Blended Length Genome Sequencing (blend-seq): Combining Short Reads with Low-Coverage Long Reads to Maximize Variant Discovery

Ricky Magner[*,1], **Fabio Cunial**[*,1], **Sumit Basu**[2], **Ron Paulsen**[2], **Scott Saponas**[2], **Megan Shand**[1], **Niall Lennon**[1], **and Eric Banks**[1]

[*]These authors contributed equally to this work
[1]Broad Institute of Harvard and MIT, Cambridge, MA, USA.
[2]Microsoft Research, Redmond, WA, USA.

We introduce blend-seq, a method for combining data from traditional short-read sequencing pipelines with low-coverage long reads, with the goal of substantially improving variant discovery for single samples without the full cost of high-coverage long reads. We demonstrate that with only 4x long read coverage augmenting 30x short reads, we can improve SNP discovery across the genome and achieve precision and recall beyond what is possible with short reads, even at very high coverage (60x). For genotype-agnostic discovery of structural variants, we see a threefold improvement in recall while maintaining precision by using the low-coverage long reads on their own. For the more specialized scenario of genotype-aware structural variant calling, we show how combining the long and short reads in a graph-based approach results in greater performance than either technology on its own. The observed gains highlight the complementary nature of short and long read technologies: long reads help with SNP discovery by better mapping to difficult regions, and they provide better performance with long insertions and deletions (structural variants) by virtue of their length, while the larger number of short-read layers help with genotyping structural variants discovered by long reads. In this way, blend-seq offers many of the benefits of long-read pipelines without incurring the cost of high-coverage long reads.

short-read sequencing | long-read sequencing | hybrid sequencing | small variants | structural variants | genotyping | hard-to-map regions | cost optimization

Correspondence: ***rmagner@broadinstitute.org***

1: Introduction

While market-leading short-read sequencing technologies (such as Illumina's Novaseq X) are excellent at capturing small variations (SNPs and INDELs) in most of the human genome, they are far less effective both for hard-to-map regions as well as for structural variants (SVs). Long-read technologies (such as Pacific Biosciences' Revio and Oxford Nanopore Technologies' nanopore-based method), on the other hand, have shown remarkable success in both of these areas; however, the significantly higher cost has left them out of reach for many applications. In this paper, we introduce *Blended Length Genome Sequencing* (**blend-seq**), a novel means of combining traditional short-read pipelines with very low coverage long reads, and demonstrate how we can make substantial gains in both short and long variant calling for single samples without the full cost of high- or medium-coverage long-read sequencing.

Blend-seq is flexible with respect to choices and coverage levels of each sequencing technology; the methods and choices we present can adapt to a wide variety of setups. A pragmatic setting given existing practices is to augment the typical 30x coverage of Illumina short reads with 4x coverage of PacBio long reads. For short variants, we show that with this setting we can combine read types to improve SNP recall (while maintaining precision) not only in hard-to-map regions but across the entire genome, even compared to very high coverage short-read performance, which levels off well before 60x. For SVs, we show that we can improve recall by a factor of three across the genome, and by even more for variants overlapping with the exome, simply by using the low-coverage long reads on their own. For settings where the genotype of structural variants matter, we show even greater gains by combining short and long reads in a graph-based approach.

The implications for the SNP results are immediate: the vast majority of known genetic diseases are identified by these single nucleotide polymorphisms, so our improvements to SNP recall have the potential to impact clinical diagnosis. Such gains apply to the discovery of new variant associations as well: greater recall implies a greater ability to associate SNPs with phenotypic profiles in biobank or cohort data. As such this improvement in SNP performance has the potential of both improving existing diagnostic procedures and the discovery of new diagnostics.

At the same time, SNPs tell only part of the story. From population genomics (1) to understanding disease mechanics (2, 3), there is a need to move beyond short variants and consider a wider set of causes for genetic disease (4). To date, the best way to address these longer variants has been through long-read technologies, such as those from Pacific Biosciences (PacBio), with reads of 15-20kb, and Oxford Nanopore Technologies (ONT), with reads of up to 2mb. These far longer reads can span much larger variants, while simultaneously mapping more accurately to the correct locations in the human genome, particularly in low-complexity regions.

Our hypothesis has been that even a small amount of long-read information could greatly augment short-read results, shedding significant additional light on these problems without incurring the full cost of high-depth long-read sequencing. More generally, we hypothesized that the complementary capabilities of varying levels of short and long reads could be balanced to best tune their joint performance for a specific application. The blend-seq approach we present in

this paper is thus to vary the choice and coverage for each technology, leveraging their individual and joint capabilities to provide gains in short variants, structural variant discovery, and structural variant genotyping. We show results using PacBio and ONT technologies for long reads, coupled with Illumina sequencing for short reads, all with varying levels of coverage. Moreover, we show that different choices lead to differential performance gains for specific regions and classes of variants. This tuning approach provides a continuous path to increasing the amount and quality of the desired genomic information per unit of cost, modulated by the "control knobs" of levels of coverage as well as the choices of specific short/long read technologies.

We demonstrate these gains with three approaches, varying in complexity and degree of integration between short- and long-read pipelines. For improvements in SNP discovery, we leverage the mapping capabilities of low-coverage long reads, then break these reads up into "virtual" short reads that can be combined with true short reads in the state-of-the-art DRAGEN calling pipeline. For SV discovery, we simply use the low-coverage long reads to call structural variants without assistance from the short reads; while some past work has found that short reads can detect some SVs that long reads miss (5) and others have found merging SV calls from long and short reads gives slight improvements (6), we saw no appreciable gains in practice using our methods. For SV genotyping, we create a sample-specific graph reference that augments the standard reference with information from the long reads, then map the short reads to this augmented reference for calling genotype-aware structural variants.

In terms of related work, we note there have been some significant prior investigations on combining short and long reads. However, our work has focused on the specific scenario of combining *standard coverage* short reads with *low coverage* long reads, and doing so using independent, unmodified pipelines using only *in silico* methods; we are not aware of other work in this setting. Prior work has instead either required high coverage for both long and short reads, or required library preparation changes to establish biochemical correspondence between the long and short reads. The closest work to ours falls in the first category: a new offering from Variantyx (7, 8) which uses high coverage long reads (12-15x) in addition to standard short read coverage to improve variant discovery for diagnostic scenarios. Beyond this, the first category (high coverage long and short reads) is dominated by methods for hybrid *de novo* assembly, in which long and short reads are integrated to improve assembly quality (see e.g. (9–15)). The second category (library preparation changes) includes a wide variety of approaches, including read clouds, synthetic long reads, and linked reads as mechanisms to improve assembly (see e.g. (16–21)). In these approaches, the correspondence between short and long molecules is established *in vitro*. Finally, since we are focused on performance with real data, we do not investigate or compare to simulated read mixes as was done in (22).

While we believe the gains we demonstrate to be significant, we recognize that this work represents only the first steps of methods development for the blend-seq scenario. To that end, all code and sequencing data used to produce our results are freely available for future use.

2: Results

We describe our results in three areas detailed in the subsections below. The first is on short variants, and shows how modifying short-read calling pipelines to leverage long-read information along with short reads can result in significant gains in SNP recall across the genome, and particularly in hard-to-map regions. The second section focuses on structural variants, and shows how we can increase the recall of SVs across the genome by a factor of three with just a few layers of long reads. The last section investigates the setting where genotyping is critical for SVs, and shows how a graph-based approach can combine long and short reads to provide genotype-aware SV calling performance that exceeds either technology in isolation.

2.1. Short Variant Calling Performance. In this first subsection, we focus on improvements to the calling of short variants (events less than 50bp in length) with blend-seq. First, we show that we can substantially improve SNP recall (while maintaining precision) across the genome and by an even greater amount in hard-to-map regions. We achieve this by breaking mapped long reads into synthetic short reads that can augment existing short reads in the DRAGEN pipeline. Next, we investigate the sources of these gains to understand what fractions are coming from hard-to-map regions vs. other scenarios. Finally, we examine the same approach for INDELs, and find that while we can achieve some gains in recall, the effect is diluted by a drop in precision.

Approach and Data. While short-read pipelines perform well where their read lengths are sufficient to map unambiguously to the genome, their performance suffers when the mapping is no longer unique; conversely, long reads are well-known to be better at mapping uniquely to these more difficult regions. We thus sought a means by which we could leverage the higher-quality mapping of long reads within a state-of-the-art short-read calling pipeline. Our simple approach was to map the long reads on their own, then break each alignment into segments of short-read size (an idea that already appeared e.g. in (23)). These "virtual" short-read alignments were then passed in to the short-read calling pipeline along with the alignments of the actual short reads, to create a hybrid BAM which was then used for variant calling.

We performed our experiments with varying levels of short-read coverage, from 10x to 60x, simulated by downsampling a single set of reads (Illumina NovaSeq X) for a single sample (human sample HG002). The Illumina reads were processed using DRAGEN v3.7.8, an Illumina short-read pipeline (24) employed in several large-scale projects, including All of Us (25); in particular, we used DRAGMAP to align the reads.

For the long reads, we used four layers (4x) of PacBio Revio: this level of coverage was selected based on the tradeoffs between coverage and discovery performance for structural variants described in Section 2.2. The analysis involved first mapping the long reads using minimap2 (26), then partitioning these alignments into disjoint chunks of length 151 to match the expectations of DRAGEN for short reads. The alignments preserved their original positions and CIGAR strings. These "virtual" short-read alignments were then added to the BAM constructed for the real short reads. We then used DRAGEN v3.7.8 to call variants from this hybrid BAM.

We compared the results of the short reads alone (with DRAGEN), the 4x long reads alone (with DeepVariant), and the hybrid regime described above, against the NIST truth dataset for HG002 over their high-confidence interval regions (27). In addition to calculating whole-genome precision and recall statistics, we also collected these over some subregions of interest as determined by NIST (28).

SNP Performance. Using our hybrid method, we found that SNP recall with blend-seq showed significant improvements over short reads alone (at any coverage), while maintaining precision, for combinations with both PacBio and ONT (see Fig. 1). The peak performance of blend-seq is with 25-30x range of short-read coverage, where it significantly outperforms short reads alone in terms of recall while matching precision, even when compared to very high levels of short-read coverage where short-read performance appears to level off (see Table 1 for detailed values). Even at relatively low levels of short-read coverage (12x), we saw recall exceeding this maximal short-read recall, despite a small loss in precision in ONT (see Table 1).

We saw an even stronger improvement in performance in regions that are hard to map for short reads (see Fig. 2), substantially improving recall compared to short reads alone while making small gains in precision for the PacBio blend (see Table 1 for detailed values). These regions are difficult due to the limitations of short reads (28); as such, we expected long reads to have less difficulty mapping unambiguously. Even with only four layers of long-read coverage, we can see the complementary nature of the short and long reads boosting performance. While improvements across the genome and in difficult regions were promising, these gains might only appear in regions of the genome where variants would have low interpretability. To investigate whether this was the case, we ran the analysis above restricted to the exome (Fig. 3). The result is that we see a similar gain in recall as we saw for the whole genome even in these more interpretable regions (see Table 1 for details). Since most SNPs used for clinical diagnosis are in the exome, this increases the potential relevance of our improvements in diagnostic settings.

Sources of SNP Improvement. To better understand the source of this boost in SNP performance, we performed a detailed analysis into characterizing the new SNPs discovered using the hybrid readset. We focused on the low-mappability regions, since these experienced the largest gains. In particular, we looked at the additional SNPs we were able to capture via our approach: SNPs which were previously labeled false negatives with short reads alone, but became true positives with the hybrid method. After restricting the data to these cases, we manually inspected a subset of sites with the Integrative Genomics Viewer (IGV) (29). We observed cases (Appendix A Fig. 1) of SNPs where the short-read support either had zero depth or all reads had mapping quality zero. In both cases, the variant caller would have no evidence to use when making a call. We also observed situations (Appendix A Fig. 2) where nearly all short reads have mapping quality zero. In these situations the short reads are disadvantaged from making a call, but not necessarily deterministically across runs. Other examples (Appendix A Fig. 3) show a more complex situation where reference-supporting reads map well, but the few reads representing a correct haplotype within a cluster of nearby SNPs have mapping quality zero.

To better understand how general these cases are, we assigned categories to these additional SNPs gained by our method in the low-mappability regions. The category labels, based on mapping depth and quality, articulate the different reasons why a SNP was missed in the pure short-read data but captured in the hybrid regime (details are provided in the Methods section). The breakdown of these labels for the SNPs in question is given by the pie chart in Fig. 4 for the 30x short-read and 4x long-read mix (30x mix experiment), with other coverages described in the Supplement.

The first four categories (Zero Reads, Zero Alt-Supporting Reads, Only MAPQ Zero Reads, Only MAPQ Zero Alt-Supporting Reads) represent the case where the pure short-read variant caller has no usable evidence to make the correct call: these are easy to interpret, as the long reads were able to add valuable supporting haplotypes – enough to salvage the site and make the right call. In the 30x mix experiment, about 37% of heterozygous sites and about 78% of homozygous sites among those "corrected" in the long-reads mix fall into these categories. The remaining categories (Low Depth, Low Alt-Supporting Depth, Low MAPQ Reads, Low BaseQ Alt-Supporting Reads, and Other) are mostly sites where the pure short-read evidence is weak but not absent. Equivalent charts for other levels of short-read coverage, 20x (Appendix A Fig. 5) and 10x (Appendix A Fig. 4), are provided as supplementary data: we see a similar fraction of gained SNPs coming from the first four categories, with long-read evidence added to pileups where pure short reads had no evidence.

INDEL Performance. INDELs are defined as insertion or deletion events of length less than 50bp. Applying the hybrid method described above over the entire genome, we found that INDEL performance has greater recall for the PacBio blend compared to short reads alone, but precision is marginally lower without any additional filtering (Appendix

SNP Performance by Region

	WholeGenome			LowMappability			Exome		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Illumina 15x (control)	.9889	.9944	.9916	.8556	.9312	.8918	.9857	.9870	.9864
Illumina 30x (control)	.9926	.9961	.9944	.8813	.9461	.9125	.9887	.9902	.9894
Illumina 60x (high depth)	.9929	.9963	.9946	.8856	.9490	.9162	.9895	.9901	.9898
blend-seq 15x IL + 4x ONT	.9956	.9927	.9942	.9449	.9178	.9311	.9927	.9864	.9895
blend-seq 15x IL + 4x PB	.9956	.9951	.9954	.9409	.9439	.9424	.9934	.9895	.9914
blend-seq 30x IL + 4x ONT	.9965	.9952	.9958	.9462	.9382	.9422	.9933	.9892	.9912
blend-seq 30x IL + 4x PB	.9963	.9964	.9963	.9413	.9521	.9467	.9935	.9910	.9922
30x PB (high depth)	**.9989**	**.9993**	**.9991**	**.9861**	**.9944**	**.9902**	**.9983**	**.9991**	**.9987**

Table 1. SNP performance over various genomic regions. In each region and performance statistic, blend-seq outperforms the short read controls. Blend-seq with 30x Illumina (IL) + 4x PacBio (PB) exceeds asymptotic (60x) Illumina performance for all metrics. Bold values are max values in column and shaded cells are those exceeding all control values in column (first three rows).

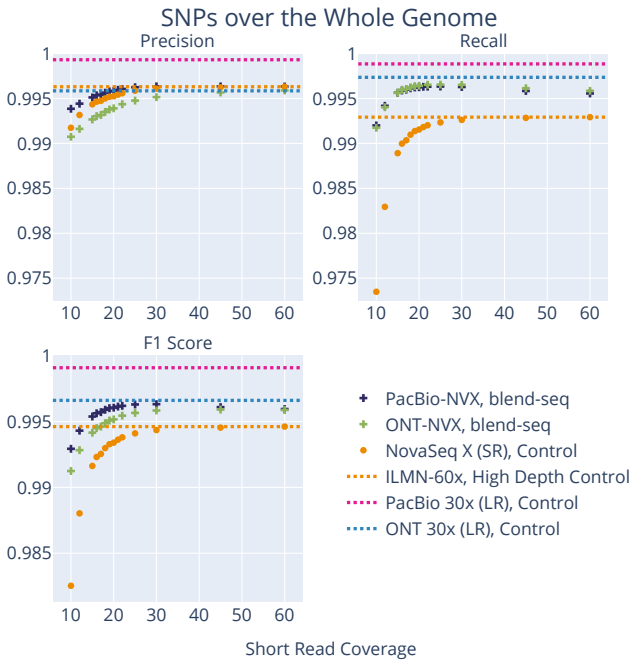


Fig. 1. The PacBio-Illumina hybrid at 4x+30x shows greater recall and precision of SNPs compared to short reads alone, even at 60x coverage; the ONT-Illumina hybrid shows greater recall but slightly lower precision. Note that even at the 4x+12x setting, both hybrids show recall exceeding short reads alone at 60x though with lower precision.

