Whole-genome long-read sequencing downsampling and its effect on variant calling precision and recall

William T. Harvey[1], Peter Ebert[2,3,4], Jana Ebler[2,4], Peter A. Audano[5], Katherine M. Munson[1], Kendra Hoekzema[1], David Porubsky[1], Christine R. Beck[5,6], Tobias Marschall[2,4], Kiran Garimella[7], Evan E. Eichler[1,8*]

1. Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA, USA
2. Institute for Medical Biometry and Bioinformatics, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany
3. Core Unit Bioinformatics, Medical Faculty, Heinrich Heine University, Düsseldorf, Germany
4. Center for Digital Medicine, Heinrich Heine University, Düsseldorf, Germany
5. The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
6. Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT 06032 USA
7. Data Sciences Platform, Broad Institute of MIT and Harvard, Cambridge, MA, USA
8. Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA

*Corresponding author: Evan E. Eichler (eee@gs.washington.edu)

Running title: Downsampling with long-read sequencing

ABSTRACT

Advances in long-read sequencing (LRS) technology continue to make whole-genome sequencing more complete, affordable, and accurate. LRS provides significant advantages over short-read sequencing approaches, including phased *de novo* genome assembly, access to previously excluded genomic regions, and discovery of more complex structural variants (SVs) associated with disease. Limitations remain with respect to cost, scalability, and platform-dependent read accuracy and the tradeoffs between sequence coverage and sensitivity of variant discovery are important experimental considerations for the application of LRS. We compare the genetic variant calling precision and recall of Oxford Nanopore Technologies (ONT) and PacBio HiFi platforms over a range of sequence coverages. For read-based applications, LRS sensitivity begins to plateau around 12-fold coverage with a majority of variants called with reasonable accuracy (F1 score above 0.5), and both platforms perform well for SV detection. Genome assembly increases variant calling precision and recall of SVs and indels in HiFi datasets with HiFi outperforming ONT in quality as measured by the F1 score of assembly-based variant callsets. While both technologies continue to evolve, our work offers guidance to design cost-effective experimental strategies that do not compromise on discovering novel biology.

INTRODUCTION

Over the last five years, long-read sequencing (LRS) technologies have transformed the landscape of genetic variant discovery in two fundamental ways. First, they have increased the sensitivity of structural variant (SV) discovery by approximately threefold by providing access to repetitive regions of genomes typically masked or excluded as part of short-read sequencing analyses (Audano et al., 2019; Chaisson et al., 2015, 2019) and by providing breakpoint resolution of variants previously inferred by indirect read-pair or read-depth approaches (R. L. Collins et al., 2020). Second, LRS has enabled the routine generation of genome assemblies (Koren et al., 2017; Shafin et al., 2020), and recent advances in sequencing technology and methods are now routinely producing phased genome assemblies fully capturing both haplotypes (Cheng et al., 2021; Lorig-Roach et al., 2023; Porubsky et al., 2021). These advances have begun to improve our understanding of mutational processes, recurrent mutations, and new variants associated with disease and adaptation (Begum et al., 2021; Dutta et al., 2019; Hsieh et al., 2021; Miller et al., 2022; Porubsky et al., 2022).

Consequently, large-scale LRS efforts have enabled the construction of improved reference genomes including pangenomic representations of species (Liao et al., 2022) and exploration of the pattern of normal and disease variation across a variety of NIH initiatives in unprecedented detail, e.g., the *All of Us* (All of Us Research Program Investigators et al., 2019) and GREGoR (Chadwick & Chris Wellington, n.d.) programs. A critical question in such large-scale projects is the tradeoff between sensitivity and specificity for variant discovery as a function of genome coverage. This is especially important given that throughput and cost are still major limitations of LRS. In this study, we attempt to address this issue by comparing two of the most common platforms, Oxford Nanopore Technologies (ONT) and PacBio HiFi sequencing, as well as commonly used read-based and assembly-based variant callers. To establish a truth set for comparison, we analyze two deeply sequenced human genomes, HG00733 and HG002, with a specific focus on the recovery of SVs. Realizing that both LRS technologies and variant callers are under continuous development, this analysis is a snapshot in time that aims at informing experimental design to achieve high sensitivity and specificity within realistic economic boundaries.

RESULTS

Because LRS data can enable phased *de novo* assembly, we distinguish two LRS approaches for variant discovery: read-based and assembly-based methods. We define read-based methodologies as those requiring alignment of individual sequencing reads to a reference genome and applying specific read-based variant-calling algorithms to these alignments to identify variants. Assembly-based methods, in contrast, first generate *ab initio* a whole-genome assembly from LRS reads without guidance from a particular reference genome, and then proceed analogously by aligning this assembly to a reference genome to call variants using assembly-based calling algorithms. Many different tools implement variant-calling algorithms and they differ in their support for sequencing technologies (PacBio, ONT, etc.), variant types (SVs, indels, etc.), or data input (assembly, reads, etc.). In this study, we limit our analysis to eight read-based callers: Clair3 [v0.1-r11] (Zheng et al., 2021), cuteSV [v1.0.13] (Jiang et al.,

2020), DeepVariant [v1.3.0] (Poplin et al., 2018), Delly [v1.0.3] (Rausch et al., 2012), PEPPER-Margin-DeepVariant [r0.8] (Shafin et al., 2021), Sniffles [v2.0.2] (Smolka et al., 2022), PBSV [v2.8.0] (*Pbsv: Pbsv - PacBio Structural Variant (SV) Calling and Analysis Tools*, n.d.), and SVIM [v1.4.2] (Heller & Vingron, 2019), and two assembly-based callers: PAV [v1.2.2] (Ebert et al., 2021) and SVIM-asm [v1.0.2] (Heller & Vingron, 2020). Assemblies were generated considering three algorithms: hifiasm [v0.16.1] (Cheng et al., 2021), PGAS [v14-dev] (Ebert et al., 2021; Porubsky et al., 2021), and Flye [v2.9] (Kolmogorov et al., 2019).

We set out to determine how variant-calling performance differs depending on the platform, depth of sequence coverage (X), and computational method. For this assessment, we generated downsampled sets of HiFi and both standard and ultra-long ONT (UL-ONT) sequence data at depths of 5, 8, 10, 12, 15, 17, 20, 25, and 30X assuming a 3.1 Gbp haploid genome size. We applied standard practice algorithms and procedures and evaluated precision and recall of each algorithm for single-nucleotide variants (SNVs), small (<50 bp) indels (insertions and deletions), and SVs with respect to the human reference genome GRCh38. We consider two publicly available human genomes that have been sequenced extensively: HG002 (the Genome in a Bottle [GIAB] Ashkenazim child reference genome) (Wagner et al., 2022) and HG00733 (a Puerto Rican reference genome from the 1000 Genomes Project). In addition to GIAB analysis of HG002 (Zook et al., 2016), both genomes have been extensively characterized for genetic variants by both the Human Genome Structural Variation Consortium (HGSVC) (Ebert et al., 2021) and Human Pangenome Reference Consortium (HPRC) (Liao et al., 2022), which has led to the availability of thoroughly vetted variant callsets (Ebert et al., 2021) that are used in this study as truth sets (referred to as HGSVC Freeze 4). Both genomes have the advantage that they are targets of telomere-to-telomere (T2T) assembly development (Rautiainen et al., 2022) and, as such, more accurate and complete variant callsets will likely be available in the future to further refine truth sets for comparison. As both of these genomes have been characterized in multiple LRS efforts, sufficiently deep and high-quality input sets are available from both ONT and PacBio. For PacBio HiFi, these sets include 78.6X/17.9 kbp (depth/N50) and 99.54X/20.6 kbp for HG002 and HG00733, respectively. ONT standard length datasets were 153.4X/30.23 kbp and 92.3X/33.6 kbp and the UL-ONT data were 33.15X/96.4 kbp and 38.11X/132.7 kbp for HG002 and HG00733, respectively (**Supplemental Table S1**).

Read-based variant calling. Read-based SNVs were called with DeepVariant and Clair3 and showed the least variability between callers and technologies out of all three variant categories. At sequence read depth below 15X, recall of PacBio HiFi-tuned algorithms consistently outperformed ONT by an average of 0.03 (**Figure 1**). In fact, at ~10X coverage (current production from a single Sequel II SMRT cell) both precision and recall for HiFi data plateau while reaching a precision of 0.96 and recall of 0.90. At 5X coverage, DeepVariant and Clair3 showed on average 0.05 higher F1 scores in PacBio compared to ONT (**Supplemental Table S2**). This was demonstrated in both precision and recall with DeepVariant performing better with respect to precision and Clair3 with respect to recall. At coverage depths above 15X, the F1 score plateaued around 0.94 with recall being consistently higher than precision for all callers and technologies. The data suggest that HiFi is generally better with regard to recall but that 12X standard ONT and HiFi perform comparably. It should be noted that SNV calling for HG002

performed by GIAB has been subjected to extensive QC and specific regions are likely undercalled as reflected in the clustering of SNVs only observed in the LRS callsets. These clusters correspond to large blocks of highly identical segmental duplications, tandem repeats, and subtelomeric repeats (**Supplemental Figure S1**). In our analysis of 30X coverage datasets, we observe 639,007 SNV calls, which were not seen in GIAB for HG002. Of these 639,007 calls, 284,760 (44.56%) were observed by both ONT and HiFi suggesting true positive calls, though missing from the current GIAB set. This may help explain the precision plateaus at 90% across technologies and algorithms.

