# 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. In addition, we demonstrate how the long reads can better phase these variants together, incorporating long-context information in the genome to substantially outperform phasing with short reads alone. 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 large 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 | phasing | 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 structural variant calling for single samples as well as phasing 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 matters, we show even greater gains by combining short and long reads in a graph-based approach. For phasing, we show performance rivaling or exceeding full coverage (30x) PacBio pipelines.

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.

Beyond the problem of finding variants, for diploid organisms there is also the need for phasing – determining which variants occur on the same vs. the opposite chromosome. For instance, correctly determining the haplotype of a gene can

be essential in determining the potential phenotypic ramifications of the observed variants. Estimating phasing relies both on high precision in calling variants and on having overlapping spans of reads covering multiple variants. While short reads excel at the former problem, the latter makes phasing particularly challenging for short reads, whereas long reads have much greater capability to span multiple variants.

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 improved phasing of these variants. 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 phasing, we leverage the benefits of blend-seq for calling variants, then use the long reads alone to bridge the variants together, combining the strengths of both to substantially exceed the performance of short reads alone.

In terms of related work, we note there have been some significant prior investigations on combining short and long reads. However, our work is 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. (16–22) and references therein). 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 did not investigate or compare to simulated read mixes as was done in (23).

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. We also investigate 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. The last section looks at phasing, and shows how the combination of short reads with a few layers of long reads dramatically increases the length of phased blocks throughout the genome.

**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 (24)). 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 (25) employed in several large-scale projects, including All of Us (26); in particular, we used DRAGMAP to map the reads.

For the long reads, we used four layers (4x) of PacBio Revio: this level of coverage was selected based on the trade-offs between coverage and discovery performance for structural variants described in Section 2.2. The analysis involved first mapping the long reads using minimap2 (27), 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 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 (28). In addition to calculating whole-genome precision and recall, we also collected these over some subregions of interest as determined by NIST (29).

***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 (29); 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) (30). We observed (Appendix A Fig. 1) 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 SNPs (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,

## 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).

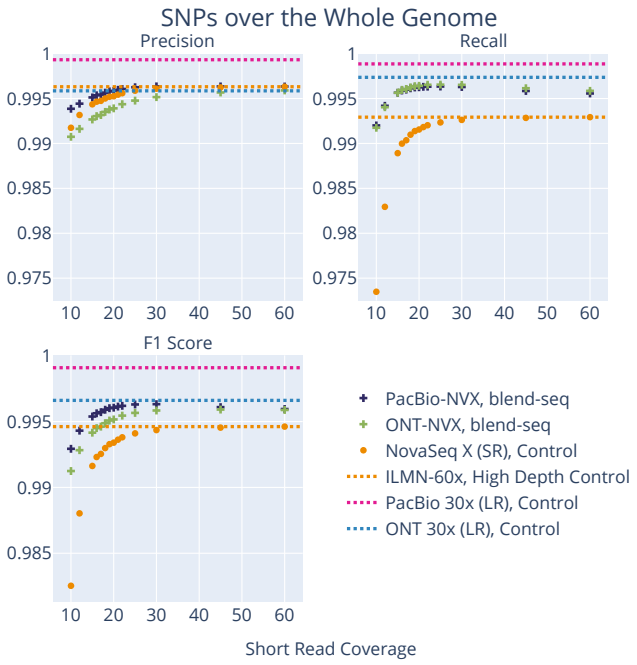



**Fig. 1.** SNP discovery performance over the whole genome. The PacBio-Illumina hybrid at 4x+30x shows greater recall and precision of SNPs compared to short reads alone, even at 60x coverage; the ONT-Illumina hybrid shows greater recall but slightly lower precision. Note that even at the 4x+12x setting,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

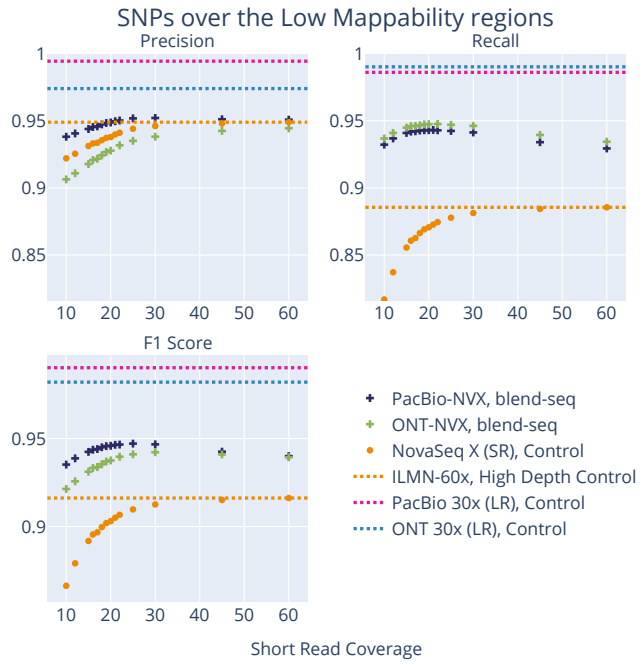



**Fig. 2.** SNP discovery performance over low mappability regions. When restricting to low-mappability regions, the boost in hybrid performance is more pronounced. Even 4x+10x long+short read mixtures outperform the 60x short reads in recall, leading to a net F1 increase for both the PacBio and ONT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

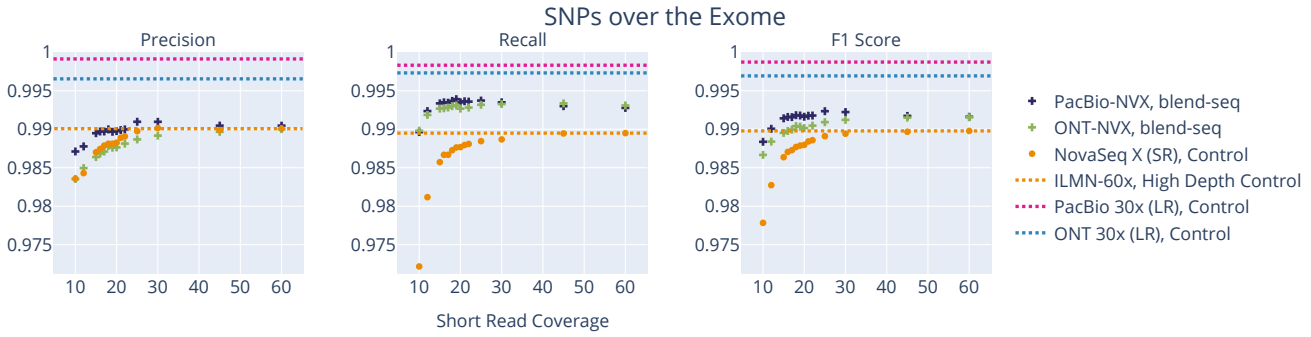



**Fig. 3.** SNP discovery over the exome. When restricting to the exome, blend-seq shows similar improvements in SNP performance as for the entire genome, with the PacBio-Illumina hybrid at 4x+30x outperforming recall and precision of short reads alone, even at 60x coverage.

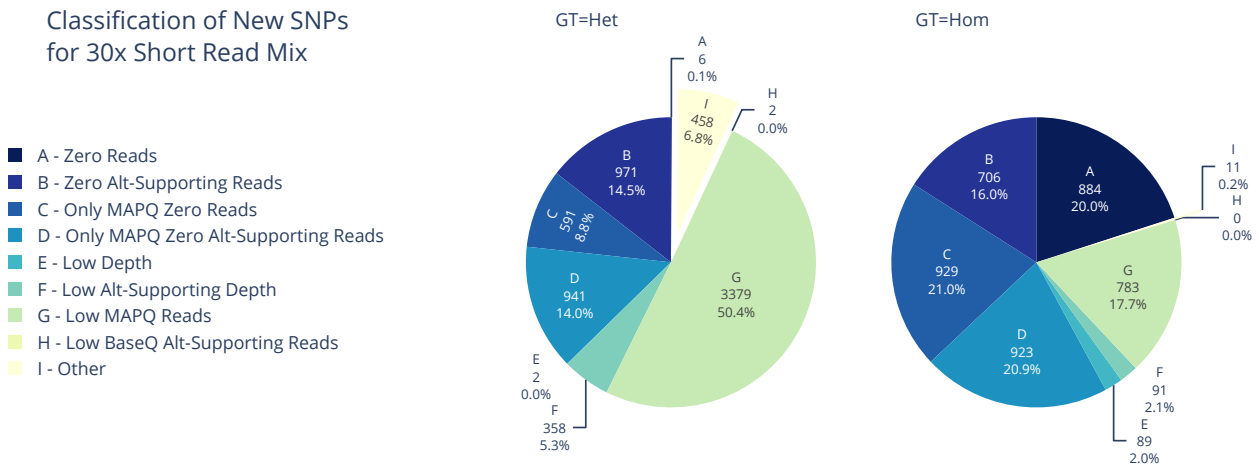



**Fig. 4.** A breakdown of the additional SNPs found by blend-seq in low-mappability regions. For the 30+4x Illumina-PacBio mix, we take the SNPs in low-mappability regions that were missed by short reads alone but discovered by blend-seq, then categorize them according to the nature of evidence supporting 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.

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 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 6 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. (31–38)). 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.

Lastly, we considered 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 applied two SV short-read genotypers, which use sample-specific graph references built from the low-coverage long reads, to better map the short reads and to perform regenotyping from the graph. We show that over easy regions, SV genotyping accuracy improves using these methods.

***Approach and Data.*** We investigated the performance of SV detection at different levels of long-read coverage, focusing on structural variants (insertions and deletions with length $\geq$50bp) in the HG002 sample discovered using Sniffles (39) and allowing even a single read of evidence to support a call (results from other state-of-the-art SV callers, e.g. (40–42), displayed similar trends and are omitted for brevity). The outputs were cleaned and benchmarked with Truvari (43) against the NIST-Q100 SV V1.1 truth (44, 45), a set

of calls derived from the highest-quality long-read assembly of HG002 available to date (46). 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 (47), 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 (48). This callset is the result of performing multi-sample SV calling with GATK-SV (49) 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 (50)), cn.MOPS and GATK-gCNV for copy-number variants (51, 52), and MELT (53) 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 investigations in this area. First, long read downsampling analyses for SV discovery have been performed in the past (see e.g. (54)), 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 mapping algorithm on SV calling require the systematic empirical investigation we present here. 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.

After generating a high-quality SV callset, we wanted to explore whether we can improve *genotype-aware* SV-calling performance by integrating short-read information. To do this, for each long-read technology we applied two short-read genotypers which employ a personalized graph reference to try and improve the genotypes of the discovered variants. The first of these is the vg toolkit (57), which we used to create an augmented graph reference from the sample-specific SV calls produced by Sniffles as described above. Afterwards, the short reads were mapped to this graph, and genotypes were updated based on the pileup information. The second genotyper is Paragraph (58), a tool by Illumina that maps the short reads to a linear reference, builds a local directed acyclic graph for each SV call in isolation, maps to the graph only the reads that mapped to a flanking region in the linear reference, and genotypes each SV breakpoint. In the above analysis, 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 for these genotyping experiments. 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 (59–63). 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.

After regenotyping the SVs, 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 (43). 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 is the output used in reporting SV performance using the long reads alone. The tool also reports TP matches *with* genotype information forced to match (43). 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."

***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

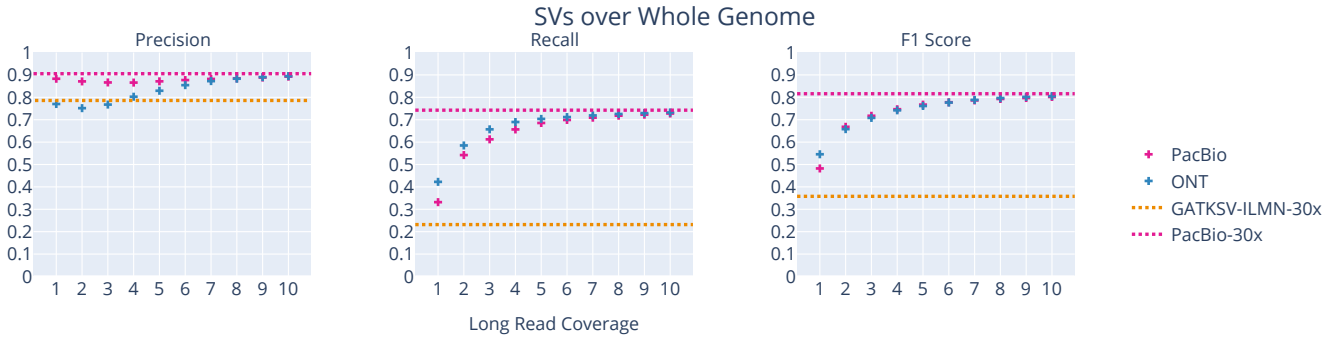



**Fig. 5.** Structural variant discovery with long reads vs. short reads. Long reads on their own, even at very low coverage, substantially outperform state-of-the-art short reads in both precision and recall. The GATKSV-ILMN-30x line represents the performance of short-read SV pipelines with Novaseq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

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 (29)), 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. 11; easy regions: Appendix A Fig. 12).

***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 there is a drop in precision for the ONT reads since short reads already 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.

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 yield better precision for smaller events even at extremely low coverage, and the ONT recall outperforms the PacBio reads for larger events, surpassing the PacBio 30x recall even with just 3x of ONT reads. In both cases, long reads outperform the short read control.

***Genotyping SVs with Short Reads.*** For the genotype-aware setting, we initially did not see any improvements overall in using the short-read genotypers on the SV calls (Appendix A Fig. 23), even when restricting to GT-aware statistics (Appendix A Fig. 24). However, we investigated performance metrics across different regions, especially where the short reads may have an advantage over the long reads with their extra depth without facing mapping difficulty. 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. This setting reduces the total 30k variants in our truth dataset to 11.5k overlapping or contained within this region (29). We see that the GT-aware statistics for blend-seq outperform the short-read and 4x long-read controls overall (Fig. 10).

Furthermore, this trend holds even when breaking down by SV deletions and SV insertions. In the SV deletion case (Fig. 11), the blend-seq mixtures achieve near perfect GT-precision when using Paragraph, beating even the high-depth long-read controls, and resulting in a GT-F1 Score that is almost on par with 30x PacBio. For SV insertions (Fig. 12), the short reads help both with GT-Precision and GT-Recall, outperforming all short-read and lower-depth long-read controls.

We remark that even outside the GT-aware setting, the short reads are useful in filtering out false positive SV calls made by the low-depth long reads, as evidenced in Appendix A Fig. 25 for deletions and Appendix A Fig. 26 for insertions, where precision increases even above the 30x PacBio control. For SV deletions, this comes with similar recall when compared to the low-depth long-read controls, making this method a net boost to SV deletion performance in easy regions. For SV insertions, the increased precision comes with lower recall, making this method only more useful where precision is more important than recall, such as some diagnostic settings.

