

## TECHNICAL NOTE

# NPSV: A simulation-driven approach to genotyping structural variants in whole-genome sequencing data

Michael D. Linderman <sup>1,\*</sup>, Crystal Paudyal <sup>1</sup>, Musab Shakeel <sup>1</sup>,  
William Kelley<sup>1</sup>, Ali Bashir <sup>2,†</sup> and Bruce D. Gelb <sup>3,†</sup>

<sup>1</sup>Department of Computer Science, Middlebury College, 14 Old Chapel Road, Middlebury, VT 05753, USA;

<sup>2</sup>Google, 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA and <sup>3</sup>Mindich Child Health and Development Institute and the Departments of Pediatrics and Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, One Gustave Levy Place, Box 1040, New York, NY 10029, USA

\*Correspondence address. Michael D. Linderman, Department of Computer Science, Middlebury College, 14 Old Chapel Road, Middlebury, VT 05753, USA. Tel/Fax: +(802) 443-5737. E-mail: [mlinderman@middlebury.edu](mailto:mlinderman@middlebury.edu)  <http://orcid.org/0000-0002-9643-7148>

<sup>†</sup>These authors contributed equally.

## Abstract

**Background:** Structural variants (SVs) play a causal role in numerous diseases but are difficult to detect and accurately genotype (determine zygosity) in whole-genome next-generation sequencing data. SV genotypers that assume that the aligned sequencing data uniformly reflect the underlying SV or use existing SV call sets as training data can only partially account for variant and sample-specific biases. **Results:** We introduce NPSV, a machine learning-based approach for genotyping previously discovered SVs that uses next-generation sequencing simulation to model the combined effects of the genomic region, sequencer, and alignment pipeline on the observed SV evidence. We evaluate NPSV alongside existing SV genotypers on multiple benchmark call sets. We show that NPSV consistently achieves or exceeds state-of-the-art genotyping accuracy across SV call sets, samples, and variant types. NPSV can specifically identify putative *de novo* SVs in a trio context and is robust to offset SV breakpoints. **Conclusions:** Growing SV databases and the increasing availability of SV calls from long-read sequencing make stand-alone genotyping of previously identified SVs an increasingly important component of genome analyses. By treating potential biases as a “black box” that can be simulated, NPSV provides a framework for accurately genotyping a broad range of SVs in both targeted and genome-scale applications.

**Keywords:** structural variants; next-generation sequencing; whole-genome sequencing

## Findings

### Background

Structural variants (SVs) play a causal role in numerous diseases [1]. However, our ability to detect and analyze disease-causing SVs in short-read whole-genome sequencing (WGS) data can be limited by inaccurate genotyping (determining zygosity) [2, 3]. While numerous tools integrate SV discovery and genotyping [4–6], our focus here is “stand-alone” genotyping of putative SVs identified by discovery tools and/or obtained

from the literature or SV catalogs [7]. Stand-alone genotyping is a critical step in ensemble pipelines that integrate multiple SV discovery tools; in clinical workflows, where we seek to accurately genotype known pathogenic SVs (e.g., from dbVar [8]) alongside detecting novel SVs; and in population studies, which generate “squared-off” genotypes for all variants in all samples [7].

SVs, defined here as variants  $\geq 50$ bp [9], are similar in size to or larger than the read length of short-read next-generation sequencers (NGS) and, thus, typically cannot be detected directly.

Received: 21 December 2020; Revised: 4 May 2021; Accepted: 7 June 2021

© The Author(s) 2021. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Instead SVs must be inferred from secondary features in the sequencing data such as split reads, discordant read-pairs, and read depth [9]. As a result, the precision and recall for detecting and genotyping SVs in NGS data can be much lower than for single-nucleotide variants and short indels [4, 7, 10–13]. Long-read sequencing (read lengths of  $\geq 10$  kb) improves the recall and precision of SV detection (the long reads span more events and can be more reliably mapped) [14–16]. However, long-read sequencing is more expensive than NGS [17], so many more samples have been and will continue to be sequenced with NGS technologies. Thus, despite the growth in long-read sequencing, there is a need to develop improved NGS SV genotyping tools, including to genotype those SVs first (and exclusively) detected with long-read sequencing.

Existing SV genotyping tools [18–26] (see Chander et al. [7] for a recent comparison) exclusively target specific variant types/sizes, use parametric (i.e., fixed-size [27]) models of SV evidence, and/or are trained on existing genome-wide call sets. These approaches assume that different SV call sets are similar and/or that the aligned sequencing data consistently and uniformly reflect the underlying variant (e.g., the read depth is proportional to copy number, alternate alleles are identified at a consistent rate, and/or consistent breakpoint features will be observed across all variants). However, these assumptions do not hold for all variants. The different types of SVs, range of SV sizes, different genomic contexts, and different sequencers/pipelines, all of which influence the available evidence for predicting the SV genotype, motivate an ensemble of approaches, each optimized for a specific subset of SVs [10, 28]. NGS simulation is a possible strategy to enable automatic ensemble creation. For example, for select SVs from the 1000 Genomes Project, Chu et al. showed that training the GINDEL SV genotyper on simulated data could achieve genotyping accuracy within a few percentage points of models trained on held-out data [29].

Here, we propose the Non-Parametric Structural Variant (NPSV) genotyper. NPSV extends current ensemble methods by automatically creating classifiers for predicting SV genotypes optimized for the specific SVs and sample under analysis and even a single specific SV. In this non-parametric approach, the number of models can grow to capture genomic region, sequencer, and pipeline-specific SV evidence. NPSV performs detailed simulation of the putative SVs to be genotyped. The simulated data, which are representative of the actual observed sequencing data, are used to train sample- and variant-specific classifiers for predicting SV genotypes. In contrast to training data sourced from existing SV call sets, by using simulation we can generate representative training data for any putative SV, not just those previously observed, with accurate sequence-resolved breakpoints and “ground truth” genotype labels.

We present a rigorous evaluation of NPSV genotyping accuracy across multiple truth sets in the HG002 and NA12878 reference samples. We compare NPSV to similar stand-alone SV genotyping tools (that accept a VCF of putative SVs and aligned reads as input and predict the SV genotype), chosen to be representative of different alignment, graph, and machine learning-based SV genotyping methods: Delly2 [18], SV-Typer [19], svviz2 [20], Paragraph [25], GraphTyper2 [21], SV2 [22], and GenomeSTRiP [26]. We show that NPSV consistently achieves similar or better genotyping accuracy across the different datasets, samples, and variant types, can sensitively and specifically identify putative *de novo* SVs in a trio context, and is robust to offsets in SV breakpoints.

## Simulation-driven SV genotyping

The NPSV dataflow is shown in Fig. 1A. The inputs are the aligned reads (BAM/CRAM file), termed the “actual” data, and a VCF file of putative sequence-resolved deletion and insertion SVs. For each putative SV and possible genotype, NPSV generates synthetic short-read datasets using an NGS simulator configured to match the actual data (bottom path in Fig. 1A). We process the simulated datasets with the same alignment pipeline as the actual data and then extract realignment, read pair, and coverage SV features from each simulated replicate. The features extracted from the simulated data are used to train sample- and variant-specific classifier(s) to predict the genotype from the SV evidence similarly extracted from the actual sequencing reads. The simulation, feature extraction, and classification approaches are described in more detail in the Methods (NPSV genotyping algorithm) with the features specifically described in Supplementary Table S1.

