data for the mock genomes. The SimLoRD PacBio sequencing runs were simulated at a depth of 5X, 15X, 30X, 60X, and 142X (the median depth of real PacBio data used to predict SVs using long-read tools). The SVIM quality score cut-off was 1 for 5X, 5 for 15X and 30X, 10 for 60X, and 15 for 142X.

REAL DATA

Data from the *Caenorhabditis elegans* Natural Diversity Resource (CeNDR) (Cook et al. 2017) were used to predict SVs in 14 *C. elegans* isolates collected from the wild. SV prediction using the SVE/FusorSV pipeline was performed using the BAM files provided for the 20200815 CeNDR release. SimuSCoP (v1.0) was used to generate the simulated DNA-sequencing data that trained the FusorSV model. DNA sequencing of the *C. elegans* N2 reference strain were used to create the sequencing profiles used by SimuSCoP (SRA run = SRR3452263, SRA run = SRR1013928, SRA run = SRR9719854). For each strain, FusorSV models trained on simulated data of similar sequencing depth and read length were used to predict variants.

BC4586, a *C. elegans* strain containing validated structural variants, was used to evaluate the ability of the SVE/FusorSV pipeline in the prediction of experimentally validated structural variants. Simulated SimuSCoP DNA sequencing data was generated using an N2 profile (SRA run = SRR14489487) and used to train the FusorSV model. SVs for BC4586 (SRA run = SRR14489485) were predicted using the SVE/FusorSV pipeline and used to identify the presence of a deletion on chromosome IV (coordinates = 9853675-9857585), a tandem duplication on chromosome IV (coordinates = 9857585-9862397), and an inversion on chromosome IV (coordinates = 9853123-9853675).

STRUCTURAL VARIATION BENCHMARKING

Variant calling resulted in multiple overlapping structural variants of the same type, which can lead to inflated performance metrics, as each prediction may be counted as a true positive when compared to the truth dataset. For overlapping SV predictions of the same type and caller, a single call was selected using the criteria described in Supplemental_Table_S13.xls. When no discriminating information was available among overlapping calls, the final SV was selected randomly.

BENCHMARKING USING SIMULATED DATA

For simulated data, each predicted variant was classified as being either a true positive (TP) or false positive (FP) using the Bedtools intersect command. Predictions that overlapped (minimum reciprocal overlap = 0.5) with at least one simulated variant in the mock genome were classified as true positives (TP), and those calls that did not were classified as false positives (FP). Simulated variants that were not predicted were classified as false negatives (FN). These classifications were used to calculate the following performance metrics:

Precision is the ratio of true positives (TP) to the total predicted variants, and was calculated as follows:

$$\text{Precision} = TP / TP + FP$$

Recall is the ration of true positives to the total number of simulated variants, and was calculated as follows:

$$\text{Recall} = TP / TP + FN$$

The F_1 score provides a measure of the prediction accuracy by taking the weighted average of the precision and recall. The F1 score was calculated as follows:

$$F_1 = 2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$$

Because the precision, recall, and F1 scores were calculated using binary classifications, they may lead to misleading benchmarks. For example, true positives are not penalized for predicting breakpoints outside of the region of the structural variant. Similarly, true positives are not penalized for predicting breakpoints within the true breakpoints of the structural variant. Therefore, the Jaccard index was calculated to measure the amount of overlap between the predicted variants and simulated variants.

The Jaccard index was calculated using the ratio of the number of base pairs in the intersection and union of the predicted and simulated variants:

$$\text{Jaccard} = \text{prediction variants} \cap \text{simulated variants} / \text{prediction variants} \cup \text{simulated variants}$$

BENCHMARKING USING REAL DATA

The agreement between different short-read and long-read SV calling approaches were assessed for the CeNDR data due to the lack of a truth set for these strains. *C. elegans* genome annotations obtained from WormBase (Harris et al. 2004) (release = WBPS14) were used to identify which genes were spanned by SVs. SVs larger than 50 Kbp were excluded due to this being the maximum size considered to be reliable in Assemblytics.

DATA ACCESS

Custom Python (v3.7) and Bash scripts are available at https://github.com/kyleLesack/sv_calling_benchmarking.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGEMENTS

This work was supported by Results Driven Agricultural Research (RDAR #2016F013R) to JDW, the Natural Sciences and Engineering Research Council of Canada (NSERC) through Discovery Grants (#06239-2015 and 04589-2020) to JDW, and an NSERC Collaborative Research and Training Experience Program (CREATE) program in Host-Parasite Interactions (#413888-2012) to JDW and others. ECA was supported by a National Science Foundation CAREER award.

REFERENCES

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974–984.

Becker T, Lee WP, Leone J, Zhu Q, Zhang C, Liu S, Sargent J, Shanker K, Mil-homens A, Cerveira E, et al. 2018. FusorSV: An algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol* **19**: 1–14.