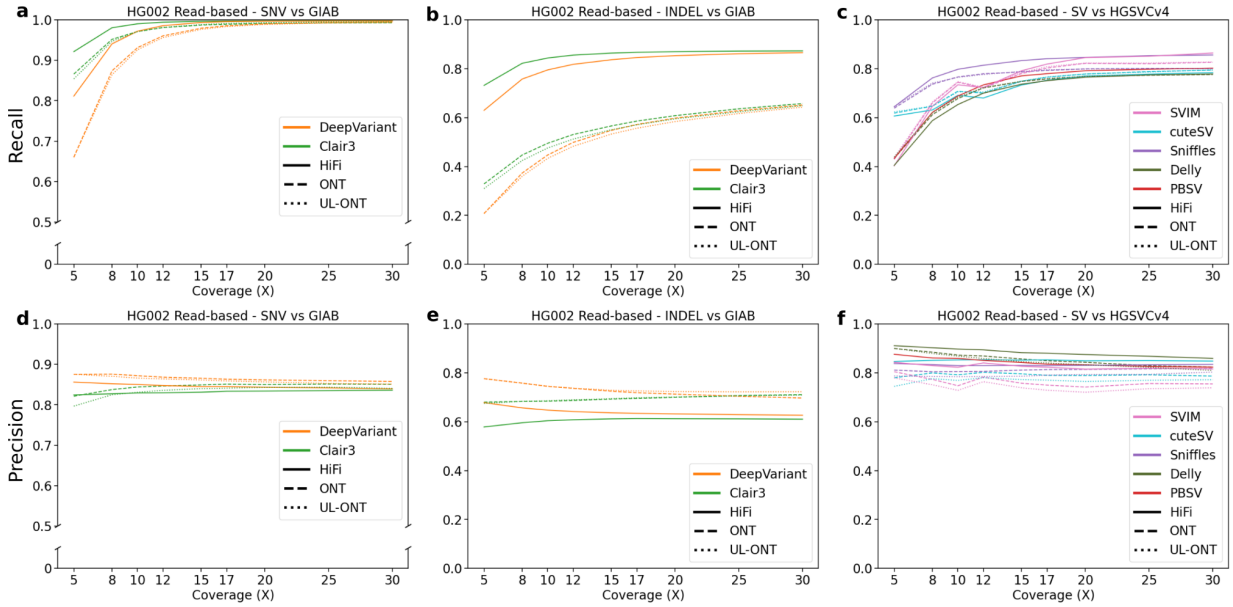

Figure 1. Precision and recall for variant classes as a function of LRS coverage using read-based algorithms for HG002. a) Recall of genome sample HG002 against Genome in a Bottle (GIAB) truth sets plotted against sequencing coverage for read-based callers Clair3 and DeepVariant. Clair3 with PacBio HiFi reaches the earliest recall plateau, while all callers show saturation by 20X. **b)** Recall against GIAB truth sets plotted against sequencing coverage for read-based callers across all algorithms capable of calling indels. Recall of both Clair3 and DeepVariant HiFi sets outperform their ONT counterparts. **c)** Recall against HGSVC truth sets plotted against sequencing coverage for read-based callers across all algorithms capable of calling structural variants (SVs). **d)** Precision as a function of sequence coverage. Single-nucleotide variant (SNV) precision remains flat beyond 10X, demonstrating the ability of callers to distinguish sequencing error from true SNVs. **e)** Precision plotted against sequencing coverage for read-based callers across all algorithms capable of calling indels. Precision values for all technologies and coverages remain flat, but here the increased precision of ONT callers is demonstrated. **f)** Precision plotted against sequencing coverage for read-based callers across all algorithms capable of calling SVs.

Indels, defined here as insertions or deletions less than 50 bp, show a similar profile. There is, once again, a characteristic plateau in F1 score around 12X sequence coverage. The greatest difference in recall is demonstrated in this subset between the HiFi and ONT platform (based on the R9 nanopore technology) (**Figure 1**). While precision remains high for ONT parameterizations of DeepVariant and Clair3 with an average of 0.82 across all measured depths, recall is noticeably lower when compared to PacBio HiFi reads, on average 0.39 less at

depths less than or equal to 12X and 0.31 above 12X (**Supplemental Table S3**). Interestingly, for this class of variant, ONT reads prepared with standard library prep consistently outperform their UL counterparts with respect to precision. We observe a mean precision difference of 0.03 at or below 12X and a 0.07 difference above 12X in favor of standard ONT. Overall, recall for indels is higher in HiFi datasets at all coverages, while ONT callers are more precise. A large amount of community development has gone into refining variant callers for ONT and has allowed these callsets to reduce noise inherent to less accurate ONT sequence reads.

For SVs, we consider only insertions and deletions greater than or equal to 50 bp. SVs show the least variability between technologies (F1 standard deviation of 0.01 between HiFi and ONT sequencing platforms (**Supplemental Table S4**)). Both sequencing platforms and various coverages converge on a set of ~12,800 SVs with each calling on average 25,634 SVs (**Figure 2**). Different read-based callers, however, show considerable variation. While recall remains low at lower sequence depth, mainly due to random sampling bias, two callers stand out as having the greatest precision: PBSV and Delly. Both callers consistently perform with high precision (mean 0.89) at low coverage depths and remain consistently high as depth increases. However, this does come with the above-mentioned tradeoff between precision and recall. As one increases, the other will decrease. In terms of recall at low-coverage sequence read depths below 12X, Sniffles performs best with a mean 0.63/0.84/0.71 precision/recall/F1 with cuteSV a close second (0.57/0.84/0.67).

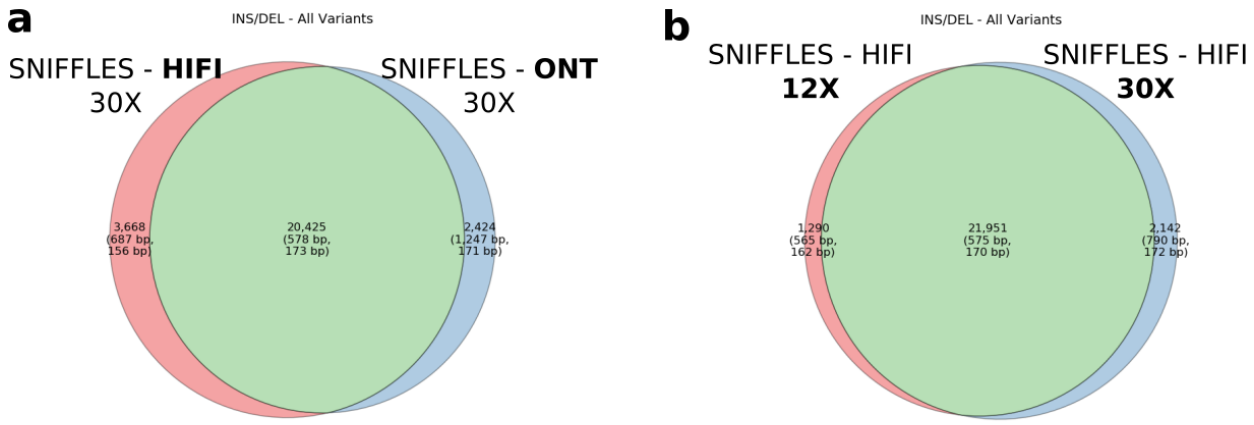

Figure 2. SV discovery. a) Venn diagram comparing Sniffles detection of SVs (both insertions and deletions) for 30X HiFi and 30X standard ONT input readsets. Of the variants unique to one technology or the other, 85% map to tandem repeat regions, which suggests breakpoint resolution rather than technology-specific bias is driving the difference. **b)** Venn diagram comparing Sniffles SV discovery at 12X and 30X HiFi callsets. A consistent set of calls is generated above 12X.

Assembly-based variant calling. Assembly-based callers have the advantage that they call variants from large contiguous haplotype blocks essentially providing access to larger and more complex forms of genetic variation and providing extended phasing for all forms of genetic variation (Wagner et al., 2021). We generated assemblies using three algorithms: hifiasm (v0.16.1), PGAS (v14-dev), and Flye (v2.9) where applicable. Hifiasm and PGAS assemblies were generated for the PacBio HiFi readsets, and Flye assemblies for the ONT reads. All

variants were called using the phased assembly variant (PAV) caller (Ebert et al., 2021) in addition to SVIM-asm specifically for SVs. It should be noted that the state of genome assembly for HiFi and ONT are not easily comparable. While HiFi reads can be assembled with numerous algorithms and assessed for phasing accuracy, ONT reads provide a greater challenge due to higher sequence error and fewer algorithms that combine both assembly and phasing. Methods such as Shasta (Shafin et al., 2020), wtdbg2 (Ruan & Li, 2020), and Canu (Koren et al., 2017) show considerable promise, but currently contiguous, haplotype-phased assemblies are not as easily generated and thus have not been utilized as frequently in current studies.