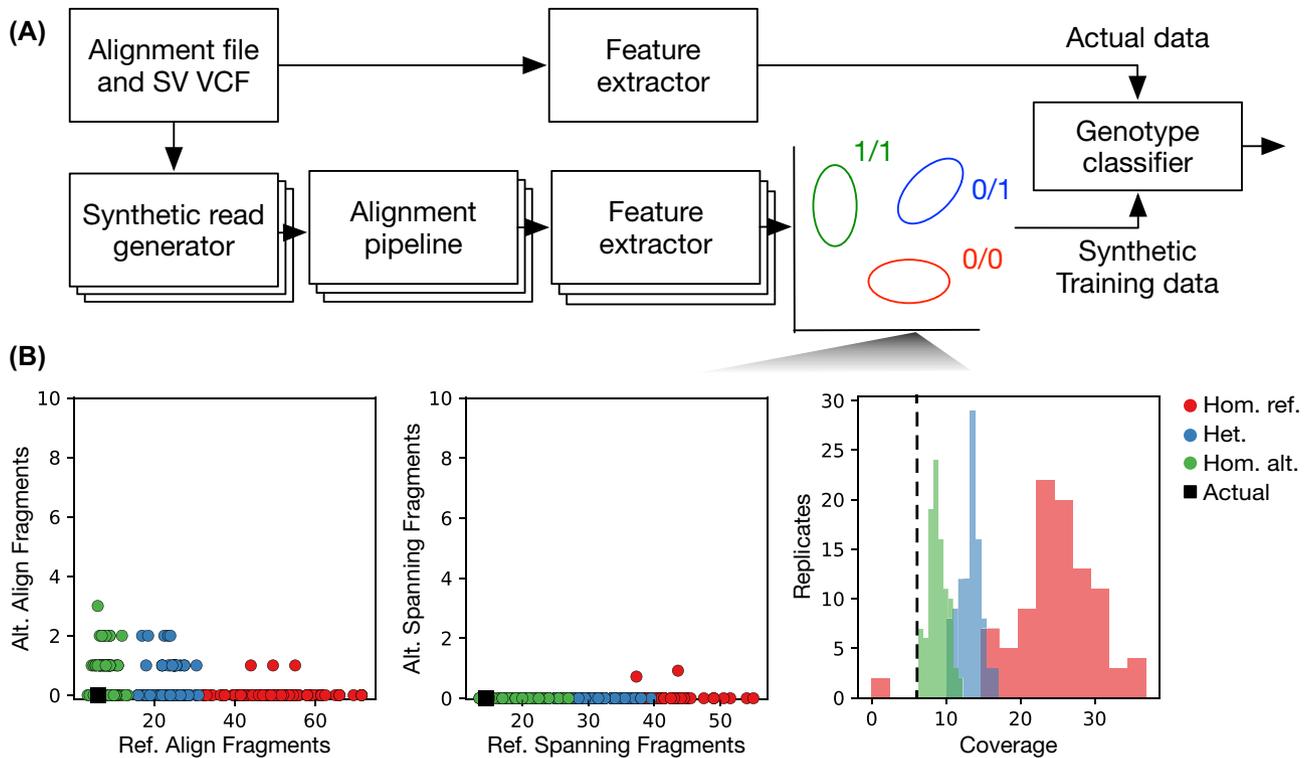
Figure 1B shows the simulated and actual SV evidence for an example homozygous alternate 822-bp deletion in the Genome in a Bottle (GIAB) HG002 call set [27], as would be generated to train a variant-specific classifier. The actual data are most consistent with the simulated homozygous alternate genotype. This SV is the deletion of 1 repeat of a tandem repeat. Owing to the underlying repetitive sequence, no reads were uniquely realigned to the SV’s alternate allele and no alternate spanning fragments were identified. The simulated data show that the absence of both of those features is consistent with the alternate allele for this SV (and pipeline) and is not an indication of a homozygous reference genotype as might otherwise be expected (indicated by the actual and simulated features in the first 2 panels “massing” on the y-axis). The NPSV variant-specific classifier correctly genotyped this variant as homozygous alternate, while genotypers that exclusively use realignment, split-read, and/or spanning read evidence did not.

NPSV implements 2 genotyping approaches: (i) a “variant” model, like that described in this section, which creates variant-specific classifiers trained on 100 replicates per variant per zygosity (i.e.,  $300n$  synthetic samples for  $n$  variants), and (ii) a “single” model that creates a single sample-specific genome-wide classifier for each variant type (e.g., deletions, insertions) trained on 1 replicate per variant per zygosity (i.e.,  $3n$  synthetic samples for  $n$  variants). The former approach is more computationally demanding but can be applied at any scale, including for just a single SV in a single sample. To reduce the computational burden, a “hybrid” model only builds variant-specific classifiers for smaller SVs (<1 kb by default) and uses the single model for larger SVs. We generally observed the hybrid model to be most accurate for deletions and the single model to be most accurate for insertions and so set that as the default configuration.

## Results

### Genotyping accuracy

We evaluated NPSV and the comparison SV genotypers with multiple SV call sets across 2 samples: the GIAB version 0.6 call set for HG002 [30] and the Polaris 2.0, Polaris 2.1, and SV-plaudit call sets for NA12878 [31, 32]. Genotype counts for each call set are reported in Supplementary Table S2. Using the call set SVs as the input, we report the genotype concordance, i.e., the fraction of predicted genotypes that exactly match the call set genotypes, and the non-reference concordance, which treats heterozygous and homozygous alternate genotypes as equivalent. The call sets and evaluation are described in more detail in the Methods.



**Figure 1:** NPSV dataflow and example SV evidence. (A) NPSV dataflow showing the matched training and prediction pipelines. For each putative SV and genotype, NPSV generates 1 or more simulated replicates. These simulated data, shown in the schematic as red, blue, and green clusters for homozygous reference (hom. ref.), heterozygous (het.), and homozygous alternate (hom. alt.) genotypes, respectively, are used to train sample- and variant-specific classifiers for predicting the SV genotype. (B) Synthetic training data (colored circles/bars) and actual data (black square/line) for a homozygous alternate 822-bp deletion in HG002. This SV is the deletion of 1 copy of a repeat and as a result of the repetitive genomic context, no fragments were uniquely realigned to the SV's alternate allele and no alternate spanning fragments were identified. The actual data are consistent with the simulated homozygous alternate data and not a homozygous reference genotype as might be expected from the absence of alternate allele realignments. This SV is successfully genotyped as homozygous alternate by NPSV when building a variant-specific classifier.

Figure 2 shows the genotyping accuracy for NPSV and comparison tools for all truth sets. As a result of randomization in the simulations, SV sampling, and classifier training, NPSV genotyping is not deterministic. In Fig. 2 we show the mean accuracy for 10 complete NPSV genotyping runs and report the mean and standard deviation in Supplementary Tables S3–S6. Figure 2 (top) (Supplementary Table S3) shows the genotyping accuracy for GIAB SVs in the high-confidence Tier 1 regions (6,449 DEL and 6,462 INS SVs) and in the Tier 1 regions combined with lower-confidence Tier 2 SVs (8,370 DEL and 8,413 INS SVs). NPSV achieves similar or better exact genotype concordance and non-reference concordance than the comparison tools for both deletions and insertions. For SVs in Tier 1 regions, NPSV improves genotype and non-reference concordance for deletions and insertions by 0.8–2.1 percentage points. Fig. 2 (bottom) (Supplementary Table S5) shows the genotyping accuracy for the NA12878 truth sets (1,143 DEL SVs in SV-plaudit, 8,073 DEL and 6,246 INS SVs in Polaris 2.0, and 20,610 DEL and 12,028 INS SVs in Polaris 2.1). NPSV generalizes across these datasets, achieving similar or better accuracy than the best comparison SV genotypers across all 3 datasets and both insertions and deletions.