Blaxter ML. 2022. Sequence locally, think globally: The Darwin Tree of Life Project. *Proc Natl Acad Sci U S A* **119**: 1–7.

Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* **10**: 1–11. http://dx.doi.org/10.1038/s41467-019-11146-4.

Cook DE, Zdraljevic S, Roberts JP, Andersen EC. 2017. CeNDR, the Caenorhabditis elegans natural diversity resource. *Nucleic Acids Res* **45**: D650–D657.

Dos Santos HG, Nunez-Castilla J, Siltberg-Liberles J. 2016. Functional diversification after gene duplication: Paralog specific regions of structural disorder and phosphorylation in p53, p63, and p73. *PLoS One* **11**: 1–27.

Fan X, Abbott TE, Larson D, Chen K. 2014. BreakDancer: Identification of genomic structural variation from paired-end read mapping. *Curr Protoc Bioinforma* 1–11.

Faria R, Johannesson K, Butlin RK, Westram AM. 2019. Evolving Inversions. *Trends Ecol Evol* **34**: 239–248. https://doi.org/10.1016/j.tree.2018.12.005.

Harris TW, Chen N, Cunningham F, Tello-Ruiz M, Antoshechkin I, Bastiani C, Bieri T, Blasiar D, Bradnam K, Chan J, et al. 2004. WormBase: A multi-species resource for nematode biology and genomics. *Nucleic Acids Res* **32**: 411–417.

Heller D, Vingron M. 2019. SVIM: Structural variant identification using mapped long reads. *Bioinformatics* **35**: 2907–2915.

Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nat Rev Genet* **21**: 171–189. http://dx.doi.org/10.1038/s41576-019-0180-9.

Hurles ME, Dermitzakis ET, Tyler-Smith C. 2008. The functional impact of structural variation in humans. *Trends Genet* **24**: 238–245.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–1326.

Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S. 2012. Cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res* **40**: 1–14.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res* **27**: 722–736.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**: 8–11.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome Biol* **15**: 1–19.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards S V., Forest F, Gilbert MTP, et al. 2018. Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U S A* **115**: 4325–4333.

Lindberg MR, Hall IM, Quinlan AR. 2015. Population-based structural variation discovery with Hydra-Multi. *Bioinformatics* **31**: 1286–1289.

Liu SJ, Yao L, Ding DL, Zhu HZ. 2010. CCL3L1 copy number variation and susceptibility to HIV-1 infection: A meta-analysis. *PLoS One* **5**: 1–7.

MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. 2014. The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res* **42**: 986–992.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–753.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**: 1–14.

Maroilley T, Li X, Oldach M, Jean F, Stasiuk SJ, Tarailo-Graovac M. 2021. Deciphering complex genome rearrangements in C. elegans using short-read whole genome sequencing. *Sci Rep* **11**: 1–11. https://doi.org/10.1038/s41598-021-97764-9.

Marques AC, Vinckenbosch N, Brawand D, Kaessmann H. 2008. Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol* **9**: 1–12.

Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, Antaki D, Shetty A, Holmans PA, Pinto D, et al. 2017. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* **49**: 27–35.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. *Trends Ecol Evol* **35**: 561–572. https://doi.org/10.1016/j.tree.2020.03.002.

Nattestad M, Schatz MC. 2016. Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics* **32**: 3021–3023.

Pacific Biosciences. PacBio structural variant (SV) calling and analysis tools. https://github.com/PacificBiosciences/pbsv.

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**: 333–339.

Santos M, Niemi M, Hiratsuka M, Kumondai M, Ingelman-Sundberg M, Lauschke VM, Rodríguez-Antona C. 2018. Novel copy-number variations in pharmacogenes contribute to interindividual differences in drug pharmacokinetics. *Genet Med* **20**: 622–629. http://dx.doi.org/10.1038/gim.2017.156.

Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nat*

Methods **15**: 461–468. http://dx.doi.org/10.1038/s41592-018-0001-7.

Storz JF, Opazo JC, Hoffmann FG. 2013. Gene duplication, genome duplication, and the functional diversification of vertebrate globins. *Mol Phylogenet Evol* **66**: 469–478. http://dx.doi.org/10.1016/j.ympev.2012.07.013.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Vicari S, Napoli E, Cordeddu V, Menghini D, Alesi V, Loddo S, Novelli A, Tartaglia M. 2019. Copy number variants in autism spectrum disorders. *Prog Neuro-Psychopharmacology Biol Psychiatry* **92**: 421–427. https://doi.org/10.1016/j.pnpbp.2019.02.012.

Zarate S, Carroll A, Mahmoud M, Krasheninina O, Jun G, Salerno WJ, Schatz MC, Boerwinkle E, Gibbs RA, Sedlazeck FJ. 2021. Parliament2: Accurate structural variant calling at scale. *Gigascience* **9**: 1–9.