SNV calling with assembly-based callers generally underperforms read-based discovery especially at lower coverages. Precision in ONT and UL-ONT assembly-based methods shows the greatest difference with an average reduction of 0.33 across all sequencing depths (**Figure 3**). This is especially true in low-coverage (<12X) scenarios, and is driven by an excess of assembly-based SNV calls in ONT datasets (mean 8.33M in ONT; mean 10.00M in UL-ONT). PacBio HiFi methods have the opposite problem in that they underreport SNVs with a mean of 3.00M calls, although that does not greatly affect precision. This undercalling in HiFi assembly-based SNV callsets is a result of far less of the genome being assembled into haplotype-resolved contigs at lower coverages (**Figure 4**). However, when coverage reaches 12X, assembly-based methods show excellent recall (mean 0.96) for SNVs across all technologies (**Supplemental Table S5**) which mirrors the plateau observed in read-based methods. Below this threshold, read-based callers recall nearly 4X more (2,551 vs. 651) SNV windows based on recovery of over 90% of variants partitioned into 1 Mbp (**Figure 4**). Overall, SNV calling in low (less than 12X) coverage assemblies is not recommended, but coverages at or above 12X provide comparable precision and recall as their read-based counterparts with an average of 0.02 lower recall and 0.10 lower precision.

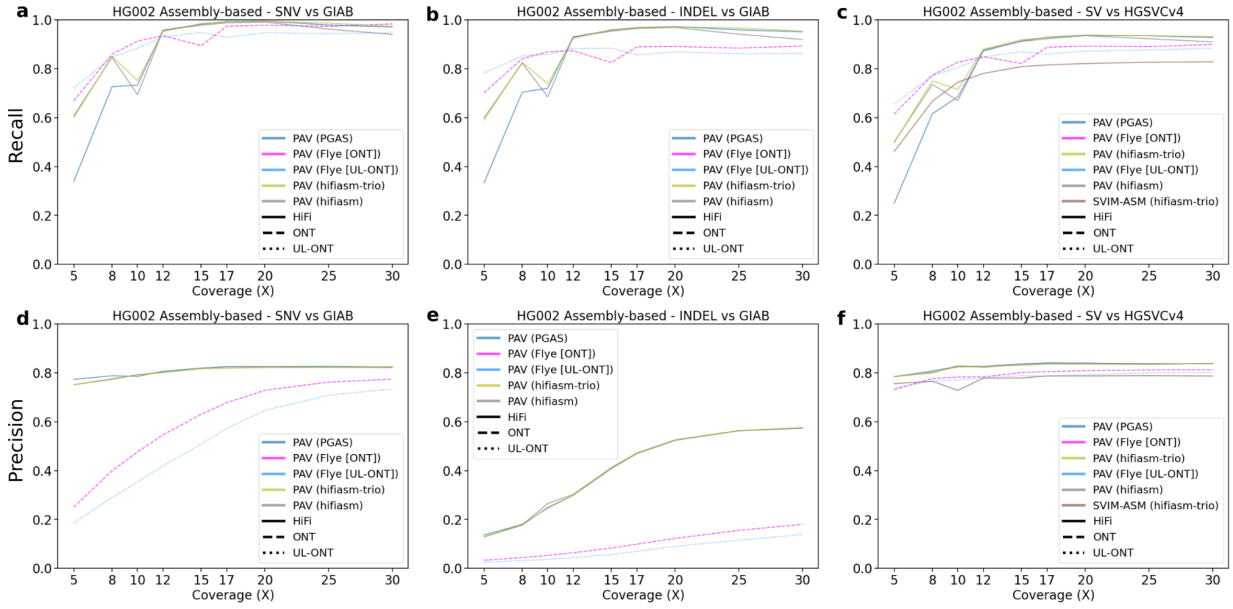

Figure 3. Precision and recall for variant classes as a function of LRS coverage using assembly-based algorithms for HG002. a) Recall for HG002 for GIAB truth sets plotted against sequencing

coverage for assembly-based callers across all algorithms capable of calling SNVs. **b)** Recall for HG002 against HGSVC truth sets plotted against sequencing coverage for assembly-based callers across all algorithms capable of calling indels. Recall in ONT assemblies performs better at low coverages before being surpassed by HiFi assemblies at 12X. **c)** Recall for HG002 against the HGSVC Freeze 4 truth set plotted against sequencing coverage for assembly-based callers across all algorithms capable of calling SVs. **d)** Precision for HG002 against HGSVC truth sets plotted against sequencing coverage for read-based callers across all algorithms capable of calling SNVs. ONT methods are comparable to HiFi precision at high coverages but are noticeably worse at coverages below 15X. **e)** Precision plotted against sequencing coverage for assembly-based callers across all algorithms capable of calling indels. Like read-based methods, values for all technologies and coverages remains low, likely due to the incomplete nature of indels in complex regions in the GIAB truth set. **f)** Precision plotted against sequencing coverage for assembly-based callers across all algorithms capable of calling SVs.

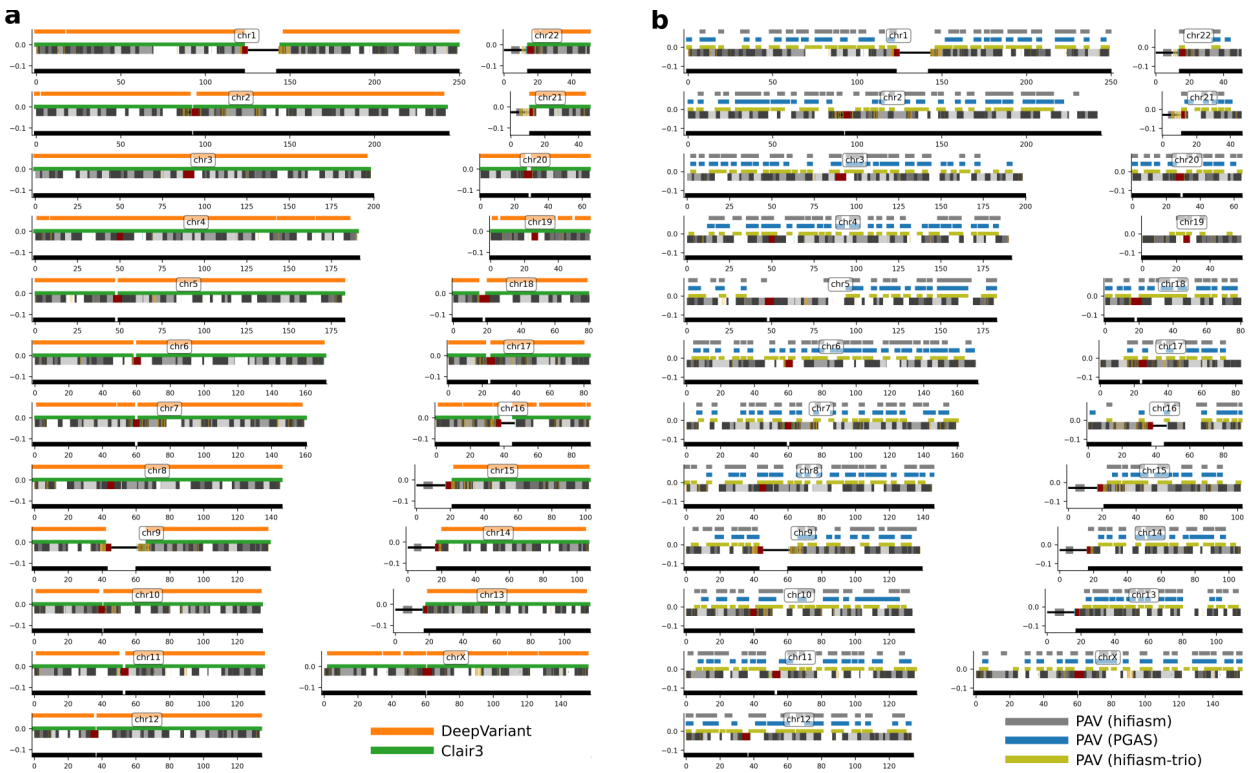

Figure 4. Ideogram comparison of autosomal SNV recall at 8X for PacBio HiFi. a) PacBio HiFi (8X) read-based recall of HG002 SNVs against GIAB truth sets. A bar over a chromosome depicts a 1 Mbp window where there was >90% SNV recall for Clair3 (green) and DeepVariant (orange) with all regions where SNV were called in black. **b)** PacBio HiFi (8X) assembly-based recall of HG002 GIAB SNV truth set using PAV. There are fewer 1 Mbp windows with >90% recall irrespective of assembly algorithm including hifiasm-trio (yellow), PGAS (blue), or hifiasm (gray). The black bar under the chromosome represents 1 Mbp windows with SNVs in the truth set.