**2.3. Phasing.** In this subsection, we demonstrate the benefits of blend-seq for phasing short variants (SNPs and IN-

## 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) | .2316 | .7860 | .3578 | .7205 | .8633 | .7855 | .1749 | .7611 | .2844 | .2168 | .5349 | .3085 |
| ONT 1x | .4225 | .7702 | .5457 | .5319 | .7255 | .6138 | .4087 | .8240 | .5464 | .4126 | .7740 | .5382 |
| PacBio 1x | .3318 | .8830 | .4824 | .4218 | .9332 | .5810 | .3201 | .8789 | .4693 | .2517 | .8395 | .3873 |
| ONT 4x | .6897 | .8026 | .7419 | .8787 | .7808 | .8268 | .6639 | .8381 | .7409 | **.7692** | .7714 | **.7703** |
| PacBio 4x | .6566 | .8658 | .7468 | .8332 | .9211 | .8750 | .6330 | .8612 | .7296 | .5315 | .8596 | .6568 |
| PacBio 30x (high depth) | **.7426** | **.9055** | **.8160** | **.9553** | **.9464** | **.9508** | **.7140** | **.8983** | **.7956** | .6643 | **.8819** | .7578 |

**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).

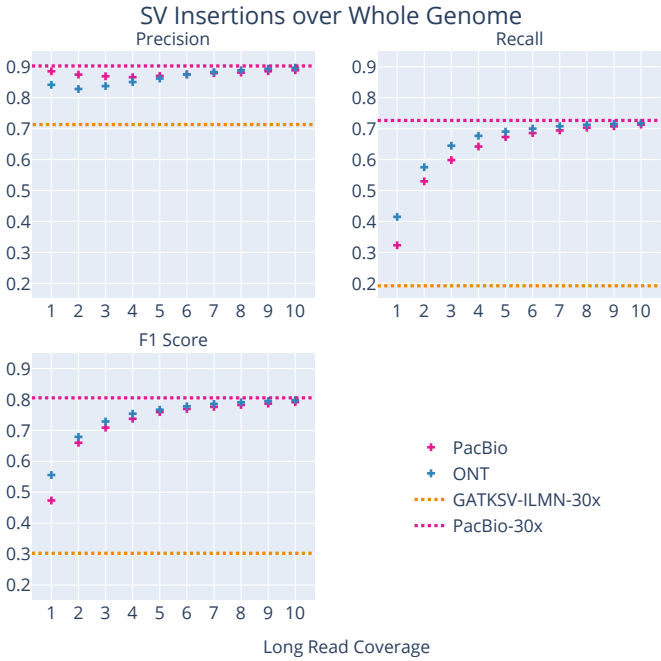



**Fig. 6.** Performance on insertion structural variants for short reads and long reads. Long reads outperform short reads even at very low coverage in this setting; this is because large insertions are difficult for short reads to resolve as they cannot be mapped to the reference. Note that these comparisons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

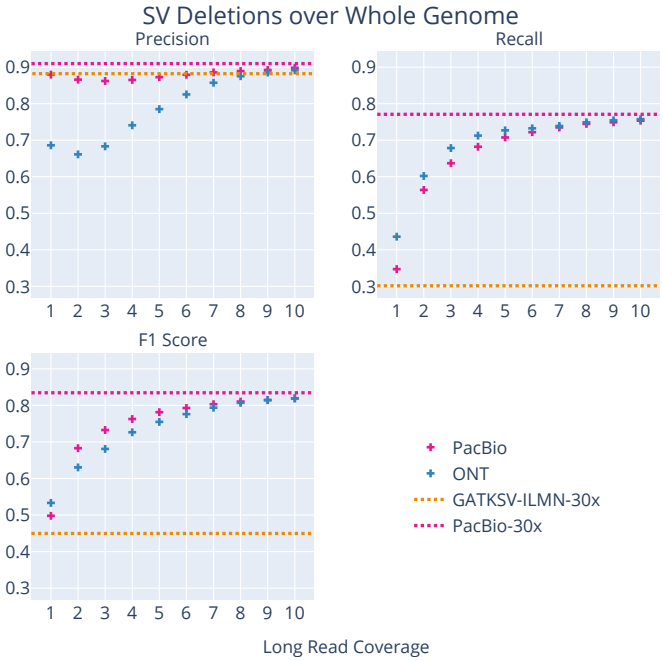



**Fig. 7.** Performance on deletion structural variants for short reads and long reads. While PacBio long reads at 4x outperform short reads substantially in terms of recall while matching precision, we note that short-read precision rivals that of long reads. This is because deletions are easier for short reads to resolve than insertions, as there is no novel sequence 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) | .1733 | .8120 | .2857 | .2371 | .7493 | .3602 |
| ONT 1x | .3987 | .7640 | .5240 | .5278 | .8745 | .6583 |
| PacBio 1x | .3193 | .8854 | .4694 | .2418 | .8735 | .3787 |
| ONT 4x | .6498 | .7992 | .7168 | **.8805** | .8304 | .8547 |
| PacBio 4x | .6317 | .8681 | .7313 | .6322 | **.8864** | .7380 |
| PacBio 30x (high depth) | **.6997** | **.9101** | **.7912** | .8307 | .8848 | **.8569** |

**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).

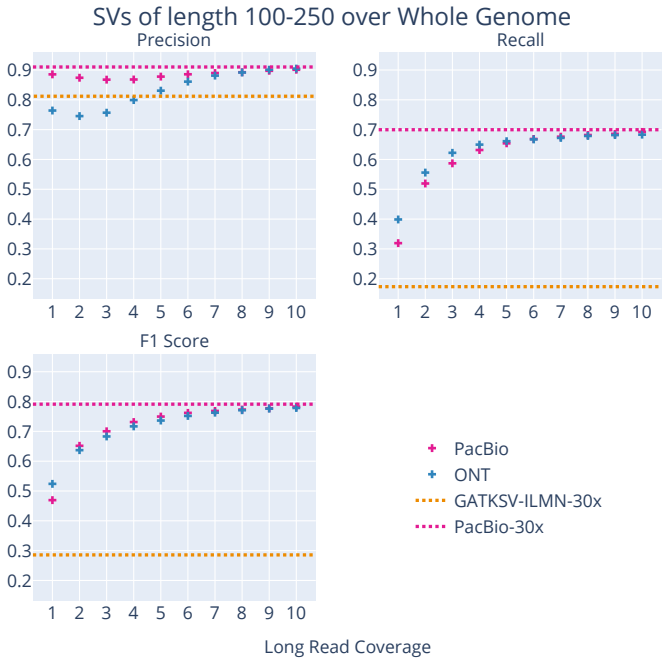



**Fig. 8.** Performance on 100-250 bp structural variants for short and long reads. These are the most common type of SVs; as such, performance trends largely follow the earlier results for all SV.

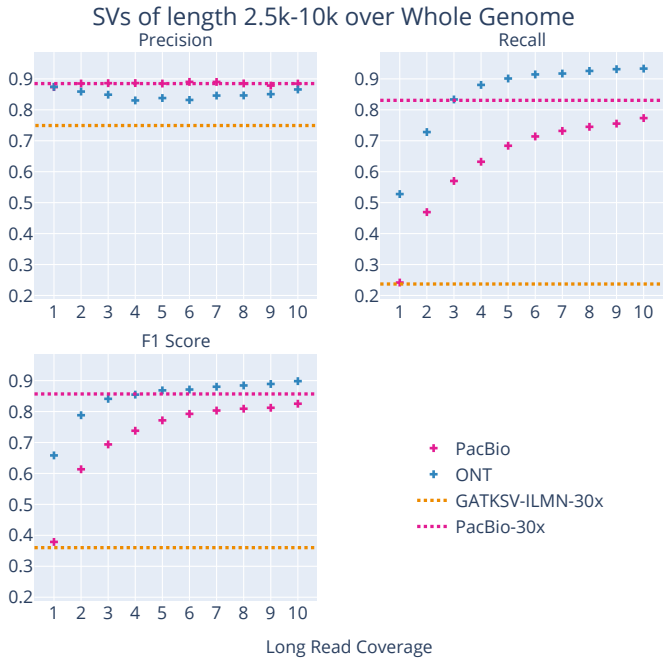



**Fig. 9.** Performance on 2.5k-10k bp structural variants for short and long reads. When restricting to these longer events, we see the ONT Ultra Long reads have better recall than PacBio.

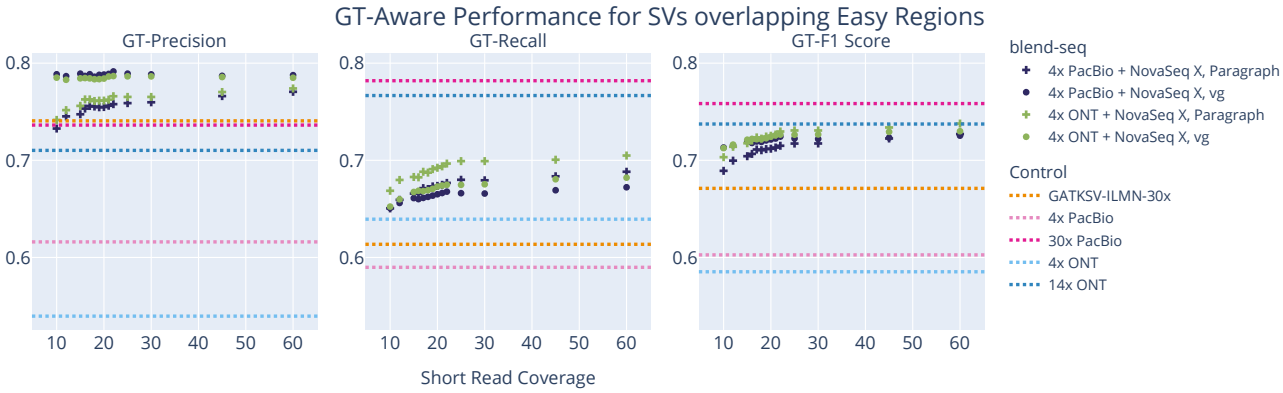



**Fig. 10.** GT-aware SV discovery on easier regions with long reads alone, short reads alone, and our hybrid method. We see the GT-aware statistics are an improvement over the long-read and short-read controls, with GT-Precision exceeding even the high-depth long-reads' performances. This results in the blend-seq experiments having better GT-F1 Scores compared to all controls powered also by an increase in GT-Recall.

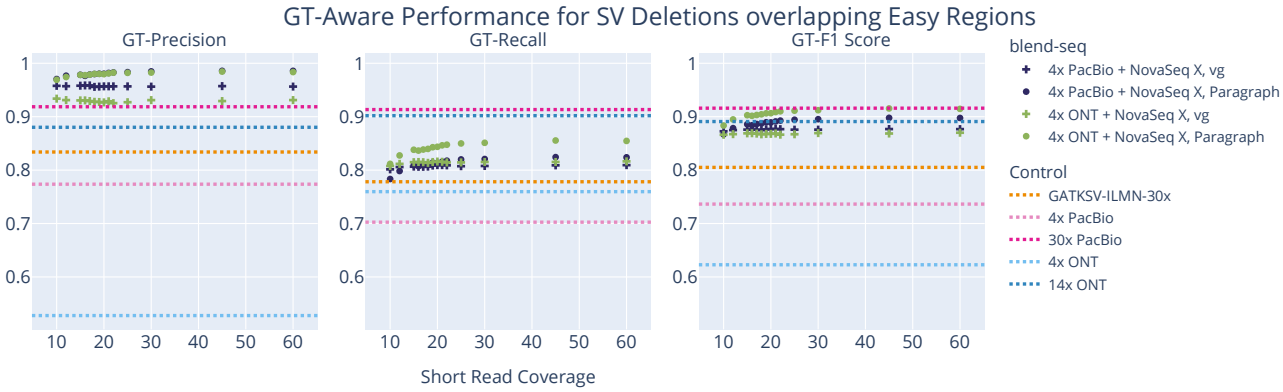



**Fig. 11.** GT-aware SV discovery for deletions on easier regions with long reads alone, short reads alone, and our hybrid method. On SV deletions, we see that Paragraph acheives near perfect GT-Precision, while also increasing GT-Recall. This puts the GT-F1 Score on par even with the 30x PacBio performance.

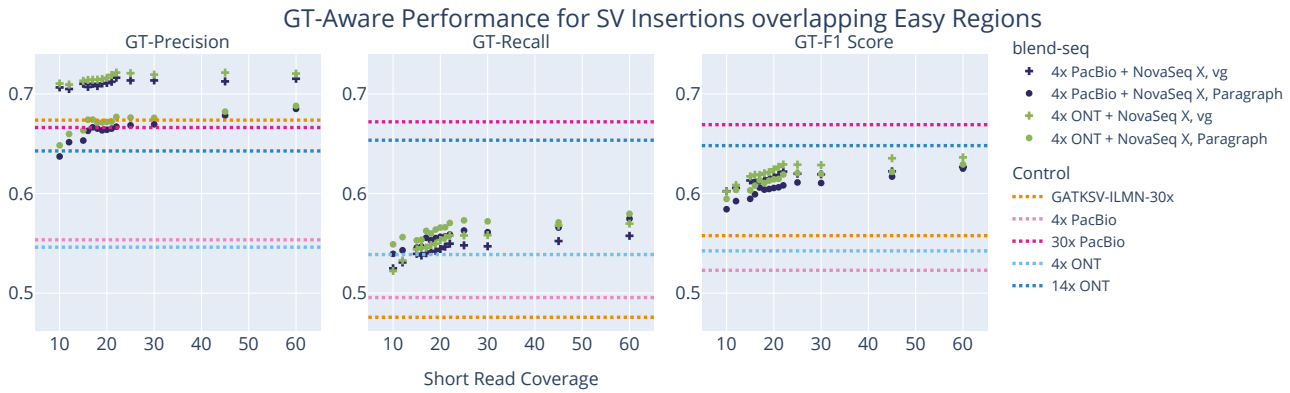



**Fig. 12.** GT-aware SV discovery for insertions on easier regions with long reads alone, short reads alone, and our hybrid method. On SV insertions, there is a bit of precision-recall trade-off happening between the genotypers, where Paragraph improves GT-Precision further at a greater cost to GT-Recall compared to vg. In both cases, the GT-F1 Scores exceed the short-read and long-read controls.

## GT-Aware Performance for SV Deletions 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) | .2548 | .7477 | .3801 | .7782 | .8339 | .8051 | .2292 | .7275 | .3486 |
| ONT 4x (LR-only) | .5518 | .5998 | .5748 | .7597 | .5277 | .6228 | .5454 | .6556 | .5955 |
| PacBio 4x (LR-only) | .5037 | .6706 | .5753 | .7025 | .7736 | .7364 | .4959 | .6699 | .5699 |
| blend-seq ONT 4x + Illumina 10x (vg) | .4607 | .7846 | .5805 | .8077 | .9338 | .8662 | .4458 | .7745 | .5659 |
| blend-seq ONT 4x + Illumina 10x (Paragraph) | .4560 | .7916 | .5786 | .8117 | .9691 | .8834 | .4403 | .7791 | .5626 |
| blend-seq PacBio 4x + Illumina 10x (vg) | .4988 | .7800 | .6085 | .8014 | .9577 | .8726 | .4860 | .7692 | .5957 |
| blend-seq PacBio 4x + Illumina 10x (Paragraph) | .4539 | .7823 | .5745 | .7833 | .9704 | .8669 | .4389 | .7695 | .5590 |
| blend-seq ONT 4x + Illumina 30x (vg) | .4802 | .7936 | .5984 | .8152 | .9310 | .8693 | .4657 | .7840 | .5843 |
| blend-seq ONT 4x + Illumina 30x (Paragraph) | .4987 | .8064 | .6163 | .8511 | .9825 | .9121 | .4836 | .7947 | .6013 |
| blend-seq PacBio 4x + Illumina 30x (vg) | .5289 | .7911 | .6340 | .8077 | .9563 | .8757 | .5171 | .7812 | .6223 |
| blend-seq PacBio 4x + Illumina 30x (Paragraph) | .4951 | .7984 | .6112 | .8207 | **.9847** | .8953 | .4805 | .7863 | .5965 |
| ONT 14x (LR-only) | .6598 | .8132 | .7285 | .9019 | .8802 | .8909 | .6501 | .8098 | .7212 |
| PacBio 30x (LR-only) | **.6731** | **.8280** | **.7425** | **.9133** | .9186 | **.9159** | **.6625** | **.8210** | **.7333** |

**Table 4.** GT-Aware performance for SV deletions after using short-read genotypers. The blend-seq short-read genotyped SVs have better GT-Precision over the whole genome compared to the short-read 30x control and the 4x long-read controls, with recall still exceeding the 30x short reads alone. Over the easy regions, the blend-seq hybrids beat out these controls in all GT-aware metrics. Even over hard regions, we see the GT-Precision for the blend-seq mixtures doing better than these controls, which causes GT-F1 Scores to outperform these controls when using 30x short reads in addition to the 4x long reads. 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).