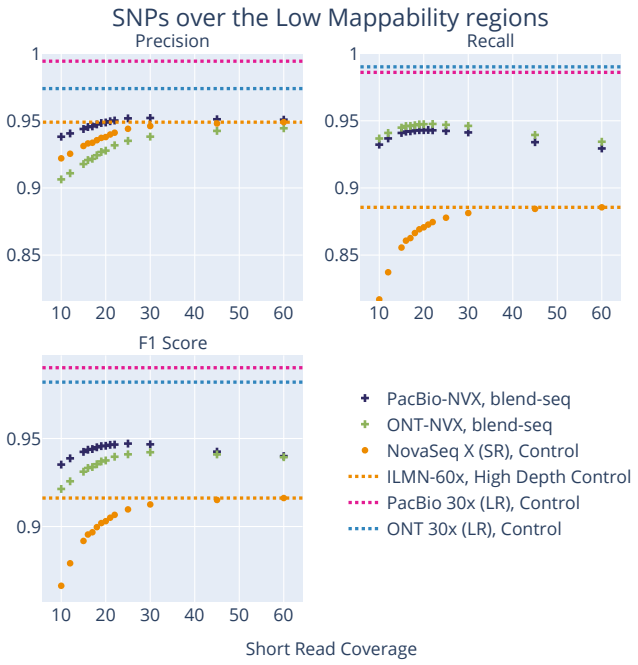


Fig. 2. When restricting to low-mappability regions, the boost in hybrid performance is more pronounced. Even 4x+10x long+short read mixtures outperform the 60x short reads in recall, leading to a net F1 increase for both the PacBio and ONT hybrids with 10x short reads.

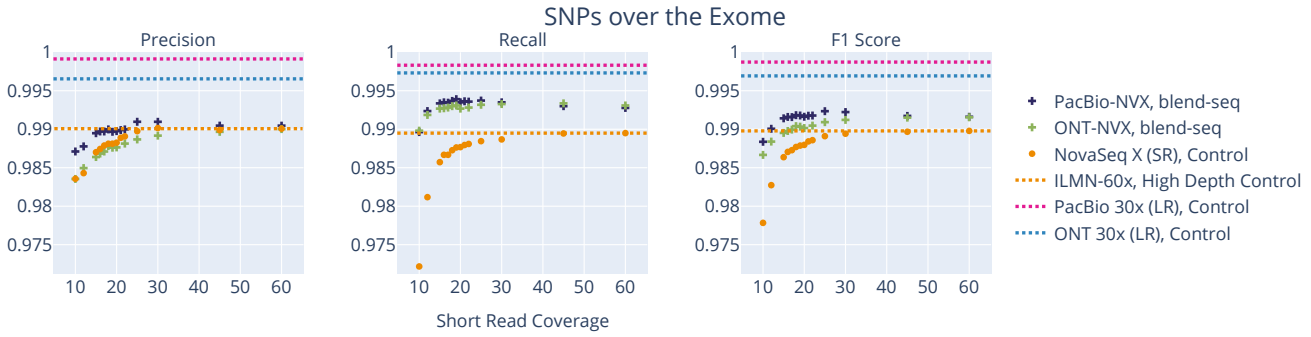


Fig. 3. Over the exome, blend-seq shows similar improvements in SNP performance as for the entire genome, with the PacBio-Illumina hybrid at 4x+30x outperforming recall and precision of short reads alone, even at 60x coverage.

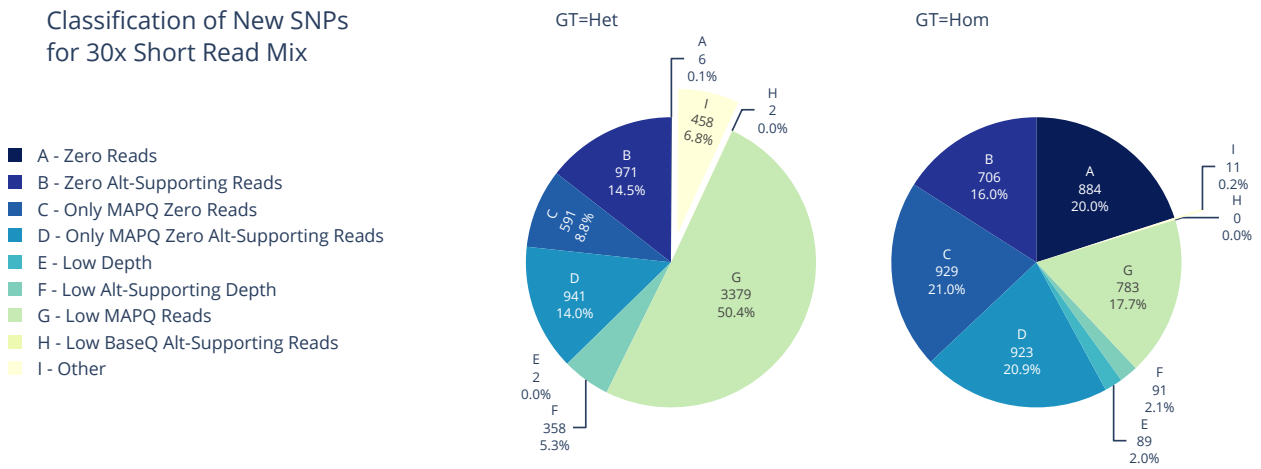


Fig. 4. A breakdown of the additional SNPs found by blend-seq in low-mappability regions. For the 30+4x Illumina-PacBio mix, we take the SNPs in low-mappability regions that were missed by short reads alone but discovered by blend-seq, then categorize them according to the nature of evidence supporting them. In the first four categories, there was no short-read evidence for the SNPs: this accounted for around 37% of the heterozygous sites and around 3/4 of the homozygous sites. The remaining labels represent sites where the short-read evidence was non-zero but limited. Detailed descriptions of each category are in the Methods section.

A Fig. 6). There was a greater boost in recall over low-mappability regions, but the overall effect was diluted by a lower precision. This is most likely due to a higher INDEL error rate in long read alignments and the fact that current tools are tuned to Illumina error rates, driving false positives for this variant type. See Table 7 for details.

2.2. Structural Variant Discovery. In this subsection, we investigate the use of blend-seq at low long-read coverage for SV discovery. This is motivated by the well-established advantage of long reads for structural variant detection as compared to short reads (see e.g. (30–37)). In this first scenario, we are interested in retrieving SVs without concern for their underlying genotype; we will consider genotype-aware retrieval in the next subsection. Unlike our earlier results for SNPs, the best performance in this setting comes from using low-coverage long reads alone with conventional long-read-calling pipelines; the short reads are not utilized at all.

We first show results over the entire genome, demonstrating an over three-fold increase in SV recall while maintaining precision. We then break down our results in two ways: first, to investigate where the gains are coming from, we study SV performance by genomic regions, including regions overlapping the exome, NIST-defined difficult regions (those that are harder to map for short reads), and easier regions (the complement of difficult regions). Second, we look at the results across SV types, to see what classes of SVs are most improved by our approach.

Approach and Data. We investigated the performance of SV detection at different levels of long-read coverage, focusing on structural variants (large insertions and deletions with length $\geq$ 50bp) in the HG002 sample discovered using Sniffles2 (38) and allowing even a single read of evidence to support a call (results from other state-of-the-art SV callers, e.g. (39–41), displayed similar trends and are omitted for brevity). The outputs were cleaned and benchmarked with Truvari (42) against the NIST-Q100 SV V1.0 truth (43, 44), a set of calls derived from the highest-quality long-read assembly of HG002 available to date (45). Default values were used for all parameters of Truvari other than sequence comparison, which was turned off. This latter setting was used so that we could fairly compare the benchmarking results with the pure short-read pipeline, which does not report inserted bases. Note that with default settings Truvari labels variants as true positive, false positive, etc. based solely on matching alleles (up to some ambiguity in position, length, etc.), and does not consider the genotype by default; we will consider genotyping accuracy in more detail in the following section.

For our long reads, we used PacBio Revio reads and ONT R10.4 reads from a public experiment (46), downsampled to 1–10x, as our experimental groups, and also included the full 30x setting to examine high-coverage long read performance. These were compared against what we refer to as the "GATK-SV callset," the HG002-specific calls from a recent Illumina 30x cohort-level callset (47). This callset is the result of performing multi-sample SV calling with GATK-SV (48) and subsequently filtering by several criteria (including restricting results to HG002). GATK-SV is a state-of-the-art ensemble short-read SV discovery pipeline that integrates the output of several specialized callers, including Manta (developed by Illumina (49)), cn.MOPS and GATK-gCNV for copy-number variants (50, 51), and MELT (52) for mobile element insertions. Note that unlike our method, which operates on individual samples in isolation, GATK-SV can only be run on cohorts, as it leverages information from multiple samples to improve its precision.

We provide a few remarks with respect to prior investiga-

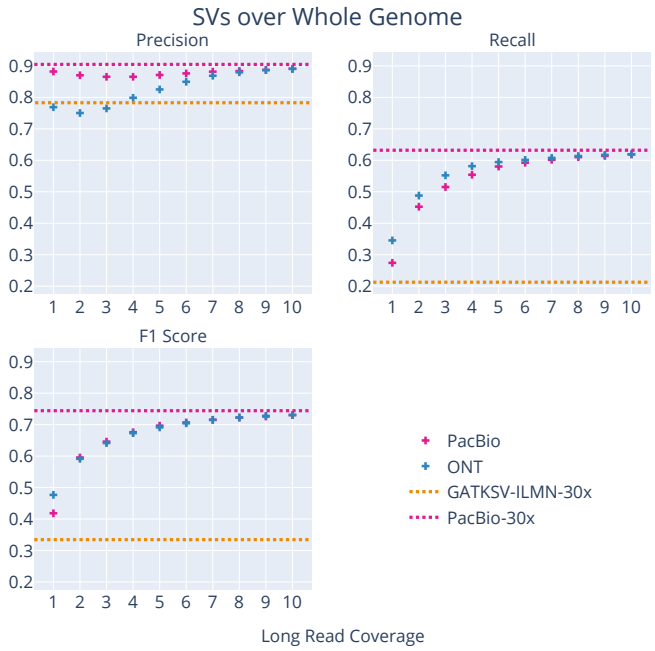


Fig. 5. Structural variant discovery with long reads vs. short reads. Long reads on their own, even at very low coverage, substantially outperform state-of-the-art short reads in both precision and recall. The GATKSV-ILMN-30x line represents the performance of short-read SV pipelines with Novaseq X reads; PacBio-30x represents full-coverage long-read performance.

tions in this area. First, long read downsampling analyses for SV discovery have been performed in the past (see e.g. (53)), but those studies either did not consider coverage levels as low as ours, used more stringent parameters in the callers, or compared to less accurate and complete truthsets. Second, although it is in principle possible to model the theoretical recall of structural variants as a function of coverage and read length (see e.g. (5)), our perspective is that practical issues like the association of SVs with repeats and the effect of sequencing errors and of read alignment algorithm on SV calling require the systematic empirical investigation we present here. Next, for the PacBio reads, we note that our sniffles2-based long read SV recall is lower than what is reported in (54), even for high levels of coverage: this is because we did not have phasing information for our hybrid data, as is further explained in the Discussion. Finally, for the ONT reads, which are known to have lower base quality, prior work has shown that short reads can be used to ameliorate this issue (see e.g. (55, 56) and references therein). We did not explore this approach in this study, and leave that for future work.

SV Performance by Genomic Region. We observed that even at just one layer of coverage, the long reads outperformed the short reads in terms of precision and recall for the genome as a whole (Fig. 5). With four layers of coverage, long reads substantially outperformed short reads in all regions. We observed diminishing returns on long-read depth beyond a small number of layers, with results approaching the 30x long-read performance at 10x coverage.

At 4x coverage for both PacBio and ONT, we close much of the gap to the full 30x using this pipeline; as such this serves as a compelling coverage/performance tradeoff. We broke down the results across regions of interest, such as the exome, difficult regions (NIST's "AllDifficult" regions, deemed to be difficult for sequencing by short-read technologies (28)), and easy regions (the complement of difficult regions). Table 2 shows details for all of these regions for short reads alone, 1x and 4x coverage for both PacBio and ONT, and 30x PacBio (a proxy for the asymptotic limit of this technology). We see similar trends in each set of regions, with 4x long reads substantially outperforming the short reads even in the easier regions where we expect short reads to perform the best.

Additional figures in the style of Fig. 5 for these regions are included in the Supplement (exome: Appendix A Fig. 7; easy regions: Appendix A Fig. 8).

Performance by SV Type. When we break down our results by SV type, we see a difference in performance improvement for insertions vs. deletions (Fig. 6, Fig. 7). Blend-seq results in large gains in recall for both categories, but for deletions the gain is smaller for precision since short reads achieve relatively high precision in this category. This is not surprising, given that short reads can gather accurate depth evidence for large deletions, but struggle to assemble large inserted sequences. Note that even if we further restrict deletions to regions where we expect short reads to be most competitive, i.e., the regions that are easiest for short reads to map ("Easy Regions"), we still find that four PacBio long read layers substantially outperform short-read performance in both precision and recall (figures and detailed numbers are in the Supplement: Appendix A Fig. 9 and Table 8).

We broke down our results by SV length, to see whether performance varies between relatively short (100-250bp) events and longer (2.5k-10k) events. Our expectation was that we would see a difference between PacBio and ONT performance: while ONT reads are longer (mean 50.7kb with standard deviation 66.9kb), PacBio reads (mean 15.6kb with standard deviation 3.9kb) have higher per-base accuracy, so we expected to see ONT do better in longer events and PacBio in shorter events. This is indeed the case, as shown in Fig. 8, Fig. 9, and Table 3. The PacBio reads show better precision for smaller events even at extremely low coverage, and the ONT recall outperforms the PacBio reads for larger events. In both cases, long reads outperform the short read control.

2.3. Genotype-Aware Structural Variant Discovery. In this subsection, we consider the setting in which we not only need to find structural variants, but to genotype them correctly as well. This is not critical for all applications, but in diagnostic scenarios, for instance, knowing the difference between a heterozygous vs. a homozygous variant could have substantial implications on a potential loss of function or other disruptions in the transcriptome. For this setting, we develop a method in which we create a sample-specific graph reference from the low-coverage long reads, then map the short reads to this reference and perform variant call-

SV Discovery by Region

	Whole Genome			Easy Regions			Hard Regions			Exome		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Illumina 30x (control)	.2126	.7830	.3344	.4030	.8606	.5490	.1759	.7576	.2855	.1096	.5147	.1807
ONT 1x	.3454	.7688	.4767	.4109	.7276	.5252	.3381	.8215	.4790	.3455	.7741	.4778
PacBio 1x	.2742	.8822	.4183	.3226	.9345	.4796	.2655	.8780	.4077	.2070	.8354	.3318
ONT 4x	.5815	.7987	.6730	.6932	.7830	.7354	.5675	.8325	.6749	**.6088**	.7716	.6806
PacBio 4x	.5542	.8656	.6757	.6487	.9236	.7621	.5391	.8606	.6630	.4627	.8532	.6000
PacBio 30x (high depth)	**.6321**	**.9048**	**.7442**	**.7541**	**.9475**	**.8398**	**.6153**	**.8973**	**.7300**	.5906	**.8789**	**.7065**

Table 2. SV discovery by region: long reads (both PacBio and ONT) outperform short reads in in recall and precision, even at 1x coverage when evaluated over the whole genome (except ONT precision), and at 4x for all individual regions (except ONT precision in the Easy Regions). Bold values are max values in column and shaded cells are those exceeding the control value in column (first row).

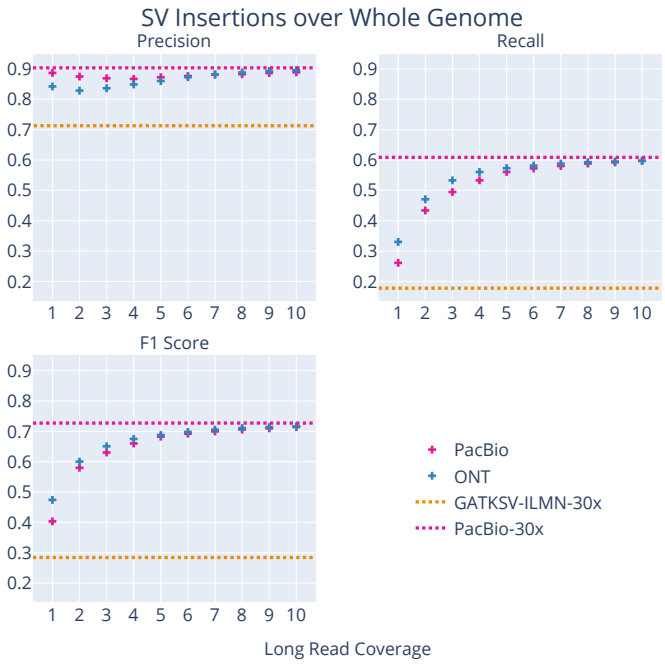


Fig. 6. Performance on insertion structural variants for short reads and long reads. Long reads outperform short reads even at very low coverage in this setting; this is because large insertions are difficult for short reads to resolve as they cannot be mapped to the reference. Note that these comparisons do not attempt to match the inserted sequence, as this data is not available for the short-read callset.