Detecting indels from assembly-based methods is especially challenging (**Figure 3**), in part due to the known LRS error profiles associated with indels of smaller motif sizes (Delahaye & Nicolas, 2021; Wenger et al., 2019). Inability to correct these errors at low sequencing depth significantly inflates indel counts (1,145,880 indel insertion calls on average in PacBio HiFi 5X vs. 444,045 indel insertion calls in PacBio 30X). As such, precision is lowest for indels called in

assemblies below 12X (**Supplemental Table S6**). In ONT datasets, this issue is exacerbated by an order of magnitude at reduced coverages (8,105,758 at 5X) and remains problematic even at high coverage (1,137,763 at 30X). Precision estimates, however, may be underestimated due to the limited capability of Illumina to detect variation in more complex regions of the genome that were not accessible to the GIAB truth set. Additional development and orthogonal validation of indels should be an active area of LRS technology development.

SVs follow the trend of assembly-based callsets in general with a steep recall curve, steady precision curve, and early plateau across sequencing depths and technologies. For low (below 8X) HiFi coverages, assembly-based methods underperform their read-based counterparts with respect to recall by an average of 0.03 (**Supplemental Table S7**). While ONT assemblies demonstrate higher recall than their read-based counterparts by 0.09 and 0.10 for standard ONT and UL-ONT, respectively. Above this coverage, all assembly-based methods outperform read-based methods by at least 0.08 for recall. The HG002 assemblies using PacBio HiFi reads at 10X sequencing depth are a clear outlier and may be attributable to a systematic failure to remove false duplications. We did not observe a similar outlier in HG00733. Although the assembly size is larger than expected, metrics such as contiguity (N50) and callable loci are consistent with other assemblies. Similar outliers may be avoidable with deeper coverage to support high-quality assembly-based callsets (Ebert et al., 2021; Liao et al., 2022).

Cross-callset comparisons. Because LRS technologies claim to access more of the genome and more complex classes of genetic variants, we first evaluate genome-wide SV callability. To assess callability across the genome, we first divided GRCh38 into 1 Mbp windows and intersected those windows with the HGSVC SV truth set for HG00733, yielding 2,679 and 2,482 windows for insertions and deletions, respectively. In order for a window to be established as callable, >90% of the calls contained in this window must be accurately recovered (**Figure 5**). At low coverages (5X), read-based methods outperform assembly-based methods for each respective technology. At these low coverages, Sniffles used with HiFi reads performs the best, recovering 1,118/2,482 (45%) windows when considering deletion calls. This is almost double the PacBio HiFi callable windows for assembly-based methods. This trend holds for insertions, but we do note that Flye assembly-based methods using UL-ONT perform better than Sniffles on HiFi reads. At 10X and above, the pattern switches with HiFi assembly methods outperforming all read-based callers with the starkest difference occurring at 15X where assembly-based methods recover an additional 500 Mbp and 383 Mbp of the genome (for insertions and deletions, respectively) than read-based methods.

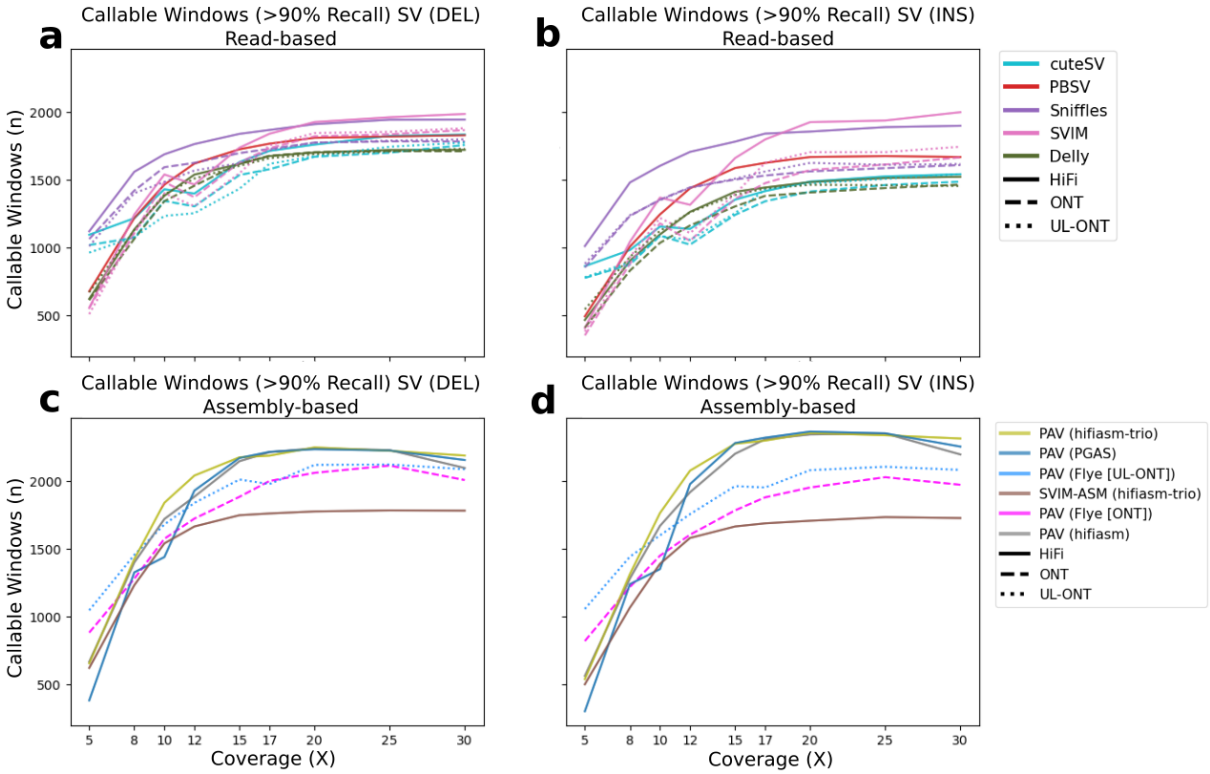

Figure 5. Evaluation of SV callable bases by technology and algorithm. Read-based callable windows for **(a)** insertions and **(b)** deletions, and assembly-based callable regions for **(c)** insertions and **(d)** deletions. Regions were compared against the HGSVC HG00733 truth set in 1 Mbp windows requiring at least 90% recall.

Clinical SVs in HG002. A list of clinically relevant SVs was released for the GIAB sample HG002 (Wagner et al., 2021) including 273 challenging genes or regions that map to repetitive and structurally complex polymorphic regions. At 30X coverage, PBSV was able to recover 97% of these SVs in clinically relevant genes (**Supplemental Table S8**). However, at the lowest coverage depths, Sniffles, once again, drastically outperformed the other callers across all technology types, but especially with PacBio HiFi reads where it reports recall of 0.87 and 0.82 for SV insertions and deletions, respectively, at just 8X sequencing coverage. Compared to read-based methods, assembly-based methods demonstrated lower recall at low coverages with a max of 0.72 for insertions and 0.79 for deletions using Flye with UL-ONT and hifiasm (non-trio binned), respectively (**Supplemental Table S9**).

Tandem repeat characterization. LRS technologies allow for more robust characterization of tandem repeats (Chaisson et al. 2015; Chaisson et al. 2019; Pendleton et al. 2015; Sedlazeck et al. 2018), the largest of which are known as variable number of tandem repeats (or VNTRs). After SNVs, tandem repeat variants are among the most abundant forms of human genetic variation comprising >20% of indels and >50% of SVs (Ebert et al., 2021) (**Supplemental Table S10**). Excluding these regions from analysis has little effect on recall, indicating that even though these regions have been difficult to characterize in prior studies, most LRS technologies and algorithms are able to detect these variants despite ambiguity in defining the exact breakpoints. However, inclusion of these regions potentially comes with a tradeoff in precision,

particularly with read-based methods where we saw precision increase when we exclude tandem repeats at a consistent rate of 0.04 compared to precision in assembly-based callers, which were more precise in lower coverage scenarios with tandem repeats included (**Supplemental Figure S2**). This indicates that even at low coverages, assembly-driven variant calling can characterize such variation.

Performance in homopolymer DNA. Accurately calling variants in homopolymer runs is challenging for both PacBio HiFi and ONT technologies (G. A. Logsdon et al., 2020; Mc Cartney et al., 2021; Shafin et al., 2021). These nonrandom error profiles impact precision and recall, especially for indel variant calls. When comparing the difference between all indel calls annotated with and without homopolymers, ONT callsets display a large difference between homopolymer and non-homopolymer DNA sequence precision and recall (**Supplemental Figure S3**). Even at high coverages, recall for insertions in homopolymer sequence is as much as 0.10 lower than when compared against the whole set. Notably, the effect that these sequence types have on precision even at higher depths is still prevalent with even 30X read-based methods showing a decrease of 0.06 between these regions. DeepVariant calls for UL-ONT reads show a decrease in homopolymer precision as sequencing depth increases. This could be due to a prior lack of training data with a ground truth for complex genomic regions uniquely aligned by this technology.