## GT-Aware Performance for SV Insertions 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) | .1435 | .5633 | .2287 | .4758 | .6736 | .5577 | .0769 | .5012 | .1334 |
| ONT 4x (LR-only) | .3770 | .5644 | .4521 | .5387 | .5461 | .5424 | .3447 | .5577 | .4261 |
| PacBio 4x (LR-only) | .3419 | .5491 | .4214 | .4955 | .5535 | .5229 | .3109 | .5303 | .3920 |
| blend-seq ONT 4x + Illumina 10x (vg) | .2535 | .6661 | .3672 | .5223 | .7104 | .6020 | .1997 | .6473 | .3053 |
| blend-seq ONT 4x + Illumina 10x (Paragraph) | .2919 | .6145 | .3958 | .5490 | .6483 | .5945 | .2405 | .5862 | .3411 |
| blend-seq PacBio 4x + Illumina 10x (vg) | .2627 | .6589 | .3756 | .5249 | .7065 | .6023 | .2105 | .6373 | .3165 |
| blend-seq PacBio 4x + Illumina 10x (Paragraph) | .3011 | .5980 | .4005 | .5394 | .6372 | .5842 | .2534 | .5665 | .3502 |
| blend-seq ONT 4x + Illumina 30x (vg) | .2903 | **.6792** | .4067 | .5582 | **.7193** | .6286 | .2368 | **.6646** | .3492 |
| blend-seq ONT 4x + Illumina 30x (Paragraph) | .3201 | .6481 | .4285 | .5720 | .6759 | .6197 | .2699 | .6225 | .3766 |
| blend-seq PacBio 4x + Illumina 30x (vg) | .3024 | .6706 | .4169 | .5470 | .7135 | .6192 | .2536 | .6523 | .3652 |
| blend-seq PacBio 4x + Illumina 30x (Paragraph) | .3318 | .6323 | .4352 | .5612 | .6694 | .6105 | .2861 | .6030 | .3881 |
| ONT 14x (LR-only) | .4418 | .6595 | .5292 | .6535 | .6427 | .6481 | .3994 | .6370 | .4910 |
| PacBio 30x (LR-only) | **.4601** | .6741 | **.5469** | **.6719** | .6662 | **.6691** | **.4176** | .6525 | **.5093** |

**Table 5.** GT-Aware performance for SV insertions after using short-read genotypers. Using the short-reads to genotype the SV insertions from the long reads, the blend-seq experiments perform well over the whole genome in GT-Precision, beating out the short-read control at 30x and the 4x long-read controls. In easy regions, the GT-F1 Scores for the blend-seq experiments outperform these controls as well, due to a mix of improvements in both GT-Precision and GT-Recall. Even in hard regions, where short reads struggle the most to accurately map, the 4x ONT and 30x Illumina blend-seq hybrid has the best GT-Precision out of all groups, including the high-depth long reads samples. 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).

DELs). The phasing problem comes from the diploid nature of the human genome: even if we have called all heterozygous variants with high accuracy, it also matters which subsets of these variants occur on the same copy of a given chromosome, particularly for those variants occurring within a single gene. A set of heterozygous variants is said to be *phased* if for each pair in the set it can be determined whether they appear on the same chromosome or not. Phase sets are organized into *phase blocks* and are characterized by genomic length, i.e., the distance between the first and last variant in the set. Because phasing algorithms rely on reads spanning multiple variants, it is reasonable to expect that longer reads lead to longer phase blocks. We thus hypothesize that blend-seq will be able to improve upon the phasing performance that can be achieved with short reads alone.

In terms of impact, the specific set of variants on each homologous chromosome is well-known to contribute to phenotypic expression and disease states (64), and short reads alone are known to be limited in their ability to phase variants to the quality required for pharmacogenetic profiling (see e.g. (65) and references therein). Another concrete application of phasing is accurate HLA typing, which has been associated with e.g. improved outcomes for transplant recipients (66).

***Approach and Data.*** Thus far we have shown how combining short reads with low-coverage long reads can enhance discovery rates and accuracy of SNPs and SVs. Given a VCF that contains all variant calls (including INDELs), we now wish to evaluate how these gains might impact phasing performance with respect to pure short or long read approaches. We consider phasing performance for the pipelines as a whole, i.e., considering how the phasing quality would change using the full blend-seq discovery and genotyping pipeline versus pure short-read or pure long-read pipelines. To this end, the data considered in this section for each experimental group contains only the variants that would be available to the respective pipeline.

For pure short-read control groups, we take the set of SNPs and INDELs in the short-read VCF and use the short reads to phase them using WhatsHap (67), a phasing tool designed for short reads. This produces longer, better phase blocks than the default phase blocks produced by DRAGEN (Appendix A Fig. 27). Note we do not use the SVs called by GATK-SV for short-read phasing, since that output does not provide insertion sequences as required by the phasing tools. For the long-read control groups, we use the full set of SNPs, INDELs, and SVs called by the long-read pipeline. We then experimented with two recent long-read phasers: HiPhase (68), designed for PacBio reads, and LongPhase (69), designed for both PacBio and ONT.

For the blend-seq scenarios, we begin with the SNP and INDEL VCFs produced following the procedure described in Section 2.1, where the long reads contribute synthetic short reads maintaining their original alignment, and are used for variant calling alongside the true short reads. Structural variants are called using the process of Section 2.2, using the low-coverage parameters of Sniffles on the long reads alone, but without applying any further regenotyping. Although LongPhase provides the capability of using short and long reads together for phasing, we found performance to be worse than using the long reads alone (Appendix A Fig. 28 and Appendix A Fig. 29). As such, all variants (SNP, INDEL, SV) from the hybrid process are combined together into one VCF and then phased together using the long reads alone. Note that while SVs are included in the phasing process, performance is only evaluated with respect to SNPs and INDELs to allow for more meaningful comparison with the short-read pipeline, where SVs are not available. The effect of including these is measured in Appendix A Fig. 30.

***Phasing Performance.*** Because the starting point of phasing is the set of called variants in a VCF, phasing performance is inherently tied to variant calling performance. This is true both with respect to the underlying variant being called and it being correctly genotyped. For instance, if a variant is correctly identified in a VCF but is called as homozygous when it is in fact heterozygous, phasing for that variant will be skipped as there is no need to phase variants that appear on both chromosomes. With this perspective in mind, we looked at some standard statistics for measuring phasing performance for the experimental groups described above.

The standard phasing performance metrics can be thought of as falling into two groups: measures of how extensively the genome can be phased by a given algorithm without regard to accuracy (which does not require a truth baseline), and measures of the accuracy of the resulting phasing (relative to a phased truth baseline). Beginning with the former, we looked at metrics borrowed from genome assembly. The NG50 statistic, computed by WhatsHap, measures the minimum phase block length such that at least half of the genome belongs to a phase block of that length or longer. A *larger* number thus means that *longer* distance haplotype information is available for most of the genome (more than 50%). For blend-seq, we measured this statistic across a wide range of short-read coverages (10-60x) paired with either PacBio or ONT long reads at a fixed 4x coverage.

For the short-read controls, NG50 is zero because less than half of the bases are covered by phase blocks (see Fig. 13). The NG50 for blend-seq using 4x PacBio hovers just under 100k (50% of reference bases are in phase blocks 100kb or greater in length), and the 4x ONT blend-seq NG50 is an order of magnitude greater at around 2-3M. This is due to the longer reads that allow us to better join phase blocks into larger pieces. It is also interesting to note that blend-seq does better than the 4x long-read controls alone: part of this can be explained by additional variants produced using the blend-seq short variant calling procedures from Section 2.1, which may help bridge otherwise distant heterozygous variants. For example, the 4x PacBio control has NG50 equal to zero as well, failing to phase over half of the genome.

In diploid organisms, a phase block represents a pair of haplotypes fully describing the individual's genome within that

region. Thus whenever comparing with a truth baseline, there are two possible matches between the true pair of haplotypes and the inferred pair of haplotypes, and tools like WhatsHap choose the best match. One can expand the notion of a correct match by allowing a switch at some point within the block: these are called switch errors, and two consecutive switch errors are labeled as one flip error. While NG50 does not account for errors within phase blocks, we compute the NGC50 (or "NG50 corrected") as computed by vcfdist (70), which first splits phase blocks at switch errors (based on the truth data) and then computes NG50 on this refined partition; the truth data use here is the fully-phased NIST-Q100 dataset from Section 2.2. These results are shown in Fig. 13 for the different experimental groups. We see the blend-seq mixtures with PacBio maintain their superior phasing capabilities over the short reads when accounting for phasing errors in this way, but the ONT hybrids' performance drops below their corresponding long-read control. This is potentially due to the blend-seq conditions calling substantially more variants, opening up the opportunity for a greater number of switch errors.

The frequency of switch errors is characterized by the *switch error rate*, defined as the number of switch errors per consecutive phased variant, i.e., the number of switch errors divided by the number of pairs where a switch error can occur (71). One benefit of this metric compared to NGC50 is that is normalized by the number of called variants, so calling a greater number of variants does not inherently risk reducing the score; however, it still does not account for the relative difficulty of those variants, so it is possible to call/phase only "easy" variants and not be penalized for calling variants that are harder to call/genotype. This error rate for the blend-seq mixtures was significantly lower (better) than the short-read control, but it was slightly higher (worse) than the 4x long-read controls (Fig. 14). For flip errors, the rate was about 3-5x higher for the ONT blend-seq experiments than the short-read control depending on the phasing algorithm used, with LongPhase being the better choice to avoid this error type.

## 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, in genotyping structural variants, and in phasing variants together. 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 (72) or the Human Genome Structural Variation Consortium (73), then leveraging assembly-based variant callers like Dipcall (74) or hapdiff (75). 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. (76–83)); 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.

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 (84) 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 (26), a large biobank funded by the U.S. government. It should be noted that recent work shows improved performance of DRAGEN in later versions (85). 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 (86), which also influenced the decision to use an earlier version.

For phasing, we used a fixed long-read coverage (4x), similar to our strategy for genotyping SVs discovered by long reads using the short reads, so that we could vary the short-read coverage in our performance plots. Phasing variants in theory requires only one read connecting them, so one would expect the performance of phasing to increase with further

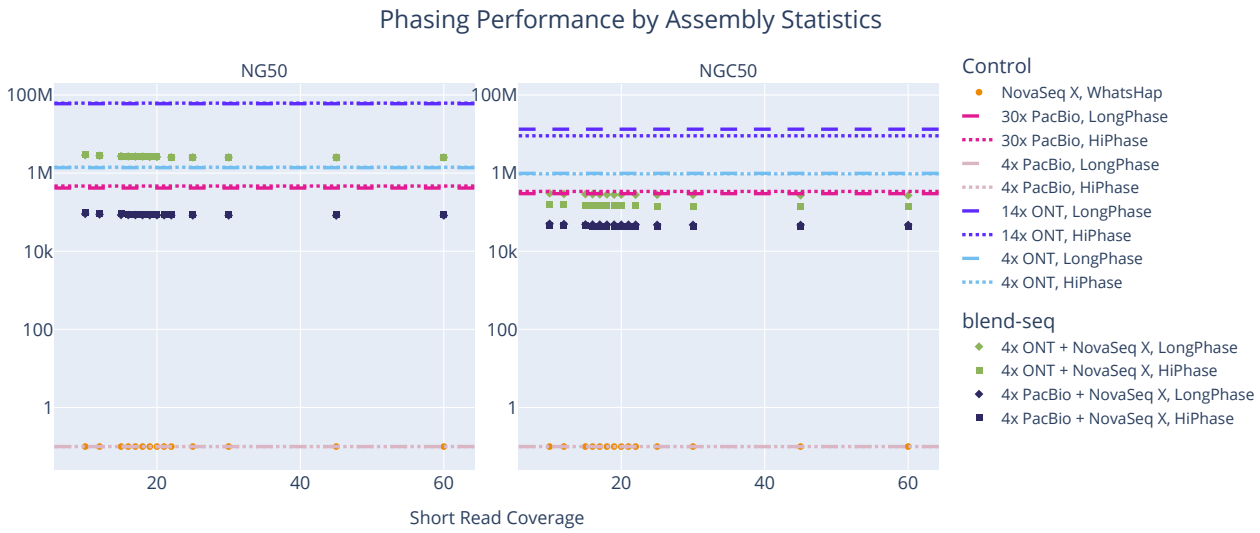



**Fig. 13.** NG50 metric (50% of the genome is within a phase block of this length) and NGC50 metric (NG50 but accounting for phasing truth data) for different settings of blend-seq and pure short and long read controls. Because the y-axis is log-scaled, NG50 values of zero (for NovaSeq X and the 4x PacBio conditions) are rescaled to 0.1. We see three distinct clusters of performance correlating with average read length. Short reads struggle to bridge distant variants and achieve an NG50 of zero, and the same is true for 4x PacBio due to missing variants at low coverage. Combining these two, blend-seq PacBio mixtures have an NG50 starting around 100k, beating out both of these controls. In the third cluster, blend-seq with ultralong ONT reads have approximately 20 times larger NG50 than the corresponding PacBio groups. For NG50, the blend-seq mixtures provide better phasing than their 4x long-read controls, likely due to better variant discovery with the additional short-read coverage. When accounting for phasing truth data in the NGC50 metric, this blend-seq improvement persists for the PacBio data, but with ONT the performance of long reads alone exceeds that of blend-seq. Note that this comparison is confounded by the fewer number of variants the 4x ONT can call (seeSection 2.1), hence giving the long-read control fewer opportunities to make a phasing error and thus potentially inflating the NGC50 score.

**Fig. 14.** Switch error rates and flip error rates for different settings of blend-seq and pure short and long read controls. In the blend-seq regimes, the switch error rate decreases with extra layers of short reads, and is significantly smaller than the short-read controls. The PacBio blend-seq switch error rate is nearly as low as the corresponding 4x long-read control. The flip error rate for blend-seq conditions stays mostly constant across short-read coverages for PacBio combinations, and increases slightly with coverage for ONT combinations. The flip error rates for PacBio blend-seq are very similar to the short-read controls; both of these curves sit slightly above the long-read controls. This figure also demonstrates the benefit of using LongPhase over HiPhase for our experiments, as it lowers both switch and flip error rates.

layers of long reads. This is explored in the supplement using long reads alone for variant calling and phasing (Appendix A Fig. 31 and Appendix A Fig. 32). However, we did not pursue replicating our phasing experiments with deeper layers of long reads combined with the short reads.

A final note with respect to SV performance presented in some prior work: because we did not want phasing performance to confound the SV performance, we could not use tools like Truvari refine (87) or vcfdist (70) to better compare the resulting haplotypes to the truth which require phased variants. This resulted in our recall being lower than what was reported by (80), which did use 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 (68) and ran Truvari refine: this resulted in a 95% F1 score (Appendix A Fig. 20), which is indeed comparable to (80).

## 4: Future Work

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

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 (88).

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 2.2 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 (89) 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 if the addition of low-coverage long reads may help with other key problems in sequencing, such as methylation. Both PacBio and ONT long-read data include information pertaining to methylation (90); it would be interesting to investigate how much of this is recoverable at low levels of coverage.

## 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 and phasing 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, and a vast performance boost in accurately phasing these variants.

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: well-mapped long reads augmented by short reads help with difficult-to-map regions for SNPs, long reads better capture long insertions and deletions (structural variants) that can then by better genotyped with short read augmentation, and long reads better phase the complete set of variants discovered by the combined approach, leading to a more complete picture of the genome. 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 among 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

**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 (26). 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 (91) 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` (92) 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 (93) 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 (91) 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 caused 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 mapped using minimap2 for ONT (94, 95) and the wrapper pbmm2 written by PacBio for the PacBio reads. The short reads were mapped using DRAGMAP as a part of the DRAGEN (96) variant calling process, using the GRCh38 human reference in both cases. The long reads (PacBio and ONT) were run through DeepVariant (97) (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 (98, 99). We used a pipeline wrapping RTG's `vcfeval` tool (100, 101) 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 ("Low Mappability", and others used in this paper) were also curated by NIST for their GIAB release (102).

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 (39) 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 (103)) 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.4.3.1 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 base-pair 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 v1.1 SV truth set for HG002 (44), which is part of the T2T-Q100 project (45), an initiative attempting to create perfect telomere-to-telomere assemblies (104) and polishing them with new tools like in the Telomere-to-Telomere project (105). 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 90% 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 (49) calls on an Illumina 30x BAM from (106), with the HG002 sample extracted with `bcftools view -s HG002` and processed with the same cleaning and benchmarking pipeline.

**Hybrid Genotyping of SVs.** Given SV calls made from the long reads, we regenotyped them using the short reads in two ways. The first, using vg v1.61.0 (107), started with 4x long reads (PacBio and ONT) mapped to the GRCh38 human reference, and created a sample-specific graph reference to map the short reads against. The full WDL pipeline steps are included in the repository described in the Code Availability section under the corresponding names. The overview of the pipeline implementation is as follows:

1. `GiraffeIndex.wdl` – The SV calls were used to augment the linear GRCh38 human reference to create a graph reference genome, using `vg construct` in SV mode, `vg index`, `vg gbwt`, and `vg minimizer`. See the source code for the specific flags used in each command.

2. `GiraffeMap.wdl` – The short reads FASTQs were mapped to this sample-specific graph reference using `vg giraffe` (89).

3. `VgCall.wdl` – The SV calls are regenotyped via the short-read graph alignments using `vg pack --min-mapq 5 --trim-ends 5` and `vg call --ploidy 2 --vcf` (108).