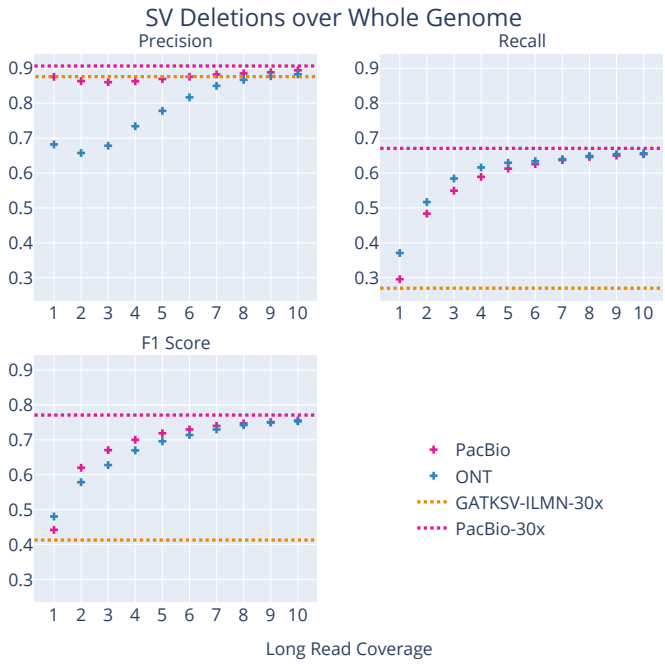


Fig. 7. Performance on deletion structural variants for short reads and long reads. While PacBio long reads at 4x outperform short reads substantially in terms of recall while matching precision, we note that short-read precision rivals that of long reads. This is because deletions are easier for short reads to resolve than insertions, as there is no novel material that cannot be mapped to the reference.

SV Discovery by Event Size

	100-250			2.5k-10.0k		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Illumina 30x (control)	.1582	.8083	.2647	.2268	.7378	.3470
ONT 1x	.3274	.7620	.4580	.4050	.8740	.5536
PacBio 1x	.2606	.8837	.4025	.1935	.8611	.3160
ONT 4x	.5503	.7933	.6499	**.6922**	.8275	**.7538**
PacBio 4x	.5337	.8672	.6607	.5050	.8783	.6412
PacBio 30x (high depth)	**.5970**	**.9093**	**.7208**	.6535	**.8784**	.7494

Table 3. SV discovery by event size: while long reads outperformed short reads in most categories, the two long read technologies showed different strengths. For shorter events, we found PacBio performed best, likely due to its higher per-base accuracy; on larger events, ONT had better recall, leveraging its longer read lengths. This was true even when 4x ONT was compared against 30x PacBio reads. Bold values are max values in column and shaded cells are those exceeding the control value in column (first row).

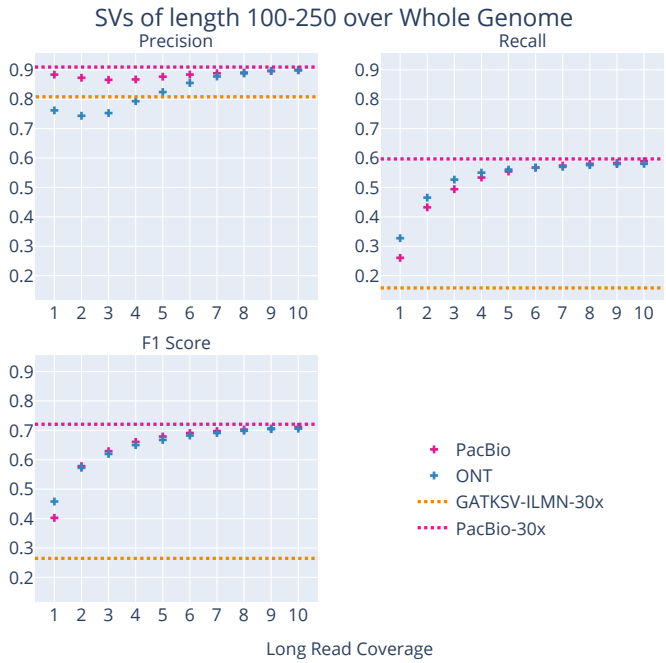


Fig. 8. Performance on 100-250 bp structural variants for short and long reads. These are the most common type of SVs; as such, performance trends largely follow the earlier results for all SV.

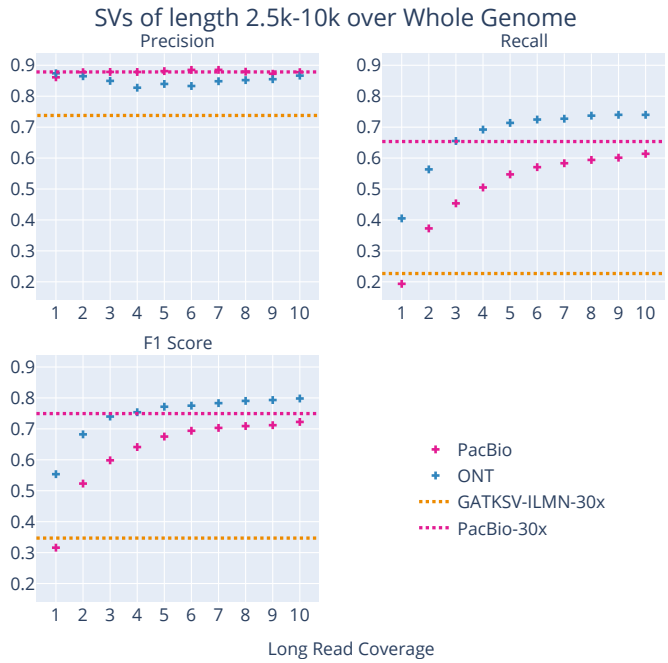


Fig. 9. Performance on 2.5k-10k bp structural variants for short and long reads. When restricting to these longer events, we see the ONT Ultra Long reads have better recall than PacBio.

ing on the graph. We present our results across the entire genome, where we show modest improvements with respect to using short or long reads alone, then present results for easier-to-map regions, where the improvement is more substantial since the short reads can provide greater assistance there. Plots over difficult regions are included in the Supplement for completeness.

Approach and Data. While we showed that we could substantially improve SV recall and precision (agnostic of genotype) in the previous section, here we wish to explore whether we can improve *genotype-aware* SV calling performance by integrating short-read information. Our approach was to create a "personalized" graph reference for the sample using the low-coverage long-read alignments, then call SVs from short reads aligned to the graph. In this way, we hoped to improve precision by avoiding calls that did not have enough short-read support, and to improve our genotyping accuracy by taking advantage of the higher-depth short reads. In the previous subsection, we found 4x long reads to be a reasonable tradeoff between coverage and SV discovery performance, and thus chose this as a fixed setting of long-read coverage in this subsection. We then varied the short-read coverage from 10 to 60x to explore the marginal benefit from additional layers of short reads.

We note that personalized graph references for short-read mapping are under active development in the bioinformatics literature, but they are typically built by subsetting a large-cohort pangenome to the parts that are most similar to the short reads for the sample at hand (57–61). As such they are unlikely to contain rare SVs carried by the sample. In our approach, on the other hand, we are building the personalized reference from sample-specific low-coverage long reads.

For each long-read technology (PacBio or ONT), we used the vg toolkit (62) to create the graph reference from the long-read alignments. This involved taking all structural variants (INDELs with length $\geq$ 50bp) from the CIGAR encoding of the alignments and writing them to a VCF file. This VCF file was used as the input to vg to create the graph; the short reads were then mapped to this. Next, vg was used to call variants from these alignments, and the resulting VCF went through a postprocessing step described in the Methods section before comparing to the NIST-Q100 HG002 SV truth set with Truvari (42). Truvari was run with all parameters set to defaults except for sequence comparisons, which were turned off. This process generated TP, FP, and FN labels, from which the precision, recall, and F1 scores were calculated.

Truvari compares events based on proximity and size. The tool allows some amount of flexibility to account for both consensus calls made from slightly different alignments as well as for reference ambiguity in repetitive regions. As opposed to traditional short-variant statistics, by default Truvari will label a “true positive” event based solely on the comparison of alleles between variants *without* genotype information taken into account – this was the output used in the previous subsection. The tool also reports TP matches *with* genotype information forced to match (42). This allows for six different performance statistics (precision, recall, and F1 for both genotype-agnostic and genotype-aware) to be calculated depending on whether the match should be forced to match genotype in addition to the site-level match. To clarify that we are referring to the genotype-aware values in these figures and tables, they are labeled as e.g. “GT-Precision” instead of

"Precision."

Performance Across the Entire Genome. For the genotype-aware setting, we see better performance with the blend-seq approach across all categories (Fig. 10). This is expected, since the short reads have much higher depth than the long reads: while the long reads lend their superior mapping capabilities and lengths to better SV haplotype discovery, the short reads can be used to help resolve the genotypes of those haplotypes. Detailed values are shown in Table 4.

Performance plots broken down by deletions vs. insertions are given in the Supplement (Appendix A Fig. 12, Appendix A Fig. 13).

Performance in Easier Regions. The boost in genotype-aware performance provided by the short reads is greatest where the short reads are best able to map: as in the last subsection, we refer to these as "Easy Regions" in our plots, comprised of the complement of NIST's "AllDifficult" regions (see Section 2.2). This setting reduces the total 30k variants in our truth dataset to 11.5k overlapping or contained within this region (28). In the genotype-aware setting, we see the hybrid method outperform 4x long reads alone in terms of precision, recall, and F1, starting around 12x of short reads when paired with ONT and 20x of short reads when paired with PacBio, with performance continuing to increase up to 60x short reads (Appendix A Fig. 14). Between the ONT and PacBio technologies, we see something of a precision-recall tradeoff: PacBio has the upper hand for precision and ONT for recall in these plots, as we would expect from our discussion in the previous subsection. Detailed numbers are supplied in Table 4).

We can break down the genotype-aware performance further in this region by SV type. As expected, we see the biggest gains are in deletions, since this is where short reads can contribute the most. The values for all statistics for deletions are recorded in Table 5. Plots across short-read coverage for easier regions are also in the Supplement (Appendix A Fig. 16). The performance boost is more modest but still visible for insertions, as seen in Table 6. Plots across read coverage for easier regions are also in the Supplement (Appendix A Fig. 15).

3: Discussion

We have demonstrated how blend-seq, by combining or selecting from standard-coverage short reads and low-coverage long reads, can provide substantial gains in SNP precision and recall, in structural variant (SV) discovery, and in genotyping structural variants. However, we recognize that we have demonstrated these results with specific choices of datasets, pipelines, and settings. In this section we discuss a few limitations of these choices, with a view towards further developments.

In terms of genomic datasets, we presented all results on a single genome, HG002, for ease of accessing high-quality truth data. To expand this in future work, we could create a proxy for ground truth using high-quality assemblies like those in the Human Pangenome Reference Consortium (63) or the Human Genome Structural Variation Consortium (64), then leveraging assembly-based variant callers like Dipcall (65) or hapdiff (66). In a similar vein, by using a complete reference genome, like the Telomere-to-Telomere (T2T) reference, we may further improve mapping quality and be able to make better variant calls using blend-seq techniques.

Further, the tools used here represent just one set of choices in a vast and evolving landscape: for example, several long-read mappers exist, as well as several SV callers and genotypers, both for long and for short reads, each with parameters that could be tuned for particular applications (see e.g. (54, 67–73) for a small sampler); the field is under constant development, particularly for long reads. One could try to run multiple long-read SV callers (including possibly some assembly-based callers) and to merge and filter their output using features that come both from long- and from short-read alignments. In addition, bespoke methods for specific hybrid regimes could be developed, both for SV and for SNP/INDEL calling.

The sequencing technologies investigated in this paper represent some of the most popular options, but are only a subsample of those that are now commercially available. In theory, any mix of technologies might provide benefits of similar magnitude to variant discovery, and different error mode combinations may provide opportunities for optimizing mixes for specific applications. In addition to the PacBio-Illumina and ONT-Illumina combinations studied here, we also investigated hybrid SV genotyping performance using Ultima short reads. The resulting performance is almost identical to hybrid Illumina, and is summarized in Section B of the Supplement.

For short variants, we emphasized the performance of SNPs alone (vs. INDELs) for two reasons: one is that the long reads we used added extra short INDELs in their alignments, which led to a greater rate of false positives than the short reads alone and reduced overall precision. Despite this, we observed a similar boost in INDEL recall, and hypothesize that with additional tuning of available filtering models (either for the alignments or for the variants) we might retain the boost in recall without sacrificing precision, similar to the SNPs. The second reason is that tabulating read support for INDELs is a more subtle problem than for SNPs, as representation differences in alignments have to be resolved, especially in repetitive regions. This too could be handled by custom methods for normalizing the alignments (74) when calculating hybrid read support.

Also in the context of short variants, the DRAGEN version used in this paper (v3.7.8) was chosen to match the version used by the All of Us Research Program (25), a large biobank funded by the U.S. government. It should be noted that recent work shows improved performance of DRAGEN

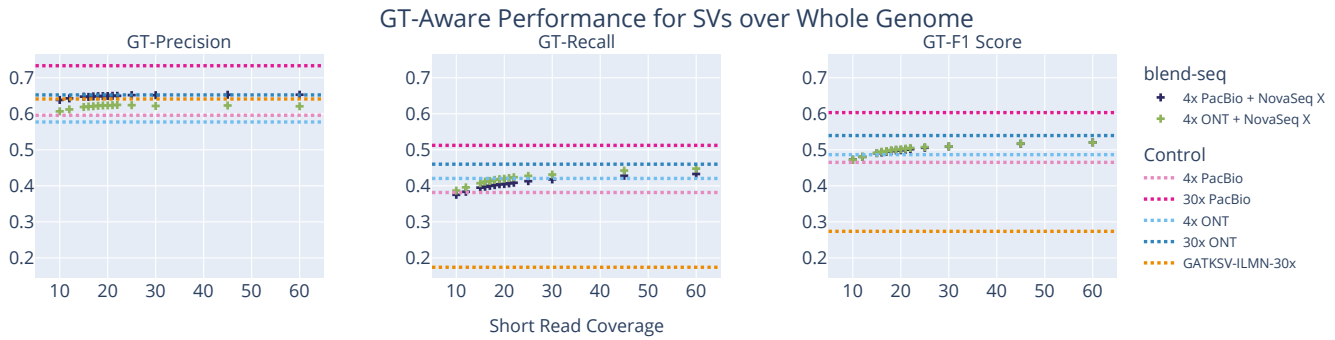


Fig. 10. GT-aware SV discovery with long reads alone, short reads alone, and our hybrid method. When we also account for GT-matches in our true positive labels, we see the GT-Precision for blend-seq outperforms the 4x long reads alone, while the GT-Recall outperforms the 30x short reads alone, leading to an overall better GT-F1 Score.

Genotyping SV Performance by Region

	Whole Genome			Easy Regions			Hard Regions		
	GT-Recall	GT-Precision	GT-F1 Score	GT-Recall	GT-Precision	GT-F1 Score	GT-Recall	GT-Precision	GT-F1 Score
Illumina 30x (control)	.1741	.6411	.2738	.3458	.7382	.4710	.1409	.6076	.2288
ONT 4x (LR-only)	.4206	.5773	.4866	.4752	.5390	.5051	.4069	.5964	.4837
PacBio 4x (LR-only)	.3815	.5959	.4652	.4314	.6155	.5073	.3672	.5858	.4514
blend-seq ONT 4x + Illumina 10x	.3868	.6068	.4724	.4738	.7244	.5729	.3693	.5908	.4545
blend-seq PacBio 4x + Illumina 10x	.3753	.6385	.4727	.4531	**.7412**	.5624	.3557	.6191	.4518
blend-seq ONT 4x + Illumina 30x	.4314	.6218	.5094	.5153	.7205	.6009	.4147	.6076	.4930
blend-seq PacBio 4x + Illumina 30x	.4179	.6517	.5092	.4858	.7386	.5861	.3995	.6341	.4902
ONT 30x (LR-only)	.4601	.6524	.5396	.5389	.6230	.5779	.4454	.6427	.5262
PacBio 30x (LR-only)	**.5124**	**.7334**	**.6033**	**.5847**	.7349	**.6512**	**.4936**	**.7198**	**.5856**

Table 4. GT-aware SV discovery performance. The 30x+4x blend-seq settings outperform the pure long- and short-read controls for the whole genome, and to an even greater degree when restricting to easier regions for short reads. This is expected, as the short reads are able to contribute more to the hybrid method in these regions. Bold values are max values in column, lightly shaded cells are those exceeding the short read control (first row), and deeply shaded cells are those where the hybrid method exceeds both short reads and long reads on their own (first three rows).