Precision, recall, and F1 scores for genotyping homozygous reference vs non-reference SVs are shown in Supplementary Tables S4 and S6. We observed NPSV to achieve similar or better F1 scores than the comparison genotypers, albeit often with increased recall and reduced precision. For the Polaris 2.1 call set, which is enriched for homozygous reference calls and thus

reflective of the common population genotyping use case, the mean DEL and INS recall are 0.939 and 0.985, respectively: the DEL and INS precision (false discovery rate) are 0.820 (0.180) and 0.945 (0.055). Given the small variance across all metrics for NPSV, for concision, the remaining analyses use a single genotyping run.

Supplementary Table S7 shows the genotype concordance for NPSV single and variant models for GIAB SVs in Tier 1 regions grouped by SV length (SVLEN), the difference in length of reference and alternate alleles. Concordance generally increases with increasing SV length as read-pair and other features become more informative and a smaller fraction of variants overlap repetitive regions (see below). For deletions  $\geq 1$ kb, the single model showed increased accuracy. Those results motivated the default 1-kb threshold for the hybrid approach, which uses the single model for larger SVs where that approach is more accurate and for which simulating an SV is more computationally demanding, and reserves the more computationally expensive but also potentially more accurate variant model for smaller variants. For GIAB insertions, the single model is more accurate than the variant model for all variant sizes. As noted above, based on these results, we set the NPSV default configuration to use the hybrid approach for deletions and single model for insertions. Supplementary Figure S3 shows the genotype concordance for all call sets and tools grouped by SV length along with the underlying SV length distributions.

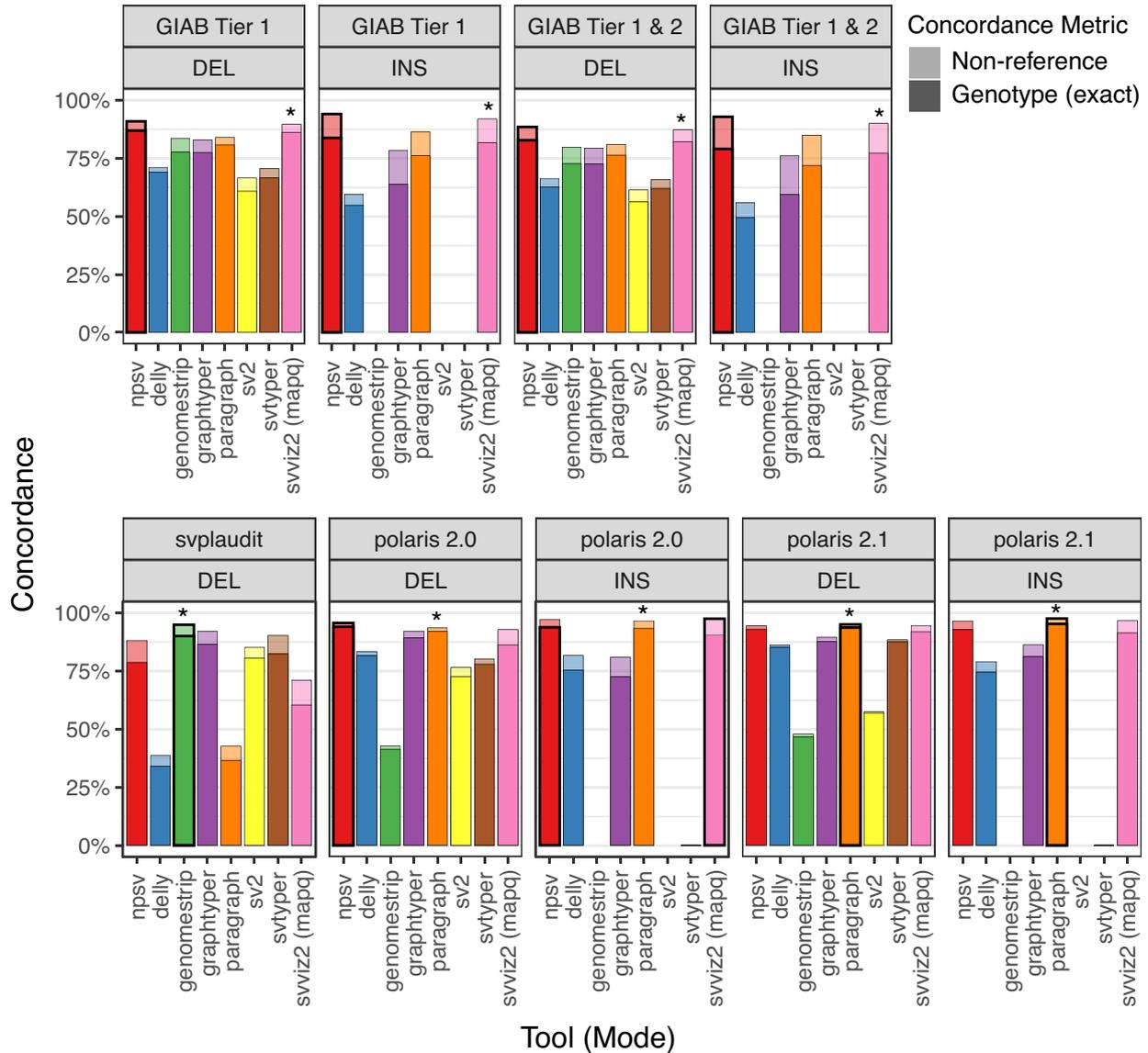


Figure 2: Genotyping accuracy for HG002 and NA12878 SVs. Top: Genotype concordance and non-reference concordance (presence or absence) for GIAB SVs (including “LongReadHomRef” SVs where “long reads supported homozygous reference for all individuals”) in high-confidence Tier 1 regions and the Tier 1 regions and lower-confidence Tier 2 SVs combined. Bottom: Concordance for NA12878 call sets. The NPSV accuracy is the mean of 10 runs. The best concordance is indicated with a black outline. The asterisk shows tools used in the construction of that call set.

Owing to the repetitive sequence, SVs in tandem repeats (TRs) are more difficult to accurately genotype. NPSV genotype concordance for GIAB DEL and INS SVs in Tier 1 regions overlapping a TR > 100 bp (as annotated by GIAB) was 77.5% and 69.0%, respectively, compared with 96.5% and 92.1% for DEL and INS SVs not overlapping a TR > 100 bp. Supplementary Figure S4 shows the genotype concordance for all NPSV modes for GIAB SVs in Tier 1 regions grouped by SV length and whether the SV overlaps a TR > 100 bp. The [50, 100) and [100, 300) size bins are enriched for SVs overlapping a TR > 100 bp, contributing to the reduced genotyping accuracy for these smaller SVs reported above.

To evaluate the use of the NPSV stand-alone genotyper with SVs identified with SV discovery tools (as opposed to benchmark call sets), we re-genotyped SVs called with Lumpy [33]/SVTyper [19] and Manta [34] in HG002. Using the discovery SVs as the input, Table 1 reports the genotyping accuracy and Supple-

mentary Table S8 shows precision, recall, and F1 scores compared to the GIAB SVs in Tier 1 regions. To focus on genotyping accuracy, SVs that were not detected (“no-calls”) were excluded from the concordance calculation. For the Lumpy call set the difference in genotyping accuracy is primarily driven by the count of putative false-positive calls, which in turn is sensitive to the criteria for matching the call set to truth set SVs. Some of the putative false-positive SVs may be true-positive SVs (i.e., non-reference) that have sufficiently different breakpoints compared to the truth set SV so as not to match during concordance analysis. When the required size similarity is relaxed from 70% to 30%, 150 more Lumpy SVs matched GIAB truth set SVs and the exact genotype concordance (non-reference concordance) for NPSV, 84.9% (91.3%), and Lumpy/SVTyper, 82.8% (91.9%), became more similar. The impact of offset or imprecise SV descriptions is described further below.