The second genotyping process used Paragraph, implemented in the `Paragraph.wdl` pipeline using the following steps:

1. Run the `CleanVCFParagraph` script to format the SV callset in ways Paragraph can handle, e.g. converting duplications into insertions, setting the reference field of symbolic deletions, and ensuring the first letter in insertion ALT field matches the reference. Calls close to the edge of chromosomes are excluded.

2. Run the `multigrmpy.py` script provided by Paragraph on the cleaned SV VCF and the short-read BAM to produce the final regenotyped SV VCF.

The output from both processes 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.

**Phasing Variants.** Phasing was performed using a collection of tools for comparison and matched appropriately to the datatype for each experimental group. The exact code used for computing this is included in the repository described in the Code Availability section. The workflows used are:

1. `PhaseVCF.wdl` – This is a wrapper around `whatshap phase` (67) after running `bcftools +fixploidy` to avoid a bug in the tool if any haploid variant calls are passed through (like on chromosomes

X and Y), even if they are not considered in the analysis. This was run on the short-read control SNP and INDEL VCFs produced via DRAGEN to improve on the default phasing outputs from DRAGEN (see Appendix A Fig. 27 for comparison), using just the short reads alone. We did not phase these together with the GATK-SV SV calls because there were missing sequences making it impossible to phase using the reads. A comparison of the effect of including SVs or not in the phasing process for the long-read experiments is included in the supplement (Appendix A Fig. 30), where a small benefit to including the SVs is observed but not as big as the difference between the short-read controls and other experimental groups.

2. `Hiphase.wdl` – This is a wrapper around HiPhase (68), a tool developed to use long reads to phase. It was run on the blend-seq experimental group SNP and INDEL VCFs merged with the SV VCFs called from their long reads, and with the long-reads-only controls using just the associated long reads on their SNP and INDEL VCFs merged with their SV VCFs. These results are labeled as using the phasing tool Hiphase in the figures.

3. `Longphase.wdl` – This is a wrapper around LongPhase (69). This was run same as HiPhase on the blend-seq experimental groups using their combined SNP, INDEL, and SV calls as input and just the long reads for phasing, as well as the long-reads-only experiments. We note that LongPhase includes a mode for hybrid phasing, but our comparisons for our experimental groups showed worse performance than just using the long reads (Appendix A Fig. 28 and Appendix A Fig. 29).

To evaluate phasing performance, the `BenchmarkPhasing.wdl` pipeline was used, provided in the repository described under Code Availability. The pipeline performs the following calculations:

1. Remove SVs if present (and used to help phase the other variants), then use `whatshap compare` with the phased VCF produced above to compare it to the fully-phased NIST Q100 V1.1 short variant VCF. This produces statistics which require a truth VCF to compute, like switch error counts. Also run `whatshap stats` to compute intrinsic statistics, like NG50.

2. A custom script using `bedtools` (109) divides the phase block bed file produced by `whatshap stats` at switch errors produced by `whatshap compare`, and then recomputes the NG50 on this refined "corrected" set of phase blocks to result in the NGC50 ("corrected NG50") statistic, similar to the method used to produce this as in vcfdist output (70).

3. To get the gene performance statistics, a bed file describing the location of human genes is first produced using the procedure described below. The vcfdist tool (70) is run ahead of time on each phased VCF alongside the truth dataset to assign TP, FP, and FN labels. This tool was used for this step as it makes use of phasing information to better harmonize variants across the input and truth datasets. A custom script inside the `CountPhasedGenes` task of the pipeline takes in the gene list, the vcfdist labels, and the WhatsHap files for switch errors and phase blocks. With this information, the script computes how many genes are fully contained in a phase block, how many of these do not contain a switch error, and how many of these without a switch error do not also contain an FP or FN site.

To create the bed file for human genes used above, we used the Gencode release (110) located at `https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_47/gencode.v47.basic.annotation.gff3.gz` (accessed Dec 8, 2024). Then `grep` and `awk` were used to extract rows which corresponded to `gene` in column 3, and has function `protein_coding`. Then `grep` was used again to remove lines beginning with names `chrX`, `chrY`, or `chrM` to produce just the list of protein-coding genes in autosomes.

**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 mapped 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 mapped using pbmm2 (111, 112) with the flags `--preset CCS --sample NA24385_1 --strip --sort --unmapped`.

- ONT: We used the set of reads from (47) which totaled around 14x, previously mapped using minimap2, and then downsampled from there, for all experiments involving "ultralong" (UL) ONT reads. This includes all of the blend-seq hybrid experimental groups using 4x ONT, in addition to the long-read structural variant discovery results. For the SNP discovery control line for 30x ONT, we found the other UL ONT reads from (47) to have dissatisfactory performance when combined to represent a reasonable control, so we used the reads from the HG002 T2T project (`https://github.com/marbl/HG002`) and downsampled to 30x.

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 structural 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

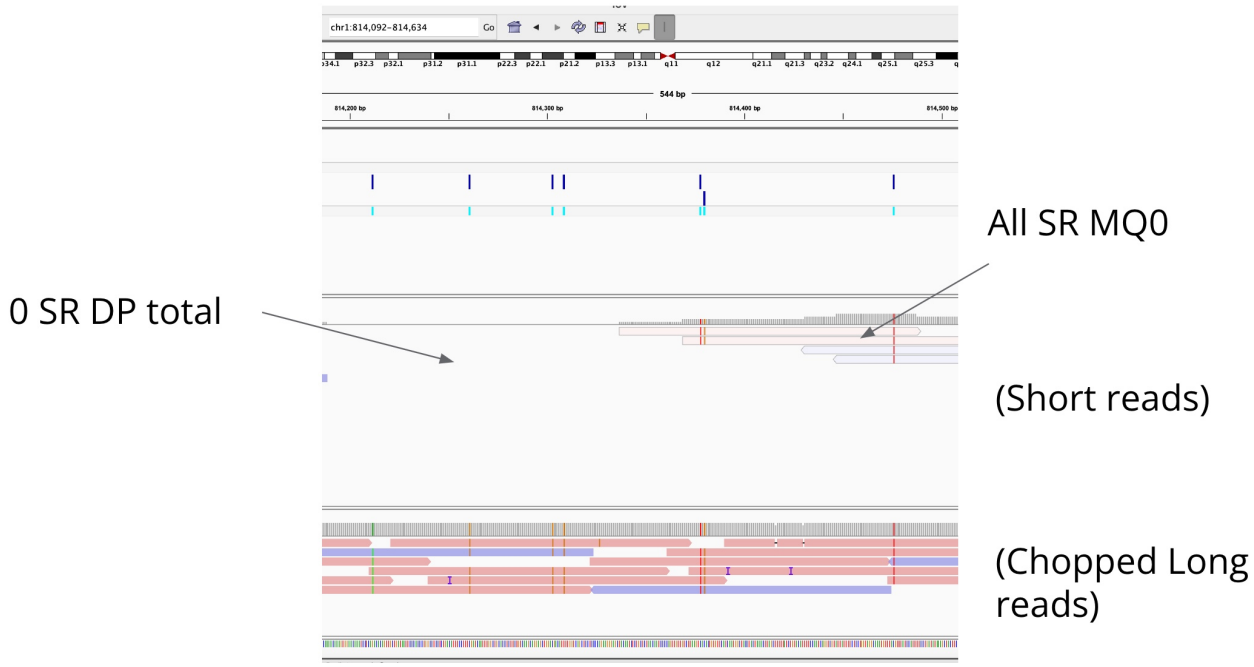
1. Peter H. Sudmant, Tobias Rausch, Eugene J. Gardner, and others. An integrated map of structural variation in 2, 504 human genomes. Nature, 526(7571):75–81, September 2015. ISSN 1476-4687. doi: 10.1038/nature15394.
2. Claudia M. B. Carvalho and James R. Lupski. Mechanisms underlying structural variant 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. Dmitrii Meleshko, Rui Yang, Salil Maharajan, and others. Blackbird: structural variant detection using synthetic and low-coverage long-reads. bioRxiv, pages 2024–11, 2024.
23. 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.
24. 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.
25. 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.
26. 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.
27. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18): 3094–3100, 2018.
28. 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.
29. 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.
30. James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, and others. Integrative genomics viewer. Nature Biotechnology, 29(1):24–26, 2011.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. Moritz Smolka, Luis F Paulin, Christopher M Grochowski, and others. Detection of mosaic and population-level structural variants with Sniffles2. Nature Biotechnology, January 2024.
40. Tao Jiang, Yongzhuang Liu, Yue Jiang, and others. Long-read-based human genomic structural variation detection with cuteSV. Genome Biology, 21:1–24, 2020.
41. David Heller and Martin Vingron. SVIM: structural variant identification using mapped long reads. Bioinformatics, 35(17):2907–2915, 2019.
42. Pacific Biosciences. pbsv, GitHub repository, 2023. github.com/PacificBiosciences/pbsv.
43. 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.
44. 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_defrabbV0.018-20240716/. Accessed: 2024-08-08.
45. T2T Consortium. Telomere-to-telomere consortium HG002 "Q100" project, 2024. https://github.com/marbl/HG002. Accessed: 2024-08-08.
46. 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.
47. 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.
48. Zan Koenig, Mary T Yohannes, Lethukuthula L Nkambule, and others. A harmonized public

resource of deeply sequenced diverse human genomes. Genome Research, 2024.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
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. 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.
58. Sai Chen, Peter Krusche, Egor Dolzhenko, and others. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biology, 20:1–13, 2019.
59. Jouni Sirén, Parsa Eskandar, Matteo Tommaso Ungaro, and others. Personalized pangenome references. Nature Methods, pages 1–7, 2024.
60. Jacob Pritt, Nae-Chyun Chen, and Ben Langmead. FORGe: prioritizing variants for graph genomes. Genome Biology, 19:1–16, 2018.
61. Chirag Jain, Neda Tavakoli, and Srinivas Aluru. A variant selection framework for genome graphs. Bioinformatics, 37(Supplement_1):i460–i467, 2021.
62. 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.
63. 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.
64. Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J. Topol, and Nicholas J. Schork. The importance of phase information for human genomics. Nature Reviews Genetics, 12(3): 215–223, February 2011. ISSN 1471-0064. doi: 10.1038/nrg2950.
65. Sylvan Manuel Caspar, Timo Schneider, Patricia Stoll, Janine Meienberg, and Gabor Matyas. Potential of whole-genome sequencing-based pharmacogenetic profiling. Pharmacogenomics, 22(3):177–190, February 2021. ISSN 1744-8042. doi: 10.2217/pgs-2020-0155.
66. Neema P. Mayor, James D. Hayhurst, Thomas R. Turner, and others. Recipients receiving better hla-matched hematopoietic cell transplantation grafts, uncovered by a novel hla typing method, have superior survival: A retrospective study. Biology of Blood and Marrow Transplantation, 25(3):443–450, March 2019. ISSN 1083-8791. doi: 10.1016/j.bbmt.2018.12.768.
67. Marcel Martin, Murray Patterson, Shilpa Garg, and others. WhatsHap: fast and accurate read-based phasing. BioRxiv, page 085050, 2016.
68. 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.
69. Jyun-Hong Lin, Liang-Chi Chen, Shu-Chi Yu, and Yao-Ting Huang. LongPhase: an ultra-fast chromosome-scale phasing algorithm for small and large variants. Bioinformatics, 38 (7):1816–1822, 2022.
70. Tim Dunn and Satish Narayanasamy. vcfdist: accurately benchmarking phased small variant calls in human genomes. Nature Communications, 14(1):8149, 2023.
71. Yongwook Choi, Agnes P. Chan, Ewen Kirkness, Amalio Telenti, and Nicholas J. Schork. Comparison of phasing strategies for whole human genomes. PLOS Genetics, 14(4): e1007308, April 2018. ISSN 1553-7404. doi: 10.1371/journal.pgen.1007308.
72. Wen-Wei Liao, Mobin Asri, Jana Ebler, and others. A draft human pangenome reference. Nature, 617(7960):312–324, 2023.
73. Glennis A Logsdon, Peter Ebert, Peter A Audano, and others. Complex genetic variation in nearly complete human genomes. bioRxiv, pages 2024–09, 2024.
74. 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.
75. Mikhail Kolmogorov. hapdiff, GitHub repository, 2021. github.com/KolmogorovLab/hapdiff.
76. Hannes P Eggertsson, Snaedis Kristmundsdottir, Doruk Beyter, and others. GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. Nature Communications, 10(1):5402, November 2019.
77. 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.
78. 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.
79. 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.
80. 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.
81. 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.
82. 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.
83. Jingwen Ren and Mark JP Chaisson. lra: A long read aligner for sequences and contigs. PLOS Computational Biology, 17(6):e1009078, 2021.
84. Adrian Tan, Gonçalo R Abecasis, and Hyun Min Kang. Unified representation of genetic variants. Bioinformatics, 31(13):2202–2204, July 2015.
85. 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.
86. 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.
87. Adam C English, Egor Dolzhenko, Helyaneh Ziaei Jam, and others. Analysis and benchmarking of small and large genomic variants across tandem repeats. Nature Biotechnology, pages 1–12, 2024.
88. 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.
89. 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.
90. 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.
91. Petr Danecek, James K Bonfield, Jennifer Liddle, and others. Twelve years of SAMtools and BCFtools. Gigascience, 10(2), February 2021.
92. 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.
93. Ricky Magner. chop_reads, 2024. https://github.com/rickymagner/chop_reads. Accessed: 2024-08-08.
94. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18): 3094–3100, September 2018.
95. Heng Li. New strategies to improve minimap2 alignment accuracy. Bioinformatics, 37(23): 4572–4574, December 2021.
96. 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.
97. 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.
98. Justin Wagner, Nathan D Olson, Lindsay Harris, and others. Benchmarking challenging small variants with linked and long reads. July 2020.
99. 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.
100. 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.
101. Real Time Genomics. rtg-tools, 2024. https://github.com/RealTimeGenomics/rtg-tools. Accessed: 2024-08-08.
102. Nathan Dwarshuis, Divya Kalra, Jennifer McDaniel, and others. The GIAB genomic stratifications resource for human reference genomes. Accessed: 2024-08-08, October 2023.
103. Broad Data Sciences Platform Methods hydro.gen Team. palantir-workflows, 2024. https://github.com/broadinstitute/palantir-workflows. Accessed: 2024-08-08.
104. Sergey Nurk, Sergey Koren, Arang Rhie, and others. The complete sequence of a human genome. Science, 376(6588):44–53, April 2022.
105. 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.
106. 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.
107. 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.
108. 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.
109. Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6):841–842, January 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq033.
110. Adam Frankish, Silvia Carbonell-Sala, Mark Diekhans, and others. Gencode: reference annotation for the human and mouse genomes in 2023. Nucleic Acids Research, 51(D1): D942–D949, November 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac1071.
111. 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.
112. Armin Töpfer et al. pbmm2, GitHub repository, 2024. https://github.com/Pacif

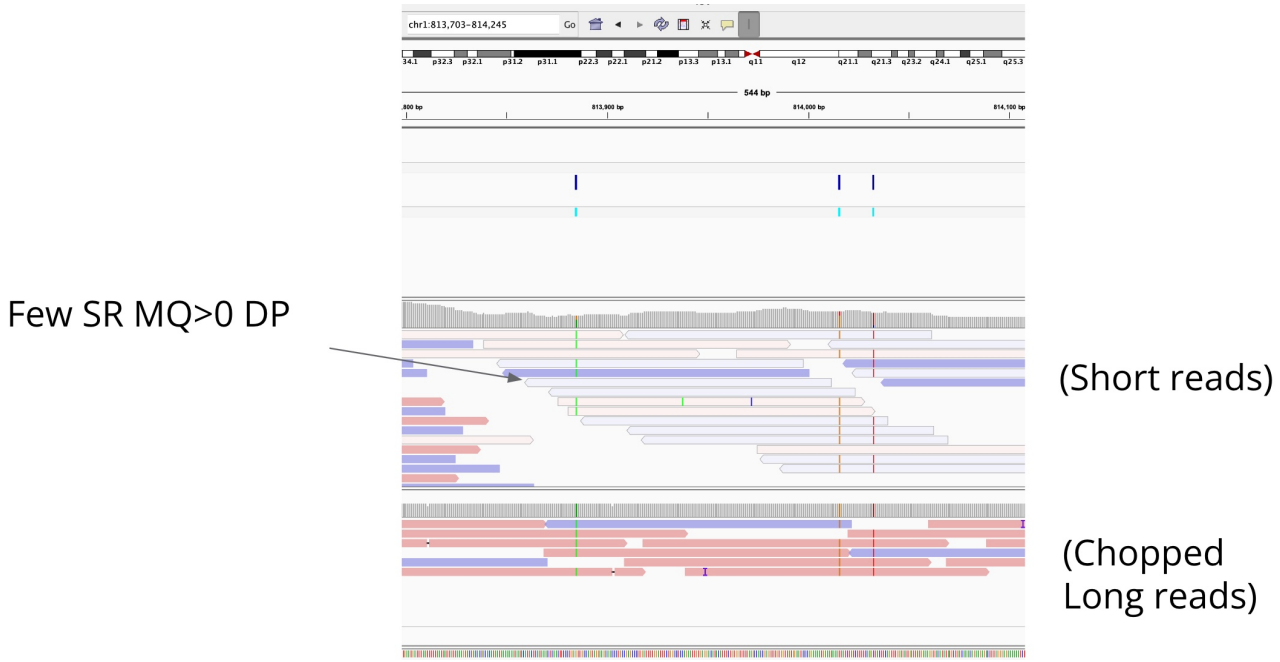
icBiosciences/pbmm2.