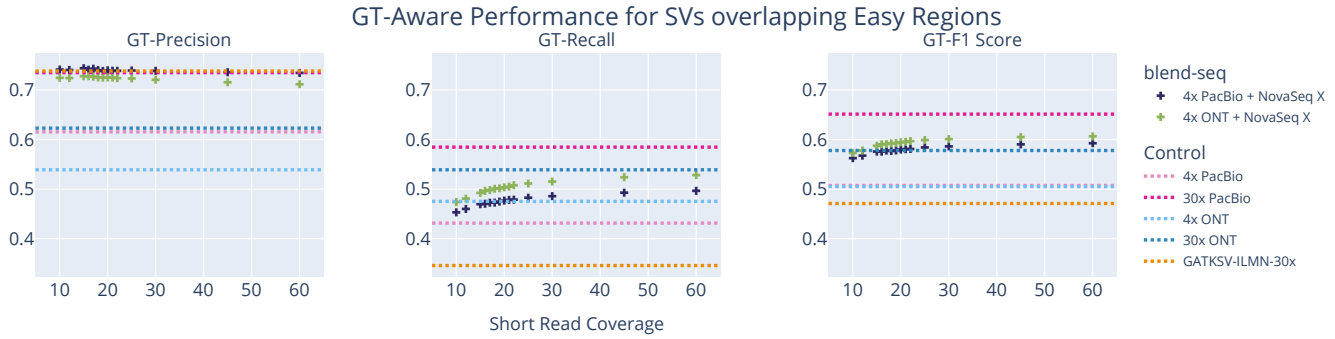


Fig. 11. GT-aware SV discovery on easier regions with long reads alone, short reads alone, and our hybrid method. The hybrid method's precision outperforms the pure long-read pipelines and is comparable to the pure short-read pipeline, with a slight boost in recall for each of the hybrid mixes when comparing to their respective pure long read recall performance. This yields an increase in GT-aware F1 scores at all layers of short-read depth for this region. Note that the 4x PacBio and 4x ONT controls visually overlap for GT-aware F1 scores with values 0.5073 and 0.5051 respectively.

in later versions (75). However, no versions past v3.7.8 have an open-source implementation, while the variant caller we used should be functionally equivalent to DRAGEN-GATK (76), which also influenced the decision to use an earlier version.

A final note with respect to SV performance presented in some prior work: because we did not have phasing information readily available for our hybrid data, we could not use tools like `truvari refine` or `vcfdist` (77) to better compare the resulting haplotypes to the truth. This resulted in our recall being lower than what was reported by (54), which did have phasing information and thus performed this processing step. To ensure this was the source of the discrepancy, we phased our 30x PacBio sample using HiPhase (78) and ran `truvari refine`: this resulted in a 95% F1 score, which is indeed comparable to (54). In future work, we could potentially adapt this evaluation to our blend-seq settings by using a long-read phasing tool compatible with multiple sequencing technologies.

4: Future Work

Beyond the immediate next steps described in the Discussion, we see a few key areas for further investigation.

Genotyping SV Deletions Performance by Region

	Whole Genome			Easy Regions			Hard Regions		
	GT-Recall	GT-Precision	GT-F1 Score	GT-Recall	GT-Precision	GT-F1 Score	GT-Recall	GT-Precision	GT-F1 Score
Illumina 30x (control)	.2292	.7436	.3503	.7688	.8289	.7977	.2054	.7233	.3199
ONT 4x (LR-only)	.5003	.5961	.5440	.7476	.5246	.6165	.4928	.6507	.5608
PacBio 4x (LR-only)	.4553	.6674	.5413	.6924	.7699	.7291	.4463	.6666	.5347
blend-seq ONT 4x + Illumina 10x	.4707	.6692	.5527	.8109	.8928	.8499	.4581	.6590	.5405
blend-seq PacBio 4x + Illumina 10x	.4651	.7102	.5621	.7707	.9332	.8442	.4529	.6987	.5496
blend-seq ONT 4x + Illumina 30x	.5093	.7051	.5914	.8217	.8971	.8577	.4976	.6963	.5804
blend-seq PacBio 4x + Illumina 30x	.4985	.7431	.5967	.7777	**.9372**	.8500	.4870	.7332	.5852
ONT 30x (LR-only)	.4942	.6797	.5723	.6569	.6466	.6517	.4848	.6759	.5646
PacBio 30x (LR-only)	**.6092**	**.8235**	**.7003**	**.9016**	.9138	**.9077**	**.5972**	**.8167**	**.6899**

Table 5. GT-aware performance for SV deletions. The 4x+30x blend-seq settings outperform short-read and long-read controls for all metrics other than genome-wide GT-Precision and in the Hard Regions. Recall is still substantially higher for blend-seq, so the GT-F1 scores significantly outperform the controls in all conditions. Bold values are max values in column, lightly shaded cells are those exceeding the short read control (first row), and deeply shaded cells are those where the hybrid method exceeds both short reads and long reads on their own (first three rows).

Genotyping SV Insertions Performance by Region

	Whole Genome			Easy Regions			Hard Regions		
	GT-Recall	GT-Precision	GT-F1 Score	GT-Recall	GT-Precision	GT-F1 Score	GT-Recall	GT-Precision	GT-F1 Score
Illumina 30x (control)	.1406	.5640	.2250	.2257	**.6733**	.3381	.1003	.5023	.1672
ONT 4x (LR-only)	**.5003**	.5961	.5440	**.7476**	.5246	.6165	**.4928**	.6507	.5608
PacBio 4x (LR-only)	.3366	.5477	.4170	.3573	.5549	.4347	.3172	.5289	.3966
blend-seq ONT 4x + Illumina 10x	.3357	.5621	.4203	.3781	.6496	.4779	.3132	.5392	.3963
blend-seq PacBio 4x + Illumina 10x	.3206	.5863	.4145	.3629	.6591	.4681	.2944	.5574	.3853
blend-seq ONT 4x + Illumina 30x	.3841	.5677	.4582	.4283	.6506	.5165	.3624	.5471	.4360
blend-seq PacBio 4x + Illumina 30x	.3688	.5918	.4545	.4029	.6616	.5008	.3444	.5658	.4281
ONT 30x (LR-only)	.4942	**.6797**	**.5723**	.6569	.6466	**.6517**	.4848	**.6759**	**.5646**
PacBio 30x (LR-only)	.4535	.6732	.5419	.4946	.6669	.5680	.4283	.6518	.5169

Table 6. GT-aware performance for SV insertions. The ONT 4x control already recovers enough SVs to beat the other categories, though we do see the blend-seq mixtures outperforming the other controls in F1 Score. Bold values are max values in column, lightly shaded cells are those exceeding the short read control (first row), and deeply shaded cells are those where the hybrid method exceeds both short reads and long reads on their own (first three rows).

The first would be to see whether the improved SNP performance we demonstrated here has any clinical significance, i.e., whether the extra SNPs we find with blend-seq have an impact in the diagnosis of diseases with known biomarkers. This would include measuring the performance of blend-seq in medically relevant genes that are challenging for short reads (79).

Moving beyond diagnostics to finding new biomarkers, we would like to assess the quality of multi-sample, joint short and long variant callsets generated on biobanks of blend-seq samples - if the results from this work generalize, we should see both new SNPs in hard to map regions as well new SVs. These could ultimately enable researchers to associate these variants with disease phenotypes without incurring the cost of high- or medium-coverage long reads.

We would also like to look more closely at specific classes of structure variants. In particular, tandem repeats and copy number variants (CNVs) were not considered independently of general SVs with any specialized algorithm, but it is likely long reads are already helping refine short-read calls to better pin down precise copy numbers and breakpoints of events in repetitive regions. Future work could look explicitly at what was happening with these events. Similarly, studying the performance of blend-seq in discovering inversions, translocations, mobile element insertions, and complex SVs could reveal new application areas for these methods.

In terms of our underlying methods, we would like to more fully investigate how the hybrid graph representation from Section 2.3 could benefit the other results presented here. For instance, we could add our short-variant calls to the personalized graph reference, potentially improving the mapping of short reads and potentially SV genotyping performance as well. In the same vein, mapping short reads to our SV-aware personalized graph reference could also improve short-variant discovery, since one could project the graph alignments to the usual linear reference (80) and feed the resulting BAM to the short read calling pipeline.

We would also like to consider even more diverse combinations of sequencing technologies. For example, combining PacBio and ONT reads may improve performance, as longer ONT reads may help discover larger SVs or span large repeated regions, while the PacBio CCS reads can provide higher base qualities for resolving smaller events.

Finally, in terms of additional applications, we wish to explore is how the addition of low-coverage long reads may help with other key problems in sequencing, such as methylation and phasing. Both PacBio and ONT long read data include information pertaining to methylation (81); it will be interesting to investigate how much of this is recoverable at low levels of coverage. For phasing, just a few high-quality

spanning long reads might be sufficient to phase nearby short variants, so we hypothesize that better phasing of short variants could be possible as well.

5: Conclusion

We have demonstrated that using blend-seq to combine (or select from) short and long read technologies leads to substantial improvements in variant calling performance in multiple scenarios. In particular, augmenting traditional short-read coverage (30x) with low-coverage long reads (4x) yields benefits that rival a high-coverage long-read pipeline: SNP recall and precision exceeding short reads even at very high coverage (60x), structural variant (SV) recall and precision far exceeding short-read capabilities and approaching full long-read performance, substantial improvements in genotype-aware SV calling precision and recall over short-read pipelines.

Ultimately these gains are a result of the complementary capabilities of short and long read technologies: while short reads can provide many layers of high-accuracy coverage at low cost, they struggle with regions that are ambiguous to map with such a short size. Long reads, on the other hand, excel at mapping to these difficult regions and can also provide high accuracy per-base information, but at a much greater cost per layer. Unsurprisingly, the greatest gains from this approach occur where this complementarity is strongest: long reads help with difficult-to-map regions for SNPs, long reads better capture long insertions and deletions (structural variants), and the many layers of short reads help provide genotyping signal for variants discovered by long reads. By judiciously combining or selecting from these technologies as we have shown here, we can achieve the best of both with only a few layers of long reads augmenting a traditional short-read run.

Finally, we note that this work is necessarily constrained by the current state of sequencing technologies: future short and long read approaches could become more affordable, and improved callers and pipeline could even increase the effectiveness of those as well as existing technologies. Our expectation, though, is that there will always be a place for leveraging the complementarity amongst those future technologies. Thus while the blend-seq of the future may involve different pipelines in different ratios and even different methods of combination, we expect the principles of this approach to remain steadfast.

6: Methods

In this section, we describe in detail the pipelines and settings used for our various analyses.

Short Variant Calling. Blend-seq short variant calls were generated by by combining preprocessed long reads with the Illumina short reads to create a hybrid Binary Alignment Map (BAM). This BAM was used as input to DRAGEN v3.7.8, with the version selected to match what is used for the All of Us Research Program (25). The tool was designed for Illumina short reads alone. Therefore, several preprocessing steps on the long reads were necessary to enable DRAGEN to run. The long read BAMs were first downsampled to 4x coverage with samtools (82) to simulate a lower coverage long read product. Then the following processing steps were performed:

1. In order to remove large scale deletion information which cannot be processed by DRAGEN which is designed for short INDELs alone, we started by converting large deletions in the long read alignments to use "N" in their CIGAR strings rather than "D" via a custom pysam script.
2. GATK's `SplitNCigarReads` (83) was used to split alignments at these large deletions to provide cleanly aligned reads as a technical intermediate to feed into the following custom tool.
3. To simulate a short read structure from the long reads, we wrote a custom `chop-reads` program (84) to chop the alignments evenly at 151 read bases and update the CIGAR strings accordingly. This allows DRAGEN to see reads of equal length regardless of which technology generated the read, while still keeping the alignment information generated from the original long reads.
4. To generate the final hybrid BAM, `samtools` was used to sort (82) and then merge these reads with the DRAGMAP aligned short reads. The header was also updated to just have one read group via `samtools reheader`, to work around a limitation in the DRAGEN version used.
5. Chopped long reads with CIGAR string equal to a pure insertion (i.e. "151I") were removed, since these are not useful for short INDEL calling and cased errors in the DRAGEN pipeline.
6. `samtools` was used to index these hybrid BAMs and then they were run through the full DRAGEN variant calling tool.

For comparison with the hybrid samples, control samples for just long and just short reads were processed using standard SNP and short INDEL calling pipelines for long and short reads respectively. The long reads were aligned using minimap2 (85, 86), and the short reads were aligned using DRAGMAP as a part of the DRAGEN (87) variant calling process, using the GRCh38 human reference in both cases. The long reads (PacBio and ONT) were run through DeepVariant (88) (using v1.3 for PacBio and v1.6 for ONT), while the short reads (Illumina) were run through DRAGEN v3.7.8.

After obtaining the hybrid DRAGEN Variant Call Format files (VCFs) at each desired mixture of coverages, we benchmarked against the NIST HG002 v4.2.1 standard Genomes

in a Bottle (GIAB) truth set (89, 90). We used a pipeline wrapping RTG's `vcfeval` tool (91, 92) which resolves haplotypes underlying variant representations to ensure accurate comparisons. From this, we obtain true positive (TP), false positive (FP), and false negative (FN) labels from which precision, recall, and F1 statistics are derived, including over different subsets of the genome. These BED file regions ("LowMappabilityAll", and others used in this paper) were also curated by NIST for their GIAB release (93).

To create the pie charts demonstrating the breakdown of types of short read evidence (or lacking) among the variants uniquely picked up by the hybrid method, we took that set of SNPs labeled with FN from `vcfeval` comparing the pure short read VCF against the NIST truth, and intersected with those labeled TP from `vcfeval` comparing the hybrid VCF against the NIST truth. Then we restricted to those contained in NIST's low mappability regions, where the largest boost in performance was observed. Each SNP was given a unique label in the following list with higher labels taking precedence, depending on how the evidence in the pure short read BAM looked:

A. Zero Reads: zero depth (DP) at the site;

B. Zero Alt-Supporting Reads: zero short reads supporting the SNP in their alignment;

C. Only MAPQ Zero Reads: all reads above site have mapping quality (MQ) 0;

D. Only MAPQ Zero Alt-Supporting Reads: all reads supporting SNP in alignment have MQ 0;

E. Low Depth: total DP is less than 2;

F. Low Alt-Supporting Depth: total DP among reads supporting SNP is less than 2;

G. Low MAPQ Reads: the average of the mean MQ reference-supporting reads and the mean MQ of the SNP-supporting reads is less than or equal to 20;

H. Low BaseQ Alt-Supporting Reads: the average base qualities of reads supporting the SNP is less than 20;

I. Other: all other SNPs.

This provides a somewhat sorted categorization from harder to call with short reads alone, to easier to call with short reads SNPs.

Structural Variant Discovery with Long Reads. In order to measure structural variant (SV) discovery at various levels of coverage, each set of long reads (PacBio and ONT) were downsampled from 1-10x, and then run through `sniffles` v2.2 (38) to call SVs. Parameters were chosen to be more sensitive at lower coverage. In particular, it was run with settings: `--minsupport 1 --qc-output-all --qc-coverage 1 --long-dup-coverage 1 --detect-large-ins True`.

A custom script was then used to clean the data (`CleanSVs.wdl` from (94)) which, splits multiallelic sites, recomputes the SVLEN annotation (SV Length) to standardized positive values and fill in missing values, strips any filters to maximize recall, and removes events smaller than 50bp which fall under the short INDEL category (see above). Following this step, the cleaned VCF was run through a pipeline wrapping `truvari` v.4.0.0 setting `dup-to-ins` and `passonly` to be true, and using `pctseq` equal to 0 in order to have a more equal comparison with the short read control which do not report basepair resolution of differences in large insertions. The SV VCF was also restricted to events with AC at least 1 using `bcftools view --min-ac 1`, so reference calls were removed before running `truvari`. These were compared to the NIST Q100 v1.0 SV truth set for HG002 (43), which is part of the T2T-Q100 project (44), an initiative attempting to create perfect telomere-to-telomere assemblies (95) and polishing them with new tools like in the Telomere-to-Telomere project (96). The truth VCF was also restricted to events of size at least 50bp, and multiallelic sites were split, in the same way that the comparison VCFs were cleaned, except filters were kept in the truth.