**Table 1:** Genotyping accuracy with discovery SVs as the input to SV genotyping and GIAB SVs in Tier 1 regions as the truth set; concordance is calculated for the subset of SVs successfully identified by the discovery tool

Caller	Type	Discovery recall, %	Caller genotyping, %		NPSV genotyper, %	
			Concordance	Non-reference concordance	Concordance	Non-reference concordance
lumpy	DEL	30.5	82.1	87.2	88.5	92.7
manta	DEL	67.9	90.1	91.8	92.2	93.6
manta	INS	25.2	87.3	93.5	89.1	93.7

### Trio analysis

We evaluated SV genotyping in a trio context using the HG002 trio. Table 2 shows the Mendelian error rate (MER) and counts of different types of Mendelian errors (MEs) for GIAB SVs in Tier 1 regions. The NPSV MERs for both deletions and insertions are greater than the MERs for some of the existing genotypers, e.g., svviz2. However, most of the NPSV MEs are variants with low-confidence genotypes and thus can be specifically filtered out on the basis of the NPSV-reported genotype quality (GQ); e.g., 94% of ME deletions in Tier 1 regions have a minimum GQ < 10 (among all trio members). The highest-quality NPSV deletion ME in the Tier 1 regions was explicitly reported by Zook et al. [30] as a “likely *de novo* deletion.” Supplementary Table S9 lists the trio genotypes, minimum GQ, and GQ ranking for that deletion and a second deletion ME reported by Zook et al. in a locus known to undergo somatic rearrangement. Consistent with Zook et al. NPSV genotyped the 2 deletions as *de novo*. NPSV in variant mode reported the 2 deletions as the most confident ME deletions in the Tier 1 regions, and the hybrid mode reported the variants among the top 4 most confident.

### Offset SV representations

As shown in Supplementary Fig. S4, the set of GIAB SVs with discordant NPSV genotypes (i.e., the NPSV genotype does not match the GIAB genotype) is enriched for variants that overlap TRs. Across all NPSV modes,  $\geq 86\%$  of GIAB discordant deletions and  $\geq 68\%$  of discordant insertions in Tier 1 regions are annotated in the GIAB call set as overlapping a TR > 100 bp, while <44% of concordant deletions and <30% of concordant insertions SVs are similarly annotated. Differences between the description of the putative SV (breakpoints and sequence change) and the true SV is one of the factors that contribute to genotyping errors for SVs in these repetitive regions (and more generally) [25]. We manually reviewed the pileup for 10 randomly selected deletions discordantly genotyped by NPSV in variant mode; 8 of 10 SVs were offset from the location indicated by long-read Pacific Biosciences (PacBio) sequencing data.

To evaluate the impact of offset breakpoints more generally, we matched the GIAB SVs (PASS variants only) in Tier 1 regions to corresponding SVs called by PBSV [35] in PacBio long-read sequencing data (4,114 of 4,203 deletions and 5,157 of 5,443 insertions successfully matched). Making the assumption that the PacBio SV calls have correct breakpoints, we infer the offset from the distance between the GIAB breakpoints and the breakpoints identified in the long-read data (modeled on the approach in Chen et al. [25]). Figure 3 shows genotype concordance for SVs grouped by the breakpoint offset (the same analysis for select comparison tools is included in Supplementary Fig. S5). For deletions, we observe an expected negative association between breakpoint offsets and genotyping accuracy; genotype concordance is  $\geq 85\%$  for offsets <10bp (and  $\geq 95\%$  for no or single base offsets), decreasing to 49.6% for SVs with breakpoint off-

sets  $\geq 50$ bp. At larger offsets, the variant model increasingly outperforms the single model, suggesting that the variant-specific classifiers may be better able to model the specific genomic context around offset deletions. For insertions we observe a similar negative association between breakpoint offsets and genotyping accuracy, although with a plateau for offsets of 1–20 bp. Much of the genotype concordance is recovered when using the long-read-derived SV calls as the input call set instead of the GIAB SVs (solid line in Fig. 3).

The challenge of offset or imprecise breakpoints is further observed in the SV-plaudit call set. The majority of SV-plaudit variants  $\geq 1$ kb (410 of 551) have imprecise breakpoints (including SVs with breakpoint confidence intervals of hundreds or thousands of bases). NPSV genotype concordance was 93.6% (132 of 141) for SVs with exact breakpoints, decreasing to 63.7% (261 of 410) otherwise.

To investigate the potential for correcting SV descriptions using only the NGS data, we experimentally extended NPSV to propose and select among possible alternate alignments for an SV. We hypothesize that the actual data are most similar to the simulated data for the correct SV description and genotype, and thus we could identify a better SV representation based on the distance between the actual and simulated SV evidence. For deletions of 1 or more copies of a TR, we proposed up to 10 different alignments of the deletion within the repetitive region, choosing the SV description where the real data are closest to the non-reference synthetic data. The SV proposal algorithm is described in more detail in the Supplementary Methods. When applied to the GIAB call set SV proposal increases the sensitivity for calling heterozygous and homozygous alternate genotypes at the cost of smaller decreases in precision; the net effect is an increase in genotyping accuracy (genotype concordance of 87.8% vs 87.2% for SVs in Tier 1 regions, 83.7% vs 82.9% for SVs in Tier 1 and 2 regions) and F1 scores (0.940 vs 0.933 for SVs in Tier 1 regions, 0.926 vs 0.914 for SVs in Tier 1 and 2 regions) (Supplementary Table S10).

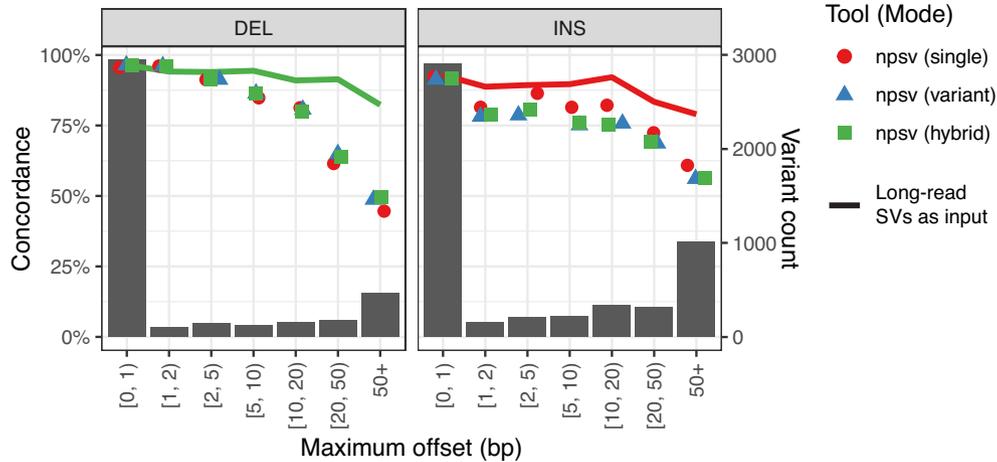
### Computational requirements

The simulation process is computationally intensive but also readily parallelizable. NPSV simulation and feature extraction are multi-threaded across variants. On a 36-core compute node (dual 18-core Intel Xeon 6140 2.3 GHz CPUs) genotyping 16,871 GIAB SVs required 21.1 hours (wall clock time as determined by the time utility); using the single, variant, and hybrid approaches exclusively required 53.3 minutes, 39.6 hours, and 31.2 hours, respectively. Maximum memory resident set size was 22.1 GB as determined by the SLURM cluster manager (17.7, 39.0, and 24.3 GB for the single, variant, and hybrid modes). The NPSV preprocessing step, which generates sequencing statistics, e.g., insert size distribution, used in simulation and feature extraction, is designed to use a combination of goleft [36], a fast alignment analysis tool, and metrics already calculated as part of a typi-