## A: Extra Figures and Tables

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



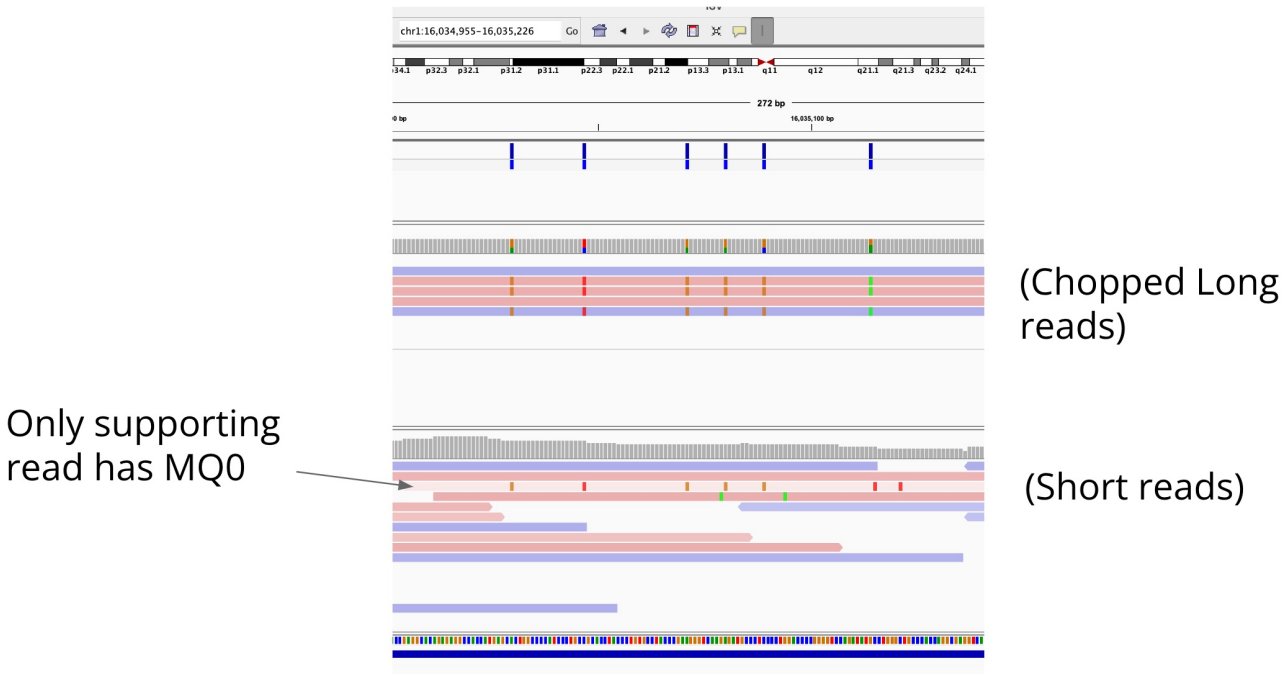

**Appendix A Supplementary Fig. 1.** Example of variants with no short read depth or all zero mapping quality. This IGV screenshot demonstrates an example of a 0-DP-SR SNP (left) and a MQ0-SR-ALL SNP (right). The top set of reads are short reads, and the bottom are PacBio. The shading demonstrates that 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.



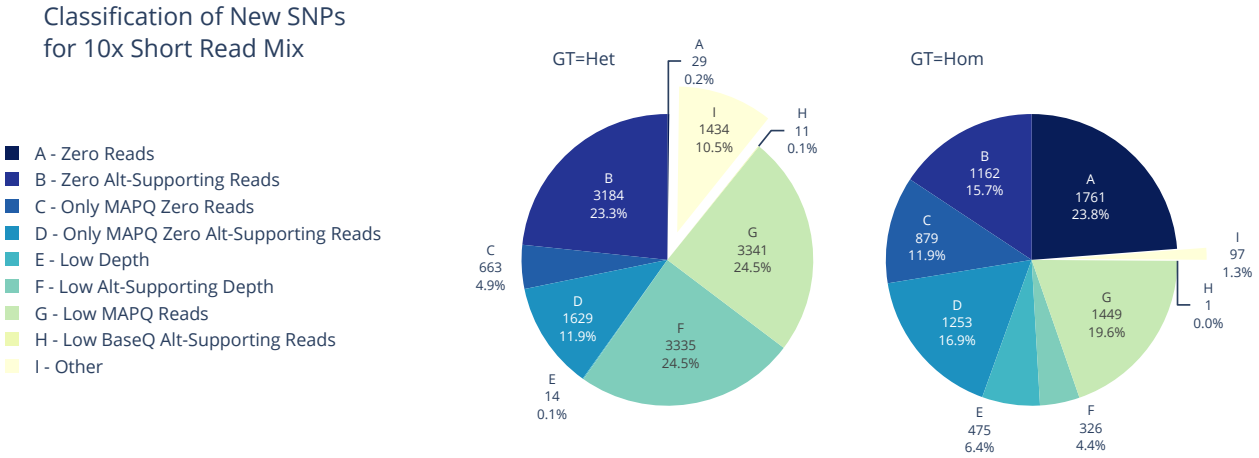

**Appendix A Supplementary Fig. 2.** Example of variants with low mapping quality in short-read evidence. This IGV screenshot includes an example of a LOWMQ-SR-ALL SNP, as we see the short reads (top) have mostly mapping quality 0 here. The long reads (bottom) have no trouble aligning here with good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

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

***A.1.1. Updated DRAGEN Variant Calling Performance.*** This section includes small variant performance curves when using DRAGEN v4.2.7 instead of v3.7.8 as in the rest of the paper. We refer to the Methods section for details on why we chose to use the latter, but include the former latest version for completion.




**Appendix A Supplementary Fig. 3.** Example of variants with only supporting short reads have mapping quality zero. 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.** The breakdown of short-read false negative and hybrid true-positive SNPs for 10x short reads and 4x PacBio long reads in low mappability regions. 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).

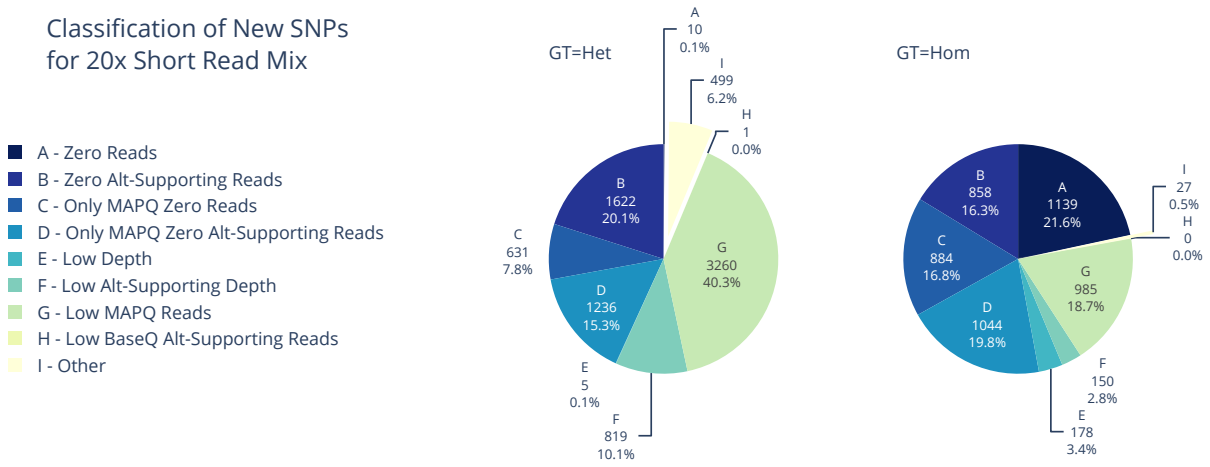
## A.2. Extra Materials for Structural Variant Discovery.

***A.2.1. Other SV Calling Choices.*** In this subsection, we present some plots demonstrating other choices for processing the data for SV discovery.

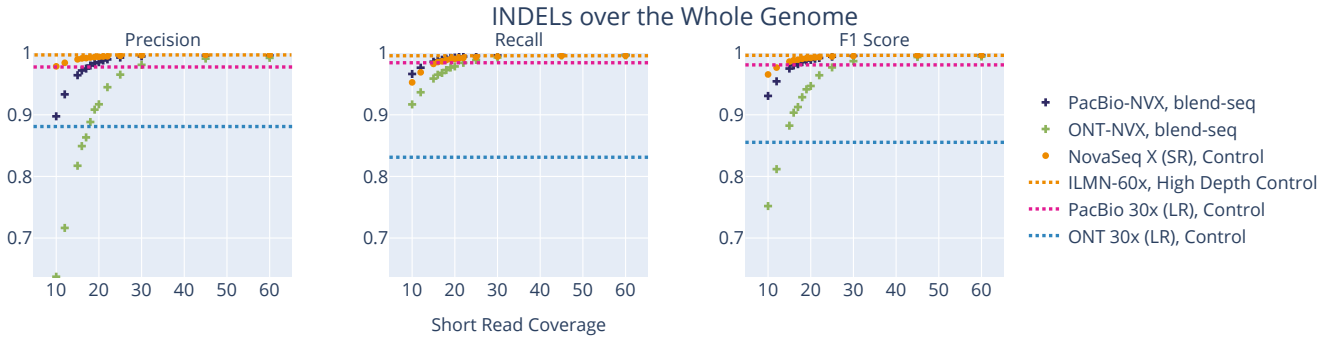
***A.2.2. Long-Read Genotyper Comparison.***

***A.2.3. Short-Read Regenotyping Structural Variants.***

## A.3. Extra Materials for Phasing Performance.




**Appendix A Supplementary Fig. 5.** The breakdown of short-read false negative and hybrid true-positive SNPs for 20x short reads and 4x PacBio long reads in low mappability regions. 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).

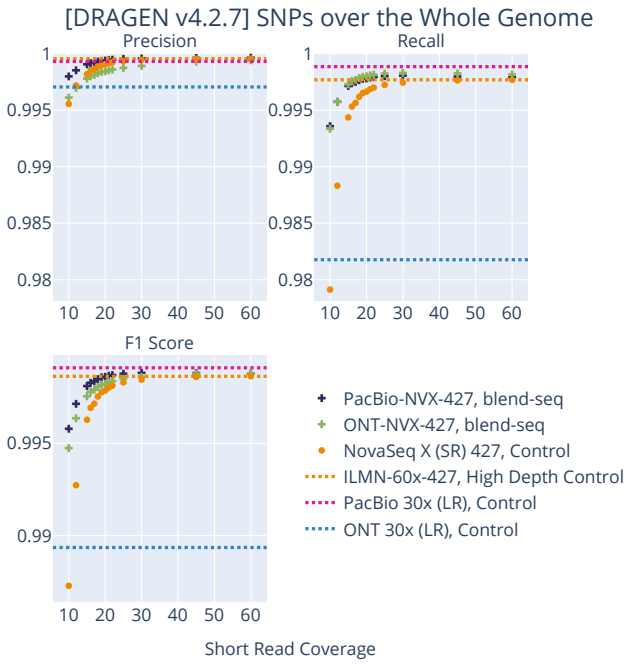



**Appendix A Supplementary Fig. 6.** INDEL performance over the whole genome. The performance for INDELs over the whole genome looks about even for the hybrids vs the pure short reads, with ONT alone standing out with poor performance as a control.

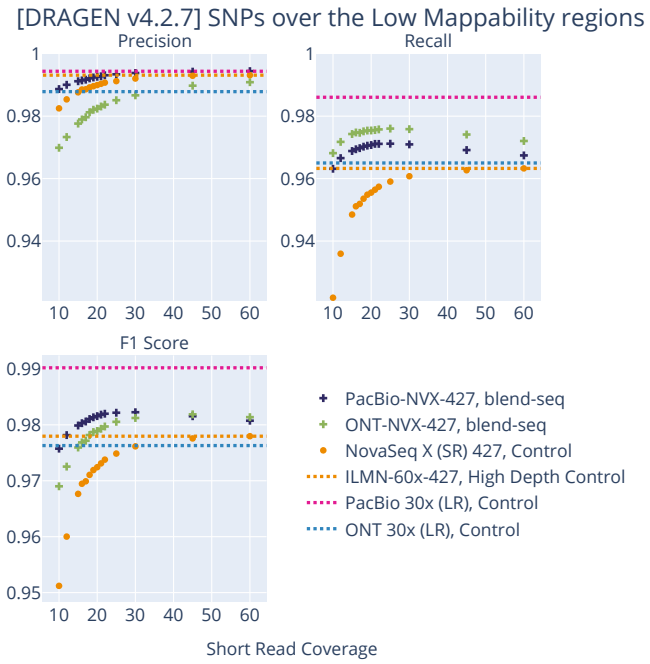
## INDEL Performance by Region

| | 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 |

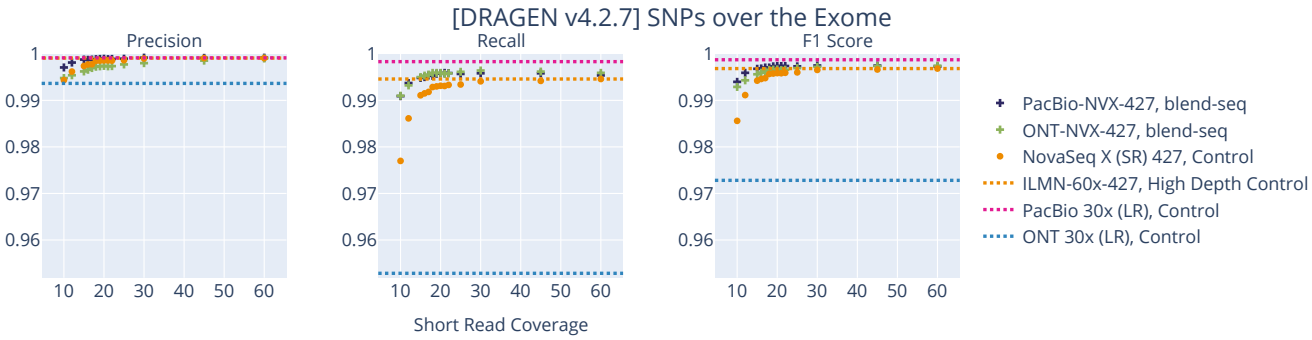
**Table 6.** INDEL performance by broken down by region. The blend-seq methods outperform the short read methods in recall overall, especially in hard-to-map regions. Precision lags a bit behind, possibly due to no hybrid filtering methods being applied to the results. Bold values are max values in column and shaded cells are those exceeding all control values in column (first three rows).