Large variant discovery. Large (>10 kbp) SVs, especially insertions within or near repeat regions, frequently evade Illumina detection (Medvedev et al., 2009). An advantage of LRS technologies is that these events can be detected directly from the sequence of the reads or the assembly themselves. We assessed each method's ability to recover large variants using the HGSVC validation set from HG00733 including 63 deletions and 40 insertions. For HiFi reads, two trends emerge: their limitation in detecting large insertions compared to ONT reads and their increased recall when assembled even at low coverages. HiFi reads consistently lag behind their ONT counterparts for large insertions, recovering only half of the insertions in standard ONT callsets and a third of the insertions detected in UL-ONT (**Supplemental Table S11**). However, by assembling these reads, HiFi datasets outperform ONT when sequence coverage exceeds 8X. Among read-based methods, UL-ONT performs the best with a minimum of 21/63 large deletions and 15/40 large insertions detected even at low sequence coverages (5X). Across all read-based algorithms, Sniffles recovers the greatest number of large events with a maximum of 0.67 and mean of 0.51 recall over all input types and coverages followed by cuteSV with 0.65 and 0.41, respectively. It should be noted that Delly failed to call any SVs above 10 kbp. HiFi assembly-driven methods perform the best overall with a maximum large variant recall of 0.87 and a mean of 0.65 when PAV is used (**Supplemental Table S12)**. Finally, it should be noted that both read-based and assembly-based methods recovered the largest (238 kbp) deletion, but only assembly-based methods identify the largest insertion of 51 kbp compared to the maximum event size in read-based methods of 32 kbp.

ONT duplex reads and Revio HiFi data. PacBio and ONT are rapidly developing new sequencing technologies that improve LRS accuracy and throughput. For example, ONT recently released an improved flowcell (R10) as well as a new "duplex" sequencing method

(*Oxford Nanopore Tech Update: New Duplex Method for Q30 Nanopore Single Molecule Reads, PromethION 2, and More*, n.d.) significantly improving individual read accuracy by sequencing both forward and complementary strands from the same single molecule (Sanderson et al., 2023). The new release of the Revio system from PacBio, in contrast, significantly increases throughput and affordability using a chemistry similar to that of the Sequel II platform (i.e., HiFi sequencing). The recent release of whole-genome sequencing (WGS) datasets from the GIAB sample HG002 allows these new emerging LRS platforms to be compared. We analyzed a 30X duplex dataset of WGS data released by ONT and compared precision and recall to standard ONT using R9.4.1 flowcells. We find that variant-calling recall for specific variant classes is substantially improved for duplex sequencing over R9 ONT variant calling at all sequence coverages and for all variant classes. The effect is most pronounced for indel recall at low coverage (≤10X) where duplex variant recall improves by 0.19 (**Figure 6**) when compared to standard ONT. Precision, however, is much more consistent with standard ONT methods. Of note, in our analysis, the precision of indel insertions actually diminishes when compared to standard ONT (an average of 0.06 reduction). This is possibly due to parameterization of variant-calling algorithms which have been largely adjusted for calling in a noisier, more error-prone, single-strand ONT signal.

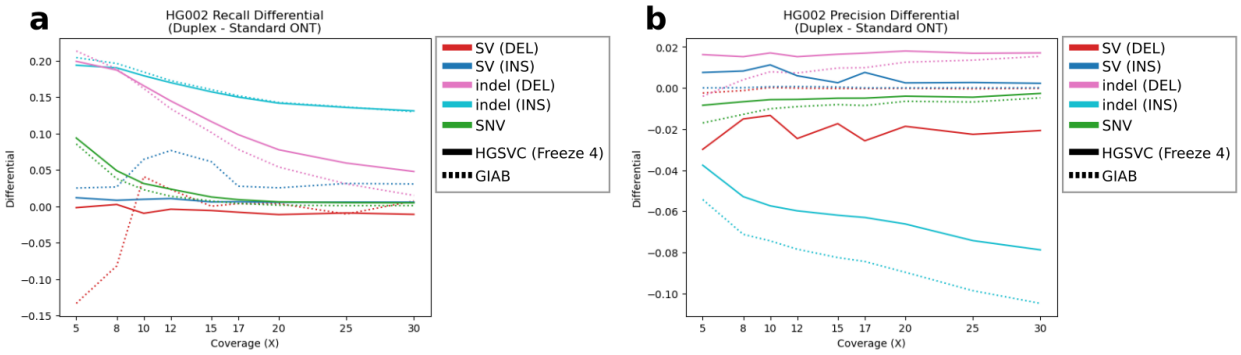

Figure 6. Comparison of precision and recall in duplex ONT variant calling versus standard ONT. Duplex ONT versus standard ONT variant calling **(a)** precision and **(b)** recall where anything above the y=0 line indicates an increase in performance compared to standard ONT and anything below the y=0 line indicates an decrease in performance compared to standard ONT.

Using 30X of WGS data from HG002 generated by the Revio system (*PacBio Revio*, 2022), we also constructed a phased human genome assembly using hifiasm. The results were nearly identical to an assembly produced from a Sequel II HiFi dataset, albeit with single flowcell. Both the contiguity (contig N50 = 44 Mbp [Revio] vs. 45 Mbp [Sequel II]) and accuracy (QV=57 [Revio] vs. 55 [Sequel II]) were virtually identical. Predictably, assembly-based variant calling were comparable for both recall (Pearson R = 0.984) and precision (Pearson R = 0.997) with some modest improvements in SNV recall (+0.02 vs. both truth sets) and small insertion precision (+0.06 vs. HGSVC Freeze 4) (**Table 1**).

Table 1. Revio versus Sequel II assembly-based callset comparison

SAMPLE	SVTYPE	TRUTH SET	RECALL (HiFi)	RECALL (REVIO)	RECALL DIFF	PRECISION (HIFI)	PRECISION (REVIO)	PRECISION DIFF
HG002	SV (ins)	Freeze 4	0.94	0.91	-0.03	0.823	0.865	0.042
HG002	SV (del)	Freeze 4	0.927	0.901	-0.026	0.869	0.859	-0.01
HG002	SNV	GIAB	0.974	0.998	0.024	0.825	0.811	-0.015
HG002	SNV	Freeze 4	0.97	0.992	0.022	0.897	0.879	-0.018
HG002	indel (ins)	GIAB	0.955	0.944	-0.012	0.549	0.584	0.036
HG002	indel (ins)	Freeze 4	0.959	0.971	0.012	0.706	0.77	0.064
HG002	indel (del)	GIAB	0.953	0.947	-0.006	0.605	0.598	-0.006
HG002	indel (del)	Freeze 4	0.965	0.986	0.022	0.776	0.789	0.014

*PAV assembly-based variant-calling comparison for WGS data generated for HG002 on a Revio system compared to the 30X downsampled HG002 generated via the Sequel II platform compared to the HGSVC truth set (Freeze 4) and Genome in a Bottle (GIAB).

DISCUSSION

Within the limits of various algorithms and sequencing platforms analyzed here, we make a few general observations and recommendations based on our analysis against current truth sets (Ebert et al., 2021; Zook et al., 2016). With respect to SNV discovery, LRS coverage in excess of 12-fold begins to show a plateau with respect to sensitivity. Read-based approaches such as Clair3 (Zheng et al., 2021) and DeepVariant (Poplin et al., 2018) significantly outperform assembly-based detection methods, such as PAV, which have been geared to improve SV discovery and breakpoint definition (Audano et al., 2019; Ebert et al., 2021). While Clair3 with PacBio HiFi performs the best for SNVs, both deep convolutional network approaches (Clair3 and DeepVariant) show excellent recall with both ONT and PacBio above 20X sequence. Irrespective of the sequencing platform, sequence coverage at 8X or lower shows significant reduction in performance and is not advised for large-scale sequencing projects dedicated to variant discovery.

By contrast, all LRS platforms currently underperform for indel variant calling and, predictably, they perform the most poorly in regions of homopolymer runs as well as short tandem repeats—precisely the regions that are most mutable for this class of variation (Willems et al., 2014). Given that caveat, we would recommend PacBio HiFi read-based methods for recall (0.69 vs. 0.61) across all read coverages and ONT for precision, although the difference is slight (0.68 vs. 0.66 mean precision for ONT vs. HiFi, respectively). A major challenge facing human genetics is the existence of a well-vetted and complete truth set for indel variants—detailed studies over the years have restricted analyses to specific regions of the genome owing to the high rate of false positives and false negatives from more mutable and difficult-to-sequence regions (Krusche et al., 2019; Olson et al., 2022; Zook et al., 2019). Our results suggested that haplotype-resolved assemblies offer some improvement for recall (an average of +0.14 across all coverages).

Completely sequenced and assembled genomes where T2T chromosomal assemblies are established along with vetted indel callsets by multiple sequencing technologies (e.g., Sanger, Illumina, ONT, and PacBio) will be required to develop a more comprehensive truth set of indels for benchmarking. Resources such as the Platinum pedigree (CEPH pedigree 14633) by Illumina will be particularly useful as they enable studying phased genome assemblies and variant calling in the context of transmission within families (Eberle et al., 2017).