**Table 2:** Mendelian error rate and Mendelian error breakdown for GIAB autosomal SVs in Tier 1 regions

Tool	DEL				INS			
	MER, % (proportion)	De novo			MER, % (proportion)	De novo		
		Heterozygous	Homozygous	Other		Heterozygous	Homozygous	Other
npsv (single)	3.60 (231/6,416)	99	7	125	4.64 (291/6,269)	58	6	227
npsv (variant)	3.09 (198/6,416)	111	1	86	5.14 (322/6,269)	74	4	244
npsv (hybrid)	3.21 (206/6,416)	106	1	99	5.10 (320/6,269)	65	6	249
delly	1.66 (92/5,535)	50	2	40	1.92 (78/4,059)	26	0	52
genomestrip	2.02 (128/6,337)	81	3	44				
graphtyper	5.53 (353/6,386)	109	16	228	10.13 (608/6,004)	86	27	495
paragraph	2.76 (175/6,351)	85	2	88	5.42 (329/6,067)	91	4	234
sv2	8.80 (536/6,089)	129	47	360				
svtyper	2.28 (145/6,349)	79	6	60				
svviz2 (mapq)	2.46 (158/6,416)	94	3	61	3.32 (208/6,269)	75	3	130

NPSV default configuration uses hybrid mode for deletions and single mode for insertions. ME: Mendelian error; MER: Mendelian error rate.



**Figure 3:** Genotype concordance for GIAB SVs with offset breakpoints. Genotype concordance for GIAB variant-only SVs in Tier 1 regions grouped by the maximum offset between the GIAB breakpoints and the breakpoints for the corresponding SV called with PBSV in PacBio long-read sequencing data. The line shows the concordance when using the PBSV SVs as the input to NPSV running the default genotyping mode (“hybrid” for deletions, “single” for insertions). The background bar chart shows the underlying distribution of offsets. The same analysis for select comparison tools is included in Supplementary Fig. S5.

cal genome analysis pipeline, e.g., with the Picard tools [37]. In that approach, <1 minute is required for preprocessing. The currently un-optimized fallback preprocessor required 2.4 hours for the 25.5× HG002 BAM file. Supplementary Table S11 lists execution time and resident set size for all of the comparison tools running on the same system.

## Discussion

NPSV is a novel stand-alone SV genotyper that simulates putative SVs to train sample and variant-specific machine learning classifiers. NPSV consistently achieved genotyping accuracy similar to or better than the comparison SV genotypers across both variant types and all truth sets (Fig. 2), including compared to the tools used to construct those truth sets (e.g., svviz2 for GIAB and Paragraph for Polaris). NPSV successfully and specifically identified the putative *de novo* SV deletions reported by GIAB. Improvements of 1–3 percentage points in genotyping accuracy translate to tens to hundreds of fewer incorrect genotypes per genome. Those incorrect genotypes leave cases unresolved, consume limited budgets for manual review and validation testing, and dilute downstream analyses.

SV call sets and reference databases can contain many SVs with incorrect or imprecise descriptions. For example, the clus-

tering of SVs with similar but unique sequence changes during the construction of the GIAB call set reduced the number of SVs 2.3-fold [30], indicating that many of the putative SVs did not have a single consensus description. Incorrect or imprecise SV descriptions can negatively affect genotyping accuracy [25]. NPSV maintains genotype concordances of  $\geq 85\%$  (DEL) and  $\geq 81\%$  (INS) for offsets <10bp, similar to or more robust than comparison tools (Fig. 3, Supplementary Fig. S5).

Making the SV features even more robust to incorrect/imprecise SV descriptions could improve genotyping accuracy. However, in a strict interpretation of the precise sequence-resolved SVs in the GIAB call set, genotyping a putative SV with an incorrect description as non-reference would be inaccurate because that specific alternate allele is absent. Ideally, we would want to identify the correct SV descriptions as part of the genotyping process. We observed substantial increases in genotyping accuracy when using SVs called in long-read sequencing data as the input call set, suggesting that there is an opportunity to further improve genotyping accuracy by refining the SV descriptions. We extended NPSV to select among alternative alignments for deletion SVs based on the similarity between the actual and simulated NGS data. The alternate representations increased the sensitivity for detecting

non-reference genotypes, with a net increase in genotyping accuracy and F1 score for detecting non-reference genotypes compared to the original SV descriptions (Supplementary Table S10). A substantial gap remains, though, between NPSV's NGS-only approach for refining SV descriptions and the accuracies observed when genotyping the long read-derived calls (Fig. 3). We are actively working on improving all aspects (SV proposal, NGS simulation fidelity, features, and the similarity metric) of the SV refinement algorithm.

Because NPSV simulates the expected alleles, it is limited to sequence-resolved SVs with discrete genotypes and does not genotype “position independent” SVs, e.g., high copy number duplications. At present, NPSV only supports biallelic deletions and insertions and treats each SV independently. It does not currently genotype inversions or other SV types. However, the underlying method can be extended to support other SV types, e.g., inversions, and more complex variants/genotypes, e.g., compound heterozygous genotypes or multiple SVs on the same haplotypes. We hypothesize that the simulation-based approach, which is not dependent on the previous generation of representative training data, may be particularly useful for complex SVs. Minimal high-quality “ground truth” data are available for these sites; GIAB, for example, largely excluded complex SVs from the benchmark call set and most of the genotype errors identified in manual review were identified as complex [30].

The simulation process can be computationally intensive, particularly when simulating many replicates to build per-variant classifiers. However, because the variant-specific classifiers can be built at the granularity of a single variant, they can be used in a targeted fashion, e.g., on SVs with low-confidence genotypes or in repetitive regions, to model the biases introduced by the genomic region, sequencer, and/or the analysis pipeline. Alternately when the same call set is being genotyped across multiple samples, the simulated training data could be reused. Preliminary experiments using the simulated training data generated for the parental samples to genotype the GIAB SVs in HG002 showed similar genotyping accuracy. When using shared training data, the simulation costs scale with the call set size, not the product of the call set and cohort sizes. Optimizing NPSV for large cohorts is an area of ongoing work. Large, highly consistent cohorts, such as gnomAD, can use other samples as the reference panel [38] but may have few and/or potentially ambiguous examples of extremely rare variants/genotypes. NPSV can effectively create a synthetic “reference panel” for all zygosity, for any variant, in any number of samples (including a single genome).

## Conclusions

Here we present NPSV, a stand-alone SV genotyper for WGS data. Instead of attempting to develop a model for the complex and interconnected effects of the genomic region, sequencer, and alignment pipeline on the observed SV evidence, NPSV uses detailed simulation of the sequencing process to train sample- and variant-specific classifiers for predicting SV genotypes. Because NPSV can generate relevant training data for any variant(s), at any granularity, it supports a range of targeted (a single variant) and large-scale (whole genome) SV genotyping applications. We showed that NPSV consistently achieves similar or improved genotyping accuracy for benchmark call sets. Looking forward, NPSV's simulation-based approach provides a framework for genotyping the important “long tail” of SVs that are rare, complex, and/or exclusively discovered with long-read

technologies, and thus lack high-quality representative training examples.