We considered only events with at least 25% overlap with NIST's truth high confidence BED file (labeled `stvar.benchmark.bed`) for our benchmarking statistics. After obtaining TP, FP, and FN labels in this way, we derived precision, recall, and F1 statistics overall, and also when restricting to events with positive percent overlap in particular regions of interest. When considering overlaps with specific regions, we followed the convention that insertions should have reference length given by the length of the inserted sequence, to avoid nearby misses due to ambiguity in the exact starting base, particularly when considering events discovered by short reads. As a control, we compared the performance to a callset curated from GATK-SV (48) calls on an Illumina 30x BAM from (97), processed with the same cleaning and benchmarking pipeline.

Hybrid Genotyping of SVs. Starting with 4x long reads (PacBio and ONT) aligned to the GRCh38 human reference, we created a sample-specific graph reference to map the short reads against using `vg` v1.54.0 (98). The full WDL pipeline is included in the repository described in the Code Availability section. The overview of the pipeline implementation is as follows:

1. Created "trivial SV calls" using long read alignments, at places where the CIGAR string has an insertion or deletion exceeding 50bp, using a custom tool. This creates a VCF of naive SV candidates from the long read alignments themselves, and is deduplicated using `bcftools norm -d` (82).

2. A graph reference was obtained by augmenting the linear GRCh38 human reference using these trivial SV calls, using `vg autoindex -workflow giraffe` (80).

3. The short read FASTQs were split using `seqkit split2` (99, 100) to parallelize the short read alignments.

4. Each short read batch was mapped to the sample-specific graph reference via `vg giraffe` with default parameters.

5. The resulting Graph Alignment Maps (GAMs) were collected and structural variants were called using `vg pack` and `vg call` (101), which is the mode for calling novel SVs rather than regenotyping a specified set.

The output was then run through the same cleaning and benchmarking process as described in the last section. The resulting performance statistics were plotted relative to their short read coverage, and compared to the pure short read and pure long read controls, the latter of which were obtained as in the last section.

6.1. Data Sources. For each technology, we started with a high coverage set of reads and then downsampled to the approximately desired coverage using `samtools view -s`. The initial set of reads for each technology came from the following sources:

- Illumina: We took two replicates of HG002 sequenced at the Broad on a NovaSeq X machine, and merged into a 68x BAM. It was aligned using DRAGMAP v3.7.8. This was first downsampled to 60x, and then from there to all the desired coverages.
- PacBio: We used a 30x HG002 run on a Revio machine sequenced at the Broad for internal validation studies. It was aligned using `pbmm2` (102, 103) with the flags `--preset CCS --sample NA24385_1 --strip --sort --unmapped`.
- ONT: We used the set of reads from (46) which totaled around 14x, and then downsampled from there.
- Ultima: Described in an appendix further, we did some experiments with Ultima reads, starting around 60x and provided by Ultima from a run performed at the Broad.

The read data will be made available on SRA soon.

ACKNOWLEDGEMENTS

This project was funded in part by Microsoft Corporation. The authors would also like to thank Broad's Data Sciences Platform (DSP) for their helpful advice and guidance for this project. In particular, the authors would like to thank Steve Huang, Ryan Lorig-Roach, and James Emery (Broad DSP) for sharing their expertise on long read data, in addition to Mark Walker (Broad DSP) for insights on the short read SV calling pipelines and on structual variants. The authors also would like to thank Kiran Garimella and Chris Kachulis (Broad DSP) for helpful conceptual discussions in early phases of the project.

COMPETING INTERESTS

N.L. has received speaking honoraria from Illumina Inc and is an advisory board member for FYR Diagnostics and Everygene. N.L has received research collaborative funding (for work unrelated to this publication) from Illumina Inc and PacBio Inc.

CODE AVAILABILITY

Custom scripts and code specifically written for this project can be found at: https://github.com/broadinstitute/blend_seq_paper. This includes: the `TrivialCaller` java tool used to translate a BAM into a VCF of variants created from INDEL CIGAR alignments, a python script for converting large deletions in alignments into "N"s in the CIGAR as part of the graph genotyping pipeline, and the full `VgCallSVs` WDL pipeline for calling SVs using a sample-specific graph from long read alignments, as described in the graph genotyping section.

DATA AVAILABILITY

Read data used in this study will be made available on SRA shortly. All other data is from public references and cited appropriately.

Bibliography

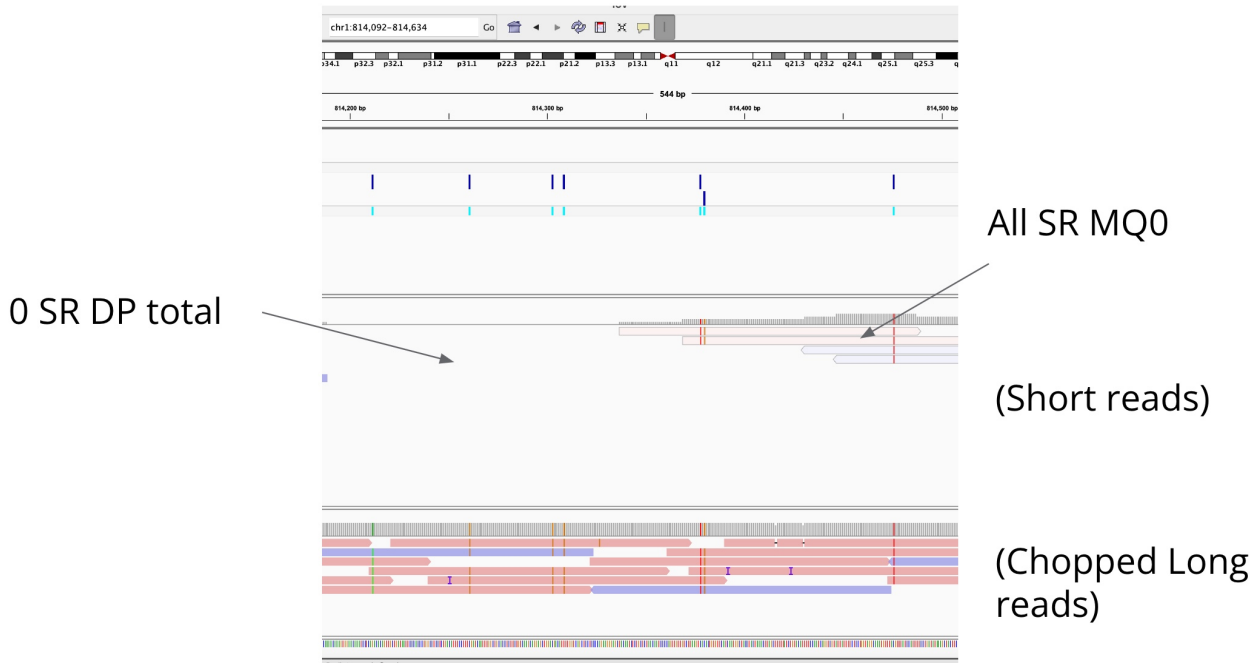
1. Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, and others. An integrated map of structural variation in 2, 504 human genomes. Nature, 526(7571):75–81, September 2015. ISSN 1476-4687. doi: 10.1038/nature15394.
2. Claudia M. B. Carvalho and James R. Lupski. Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics, 17(4):224–238, February 2016. ISSN 1471-0064. doi: 10.1038/nrg.2015.25.
3. Kijong Yi and Young Seok Ju. Patterns and mechanisms of structural variations in human cancer. Experimental amp; Molecular Medicine, 50(8):1–11, August 2018. ISSN 2092-6413. doi: 10.1038/s12276-018-0112-3.
4. Medhat Mahmoud, Nastassia Gobet, Diana Ivette Cruz-Dávalos, and others. Structural variant calling: the long and the short of it. Genome Biology, 20(1), November 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1828-7.
5. Medhat Mahmoud, Yongqing Huang, Kiran Garimella, and others. Utility of long-read sequencing for All of Us. Nature Communications, 15(1):837, 2024.
6. Kez Cleal and Duncan M Baird. Dysgu: efficient structural variant calling using short or long reads. Nucleic Acids Research, 50(9):e53–e53, 2022.
7. Huanjia Zhang. Variantyx combines long nanopore, short illumina reads for rare disease whole-genome diagnostic test, October 2024. https://www.genomeweb.com/sequencing/variantyx-combines-long-nanopore-short-illumina-reads-rare-disease-whole-genome. Accessed: 2024-10-30.
8. Ludmila Kaplun, Greice Krautz-Peterson, Nir Neerman, and others. Ont long-read wgs for variant discovery and orthogonal confirmation of short read wgs derived genetic variants in clinical genetic testing. Frontiers in Genetics, 14, April 2023. doi: 10.3389/fgene.2023.1145285.
9. Dmitry Antipov, Anton Korobeynikov, Jeffrey S McLean, and Pavel A Pevzner. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics, 32(7):1009–1015, 2016.
10. Xian Fan, Mark Chaisson, Luay Nakhleh, and Ken Chen. HySA: a hybrid structural variant assembly approach using next-generation and single-molecule sequencing technologies. Genome Research, 27(5):793–800, 2017.
11. Jason R Miller, Peng Zhou, Joann Mudge, and others. Hybrid assembly with long and short reads improves discovery of gene family expansions. BMC Genomics, 18:1–12, 2017.
12. Alaina Shumate, Brandon Wong, Geo Pertea, and Mihaela Pertea. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Computational Biology, 18(6):e1009730, 2022.
13. Ryan R Wick, Louise M Judd, Claire L Gorrie, and Kathryn E Holt. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Computational Biology, 13(6):e1005595, 2017.
14. Denis Bertrand, Jim Shaw, Manesh Kalathiyappan, and others. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nature Biotechnology, 37(8):937–944, 2019.
15. Aleksey V Zimin, Guillaume Marçais, Daniela Puiu, and others. The MaSuRCA genome assembler. Bioinformatics, 29(21):2669–2677, 2013.
16. Grace XY Zheng, Billy T Lau, Michael Schnall-Levin, and others. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nature Biotechnology, 34(3):303–311, 2016.
17. Neil I Weisenfeld, Vijay Kumar, Preyas Shah, Deanna M Church, and David B Jaffe. Direct determination of diploid genome sequences. Genome Research, 27(5):757–767, 2017.
18. Karen HY Wong, Michal Levy-Sakin, and Pui-Yan Kwok. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. Nature Communications, 9(1):3040, 2018.
19. Li C Xia, John M Bell, Christina Wood-Bouwens, and others. Identification of large rearrangements in cancer genomes with barcode linked reads. Nucleic Acids Research, 46(4):e19–e19, 2018.
20. Patrick Marks, Sarah Garcia, Alvaro Martinez Barrio, and others. Resolving the full spectrum of human genome variation using Linked-Reads. Genome Research, 29(4):635–645, 2019.
21. Li Fang, Charlly Kao, Michael V Gonzalez, and others. LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data. Nature Communications, 10(1):5585, 2019.
22. Jiang Du, Robert D Bjornson, Zhengdong D Zhang, and others. Integrating sequencing technologies in personal genomics: optimal low cost reconstruction of structural variants. PLoS Computational Biology, 5(7):e1000432, 2009.
23. Victoria Popic, Chris Rohlicek, Fabio Cunial, and others. Cue: a deep-learning framework for structural variant discovery and genotyping. Nature Methods, 20(4):559–568, 2023.