Both ONT and HiFi PacBio excel at SV detection, routinely detecting >20,000 SVs and consistently calling the same variants when sequence coverage exceeds 12X (**Figure 3**). In fact, approximately 85% of SVs in 30X datasets that are unique to one platform over another map to tandem repeat regions but are in close proximity (<10 kbp) and their size overlap suggests that differences in alignment and breakpoint definition are still potentially more rate-limiting as opposed to platform differences in sensitivity. The advance of LRS for SV detection when compared to Illumina WGS has been well established over the years (Chaisson et al., 2015, 2019; Sedlazeck et al., 2018; Shafin et al., 2021) and more sophisticated callers as well as computational pipelines continue to be developed to discover and characterize SVs as part of routine callsets (Kolmogorov et al., 2023). While ONT, and especially UL-ONT, performed well for detecting large insertions (**Supplemental Table S11**), overall, assembly-based approaches (especially hifiasm) showed the greatest specificity and precision when calling large SVs (>50 kbp) (**Supplemental Table S12**). Because large SVs are much more likely to have phenotypic consequence and precise breakpoints are relevant to the effect of this consequence, assembly-based strategies should strongly be considered when applying LRS to solving cases of Mendelian and *de novo* disease (Miller et al., 2021). However, generation of phased genome assemblies requires deeper sequencing coverage (at least 15-20X) and, as such, is still a more expensive option.

In summary, when deciding LRS depth targets, the intended purpose of the project must be considered. If the goal is recovery and characterization of SNVs at a population scale, low-depth read-based methods will provide the right balance of maximizing discovery and number of samples in the study. However, if the goal is sequence resolution of large and complex variants at the level of individual patients, assembly-based methods, in particular hifiasm, are currently one of the most accurate strategies for building phased genome assemblies but require greater investment in terms of sequence coverage (well beyond 15X) and computational processing. Importantly, the LRS platforms continue to rapidly evolve in terms of accuracy (ONT) and throughput (PacBio). Improved modeling of the platform-dependent errors as well as newer pores, or techniques (duplex sequencing) for ONT show considerable promise with suggestions that sequencing accuracy may in fact rival or surpass that of Illumina (Kolmogorov et al., 2023). Changes such as duplex sequencing with the R10 pore, however, currently come at a cost of lower throughput (Sanderson et al., 2023) and, as a result, added expense to achieve deep coverage. For the last three years, PacBio HiFi has dominated the field with respect to accuracy in large part due to the advent of circular consensus sequencing (CCS); however, multiple flowcells have been required to achieve deep sequence. The release of the new Revio platform earlier this year significantly increases throughput and decreases costs which will aid production of high quality and contiguous assemblies comparable to that of those generated previously by

multiple Sequel II flowcells. Both platforms are currently highly complementary. Recently, algorithms that aim to incorporate the strengths of both PacBio HiFi and ONT reads to generate *de novo* T2T assemblies have shown very promising results (Rautiainen et al., 2022). Such hybrid technology approaches have the potential to supplant any single LRS technology as soon as the costs drop and the production of LRS assemblies become routine. The benefit of complete T2T variant discovery should not be underestimated.

METHODS

LRS datasets and data availability

ONT data generation: UL-ONT data were generated from the HG00733 lymphoblastoid cell line according to a previously published protocol (G. Logsdon, 2022). Briefly, 3-5 x 107 cells were lysed in a buffer containing 10 mM Tris-Cl (pH 8.0), 0.1 M EDTA (pH 8.0), 0.5% w/v SDS, and 20 ug/mL RNase A (Qiagen, 19101) for 1 hour at 37°C. 200 ug/mL Proteinase K (Qiagen, 19131) was added, and the solution was incubated at 50°C for 2 hours. DNA was purified via two rounds of 25:24:1 phenol-chloroform-isoamyl alcohol extraction followed by ethanol precipitation. Precipitated DNA was solubilized in 10 mM Tris (pH 8.0) containing 0.02% Triton X-100 at 4°C for two days. Libraries were constructed using the Ultra-Long DNA Sequencing Kit (ONT, SQK-ULK001) with modifications to the manufacturer's protocol. Specifically, ~40 ug of DNA was mixed with FRA enzyme and FDB buffer as described in the protocol and incubated for 5 minutes at RT, followed by a 5-minute heat-inactivation at 75°C. RAP enzyme was mixed with the DNA solution and incubated at RT for 1 hour before the clean-up step. Clean-up was performed using the Nanobind UL Library Prep Kit (Circulomics, NB-900-601-01) and eluted in 225 uL EB. 75 uL of library was loaded onto a primed FLO-PRO002 R9.4.1 flowcell for sequencing on the PromethION, with two nuclease washes and reloads after 24 and 48 hours of sequencing.

PacBio HiFi data generation: PacBio HiFi data were generated from the HG00733 lymphoblastoid cell line as previously described (G. A. Logsdon et al., 2021) with modifications. Briefly, DNA was extracted from 4.3x10^6 cells using the Monarch HMW DNA Extraction Kit for Cells and Blood (New England Biolabs) with 1400 rpm lysis speed. After UV absorption and fluorometric quantification (Qubit High Sensitivity DNA kit, Thermo Fisher) on the DS-11 FX instrument (Denovix) and evaluation of DNA integrity on FEMTO Pulse (Agilent), 12 µg of DNA was prepared for sequencing using Megaruptor 3 shearing (Diagenode, settings 19/31) and the Express Template Prep Kit v2 and SMRTbell Cleanup Kit v2 (PacBio). The library was size-selected on a PippinHT instrument (Sage Science) using a 15 kbp high-pass cut. Five SMRT Cell 8Ms were run on a Sequel II instrument using Sequel II chemistry C2.0/P2.2 with 30-hour movie times, 2-hour pre-extension, and adaptive loading targets of 0.8-0.85 (PacBio). Circular consensus calling was performed with CCS version 6.0.0 (SMRT Link v.10.1) and reads with estimated quality scores ≥Q20 were selected for downstream analysis.

Reference genome and reliable regions: To support long-read mapping, only the primary GRCh38 assembly was used, which includes chromosome scaffolds, the mitochondrial assembly, unplaced contigs, and unlocalized contigs. No alts, patches, or decoys were present in the assembly during the alignment stages. This reference was used previously (Audano et al., 2019; Ebert et al., 2021) and is available for download here: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/reference/20200513_hg38_NoALT/. Whole-genome analysis was restricted to regions outside centromeres, pericentromeric repeats, and the mitochondrial chromosome where variant calls were previously determined to be less reproducible (Audano et al., 2019; Ebert et al., 2021). This is available here: http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/technical/filter/20210127_LowConfidenceFilter/

Downsampling: In-house python scripts were utilized to read in indexes for our input datasets and subsample reads randomly up to the desired threshold. We then used SAMtools fqidx to extract the desired reads from our larger sets and partitioned them into individual bins.

Whole-genome alignment: ONT and PacBio reads were aligned with minimap2 v2.21 (Li, 2018). Specifically, ONT reads were aligned with:

```
minimap2 -ax map-ont --MD --secondary=no --eqx -x -I 8G {input.ref} {input.read}
```

PacBio HiFi reads were aligned with:

```
minimap2 -ax map-pb -I 8G {input.ref} {input.read}
```

Assemblies: We employed two approaches to generate phased whole-genome assemblies for all PacBio HiFi sampling depths: we used the PGAS pipeline as previously described (parameter settings v14-dev, (Ebert et al., 2021; Porubsky et al., 2021), hifiasm v0.16.1), which does not rely on parental data to derive genome-wide phase information. Additionally, we executed hifiasm v0.16.1 (Cheng et al., 2021) with default parameters in trio-binning mode, leveraging parental short reads to obtain phase information. For the ONT and UL-ONT readsets, we implemented a two-step process employing first the Flye assembler v2.9 (Kolmogorov et al., 2019) to generate unphased whole-genome assemblies with default parameters (preset “--nano-hq” and “--genome-size” of 3.1 Gbp). Next, these assemblies were converted into diploid assemblies using the HapDup v0.6 tool (Kolmogorov et al., 2019; Shafin et al., 2020) with default parameters (preset “ont”).

Read-based variant calling: We used Clair3 [v0.1-r11] (Zheng et al., 2021) cuteSV [v1.0.13] (Jiang et al., 2020), DeepVariant [v1.3.0] (Poplin et al., 2018), Delly [v1.0.3] (Rausch et al.,

2012), PBSV [v2.8.0] (*Pbsv: Pbsv - PacBio Structural Variant (SV) Calling and Analysis Tools*, n.d.), PEPPER-Margin-DeepVariant [r0.8] (Shafin et al., 2021), Sniffles2 [v2.0.2] (Smolka et al., 2022), and SVIM [v1.4.2] (Heller & Vingron, 2019) in order to call SVs from the aligned PacBio HiFi, ONT, and UL-ONT reads at the different coverage levels.