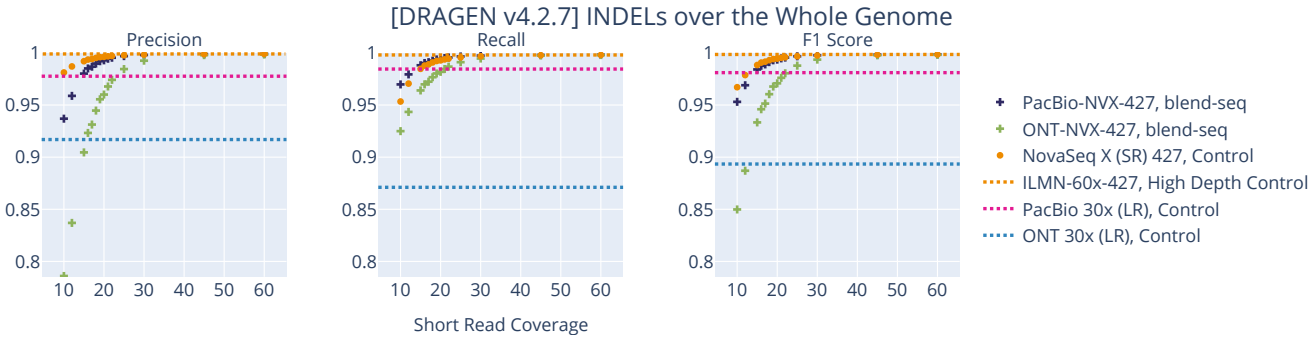

**Appendix A Supplementary Fig. 7.** DRAGEN v4.2.7 SNP performance. When using the closed-source DRAGEN v4.2.7, we see the SNP performance on the whole genome still has higher recall for the blend-seq experiments, with higher performance over the high depth 60x short read control around 20x short reads + 4x long reads.



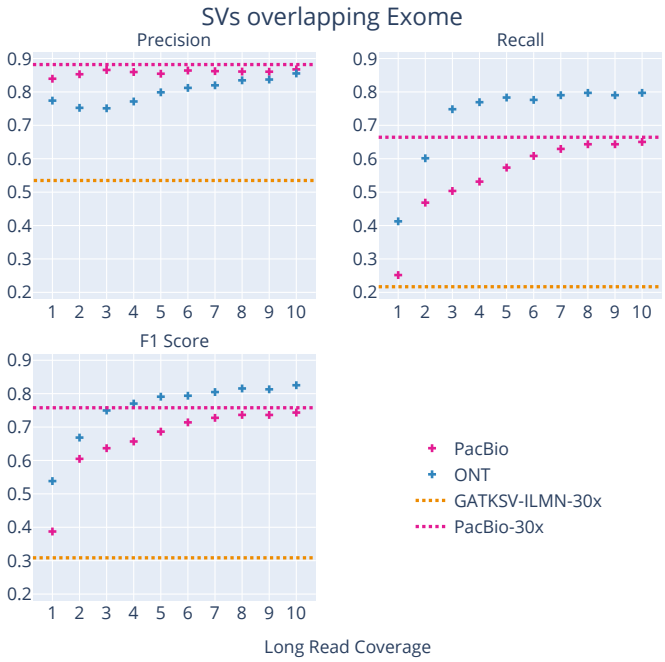

**Appendix A Supplementary Fig. 8.** DRAGEN v4.2.7 SNP performance in low mappability regions. When using the closed-source DRAGEN v4.2.7, we see the biggest gains occur again in the low mappability regions for the blend-seq experiments, with them outperforming the short controls even at the highest depth for recall, for every combination of depths considered. The F1 scores look better than the high depth 60x short control for blend-seq starting around 15x short reads + 4x PacBio reads and 20x short reads + 4x ONT reads.



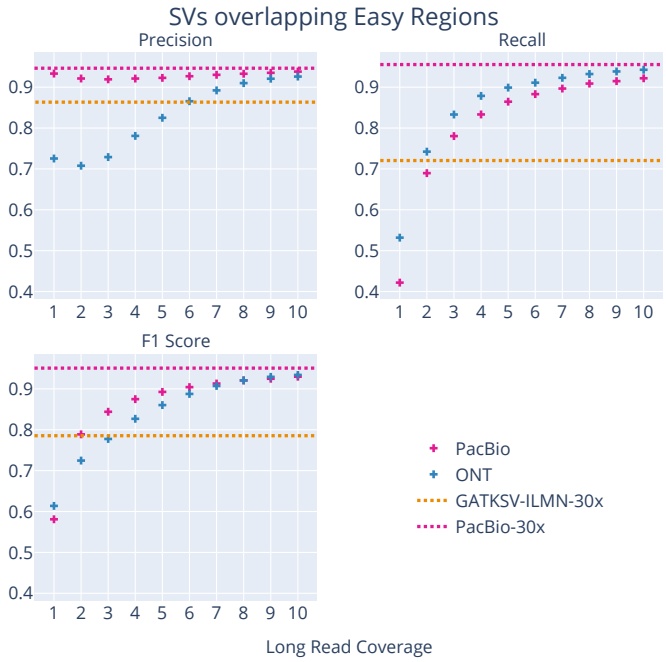

**Appendix A Supplementary Fig. 9.** DRAGEN v4.2.7 SNP performance over the exome. When using the closed-source DRAGEN v.4.2.7, blend-seq still outperforms in recall around 15x short reads + 4x long reads.




**Appendix A Supplementary Fig. 10.** DRAGEN v4.2.7 INDEL performance. When using the closed-source DRAGEN v4.2.7, blend-seq shows a slight boost in recall over the short-read controls, but at a larger sacrifice in precision, suggesting a greater need for filtering in this setting.




**Appendix A Supplementary Fig. 11.** SVs overlapping the exome. The SV discovery performance boost for the pure long reads is maintained when restricting to events overlapping the exome, with the ONT reads having better recall than the PacBio reads.

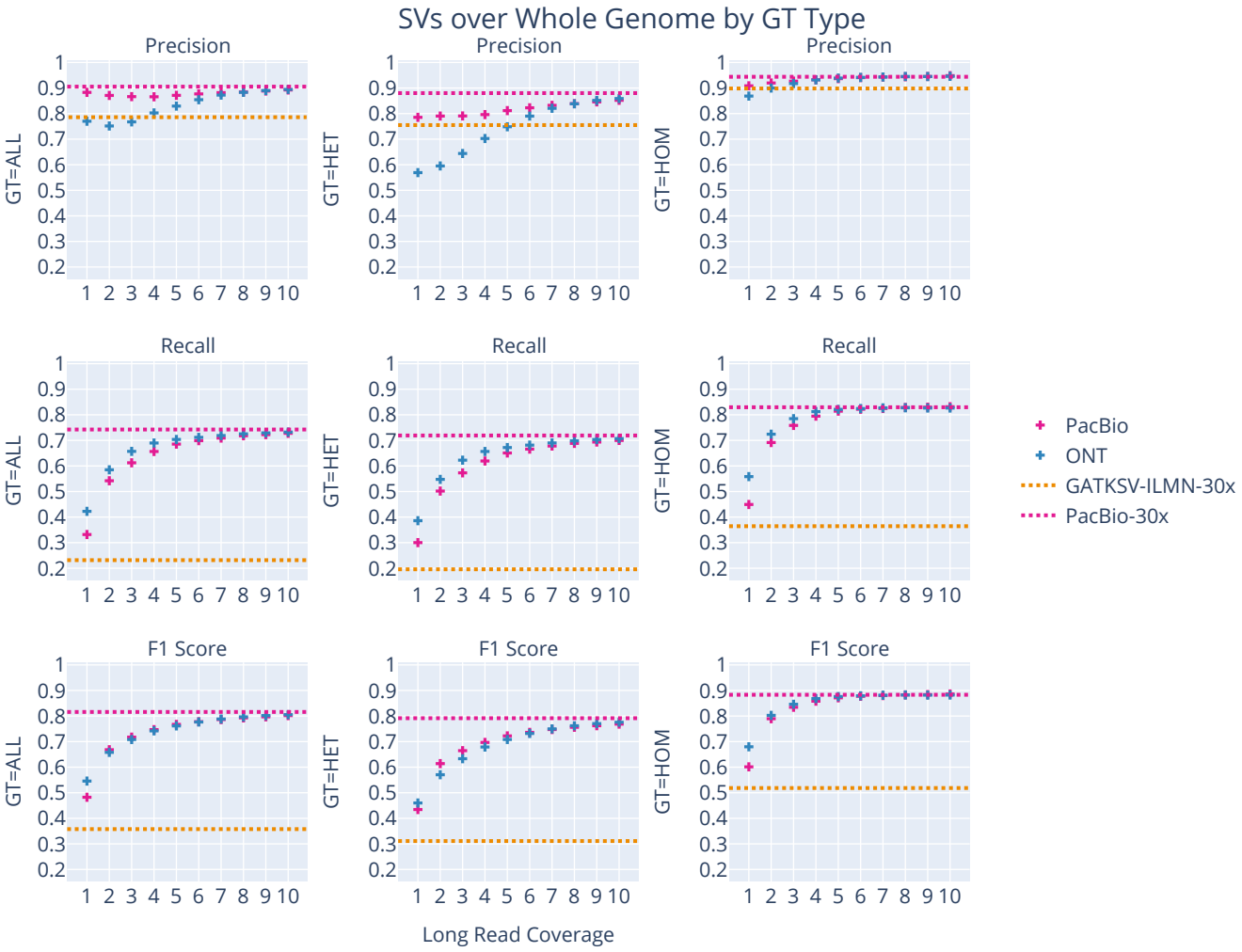



**Appendix A Supplementary Fig. 12.** SVs overlapping easy regions. When restricting to the easy regions, we give the short reads the best region for them to perform well, and still see long reads outperforming in F1 starting at only 2-3x.

## SV Deletion Discovery Performance by Region

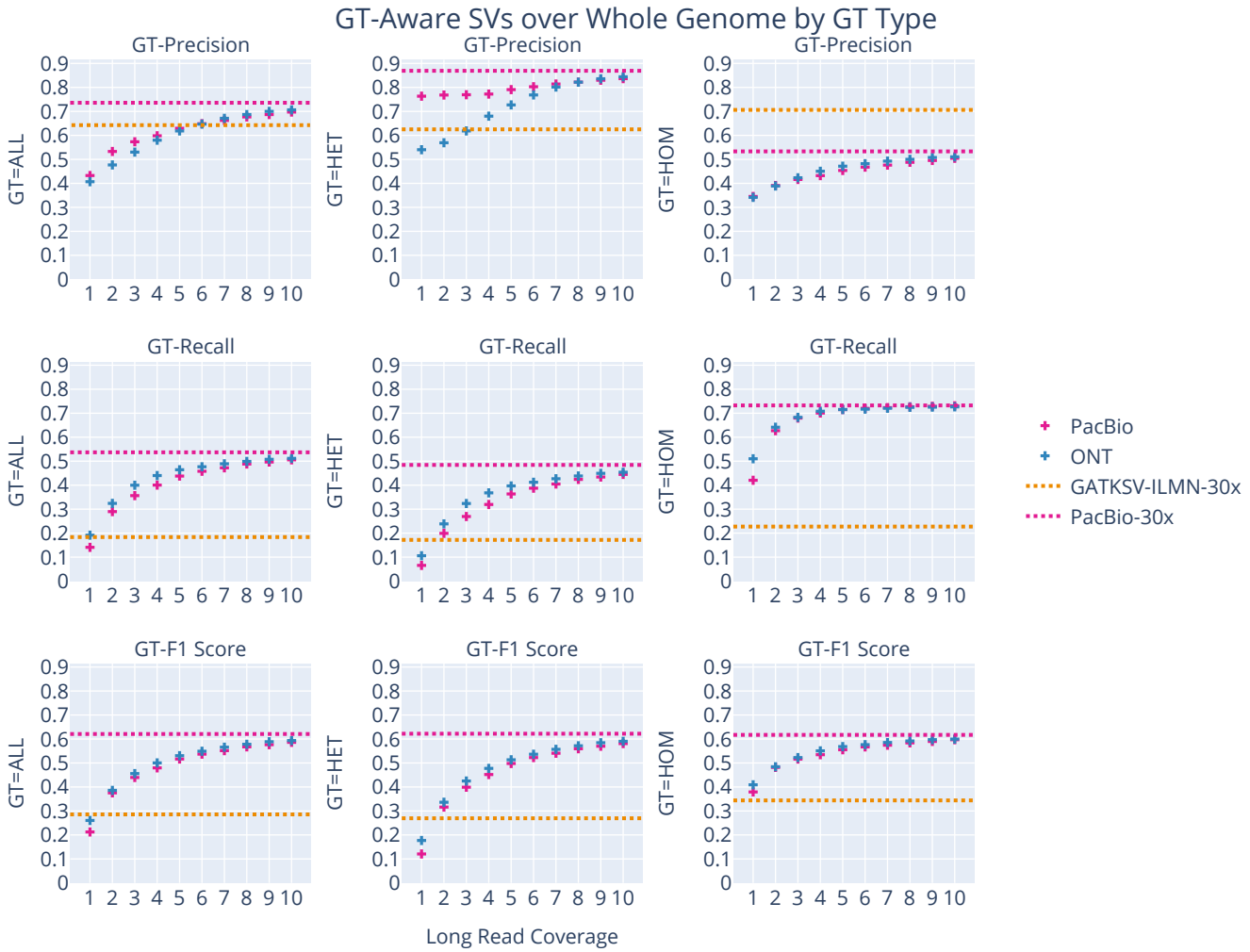
| | Whole Genome | | | Easy Regions | | | Hard Regions | | | Exome | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recall | Precision | F1 Score | Recall | Precision | F1 Score | Recall | Precision | F1 Score | Recall | Precision | F1 Score |
| Illumina 30x (control) | .3016 | **.8821** | .4495 | .8589 | .9160 | **.8866** | .2759 | .8721 | .4192 | .3333 | .6000 | .4286 |
| PacBio 3x | .6370 | .8621 | .7326 | .7841 | .9045 | .8400 | .6312 | .8689 | .7312 | .5067 | .7647 | .6095 |
| PacBio 4x | .6821 | .8650 | .7627 | .8377 | **.9180** | .8760 | .6766 | .8685 | .7606 | .5600 | **.7818** | .6526 |
| ONT 6x | .7325 | .8253 | .7761 | .8995 | .7608 | .8244 | .7275 | .8557 | .7864 | .7600 | .6543 | .7032 |
| ONT 7x | **.7390** | .8573 | **.7937** | **.9125** | .8247 | .8664 | **.7334** | **.8748** | **.7979** | **.7733** | .7397 | **.7562** |

**Table 7.** SV deletion discovery performance by region. 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).

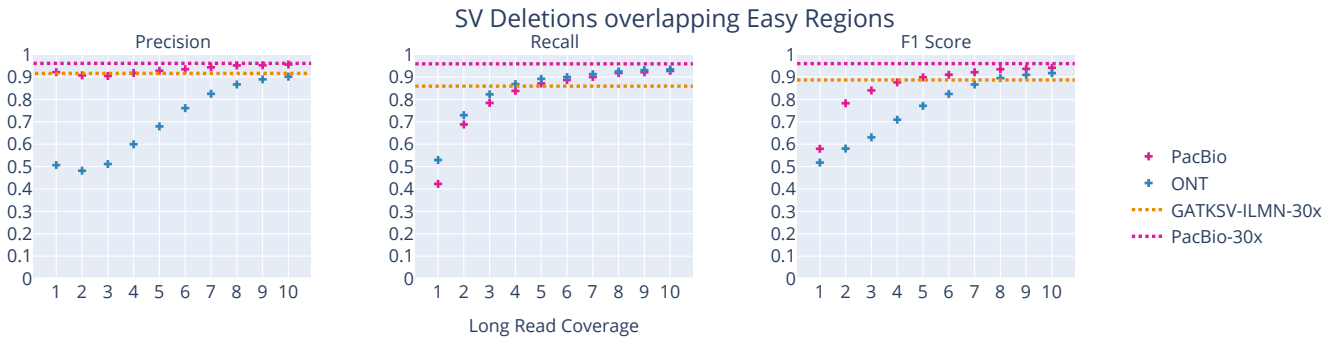



**Appendix A Supplementary Fig. 13.** Breakdown of SV performance by GT type. When breaking down performance based on genotype category, we see performance slightly better for homozygous (HOM) variants than compared to heterozygous (HET) variants. For precision, we consider HETs in the query VCF, whereas for recall we consider HOMs in the truth VCF. It should be remarked that about 20% of variants in the confidence regions for the truth file are homozygous, so the overall trends skew slightly better than the HET categories. Nevertheless, the blend-seq curves sit above the control for recall and F1 already at 1x, and for precision the PacBio reads outperform at all depths while ONT breaks even around 5x. It is worth remarking that the HOM variant recall plateaus already around 4-5x for both technologies compared to the PacBio 30x "asymptote."