24. Amit Goyal, Hyuk Jung Kwon, Kichan Lee, and others. Ultra-fast next generation human genome sequencing data processing using DRAGEN™ bio-IT processor for precision medicine. Open Journal of Genetics, 7(1):9–19, 2017.
25. Alexander G. Bick, Ginger A. Metcalf, Kelsey R. Mayo, and others. Genomic data in the All of Us Research Program. Nature, 627(8003):340–346, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06957-x.
26. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18): 3094–3100, 2018.
27. Justin M Zook, Jennifer McDaniel, Nathan D Olson, and others. An open resource for accurately benchmarking small variant and reference calls. Nature Biotechnology, 37(5): 561–566, 2019.
28. Peter Krusche, Len Trigg, Paul C Boutros, and others. Best practices for benchmarking germline small-variant calls in human genomes. Nature Biotechnology, 37(5):555–560, 2019.
29. James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, and others. Integrative genomics viewer. Nature Biotechnology, 29(1):24–26, 2011.
30. Peter Ebert, Peter A Audano, Qihui Zhu, and others. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science, 372(6537):eabf7117, 2021.
31. Mark JP Chaisson, Ashley D Sanders, Xuefang Zhao, and others. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications, 10 (1):1784, 2019.
32. Xuefang Zhao, Ryan L Collins, Wan-Ping Lee, and others. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. The American Journal of Human Genetics, 108(5):919–928, 2021.
33. Boris Noyvert, A Mesut Erzurumluoglu, Dmitriy Drichel, and others. Imputation of structural variants using a multi-ancestry long-read sequencing panel enables identification of disease associations. medRxiv, pages 2023–12, 2023.
34. Jonas A Gustafson, Sophia B Gibson, Nikhita Damaraju, and others. Nanopore sequencing of 1000 Genomes Project samples to build a comprehensive catalog of human genetic variation. medRxiv, 2024.
35. Siegfried Schloissnig, Samarendra Pani, Bernardo Rodriguez-Martin, and others. Long-read sequencing and structural variant characterization in 1,019 samples from the 1000 Genomes Project. bioRxiv, pages 2024–04, 2024.
36. Doruk Beyter, Helga Ingimundardottir, Asmundur Oddsson, and others. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. Nature Genetics, 53(6):779–786, 2021.
37. Peter A Audano, Arvis Sulovari, Tina A Graves-Lindsay, and others. Characterizing the major structural variant alleles of the human genome. Cell, 176(3):663–675, 2019.
38. Moritz Smolka, Luis F Paulin, Christopher M Grochowski, and others. Detection of mosaic and population-level structural variants with Sniffles2. Nature Biotechnology, January 2024.
39. Tao Jiang, Yongzhuang Liu, Yue Jiang, and others. Long-read-based human genomic structural variation detection with cuteSV. Genome Biology, 21:1–24, 2020.
40. David Heller and Martin Vingron. SVIM: structural variant identification using mapped long reads. Bioinformatics, 35(17):2907–2915, 2019.
41. Pacific Biosciences. pbsv, GitHub repository, 2023. github.com/PacificBiosciences/pbsv.
42. Adam C English, Vipin K Menon, Richard A Gibbs, Ginger A Metcalf, and Fritz J Sedlazeck. Truvari: refined structural variant comparison preserves allelic diversity. Genome Biology, 23(1):271, 2022.
43. Nathan D. Olson Justin M. Zook. 2024-02-15 GIAB HG002 GRCh38 assembly-based small and structural variants draft benchmark sets, 2024. https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_HG002_DraftBenchmark_defrabbV0.015-20240215/. Accessed: 2024-08-08.
44. T2T Consortium. Telomere-to-telomere consortium HG002 "Q100" project, 2024. https://github.com/marbl/HG002. Accessed: 2024-08-08.
45. Ann M Mc Cartney, Kishwar Shafin, Michael Alonge, and others. Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. Nature Methods, 19(6):687–695, 2022.
46. Ting Wang, Lucinda Antonacci-Fulton, Kerstin Howe, and others. The Human Pangenome Project: a global resource to map genomic diversity. Nature, 604(7906):437–446, April 2022.
47. Zan Koenig, Mary T Yohannes, Lethukuthula L Nkambule, and others. A harmonized public resource of deeply sequenced diverse human genomes. Genome Research, 2024.
48. Ryan L Collins, Harrison Brand, Konrad J Karczewski, and others. A structural variation reference for medical and population genetics. Nature, 581(7809):444–451, May 2020.
49. Xiaoyu Chen, Ole Schulz-Trieglaff, Richard Shaw, and others. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics, 32(8):1220–1222, 2016.
50. Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, and others. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Research, 40(9):e69–e69, 2012.
51. Mehrtash Babadi, Jack M Fu, Samuel K Lee, and others. GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. Nature Genetics, 55 (9):1589–1597, 2023.
52. Eugene J Gardner, Vincent K Lam, Daniel N Harris, and others. The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Research, 27(11):1916–1929, 2017.
53. William T Harvey, Peter Ebert, Jana Ebler, and others. Whole-genome long-read sequencing downsampling and its effect on variant-calling precision and recall. Genome Research, 33(12):2029–2040, 2023.
54. Christopher T Saunders, James M Holt, Daniel N Baker, and others. Sawfish: Improving long-read structural variant discovery and genotyping with local haplotype modeling. August 2024. doi: 10.1101/2024.08.19.608674.
55. Xiongbin Kang, Jialu Xu, Xiao Luo, and Alexander Schönhuth. Hybrid-hybrid correction of errors in long reads with HERO. Genome Biology, 24(1):275, 2023.
56. Joshua Casey Darian, Ritu Kundu, Ramesh Rajaby, and Wing-Kin Sung. Constructing telomere-to-telomere diploid genome by polishing haploid nanopore-based assembly. Nature Methods, 21(4):574–583, 2024.
57. Jouni Sirén, Parsa Eskandar, Matteo Tommaso Ungaro, and others. Personalized pangenome references. Nature Methods, pages 1–7, 2024.
58. Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. Genome Biology, 19:1–16, 2018.
59. Chirag Jain, Neda Tavakoli, and Srinivas Aluru. A variant selection framework for genome graphs. Bioinformatics, 37(Supplement_1):i460–i467, 2021.
60. Neda Tavakoli, Daniel Gibney, and Srinivas Aluru. Haplotype-aware variant selection for genome graphs. In Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pages 1–9, 2022.
61. Neda Tavakoli, Daniel Gibney, and Srinivas Aluru. GraphSlimmer: Preserving read mappability with the minimum number of variants. Journal of Computational Biology, 31(7): 616–637, 2024.
62. Erik Garrison, Jouni Sirén, Adam M Novak, and others. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology, 36 (9):875–879, 2018.
63. Wen-Wei Liao, Mobin Asri, Jana Ebler, and others. A draft human pangenome reference. Nature, 617(7960):312–324, 2023.
64. Glennis A Logsdon, Peter Ebert, Peter A Audano, and others. Complex genetic variation in nearly complete human genomes. bioRxiv, pages 2024–09, 2024.
65. Heng Li, Jonathan M. Bloom, Yossi Farjoun, and others. A synthetic-diploid benchmark for accurate variant-calling evaluation. Nature Methods, 15(8):595–597, July 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0054-7.
66. Mikhail Kolmogorov. hapdiff, GitHub repository, 2021. github.com/KolmogorovLab/hapdiff.
67. Hannes P Eggertsson, Snaedis Kristmundsdottir, Doruk Beyter, and others. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nature Commununications, 10(1):5402, November 2019.
68. Sai Chen, Peter Krusche, Egor Dolzhenko, and others. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biology, 20(1):291, December 2019.
69. Tobias Rausch, Thomas Zichner, Andreas Schlattl, and others. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics, 28(18):i333–i339, 2012.
70. Sandra Romain and Claire Lemaitre. SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph. Bioinformatics, 39(Supplement_1):i270–i278, 2023.
71. Chirag Jain, Arang Rhie, Nancy F Hansen, Sergey Koren, and Adam M Phillippy. Long-read mapping to repetitive reference sequences using Winnowmap2. Nature Methods, 19 (6):705–710, 2022.
72. Fritz J Sedlazeck, Philipp Rescheneder, Moritz Smolka, and others. Accurate detection of complex structural variations using single-molecule sequencing. Nature Methods, 15(6): 461–468, 2018.
73. Jingwen Ren and Mark JP Chaisson. lra: A long read aligner for sequences and contigs. PLOS Computational Biology, 17(6):e1009078, 2021.
74. Adrian Tan, Gonçalo R Abecasis, and Hyun Min Kang. Unified representation of genetic variants. Bioinformatics, 31(13):2202–2204, July 2015.
75. Sairam Behera, Severine Catreux, Massimiliano Rossi, and others. Comprehensive genome analysis and variant detection at scale using dragen. Nature Biotechnology, October 2024. ISSN 1546-1696. doi: 10.1038/s41587-024-02382-1.
76. Derek Caetano-Anolles. Full release of open-source dragen-gatk 1.0, 2024. https://gatk.broadinstitute.org/hc/en-us/articles/4411716682011-Full-release-of-open-source-DRAGEN-GATK-1-0. Accessed: 2024-10-31.
77. Tim Dunn and Satish Narayanasamy. vcfdist: accurately benchmarking phased small variant calls in human genomes. Nature Communications, 14(1):8149, 2023.
78. James M Holt, Christopher T Saunders, William J Rowell, and others. HiPhase: jointly phasing small, structural, and tandem repeat variants from hifi sequencing. Bioinformatics, 40(2):btae042, 2024.
79. Justin Wagner, Nathan D Olson, Lindsay Harris, and others. Curated variation benchmarks for challenging medically relevant autosomal genes. Nature Biotechnology, 40(5):672–680, 2022.
80. Jouni Sirén, Jean Monlong, Xian Chang, and others. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science, 374(6574):abg8871, December 2021.
81. Brynja D. Sigurpalsdottir, Olafur A. Stefansson, Guillaume Holley, and others. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. Genome Biology, 25(1), March 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03207-9.
82. Petr Danecek, James K Bonfield, Jennifer Liddle, and others. Twelve years of SAMtools and BCFtools. Gigascience, 10(2), February 2021.
83. Aaron McKenna, Matthew Hanna, Eric Banks, and others. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9):1297–1303, September 2010.
84. Ricky Magner. chop_reads, 2024. https://github.com/rickymagner/chop_reads. Accessed: 2024-08-08.
85. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18): 3094–3100, September 2018.
86. Heng Li. New strategies to improve minimap2 alignment accuracy. Bioinformatics, 37(23): 4572–4574, December 2021.
87. Sairam Behera, Severine Catreux, Massimiliano Rossi, and others. Comprehensive and accurate genome analysis at scale using DRAGEN accelerated algorithms. bioRxiv, January 2024. Accessed: 2024-08-08.

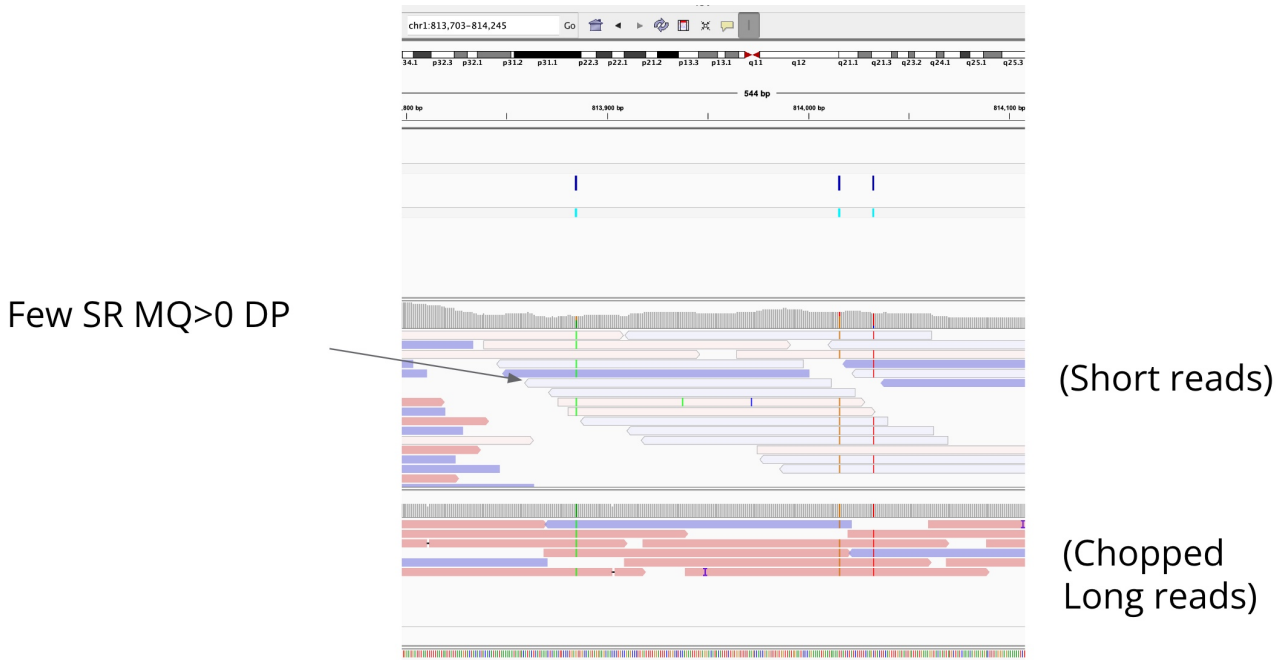
88. Ryan Poplin, Pi-Chuan Chang, David Alexander, and others. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology, 36(10):983–987, November 2018.
89. Justin Wagner, Nathan D Olson, Lindsay Harris, and others. Benchmarking challenging small variants with linked and long reads. July 2020.
90. NIST. Nist genomes in a bottle release v4.2.1, 2024. https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/. Accessed: 2024-08-08.
91. John G Cleary, Ross Braithwaite, Kurt Gaastra, and others. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. August 2015.
92. Real Time Genomics. rtg-tools, 2024. https://github.com/RealTimeGenomics/rtg-tools. Accessed: 2024-08-08.
93. Nathan Dwarshuis, Divya Kalra, Jennifer McDaniel, and others. The GIAB genomic stratifications resource for human reference genomes. Accessed: 2024-08-08, October 2023.
94. Broad Data Sciences Platform Methods hydro.gen Team. palantir-workflows, 2024. https://github.com/broadinstitute/palantir-workflows. Accessed: 2024-08-08.
95. Sergey Nurk, Sergey Koren, Arang Rhie, and others. The complete sequence of a human genome. Science, 376(6588):44–53, April 2022.
96. Mikko Rautiainen, Sergey Nurk, Brian P Walenz, and others. Telomere-to-telomere assembly of diploid chromosomes with verkko. Nature Biotechnology, 41(10):1474–1482, October 2023.
97. Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, and others. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. Cell, 185(18):3426–3440.e19, September 2022.
98. Erik Garrison, Jouni Sirén, Adam M Novak, and others. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nature Biotechnology, 36(9):875–879, October 2018.
99. Wei Shen, Shuai Le, Yan Li, and Fuquan Hu. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One, 11(10):e0163962, October 2016.
100. Wei Shen, Botond Sipos, and Liuyang Zhao. SeqKit2: A Swiss army knife for sequence and alignment processing. Imeta, 3(3):e191, June 2024.
101. Glenn Hickey, David Heller, Jean Monlong, and others. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biology, 21(1):35, February 2020.
102. Ting Hon, Kristin Mars, Greg Young, and others. Highly accurate long-read HiFi sequencing data for five complex genomes. Sci. Data, 7(1):399, November 2020.
103. Armin Töpfer et al. pbmm2, GitHub repository, 2024. https://github.com/PacificBiosciences/pbmm2.

A: Extra Figures and Tables

This section contains a collection of extra figures referenced in the main text.



Appendix A Supplementary Fig. 1. This IGV screenshot demonstrates an example of a 0-DP-SR SNP (left) and a MQ0-SR-ALL SNP (right). The top set of reads are short reads, and the bottom are PacBio. The shading demonstrates that the long reads have good mapping quality, and we see by the mismatch colorings they can correctly pick up the actual SNPs around this locus.

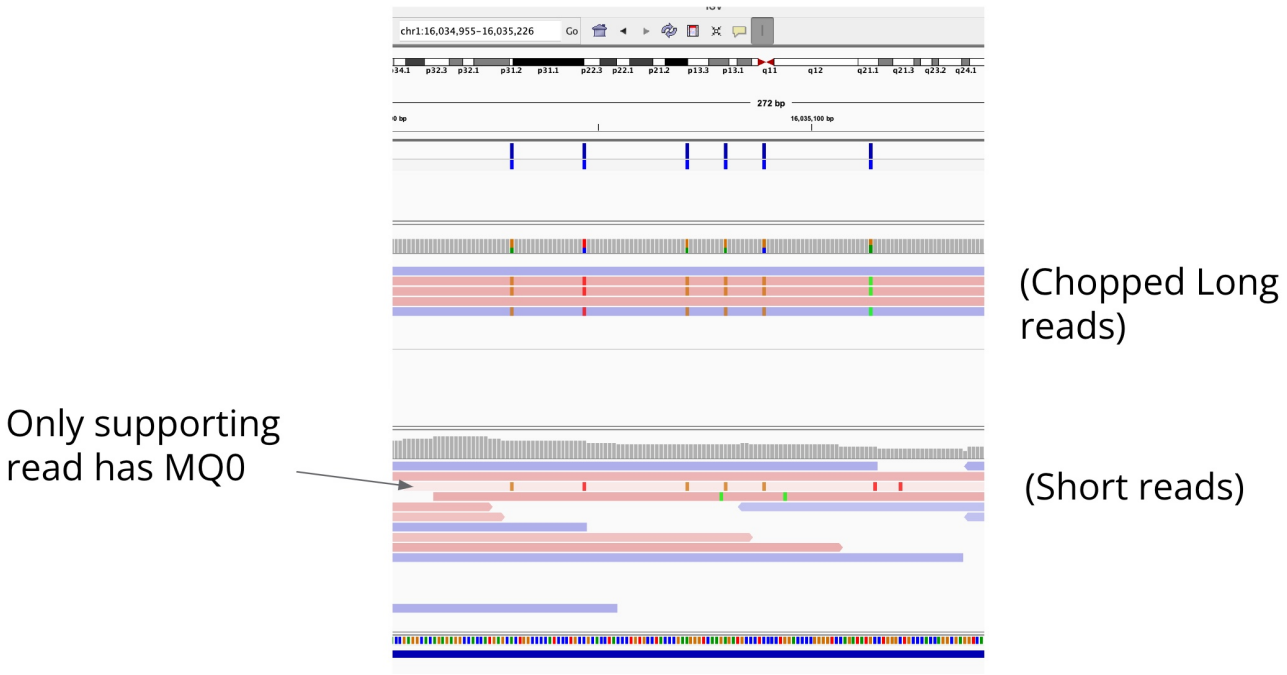


Appendix A Supplementary Fig. 2. This IGV screenshot includes an example of a LOWMQ-SR-ALL SNP, as we see the short reads (top) have mostly mapping quality 0 here. The long reads (bottom) have no trouble aligning here with good mapping quality, and provide majority evidence for the SNP call in the hybrid.

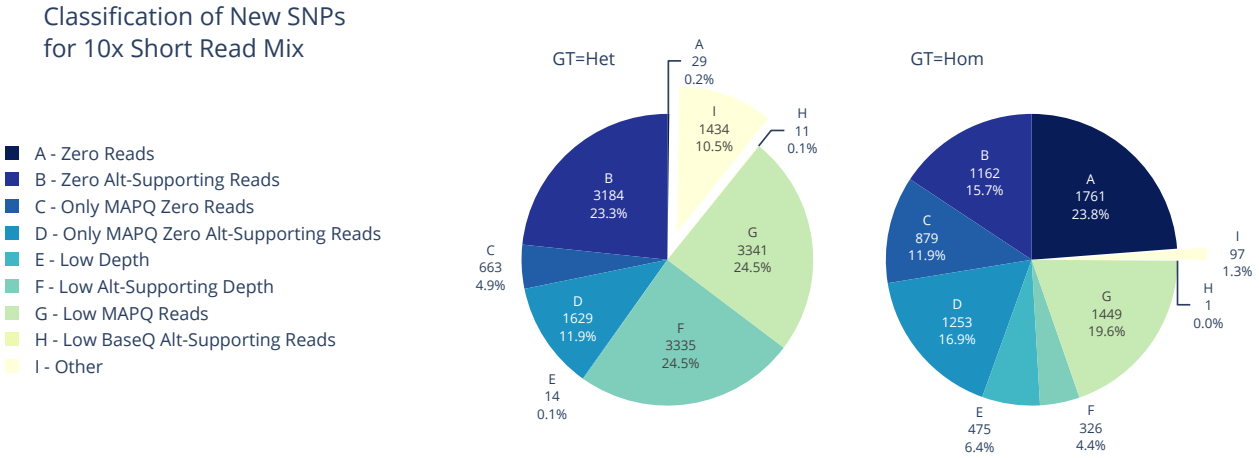
A.1. Extra Materials for Short Variant Calling Performance.

A.2. Extra Materials for Structural Variant Discovery.

A.3. Extra Materials for Genotype-Aware Structural Variant Discovery.



Appendix A Supplementary Fig. 3. This IGV screenshot provides an example of a MQ0-SR-SUPP SNP, as the short reads (bottom) are split between mapping quality 0 and good mapping quality for the SNP-supporting and reference-supporting alignments respectively. We presume both the higher-than-usual concentration of SNPs in the sample plus low mappability of the region overall contribute to making the short read alignment for the SNP-supporting read zero. The long reads (top) have no trouble aligning here, even for the reads with a cluster of (real) SNPs.