The commands used for each caller and technology are listed below:

Clair3:

(PacBio HiFi)

```
run_clair3.sh --bam_fn={input.merged_bam} --sample_name={sample} --
ref_fn={input.ref} --threads={threads} --platform=hifi --
model_path=$(dirname $( which run_clair3.sh ) )/models/hifi --
output=$(dirname {output.vcf}) --enable_phasing
```

(ONT|UL-ONT)

```
run_clair3.sh --bam_fn={input.merged_bam} --sample_name={sample} --
ref_fn={input.ref} --threads={threads} --platform=ont --
model_path=$(dirname $( which run_clair3.sh ) )/models/ont_guppy5 --
output=$(dirname {output.vcf}) --enable_phasing
```

cuteSV:

(PacBio HiFi)

```
cuteSV -t {threads} -S {sample} --max_cluster_bias_INS 1000 --
diff_ratio_merging_INS 0.9 --max_cluster_bias_DEL 1000 --
diff_ratio_merging_DEL 0.5 {input.reference} {output.vcf} --genotype
-l 50 -s {params.min_supp} {params.outputdir}
```

(ONT|UL-ONT)

```
cuteSV -t {threads} -S {sample} --max_cluster_bias_INS 100 --
diff_ratio_merging_INS 0.3 --max_cluster_bias_DEL 100 --
diff_ratio_merging_DEL 0.3 {input.reference} {output.vcf} --genotype
-l 50 -s {params.min_supp} {params.outputdir}
```

In addition, we filtered the cuteSV calls based on the minimum read support reported in the output VCF, as it generated unfiltered calls. Similarly, we filtered the SVIM calls based on the reported quality. In both cases, we used value 2 for coverages ≤5; 3 for coverages ≤10; 4 for

coverages ≤20; 5 for coverages ≤25; and 10 for coverages >30. These values were selected such that they result in the highest F-scores when comparing the filtered calls to the GIAB medically relevant SVs for HG002. The pipeline used for SV calling with cuteSV, Sniffles2, and SVIM can be found here: https://github.com/eblerjana/lrs-sv-calling.

DeepVariant:

(PacBio HiFi)

```
run_deepvariant --model_type=PACBIO --ref={ref} --reads={aln} --
output_vcf={sample}.vcf.gz --output_gvcf={sample}.gvcf --
novcf_stats_report --intermediate_results_dir=/dv_tmp/ --
num_shards={threads}
```

(ONT-duplex)

```
run_deepvariant --model_type=ONT_R10 --ref={ref} --reads={aln} --
output_vcf={sample}.vcf.gz --output_gvcf={sample}.gvcf --
novcf_stats_report --intermediate_results_dir=/dv_tmp/ --
num_shards={threads}
```

Delly:

(PacBio HiFi)

```
delly lr -y pb -g {input.ref} -x {input.exc} -o {output.bcf} {input.bam}
```

(ONT|UL-ONT)

```
delly lr -y ont -g {input.ref} -x {input.exc} -o {output.bcf} {input.bam}
```

Excluded regions for Delly can be found here:
https://github.com/dellytools/delly/blob/main/excludeTemplates/human.hg38.excl.tsv

PBSV:

(PacBio HiFi)

```
pbsv discover --tandem-repeats {input.trf} {input.bam} {output.svsig}
pbsv call -j {threads} --ccs --types DEL,INS,INV {input.ref} {input.svsig}
{output.vcf}
```

PEPPER-Margin-DeepVariant:

(ONT|UL-ONT)

```
run_pepper_margin_deepvariant call_variant -b {bam} -f {ref} -o {out_dir} -t {threads} --ont_r9_guppy5_sup
```

Sniffles:

(PacBio HiFi|ONT|UL-ONT)

```
sniffles -t {threads} -i {input.bam} -v {output} --reference {input.reference} --minsvlen 50
```

SVIM:

(PacBio HiFi|ONT|UL-ONT)

```
svim alignment {params.outdir} {input.bam} {input.reference} --min_sv_size 50
```

Assembly-based variant calls: PAV (Ebert et al., 2021) was applied to phased assemblies using default parameters. Briefly, assemblies were mapped to the GRCh38 reference genome with minimap2 2.17 (Li, 2018), alignment trimming was performed to eliminate redundantly mapped bases, and variant calling was performed to detect variants within alignments as well as large SVs that fragmented alignment records into multiple parts.

Variant merging and annotations: Variant call comparisons were performed using svpop. SNV-based comparisons were performed using the overlap feature nrid (nonredundant ID match), which requires variants to have the same SNV ID (#CHROM-POS-SNV-REF-ALT) to be called the same. Additionally, indels and SVs were matched using szro-50-200, which first matches variants on ID (#CHROM-POS-SVTYPE-SVLEN), then 50% reciprocal overlap, and then finally variants of the same type that are within 200 bp of each other and have reciprocal size overlap of 50%. This strategy allows for increased accuracy in complex regions of the genome where alignments can be biologically ambiguous.

Reference-based annotations for genomic sequence content (e.g., homopolymer, TRF) are taken directly from the UCSC Genome Browser and the UCSC GoldenPath. This is a built-in functionality of SVPOP for GRCh38.

F1 score: F1 score is defined as the harmonic mean between precision and recall and seeks to represent precision and recall in one metric.

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}$$

DATA ACCESS

HG002 HiFi data was acquired as part of the HPRC and is available here: https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=T2T/scratch/HG002/sequencing/hifi/. HG002 ONT, UL-ONT, and duplex ONT data were acquired from the EPI2ME project (*EPI2ME*™, n.d.). HG002 Revio data was acquired directly from PacBio and is available here: https://downloads.pacbcloud.com/public/revio/2022Q4/. HG00733 HiFi, ONT, and UL-ONT data were generated in house and have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA966152.

COMPETING INTEREST STATEMENT

E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc.

ACKNOWLEDGMENTS

This work was supported, in part, by US National Institutes of Health (NIH) grants R01HG010169, U24HG007497, and U01HG010971 to E.E.E. Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich Heine University Düsseldorf. E.E.E. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

W.T.H. conducted assembly generation, variant calling, variant annotation, merging, and data analysis and visualization in addition to writing the text. P.E. produced assemblies and variant calls. J.E. produced variant calls. P.A.A. assisted with variant calling and variant merging. K.M.M. produced PacBio HiFi data for HG00733. K.H. produced ONT data for HG00733. D.P. helped with assembly analysis. C.R.B. provided structural guidance. T.M. provided assistance with evaluating precision and recall and experimental design. K.G. assisted with experimental design and caller parameterization. E.E.E. provided project oversight, biological insight, and major text additions. All authors read and approved the final manuscript.

REFERENCES

All of Us Research Program Investigators, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., & Dishman, E. (2019). The "All of Us" Research Program. *The New England Journal of Medicine*, *381*(7), 668–676. https://doi.org/10.1056/NEJMsr1809937

Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., Welch, A. E., Dougherty, M. L., Nelson, B. J., Shah, A., Dutcher, S. K., Warren, W. C., Magrini, V., McGrath, S. D., Li, Y. I., Wilson, R. K., & Eichler, E. E. (2019). Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*, *176*(3), 663–675.e19. https://doi.org/10.1016/j.cell.2018.12.019

Begum, G., Albanna, A., Bankapur, A., Nassir, N., Tambi, R., Berdiev, B. K., Akter, H., Karuvantevida, N., Kellam, B., Alhashmi, D., Sung, W. W. L., Thiruvahindrapuram, B., Alsheikh-Ali, A., Scherer, S. W., & Uddin, M. (2021). Long-Read Sequencing Improves the Detection of Structural Variations Impacting Complex Non-Coding Elements of the Genome. *International Journal of Molecular Sciences*, *22*(4). https://doi.org/10.3390/ijms22042060

Chadwick, L. H., & Chris Wellington, B. S. (n.d.). *The GREGoR consortium*. Genome.gov. Retrieved September 15, 2022, from https://www.genome.gov/Funded-Programs-Projects/GREGOR-Consortium

Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

Chaisson, M. J. P., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., Rausch, T., Gardner,

E. J., Rodriguez, O. L., Guo, L., Collins, R. L., Fan, X., Wen, J., Handsaker, R. E., Fairley, S., Kronenberg, Z. N., Kong, X., Hormozdiari, F., Lee, D., Wenger, A. M., … Lee, C. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, *10*(1), 1784. https://doi.org/10.1038/s41467-018-08148-z

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Collins, D. W., & Jukes, T. H. (1994). Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics*, *20*(3), 386–396. https://doi.org/10.1006/geno.1994.1192

Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., Francioli, L. C., Khera, A. V., Lowther, C., Gauthier, L. D., Wang, H., Watts, N. A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C. W., Huang, Y., Brookings, T., … Talkowski, M. E. (2020). A structural variation reference for medical and population genetics. *Nature*, *581*(7809), 444–451. https://doi.org/10.1038/s41586-020-2287-8

Delahaye, C., & Nicolas, J. (2021). Sequencing DNA with nanopores: Troubles and biases. *PloS One*, *16*(10), e0257521. https://doi.org/10.1371/journal.pone.0257521

Dutta, U. R., Rao, S. N., Pidugu, V. K., S, V., V., Bhattacherjee, A., Bhowmik, A. D., Ramaswamy, S. K., Singh, K. G., & Dalal, A. (2019). Breakpoint mapping of a novel de novo translocation t(X;20)(q11.1;p13) by positional cloning and long read sequencing. *Genomics*, *111*(5), 1108–1114. https://doi.org/10.1016/j.ygeno.2018.07.005