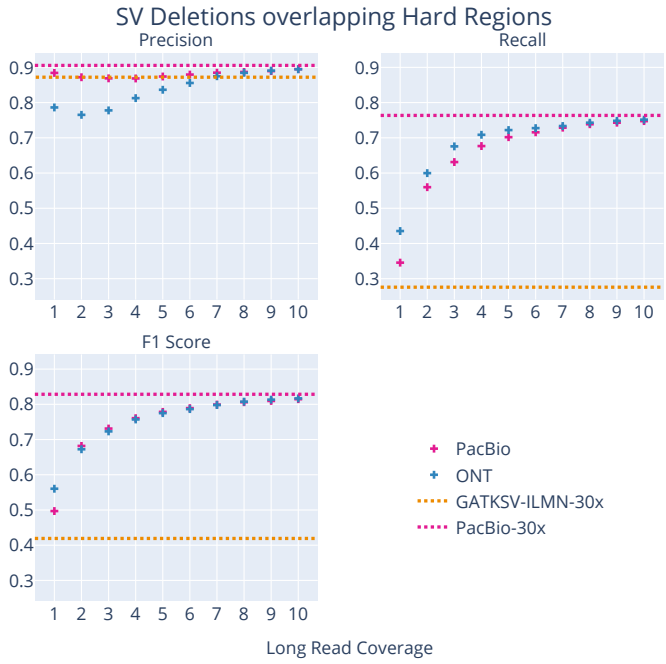
***A.3.1. Long-Read-Only Phasing.*** This subsection contains some plots of phasing performance stratified by varying long-read coverage, with no short reads used.



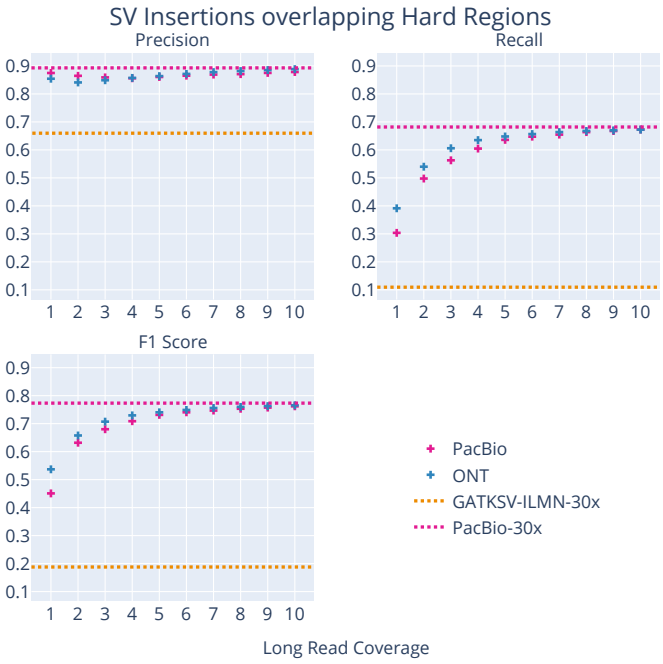

**Appendix A Supplementary Fig. 14.** Breakdown of SV GT-aware performance by GT type. When looking at GT-aware performance by genotype category, we see a similar story as the non-GT-aware breakdown. The HOM variants yield better performance unsurprisingly, while the HETs lag a little behind. Here, the GT-Precision is always better for the short read controls for HOMs, balanced out by the much lower GT-Recall, shaking out to a better GT-F1 curve for the long reads when depth is at least 2x across all variant types.




**Appendix A Supplementary Fig. 15.** Performance of SV deletion discovery in easy regions. 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.

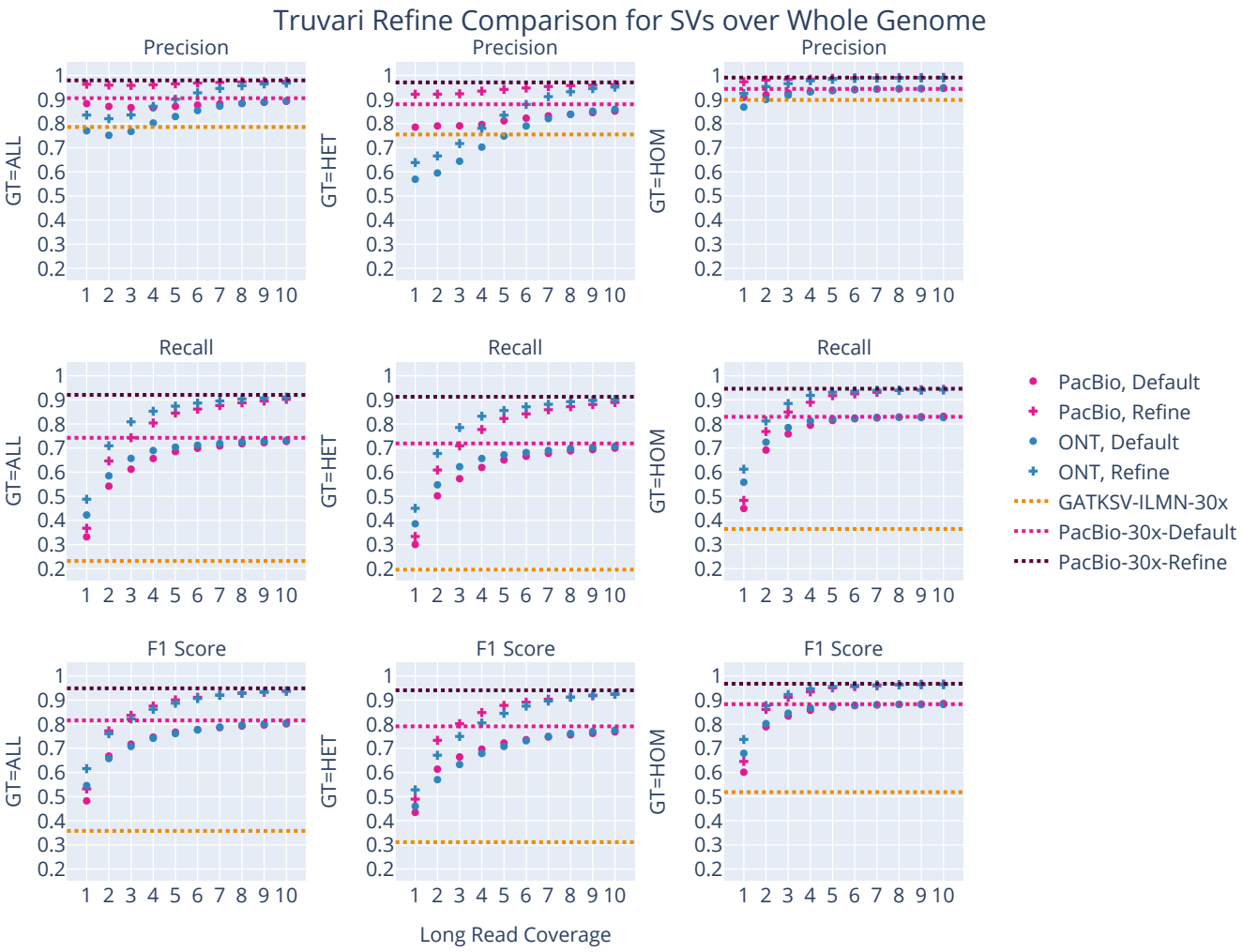



**Appendix A Supplementary Fig. 16.** Performance of SV deletion discovery in hard regions. For events overlapping difficult regions, the long reads perform best when measured by F1 even for deletions. The precision is best with PacBio, followed by Illumina, and then ONT, but recall is much higher for long reads even at 1x coverage.

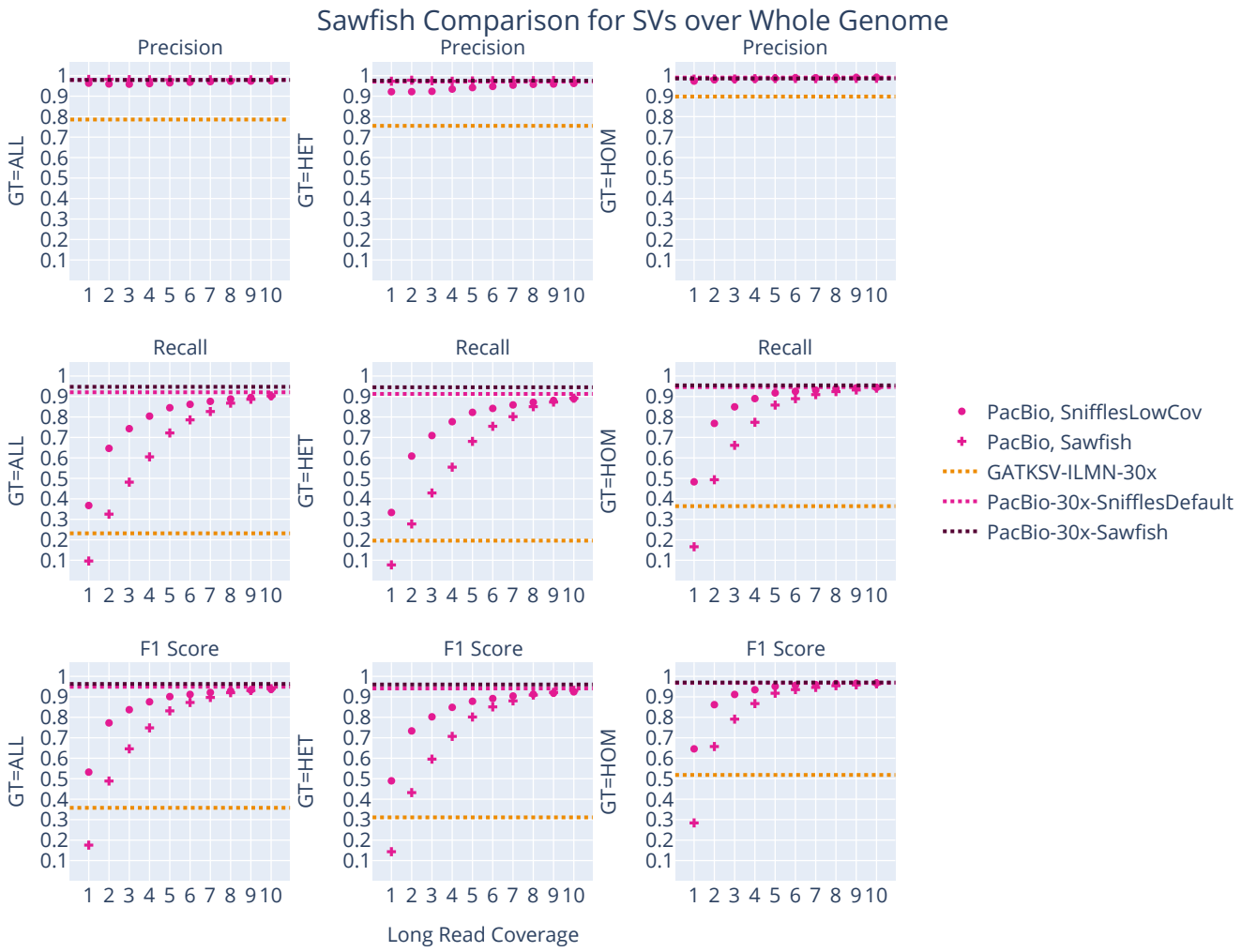



**Appendix A Supplementary Fig. 17.** Performance of SV insertion discovery in hard regions. Insertions in difficult regions are the hardest for short reads to discover, and we observe long read discovery methods outperforming in every statistical category.

**Appendix A Supplementary Fig. 18.** Comparison of Sniffles low-coverage mode to its default mode. We compared using Sniffles with its "low coverage" flags (cross symbols), as described in the Methods section, to running with its default parameters (circles) at the depths considered for SV discovery, with the 30x PacBio control run using the default parameters. The low-coverage flags outperform at lower depths for recall with a small loss in precision, which leads to an overall better F1. The gap between them shrinks at around 8x, and even sooner for ONT recall on HET sites. The drop in precision for using the low-coverage flags for PacBio still places it above the short read performance and is justified by the large boost in recall.




**Appendix A Supplementary Fig. 19.** Truvari refine compared to default Truvari parameters. We compared two different methods of using Truvari to compute benchmarking statistics. The "Default" one labeled here reflects the techniques outlined in the Methods section, whereas the "Refine" one uses the Truvari refine command as an intermediate, as described in (80). We see the Refine version has better statistics, as was observed in ibid., but emphasize that this is a process which harmonizes the query and truth variants together, so is not reflective of a processing step that can be done in the absense of truth data. Despite this, we see similar trends in our benchmarking statistics using our "Default" evaluation although the exact numbers differ.

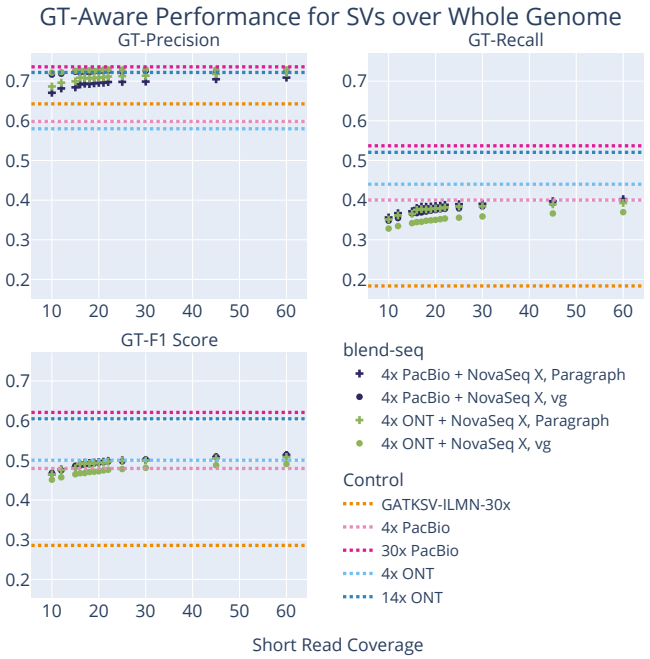



**Appendix A Supplementary Fig. 20.** Sawfish compared to Sniffles in low coverage mode. We compared the performance of Sawfish (80), an SV caller written by PacBio for their reads, to the Sniffles low-coverage settings used in our paper. The version of Sawfish that was available at the time of our experiment did not allow changing core parameters like the number of supporting reads on the command line. Although (80) investigates the change in performance by coverage, it did not analyze the lower depths we investigated here. The two methods converge at higher depth, but at lower depth Sniffles with low-coverage parameters outperforms in recall for a small drop in precision. This becomes better F1 performance, especially at the lowest 1-7x depths.

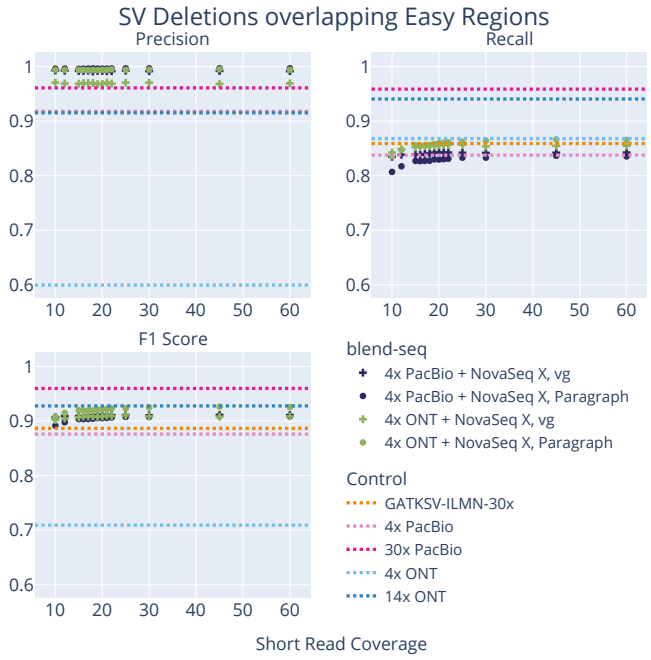
**Appendix A Supplementary Fig. 21.** Comparison of long-read genotypers SVJedi and kanpig to Sniffles low-coverage outputs, for PacBio reads. Using the long-read SV genotypers SVJedi and kanpig on PacBio data yields very little difference over using the Sniffles low-coverage calls directly, except perhaps a small boost in precision for heterozygous sites. This tradeoff with recall is barely noticeable in the F1 statistics, justifying omitting these genotyper tools from the pipeline.

**Appendix A Supplementary Fig. 22.** Comparison of long-read genotypers SVJedi and kanpig to Sniffles low-coverage outputs, for ONT reads. Using the long-read SV genotypers SVJedi and kanpig on ONT data reflects the fact they were not designed for this data-type, as performance drops dramatically as depth increases. This justifies using the Sniffles low-coverage calls directly rather than regenotyping them.