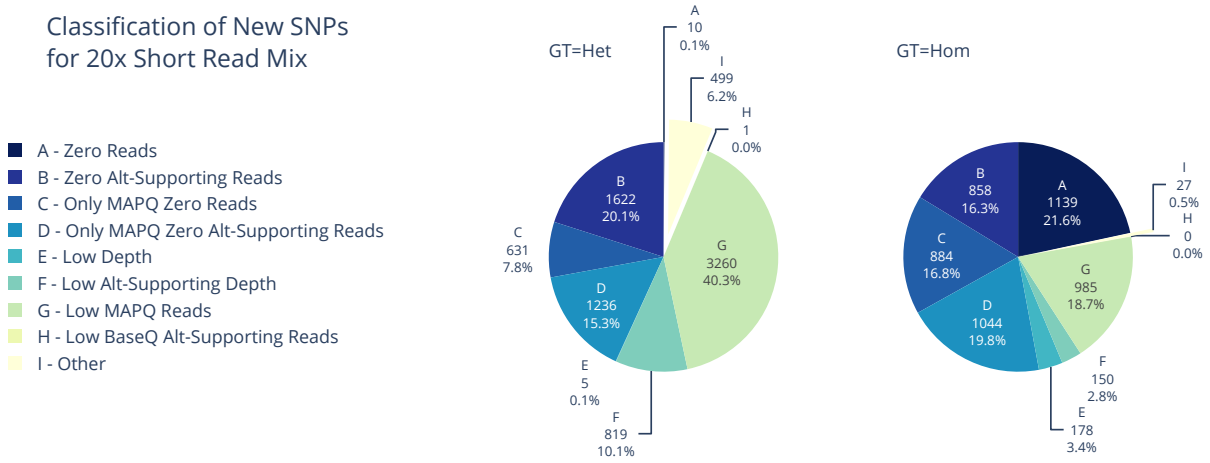


Appendix A Supplementary Fig. 4. Here is the breakdown of Short-FN+Hybrid-TP SNPs for 10x short reads + 4x PacBio long reads. We still see a large chunk of SNPs gained from the hybrid in low mappability regions are explained by lack of short read evidence (40% heterozygous and 2/3 of the homozygous sites).

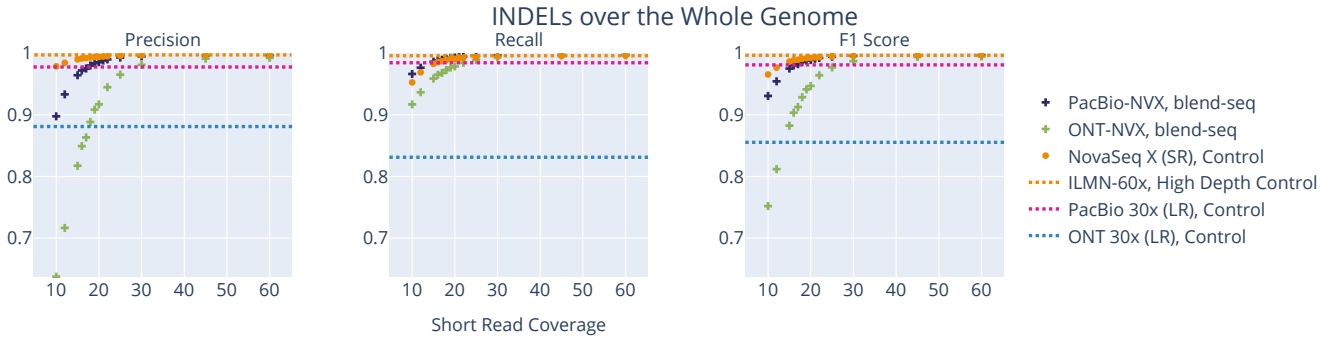
B: Ultima

We performed a similar investigation using Ultima short reads, for the hybrid SV genotyping experiments. For this, we took a 60x collection of Ultima reads for HG002, and downsampled to the same depths as the Illumina case plus 45x and 60x. We record the performance curves for genotyping SVs discovered using PacBio and ONT (Appendix B Fig. 1), manufactured the same way as the Illumina hybrids. We can investigate the same trends as before, also looking at the GT-aware versions of the performance statistics (Appendix B Fig. 2).

We see a similar pattern when restricting to easier regions (Appendix B Fig. 3), and when breaking down by SV type (Appendix B Fig. 4, Appendix B Fig. 5).



Appendix A Supplementary Fig. 5. Here is the breakdown of Short-FN+Hybrid-TP SNPs for 20x short reads + 4x PacBio long reads. We still see a large chunk of SNPs gained from the hybrid in low mappability regions are explained by lack of short read evidence (43% heterozygous and 75% homozygous).

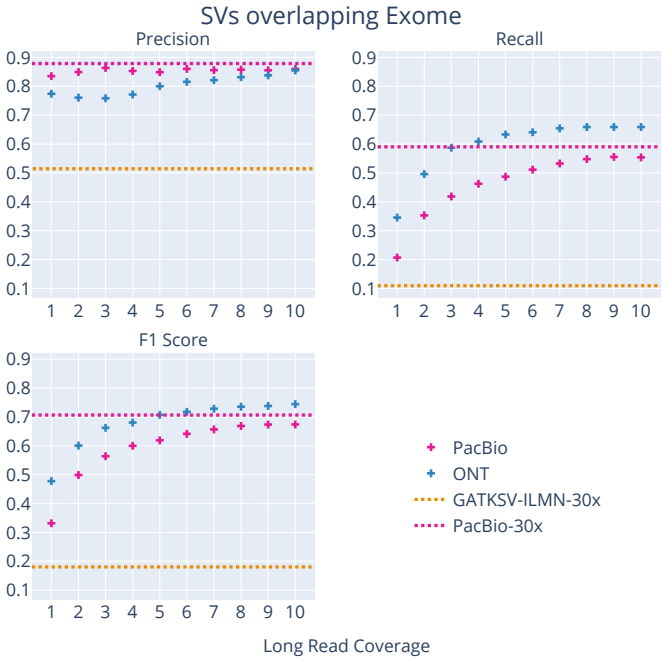


Appendix A Supplementary Fig. 6. The performance for INDELs over the whole genome looks about even for the hybrids vs the pure short reads, with ONT alone standing out with poor performance as a control.

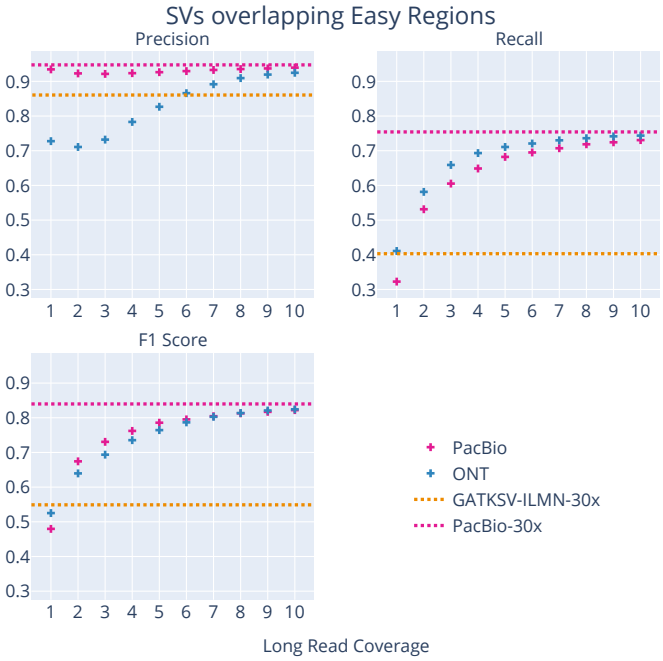
INDEL Performance by Region

	WholeGenome			LowMappability			Exome		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Illumina 15x (control)	.9826	.9900	.9863	.8258	.9168	.8689	.9813	.9659	.9735
Illumina 30x (control)	.9946	.9964	.9955	.8593	.9354	.8957	.9938	.9779	.9858
Illumina 60x (high depth)	.9959	**.9972**	**.9965**	.8647	.9364	.8991	.9958	**.9780**	**.9868**
blend-seq 15x IL + 4x ONT	.9588	.8173	.8824	.9171	.5952	.7218	.9875	.4681	.6351
blend-seq 15x IL + 4x PB	.9861	.9643	.9751	.9197	.8368	.8763	.9917	.8602	.9213
blend-seq 30x IL + 4x ONT	.9932	.9811	.9871	.9248	.7253	.8130	.9938	.8665	.9258
blend-seq 30x IL + 4x PB	**.9965**	.9949	.9957	.9241	.8889	.9062	.9938	.9625	.9779
30x PB (high depth)	.9845	.9776	.9810	**.9811**	**.9869**	**.9840**	**.9979**	.9273	.9613

Table 7. The blend-seq methods outperform the short read methods in recall overall, especially in hard-to-map regions. Precision lags a bit behind, possibly due to no hybrid filtering methods being applied to the results. Bold values are max values in column and shaded cells are those exceeding all control values in column (first three rows).



Appendix A Supplementary Fig. 7. The SV discovery performance boost for the pure long reads is maintained when restricting to events overlapping the exome, with the ONT reads having a slightly better recall than the PacBio reads.

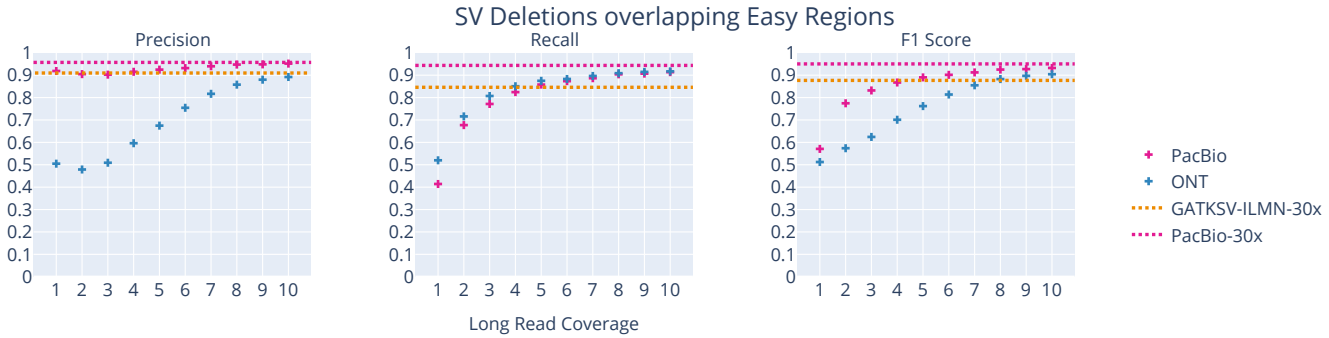


Appendix A Supplementary Fig. 8. When restricting to the easy regions, we give the short reads the best region for them to perform well, and still see long reads outperforming even at very low depth.

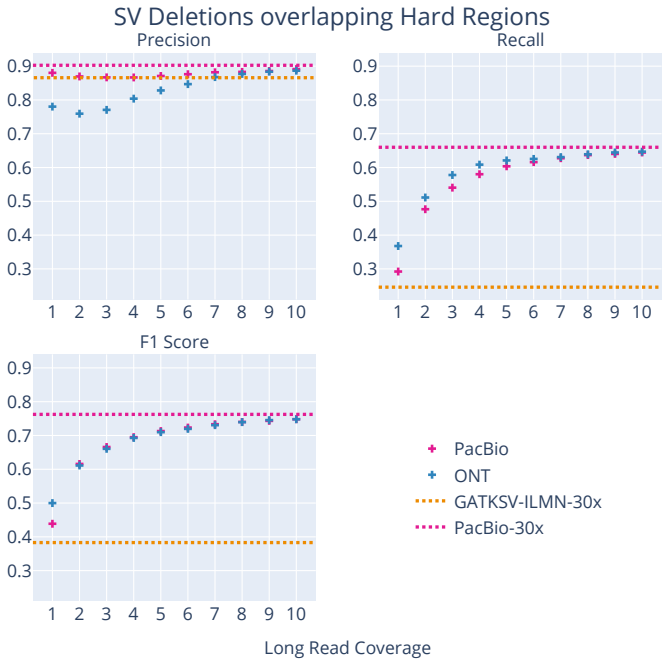
SV Deletion Discovery Performance by Region

	Whole Genome			Easy Regions			Hard Regions			Exome		
	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score	Recall	Precision	F1 Score
Illumina 30x (control)	.2700	**.8762**	.4128	.8460	.9097	**.8767**	.2459	.8658	.3830	.3000	.5745	.3942
PacBio 3x	.5493	.8602	.6705	.7715	.9024	.8319	.5405	.8669	.6658	.4875	.7692	.5968
PacBio 4x	.5889	.8632	.7002	.8244	**.9152**	.8674	.5802	.8667	.6951	.5375	**.7857**	.6383
ONT 6x	.6339	.8169	.7139	.8831	.7544	.8137	.6255	.8468	.7196	.7125	.6463	.6778
ONT 7x	**.6397**	.8497	**.7299**	**.8966**	.8168	.8548	**.6308**	**.8676**	**.7305**	**.7250**	.7297	**.7274**

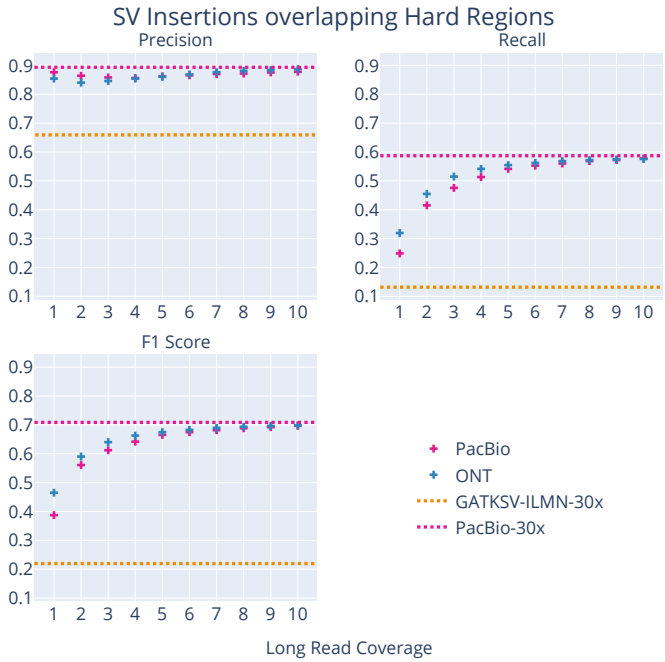
Table 8. With just a few layers, pure long reads vastly outperform short reads for SV discovery in terms of recall. This is powered by extra discovery coming from more difficult regions, although the short reads perform well in the Easy Regions. Bold values are max values in column and shaded cells are those exceeding the control value in column (first row).



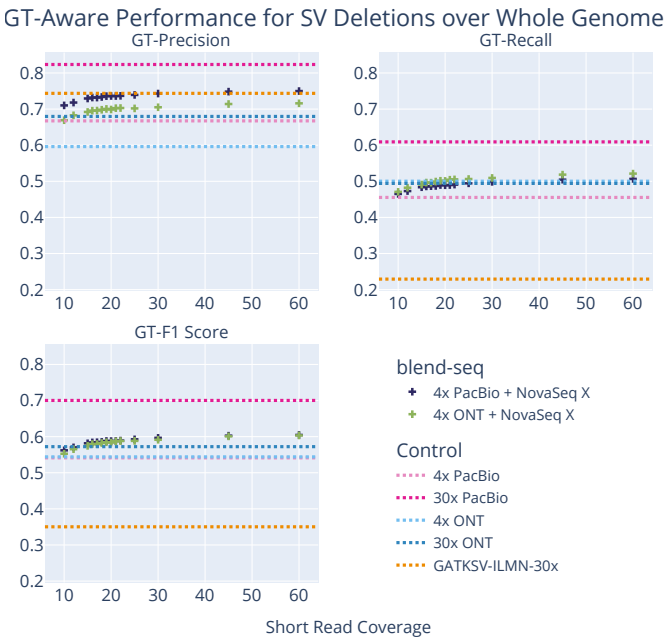
Appendix A Supplementary Fig. 9. In the easiest possible regime for short reads (overlapping the easy regions and just considering large deletions), we see that ONT precision is the worst, but the PacBio precision still beats out short reads alone. The recall is also better for long reads at just around 4-5x depth.