## Methods

### NPSV genotyping algorithm

NPSV is a Python-based tool for stand-alone genotyping of sequence-resolved SV insertions and deletions. The inputs are the aligned reads (BAM/CRAM file), termed the “actual” data, and a VCF file of putative SVs. NPSV produces a copy of the input VCF with predicted SV genotypes.

Prior to genotyping, a preprocessing step estimates the mean, per-chromosome, and per-GC fraction coverage and the insert size distribution. Those statistics inform the simulation and feature extraction. Many of those metrics are often already generated as part of the genome analysis pipeline and so do not need to be recomputed. In this evaluation, NPSV internally runs `indexcov` [36] and uses metrics previously computed with the Picard tools [37]. NPSV can also generate those statistics directly if needed using `bedtools` [39], `SAMtools` [40], and `indexcov` [36]. For each putative SV and possible genotype, NPSV generates 1 or more synthetic short-read datasets (termed replicates) using the ART NGS simulator [41] configured to model the actual sequencing data, i.e., sequencer error model, read length, insert size distribution, and coverage. In this evaluation we align the simulated WGS data with `BWA-MEM` [42] and mark duplicates with `sambaster` [43] to mimic the BCBio pipeline [44] used to align the actual data (along with `SAMtools` [40] and `sambamba` [45] for format conversion and sorting). The SV features extracted from the simulated replicates (and randomly simulated regions in the genome, see below) are used to train sample- and variant-specific classifier(s). The SV features extracted from the actual data for putative SVs are only used to predict the genotypes (and not for training).

Features extracted from the simulation of the homozygous reference genotype, i.e., the absence of the putative SV, can exhibit low variance, negatively affecting genotyping accuracy. To generate a more realistic “null” model, by default, NPSV generates the training data for homozygous reference genotypes by extracting features from the actual alignments for size-matched variants randomly sampled from the genome [46]. For haploid sex chromosomes, size-matched variants are sampled from the sex chromosomes; otherwise variants are sampled from the autosome (and the X chromosome for SVs called on a diploid X chromosome).

NPSV extracts or derives the allele, spanning read, and coverage SV features listed in Supplementary Table S1 and described in more detail in the Supplementary Methods. NPSV counts the reference and alternate reads by locally realigning read pairs to the reference and alternate sequences (derived from the putative SV description) using `BWA` [42] (via `SeqLib` [47]) and a read pair-aware alignment-scoring metric adapted from `svviz2` [20]. Only reads originally aligned within some flanking distance (default of 99th percentile of the insert size) of the putative SV breakpoints are realigned. We extract insert size probability-weighted counts of spanning reads (adapted from `SVTyper` [19]), counts of clipped reads (adapted from `SMRT-SV2` [23]), and the mean event depth relative to flank regions, chromosome, and regions of similar GC coverage features (`DHFFC`, `DHFC`, `DHBFC` adapted from `duphold` [48]) from the actual (original) alignments.

NPSV currently implements a Support Vector Machine (SVM) classifier for the single model and a random forest (RF) classi-

fier for the variant model using the scikit-learn framework [49]. The specific features used with each classifier are listed in Supplementary Table S1. We observed this combination to achieve consistently high accuracy across variant types and call sets, although the differences in accuracy between classifier algorithms were typically small (1–1.5 percentage points). Data are centered and normalized to unit variance (using `StandardScaler`) prior to training the SVM (using a radial basis function kernel) with the same scaling used for genotyping. When training the single SVM model, NPSV can perform a grid search of the  $C$  (1, 10, 100, 500, 1,000, 5,000, 10,000) and  $\gamma$  (“scale,” 0.001, 0.0055, 0.01, 0.055, 0.1, 0.55) hyperparameters with 5-fold cross-validation [23]; however, we did not observe consistently improved accuracy over default parameters for the GIAB call set and so disable the parameter sweep by default to reduce the training time. When training the single-model classifier, the training data are optionally filtered by genomic region. For the GIAB call set we excluded data outside the GIAB Tier 1 regions. When training the per-variant classifiers, observations with features  $>5$  standard deviations from the mean are excluded. To reduce execution time for the per-variant model, by default, we do not implement parameter sweeps during training. Evaluation of the RF-based variant model on the GIAB call set with different numbers of trees (10, 50, 100, 200) and maximum tree depth (variable, 3) indicated that use of the default parameters (100 trees with variable depth) achieves high accuracy with reasonable execution time. The final genotypes and GQ are determined from the label and class probabilities predicted by scikit-learn.

### Truth sets

We evaluated NPSV and the comparison SV genotypers with deletion and insertion SVs in the GIAB version 0.6 call set (GRCh37) for HG002, and the Polaris 2.0 (GRCh37), Polaris 2.1 (GRCh38), and SV-plaudit (GRCh37) call sets for NA12878. The truth sets were obtained from the GIAB FTP site [50], Polaris repository [31], and SV-Plaudit Supplementary materials [51]. GIAB SVs smaller than 50 bp or larger than 15 Mb, SVs outside the GIAB Tier 1 and 2 regions, SVs without genotypes, and filtered (i.e., not PASS) SVs, except for those variants filtered as “LongReadHomRef” (i.e., “long reads supported homozygous reference for all individuals”), were excluded. SV-plaudit and Polaris SVs smaller than 50 bp or larger than 15 Mb, SVs without genotypes, and filtered SVs were similarly excluded. In the SV-plaudit report [32] 9 researchers manually inspected SVs called in NA12878 by the 1000 Genomes Project [10]. The researchers were shown visualizations of data for the NA12878 trio and asked, “Does the sample support the variant type shown? [...],” with the possible answers “True,” “False,” or “denovo.” Only SVs for which  $>50\%$  of the curators reported that the sample supports the variant were retained. Almost all the curated SVs were deletions, so we limited the SV-plaudit analysis to deletions. Supplementary Table S2 lists the counts of each genotype in the different truth sets.

### Short-read sequencing data and SV discovery

We genotyped the GIAB SVs in a subset of the NIST Illumina HiSeq 2500  $2\times 148$  PCR-free WGS data [52–55] with coverage representative of typical WGS (mean coverage of  $25.5\times$ ,  $20.4\times$ , and  $24.7\times$  for HG002, HG003, and HG004, respectively). We aligned the WGS reads to GRCh37 and performed point variant calling and SV discovery (using Lumpy [33]/SVTyper [19] via smooove [56] and Manta [34]) with version 1.0.9 of the BCBio pipeline using the

default BWA and GATK-based configuration [44]. We genotyped the NA12878 SVs in the Illumina Platinum Genomes  $2\times 100$  WGS data [57, 58] (mean coverage of  $50.5\times$ ). We aligned the NA12878 WGS reads to GRCh37 and GRCh38 with version 1.2.3 of the BCBio pipeline.

### Comparison tools

We compared NPSV to a representative set of stand-alone SV genotyping tools. The Delly2 (v0.8.3) genotyping module [18], SVTyper (v0.7.1) [19], and GenomeSTRiP (v2.00.1958) [26] predict the genotype using a parameterized model incorporating multiple forms of evidence, e.g., depth, split-reads, and read pairs, extracted from original alignments. The svviz2 (commit b2c5126) [20] reporting module predicts the genotype assuming a binomial model for counts of reads realigned to the SV alleles with BWA. Paragraph (v2.4a) [25] and GraphTyper2 (v2.5.1) [21] employ a parametric model of reads realigned to a graph representation of the SV. SV2 (v1.5) [22] uses an SVM classifier trained on features extracted from 1000 Genomes Project data.