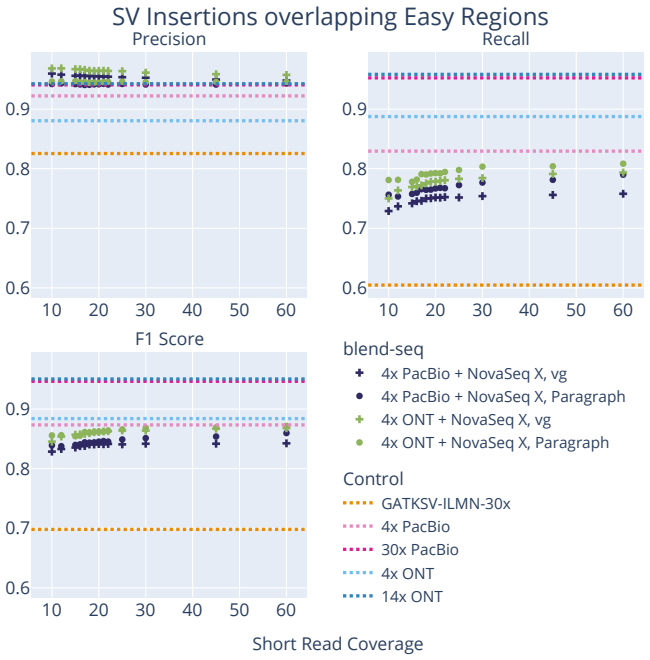
**Appendix A Supplementary Fig. 23.** Short-read SV genotyper performance for SVs overall. Although the genotypers improve SV precision overall, it comes at a large cost in decreasing recall, leading to a net decrease in performance compared to the 4x long-read controls.



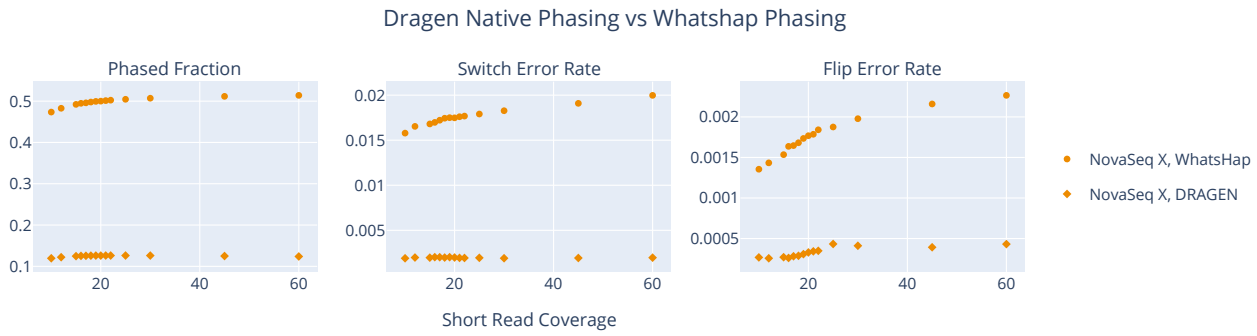

**Appendix A Supplementary Fig. 24.** Short-read SV genotyper GT-aware performance for SVs overall. Using the short-read genotypers leads to a precision-recall trade-off overall, improving GT-Precision but decreasing GT-Recall, which leads to similar GT-F1 Scores when compared to the 4x long-read controls.



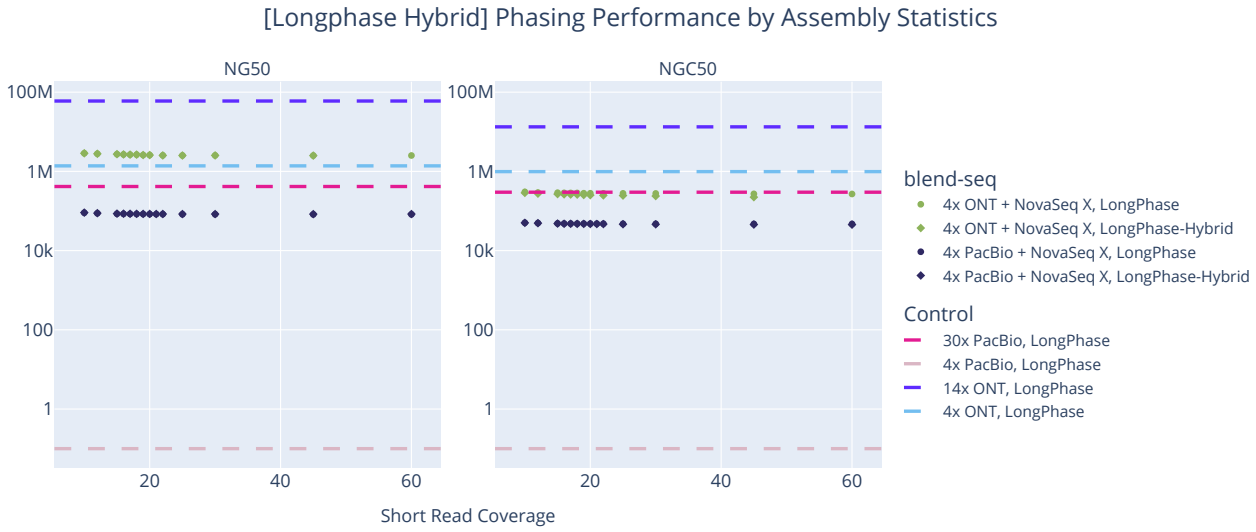

**Appendix A Supplementary Fig. 25.** Short-read SV genotyper performance for SV deletions in easy regions. Precision for SV deletions increases over the 4x long-read controls for the blend-seq experiments when using the genotypers, without sacrificing much recall, especially when short read coverage is at least 15x. This leads to a net increase in F1 scores for SV deletions in easy regions compared to the 4x long-read controls.



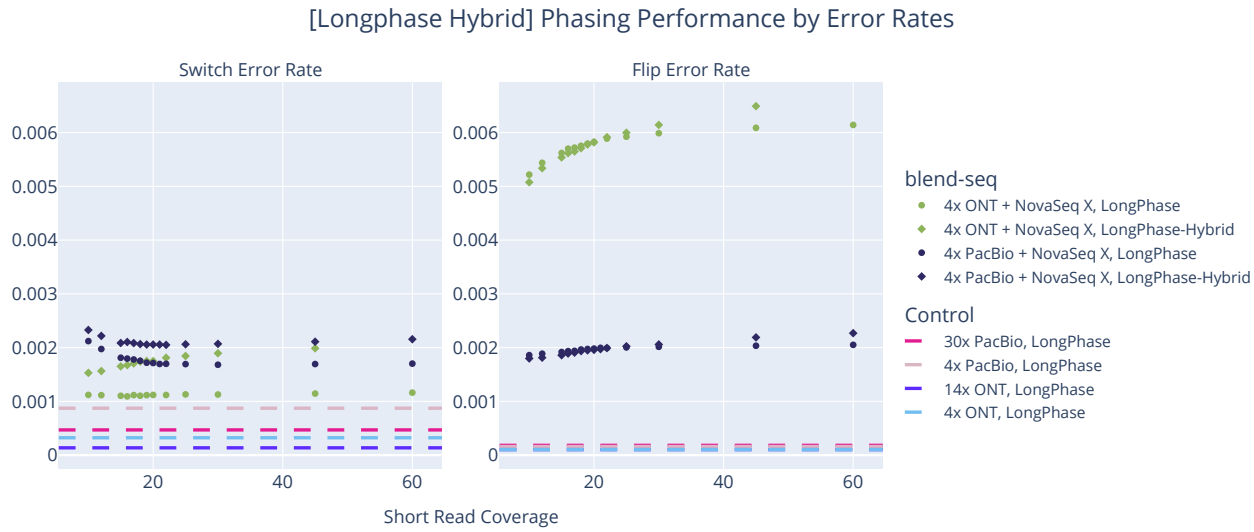

**Appendix A Supplementary Fig. 26.** Short-read SV genotyper performance for SV insertions in easy regions. Precision for SV insertions increases over the 4x long-read controls for the blend-seq experiments when using the genotypers, but comes at the cost of losing recall. This leads to a net decrease in F1 scores for SV insertions in easy regions compared to the 4x long-read controls.



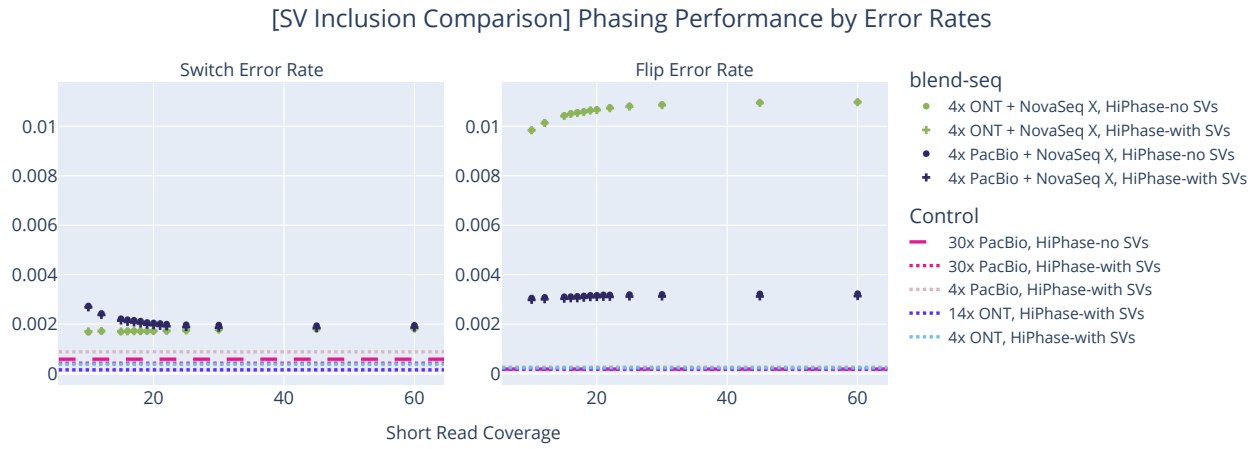

**Appendix A Supplementary Fig. 27.** Comparing DRAGEN native phasing performance to using WhatsHap phase. We see that WhatsHap phases 50% of the heterozygous variants (compared to DRAGEN's 12%). In addition, WhatsHap produces hundreds of fully-phased genes compared to the zero fully-phased by DRAGEN. Despite the higher rate of switch and flip errors, these are expected to occur more often with longer phase blocks, and the statistics for WhatsHap demonstrate a better attempt at phasing using the read data than the DRAGEN outputs.



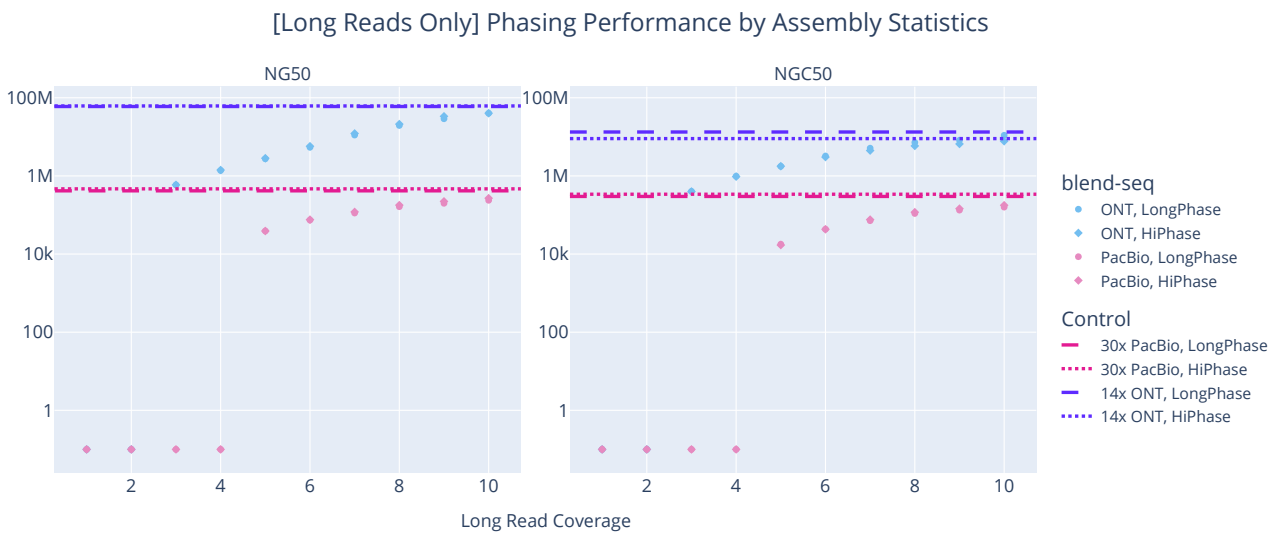

**Appendix A Supplementary Fig. 28.** LongPhase hybrid mode assembly phasing statistics comparison. Using the LongPhase hybrid mode, we see that the NG50 is mostly unchanged. However, the additional switch errors added (see Appendix A Fig. 29) causes the NGC50 to drop slightly. Values of 0 are rescaled to 0.1 to be visible on the log-scaled y-axis.



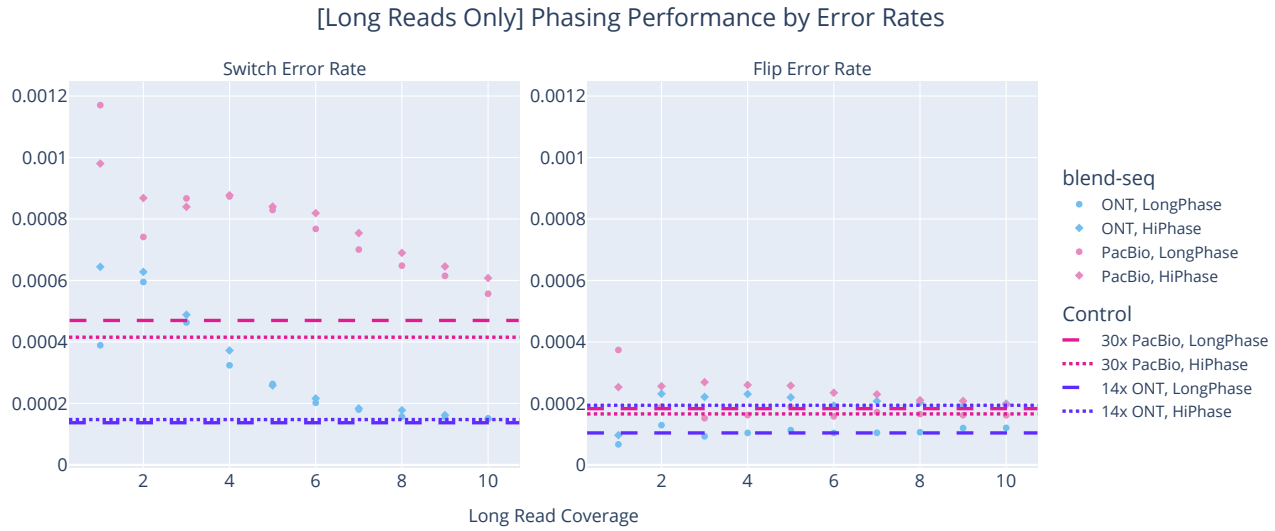

**Appendix A Supplementary Fig. 29.** LongPhase hybrid mode phasing error rates comparison. Using the LongPhase hybrid mode, we see that it provides more phased variants, but at the cost of significantly more switch errors. This supports the fact that the short reads are adding extra shorter blocks or extending existing ones but with a higher degree of inaccuracy.




**Appendix A Supplementary Fig. 30.** Comparison of phasing performance when including or excluding SVs from phasing process. We investigated the effects on phasing performance when including or excluding SVs from the input VCF for HiPhase (no difference was observed for LongPhase). There is very little difference in the error rates, but it is still a small improvement when including the SVs (especially when using the PacBio reads).




**Appendix A Supplementary Fig. 31.** Comparison of phasing assembly statistics performance using long reads alone at varying coverages. On the log-scaled shared axis, the NG50 and NGC50 statistics generally increase linearly in each layer of PacBio reads and expontentially for each layer of ONT reads, reaching close to saturation by 10x. Values of 0 are rescaled to 0.1 to be visible on the log-scaled y-axis.




**Appendix A Supplementary Fig. 32.** Comparison of phasing error rates using long reads alone at varying coverages. The switch error rate clearly decreases with extra layers of long reads, while the flip error rate follows a less defined trend either slightly downward (HiPhase) or roughly stable (LongPhase).