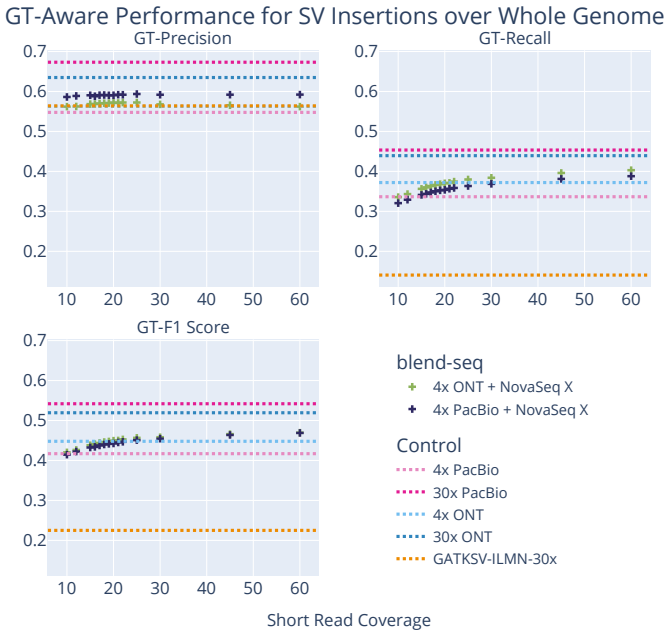
Appendix A Supplementary Fig. 10. For events overlapping difficult regions, the long reads perform best when measured by F1 even for deletions. The precision is best with PacBio, followed by Illumina, and then ONT, but recall is much higher for long reads even at 1x coverage.



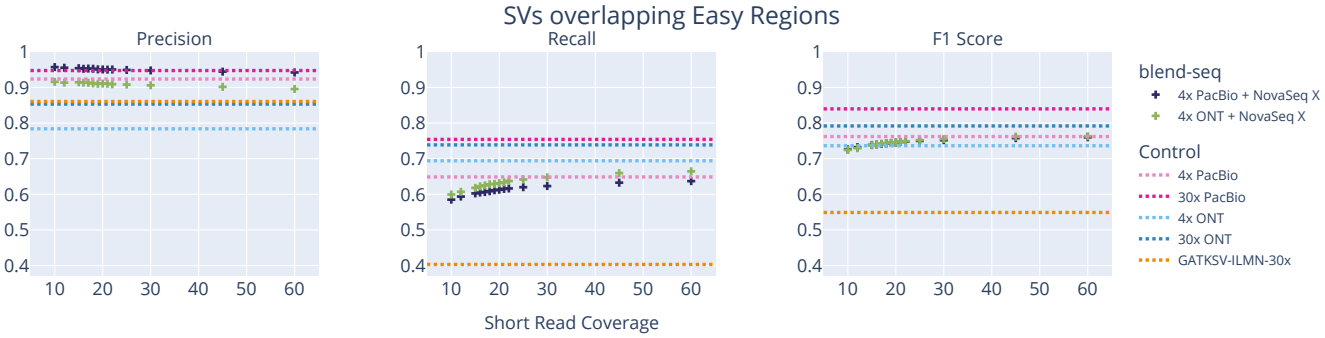
Appendix A Supplementary Fig. 11. Insertions in difficult regions are the hardest for short reads to discover, and we observe long read discovery methods outperforming in every statistical category.



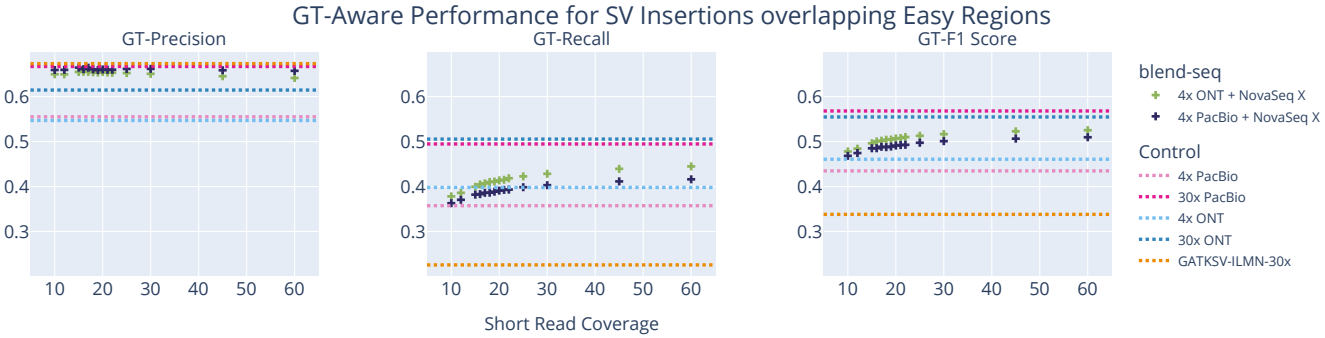
Appendix A Supplementary Fig. 12. GT-aware performance for SV deletions offers a modest improvement for blend-seq compared to pure long reads when using at least 15x short reads.



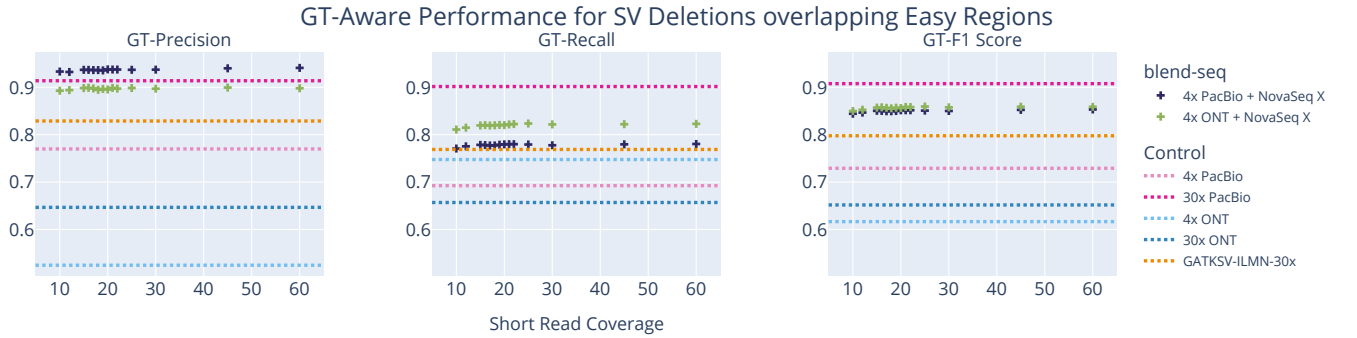
Appendix A Supplementary Fig. 13. GT-aware performance for SV insertions is slightly better for blend-seq when using at least 30x short reads.



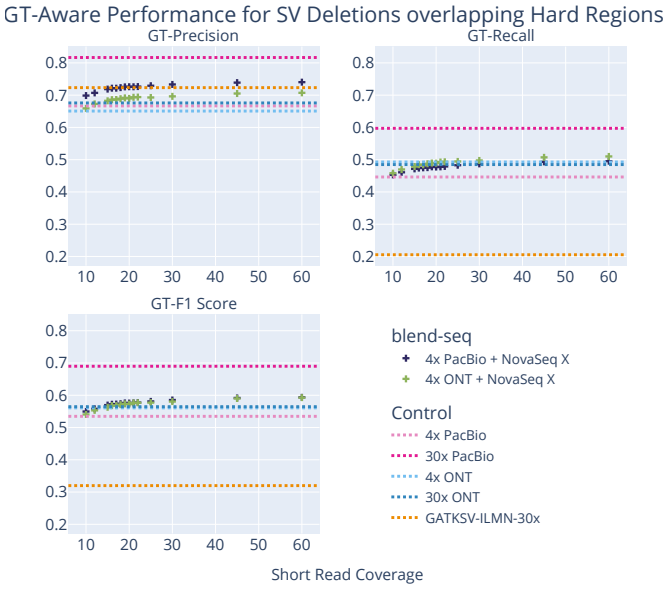
Appendix A Supplementary Fig. 14. We see little to no improvement in the SV discovery performance for the hybrid methods over the pure long read when considering only allele-level matching.



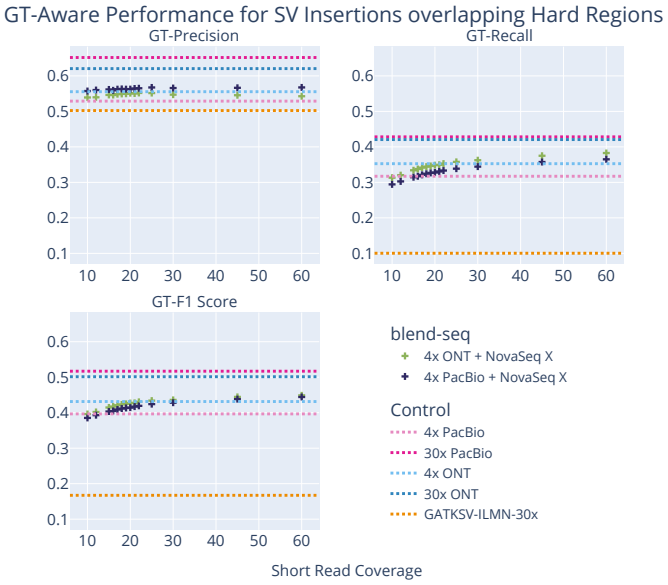
Appendix A Supplementary Fig. 15. When considering GT-aware SV discovery performance over the easy regions but for SV insertions, we still see a boost in precision relative to the pure long reads, but recall benefits are minor at lower short read depths.



Appendix A Supplementary Fig. 16. Restricting further to GT-aware performance for large deletions in the easy regions, the hybrid boost on GT-aware F1 is even more pronounced, again beating out both of the pure control alternatives. In this set-up, the hybrid curves also have higher GT-aware recall at all levels of short read depth.

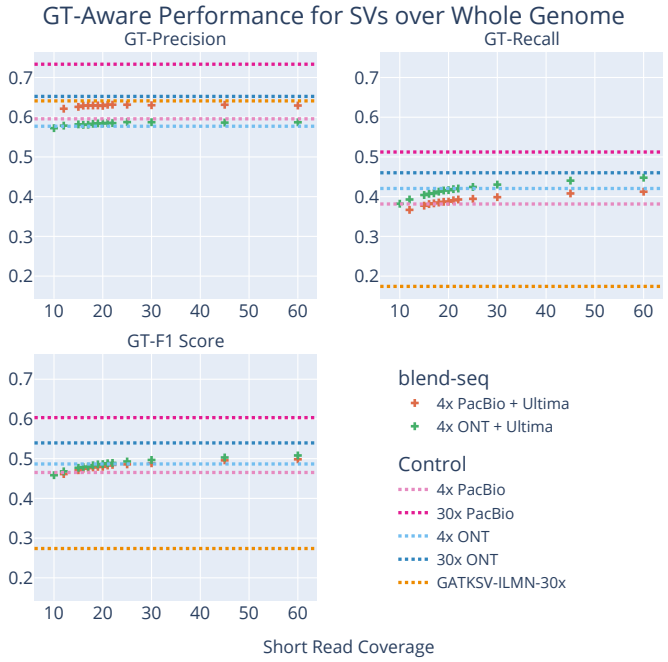


Appendix A Supplementary Fig. 17. Over hard regions, GT-aware performance of the hybrid method is on par with long reads alone, even for the easier mode of just restricting to deletions.

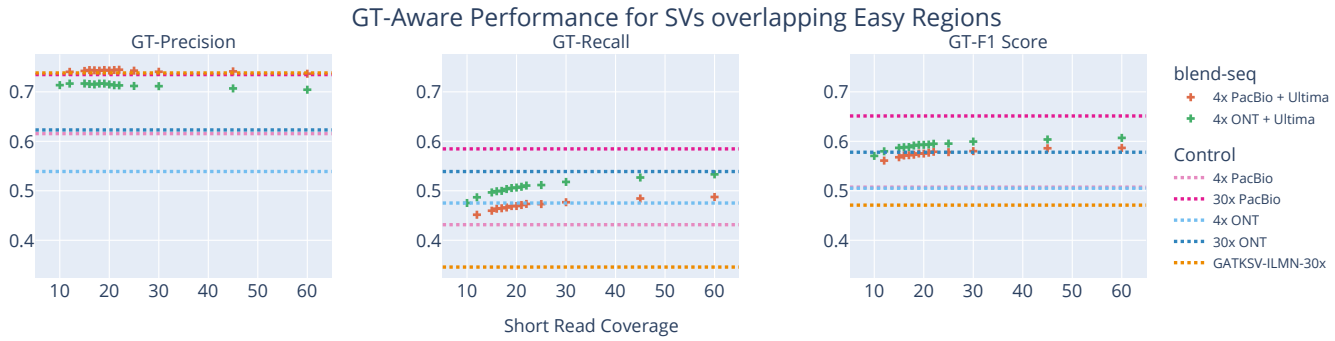


Appendix A Supplementary Fig. 18. Over hard regions, GT-aware performance of the hybrid method is on par with long reads alone, when restricting to just insertions.

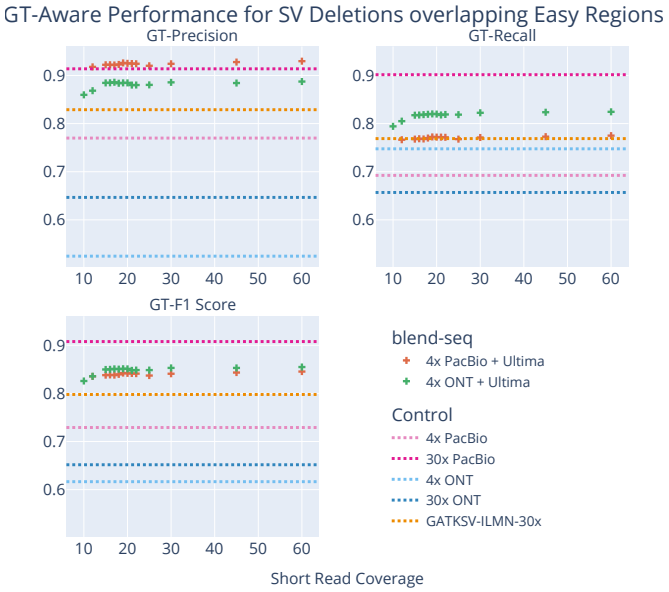
Appendix B Supplementary Fig. 1. For just allele-matching performance, we see the Ultima reads don't help for SV discovery beyond what the long reads already have.



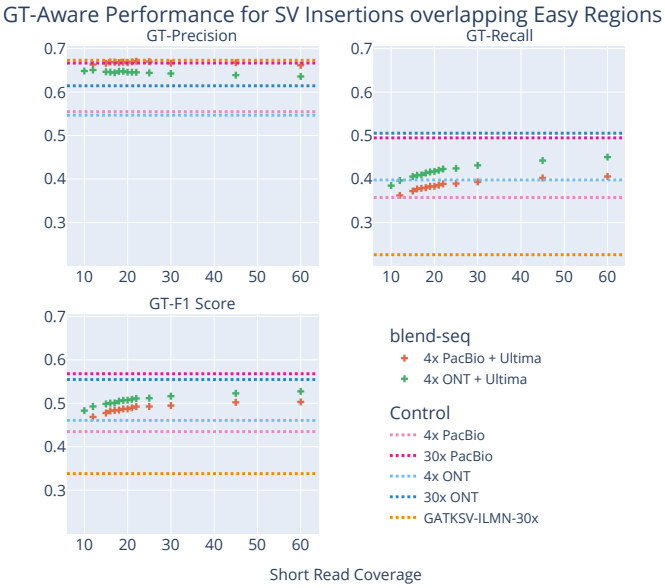
Appendix B Supplementary Fig. 2. When considering GT-aware performance, we see a boost in precision by adding the Ultima reads to the PacBio long reads, but the recall gains are less significant, and only appear at higher levels of Ultima depth.



Appendix B Supplementary Fig. 3. Over the easy regions, the PacBio hybrid outperforms in GT-precision when compared to both the GATK-SV short read control and the pure long read controls. This translates into a similar performance improvement in F1, with moderate GT-recall boosts for the hybrids over their respective long read controls.



Appendix B Supplementary Fig. 4. For SV deletions, which are typically easier for short reads, the story of an F1 and precision jump for GT-aware performance holds, and even the recall for the ONT hybrid increases at the lowest level of Ultima read depth.



Appendix B Supplementary Fig. 5. For SV insertions, we still see the GT-precision bump relative to the pure long reads in favor of the Ultima hybrid, but the GT-recall is less dramatic. Still, the GT-F1 Score comes out in favor of the hybrids, even at low Ultima read depth.