Unless otherwise noted, all tools were run with the truth set VCFs and BAMs produced by the BCBio pipeline as the inputs. SV2, SVTyper, and GenomeSTRiP do not support the insertion SVs in the GIAB and Polaris call sets and so were evaluated on the deletion SVs only. Prior to genotyping with Paragraph, we normalized the VCF to add a padding base for complex variants. The svviz2 genotypes were extracted from the “GT\_mapq” field in the report to generate a genotyped VCF (we observed the “mapq” genotypes to generally be the most accurate for the GIAB call set). For SV2, variants called by GATK haplotype caller (as implemented in the BCBio pipeline) were used as the “SNV” input. For GraphTyper2, the GIAB Tier 1 and 2 BED file was used to generate the regions for genotyping the GIAB HG002 call set, while the entire chromosomes were used as the regions for the NA12878 call sets; the “AGGREGATE” model was used as the output genotypes. GraphTyper2 converts insertions to duplications; those SVs were converted back to the call set representation to facilitate concordance analysis. Delly modifies the representation of some indel SVs such that the modified SV is no longer matched to the corresponding SV in the truth set during evaluation, reducing the reported concordance by up to 0.3 percentage points. Each tool was run with its default parameters, and thus the results presented here may not represent the best possible performance that could be achieved with expert tuning of the available configuration parameters. For example, GenomeSTRiP’s high rate of “no-calls” (./.) for some smaller Polaris SVs can be affected by the “minimum length to include depth-based genotype likelihoods” `depth.effectiveLengthThreshold` parameter (default of 200) [29]. The VCF FILTER annotations introduced by Delly, GraphTyper, Paragraph, and SV2 reduced genotyping accuracy (filtered genotypes are treated as “no calls” during concordance analysis) and so were ignored in all evaluations.

This evaluation does not exercise all of the capabilities of the different comparison tools, which may support other variant types, e.g., inversions, not yet implemented in NPSV; provide other features, such as visualization; or are explicitly designed for efficient population-scale genotyping as opposed to the single sample and trio analyses performed here.

### Evaluation

We measured genotyping accuracy using Truvari [59], modified to report the genotype confusion matrix [60]. Supplementary Fig. S1A and B shows the definitions of concordance metrics calcu-

lated from the confusion matrix when using the “truth” SVs as the input to SV genotyping. Supplementary Fig. S1C and D shows the definition of the concordance metrics when using the output of an SV discovery tool as the input to the SV genotyper.

MEs were identified in autosomal regions using BCFTools [61]. We categorized MEs as a heterozygous or homozygous *de novo*, or other (e.g., homozygous alternate proband with a homozygous reference parent).

To evaluate the impact of imprecise breakpoints, we computed the genotype concordance for GIAB deletion SVs in Tier 1 regions grouped by the maximum offset between the GIAB SV breakpoints and the corresponding SV breakpoints called in long-read sequencing data [25]. We used SV calls generated by PBSV 2.2.1 in PacBio CCS reads (obtained from the GIAB FTP repository). We matched the GIAB and PBSV calls with Truvari configured to match SVs within a 2,000-bp window, with 70% size and sequence similarity [30], and extracted the offsets from the Truvari annotations.

## Availability of Source Code and Requirements

Project name: npsv

Project home page: <https://github.com/mlinderm/npsv>

Operating system(s): Linux

Programming language: Python, C++, BASH

License: MIT

RRID:SCR\_020984

## Data Availability

Supporting data are available via the *GigaScience* database [62]. The GIAB SV call set is available in the GIAB FTP repository [50] and the sequencing data for HG002, HG003, and HG004, respectively, at [30, 52–55].

The SV-plaudit call set is available in the Supplementary materials at [32, 51]. The Polaris call sets are available via GitHub [31]. The NA12878 sequencing data are available in the European Nucleotide Archive under project PRJEB3381 [57, 58].

## Additional Files

**Supplementary Figure S1.** Accuracy metrics calculation from the contingency table

**Supplementary Figure S2.** Dataflow for proposing and selecting alternate variant SV descriptions

**Supplementary Figure S3.** Genotype concordance for different SV lengths

**Supplementary Figure S4.** Genotype concordance for GIAB SVs in Tier 1 regions grouped by SV length and TR overlap

**Supplementary Figure S5.** Genotype concordance for GIAB SVs with offset breakpoints

**Supplementary Table S1.** Features extracted from real and simulated NGS data

**Supplementary Table S2.** Counts of genotypes for SV truth sets in concordance evaluation

**Supplementary Table S3.** Genotyping accuracy for GIAB SVs

**Supplementary Table S4.** Recall, precision and F1 for genotyping homozygous reference vs. non-reference GIAB SVs

**Supplementary Table S5.** Genotyping accuracy for NA12878 call sets

**Supplementary Table S6.** Recall, precision and F1 for genotyping homozygous reference vs. non-reference for NA12878 call sets

**Supplementary Table S7.** Genotype concordance for GIAB SVs in tier 1 regions grouped by SV length

**Supplementary Table S8.** Recall, precision and F1 for genotyping homozygous reference vs. non-reference with discovery SVs as the input to SV genotyping

**Supplementary Table S9.** Genotyping Mendelian error SVs reported in Zook et al.

**Supplementary Table S10.** Genotyping accuracy and recall, precision and F1 for genotyping homozygous reference vs. non-reference SVs for original GIAB SVs and proposed alternative SVs

**Supplementary Table S11.** Execution time and memory usage for genotyping GIAB SVs

**Supplementary Methods.**

## Abbreviations

bp: base pairs; BWA: Burrows-Wheeler Aligner; CPU: central processing unit; GATK: Genome Analysis Toolkit; GIAB: Genome in a Bottle; GQ: genotype quality; kb: kilobase pairs; Mb: megabase pairs; ME: Mendelian error; MER: Mendelian error rate; NGS: next-generation sequencing; PacBio: Pacific Biosciences; RF: random forest; SV: structural variant; SVM: support vector machine; TR: tandem repeat; WGS: whole-genome sequencing.

## Competing Interests

The authors declare that they have no competing interests.

## Funding

Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) under Grant No. P20GM103449, awards UM1HL098123 and U01HL153009 from the National Heart, Lung, and Blood Institute (NHLBI) of the NIH, and the National Science Foundation (NSF) under Grant No. 1,827,373. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS, NHLBI, NIH, or the NSF.

## Authors' Contributions

M.D.L., A.B., and B.D.G. conceived of the project. M.D.L., C.P., M.S., and W.K. developed the software and performed the evaluation. M.D.L., A.B., and B.D.G. wrote the manuscript. All authors read and approved the final manuscript.

## Acknowledgements

## References

1. Weischenfeldt J, Symmons O, Spitz F, et al. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 2013;14:125–38.
2. Brandler WM, Antaki D, Gujral M, et al. Frequency and complexity of *de novo* structural mutation in autism. *Am J Hum Genet* 2016;98:667–79.
3. Kloosterman WP, Francioli LC, Hormozdiari F, et al. Characteristics of *de novo* structural changes in the human genome. *Genome Res* 2015;25:792–801.
4. Guan P, Sung W-K. Structural variation detection using next-generation sequencing data. *Methods* 2016;102:36–49.

5. Kosugi S, Momozawa Y, Liu X, et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* 2019;**20**:117.
6. Mahmoud M, Gobet N, Cruz-Dávalos DI, et al. Structural variant calling: the long and the short of it. *Genome Biol* 2019;**20**:246.
7. Chander V, Gibbs RA, Sedlazeck FJ. Evaluation of computational genotyping of structural variation for clinical diagnoses. *Gigascience* 2019;**8**:doi:10.1093/gigascience/giz110.
8. Lappalainen I, Lopez J, Skipper L, et al. dbVar and DGVA: public archives for genomic structural variation. *Nucleic Acids Res* 2012;**41**:D936–41.
9. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**:363–76.
10. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature* 2015;**526**:75–81.
11. Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature* 2011;**470**:59–65.
12. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. *Front Bioeng Biotechnol* 2015;**3**:92.
13. Teo SM, Pawitan Y, Ku CS, et al. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* 2012;**28**:2711–8.
14. Sedlazeck FJ, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;**15**:461–8.
15. Huddleston J, Chaisson MJP, Steinberg KM, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* 2017;**27**:677–85.
16. English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* 2014;**15**:180.
17. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.
18. Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 2012;**28**:i333–9.
19. Chiang C, Layer RM, Faust GG, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* 2015;**12**:966–8.
20. Spies N, Zook JM, Salit M, et al. svviz: a read viewer for validating structural variants. *Bioinformatics* 2015;**31**(24):3994–6.
21. Eggertsson HP, Kristmundsdóttir S, Beyter D, et al. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat Commun* 2019;**10**:5402.
22. Antaki D, Brandler WM, Sebat J. SV2: accurate structural variation genotyping and de novo mutation detection from whole genomes. *Bioinformatics* 2018;**34**(10):1774–7.
23. Audano PA, Sulovari A, Graves-Lindsay TA, et al. Characterizing the major structural variant alleles of the human genome. *Cell* 2019;**176**:663–75.e19.
24. Hickey G, Heller D, Monlong J, et al. Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol* 2020;**21**(1):35.
25. Chen S, Krusche P, Dolzhenko E, et al. Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol* 2019;**20**:291.
26. Handsaker RE, Van Doren V, Berman JR, et al. Large multiallelic copy number variations in humans. *Nat Genet* 2015;**47**:296–303.
27. Russell SJ, Norvig P, Davis E. *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ: Prentice Hall, 2010.
28. Mohiyuddin M, Mu JC, Li J, et al. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* 2015;**31**:2741–4.
29. Chu C, Zhang J, Wu Y. GINDEL: accurate genotype calling of insertions and deletions from low coverage population sequence reads. *PLoS One* 2014;**9**:e113324.
30. Zook JM, Hansen NF, Olson ND, et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* 2020;**38**(11):1347–55.
31. Polaris. <https://github.com/Illumina/Polaris>. Accessed 10 July 2020.
32. Belyeu JR, Nicholas TJ, Pedersen BS, et al. SV-plaudit: A cloud-based framework for manually curating thousands of structural variants. *Gigascience* 2018;**7**:doi:10.1093/gigascience/giy064.
33. Layer RM, Chiang C, Quinlan AR, et al. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* 2014;**15**:R84.
34. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;**32**:1220–2.
35. pbsv (2020) pbsv. <https://github.com/PacificBiosciences/pbsv>. Accessed: 10 July 2020.
36. Pedersen BS, Collins RL, Talkowski ME, et al. Indexcov: fast coverage quality control for whole-genome sequencing. *Gigascience* 2017;**6**:doi:10.1093/gigascience/gix090.
37. Picard toolkit(2020). Picard Toolkit (Version 2.22.7). <http://broadinstitute.github.io/picard>.
38. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. *Nature* 2019;**581**:444–51.
39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
40. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
41. Huang W, Li L, Myers JR, et al. ART: a next-generation sequencing read simulator. *Bioinformatics* 2012;**28**:593–4.
42. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013, <http://arxiv.org/abs/1303.3997>.
43. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* 2014;**30**:2503–5.
44. Chapman B, Kirchner R, Pantano L, et al.(2020). bcbio/bcbio-nextgen: v1.2.3. doi:10.5281/ZENODO.3743344.
45. Tarasov A, Vilella AJ, Cuppen E, et al. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* 2015;**31**:2032–4.
46. Parikh H, Mohiyuddin M, Lam HYK, et al. svclassify: a method to establish benchmark structural variant calls. *BMC Genomics* 2016;**17**:64.
47. Wala J, Beroukhir R. SeqLib: a C++ API for rapid BAM manipulation, sequence alignment and sequence assembly. *Bioinformatics* 2016;**33**:751–3.
48. Pedersen BS, Quinlan AR. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *Gigascience* 2019;**8**:doi:10.1093/gigascience/giz040.

49. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;12:2825–30.
50. Genome in a Bottle Sequence-Resolved SV Calls v0.6. [ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002\\_NA24385.son/NIST\\_SV\\_v0.6](ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385.son/NIST_SV_v0.6). Accessed 18 June 2019.
51. Belyeu J, Nicholas T, Pedersen B, et al. Supporting data for “SV-plaudit: A cloud-based framework for manually curating thousands of structural variants.” *GigaScience Database* 2018. <http://dx.doi.org/10.5524/100450>.
52. Zook JM, Catoe D, McDaniel J, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 2016;3:160025.
53. Genome in a Bottle HG002 Sequencing Data. [ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002\\_NA24385.son/NIST\\_HiSeq\\_HG002\\_Homogeneity-10953946/HG002\\_HiSeq300x.fastq/140528\\_D00360\\_0018\\_AH8VC6ADX](ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385.son/NIST_HiSeq_HG002_Homogeneity-10953946/HG002_HiSeq300x.fastq/140528_D00360_0018_AH8VC6ADX). Accessed 4 June 2019.
54. Genome in a Bottle HG003 Sequencing Data. [ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003\\_NA24149.father/NIST\\_HiSeq\\_HG003\\_Homogeneity-12389378/HG003\\_HiSeq300x.fastq/140721\\_D00360\\_0044\\_AHA66RADXX](ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG003_NA24149.father/NIST_HiSeq_HG003_Homogeneity-12389378/HG003_HiSeq300x.fastq/140721_D00360_0044_AHA66RADXX). Accessed 4 June 2019.
55. Genome in a Bottle HG004 Sequencing Data. [ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004\\_NA24143.mother/NIST\\_HiSeq\\_HG004\\_Homogeneity-14572558/HG004\\_HiSeq300x.fastq/140818\\_D00360\\_0046\\_AHA5R5ADXX](ftp://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG004_NA24143.mother/NIST_HiSeq_HG004_Homogeneity-14572558/HG004_HiSeq300x.fastq/140818_D00360_0046_AHA5R5ADXX). Accessed 4 June 2019.
56. smooove(2020). smooove. <https://github.com/brentp/smoove>. Accessed 25 June 2020.
57. Eberle MA, Fritzilas E, Krusche P, et al. A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* 2017;27:157–64.
58. Whole genome sequencing and variant calls for the Coriell CEPH/UTAH 1463 family to create a “platinum” standard comprehensive set for variant calling improvement. <ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194147/ERR194147.fastq.gz>. Accessed 29 May 2018.
59. Truvari(2020). Truvari. <https://github.com/spiralgenetics/truvari>. Accessed 28 May 2020.
60. Truvari fork. [https://github.com/mlinderm/truvari/tree/genotype\\_stats](https://github.com/mlinderm/truvari/tree/genotype_stats). Accessed 24 August 2020.
61. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10:doi:10.1093/gigascience/giab008.
62. Linderman M, Paudyal C, Shakeel M, et al.(2021). Supporting data for “NPSV: A simulation-driven approach to genotyping structural variants in whole-genome sequencing data.” *GigaScience Database*. <http://dx.doi.org/10.5524/100908>.