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H.-Y., Humphray, S. J., Halpern, A. L., Kruglyak, S., Margulies, E. H., McVean, G., & Bentley, D. R. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, *27*(1), 157–164. https://doi.org/10.1101/gr.210500.116

Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Serra Mari, R., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, *372*(6537). https://doi.org/10.1126/science.abf7117

EPI2ME™. (n.d.). Retrieved April 25, 2023, from https://epi2me.nanoporetech.com/

Heller, D., & Vingron, M. (2019). SVIM: structural variant identification using mapped long reads. *Bioinformatics*, *35*(17), 2907–2915. https://doi.org/10.1093/bioinformatics/btz041

Heller, D., & Vingron, M. (2020). SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/btaa1034

Hsieh, P., Dang, V., Vollger, M. R., Mao, Y., Huang, T.-H., Dishuck, P. C., Baker, C., Cantsilieris, S., Lewis, A. P., Munson, K. M., Sorensen, M., Welch, A. E., Underwood, J. G., & Eichler, E. E. (2021). Evidence for opposing selective forces operating on human-specific duplicated TCAF genes in Neanderthals and humans. *Nature Communications*, *12*(1), 5118. https://doi.org/10.1038/s41467-021-25435-4

Jiang, T., Liu, Y., Jiang, Y., Li, J., Gao, Y., Cui, Z., Liu, Y., Liu, B., & Wang, Y. (2020). Long-read-based human genomic structural variation detection with cuteSV. *Genome Biology*, *21*(1), 189. https://doi.org/10.1186/s13059-020-02107-y

Kolmogorov, M., Billingsley, K. J., Mastoras, M., Meredith, M., Monlong, J., Lorig-Roach, R., Asri, M., Alvarez Jerez, P., Malik, L., Dewan, R., Reed, X., Genner, R. M., Daida, K., Behera, S., Shafin, K., Pesout, T., Prabakaran, J., Carnevali, P., North American Brain Expression Consortium (NABEC), … Paten, B. (2023). Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *bioRxiv*, https://doi.org/10.1101/2023.01.12.523790

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), 540–546.

https://doi.org/10.1038/s41587-019-0072-8

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Krusche, P., Trigg, L., Boutros, P. C., Mason, C. E., De La Vega, F. M., Moore, B. L., Gonzalez-Porta, M., Eberle, M. A., Tezak, Z., Lababidi, S., Truty, R., Asimenos, G., Funke, B., Fleharty, M., Chapman, B. A., Salit, M., Zook, J. M., & Global Alliance for Genomics and Health Benchmarking Team. (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, *37*(5), 555–560. https://doi.org/10.1038/s41587-019-0054-x

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., … Paten, B. (2022). A Draft Human Pangenome Reference. In *bioRxiv* (p. 2022.07.09.499321). https://doi.org/10.1101/2022.07.09.499321

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Lin, Y.-L., Chang, P.-C., Hsu, C., Hung, M.-Z., Chien, Y.-H., Hwu, W.-L., Lai, F., & Lee, N.-C. (2022). Comparison of GATK and DeepVariant by trio sequencing. *Scientific Reports*, *12*(1), 1809. https://doi.org/10.1038/s41598-022-05833-4

Logsdon, G. (2022). *HMW gDNA purification and ONT ultra-long-read data generation v3*. https://doi.org/10.17504/protocols.io.b55tq86n

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews. Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Logsdon, G. A., Vollger, M. R., Hsieh, P., Mao, Y., Liskovykh, M. A., Koren, S., Nurk, S.,

Mercuri, L., Dishuck, P. C., Rhie, A., de Lima, L. G., Dvorkina, T., Porubsky, D., Harvey, W. T., Mikheenko, A., Bzikadze, A. V., Kremitzki, M., Graves-Lindsay, T. A., Jain, C., … Eichler, E. E. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, *593*(7857), 101–107. https://doi.org/10.1038/s41586-021-03420-7

Lorig-Roach, R., Meredith, M., Monlong, J., Jain, M., Olsen, H., McNulty, B., Porubsky, D., Montague, T., Lucas, J., Condon, C., Eizenga, J., Juul, S., McKenzie, S., Simmonds, S. E., Park, J., Asri, M., Koren, S., Eichler, E., Axel, R., … Paten, B. (2023). Phased nanopore assembly with Shasta and modular graph phasing with GFAse. *bioRxiv*, https://doi.org/10.1101/2023.02.21.529152

Mc Cartney, A. M., Shafin, K., Alonge, M., Bzikadze, A. V., Formenti, G., Fungtammasan, A., Howe, K., Jain, C., Koren, S., Logsdon, G. A., Miga, K. H., Mikheenko, A., Paten, B., Shumate, A., Soto, D. C., Sović, I., Wood, J. M. D., Zook, J. M., Phillippy, A. M., & Rhie, A. (2021). Chasing perfection: validation and polishing strategies for telomere-to-telomere genome assemblies. In *bioRxiv* (p. 2021.07.02.450803). https://doi.org/10.1101/2021.07.02.450803

Medvedev, P., Stanciu, M., & Brudno, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, *6*(11 Suppl), S13–S20. https://doi.org/10.1038/nmeth.1374

Miller, D. E., Hanna, P., Galey, M., Reyes, M., Linglart, A., Eichler, E. E., & Jüppner, H. (2022). Targeted Long-Read Sequencing Identifies a Retrotransposon Insertion as a Cause of Altered GNAS Exon A/B Methylation in a Family With Autosomal Dominant Pseudohypoparathyroidism Type 1b (PHP1B). *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research*. https://doi.org/10.1002/jbmr.4647

Miller, D. E., Sulovari, A., Wang, T., Loucks, H., Hoekzema, K., Munson, K. M., Lewis, A. P., Fuerte, E. P. A., Paschal, C. R., Walsh, T., Thies, J., Bennett, J. T., Glass, I., Dipple, K. M.,

Patterson, K., Bonkowski, E. S., Nelson, Z., Squire, A., Sikes, M., … Eichler, E. E. (2021). Targeted long-read sequencing identifies missing disease-causing variation. *American Journal of Human Genetics*, *108*(8), 1436–1449. https://doi.org/10.1016/j.ajhg.2021.06.006

Olson, N. D., Wagner, J., McDaniel, J., Stephens, S. H., Westreich, S. T., Prasanna, A. G., Johanson, E., Boja, E., Maier, E. J., Serang, O., Jáspez, D., Lorenzo-Salazar, J. M., Muñoz-Barrera, A., Rubio-Rodríguez, L. A., Flores, C., Kyriakidis, K., Malousi, A., Shafin, K., Pesout, T., … Zook, J. M. (2022). PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. *Cell Genomics*, *2*(5). https://doi.org/10.1016/j.xgen.2022.100129

Oxford Nanopore Tech Update: new Duplex method for Q30 nanopore single molecule reads, PromethION 2, and more. (n.d.). Oxford Nanopore Technologies. Retrieved April 8, 2023, from https://nanoporetech.com/about-us/news/oxford-nanopore-tech-update-new-duplex-method-q30-nanopore-single-molecule-reads-0

PacBio revio. (2022, October 26). PacBio. https://www.pacb.com/revio/

pbsv: pbsv - PacBio structural variant (SV) calling and analysis tools. (n.d.). Github. Retrieved April 7, 2023, from https://github.com/PacificBiosciences/pbsv

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., Dai, H., Fritz, M. H.-Y., Cao, H., Cohain, A., Deikus, G., Durrett, R. E., Blanchard, S. C., Altman, R., Chin, C.-S., … Bashir, A. (2015). Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, *12*(8), 780–786. https://doi.org/10.1038/nmeth.3454

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P. T., Gross, S. S., Dorfman, L., McLean, C. Y., & DePristo, M. A. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987. https://doi.org/10.1038/nbt.4235

Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., Marijon, P., Ebler, J.,

Munson, K. M., Sorensen, M., Sulovari, A., Haukness, M., Ghareghani, M., Human Genome Structural Variation Consortium, Lansdorp, P. M., Paten, B., Devine, S. E., Sanders, A. D., Lee, C., Chaisson, M. J. P., … Marschall, T. (2021). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, *39*(3), 302–308. https://doi.org/10.1038/s41587-020-0719-5

Porubsky, D., Höps, W., Ashraf, H., Hsieh, P., Rodriguez-Martin, B., Yilmaz, F., Ebler, J., Hallast, P., Maria Maggiolini, F. A., Harvey, W. T., Henning, B., Audano, P. A., Gordon, D. S., Ebert, P., Hasenfeld, P., Benito, E., Zhu, Q., Human Genome Structural Variation Consortium (HGSVC), Lee, C., … Korbel, J. O. (2022). Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell*, *185*(11), 1986–2005.e26. https://doi.org/10.1016/j.cell.2022.04.017

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Rautiainen, M., Nurk, S., Walenz, B. P., Logsdon, G. A., Porubsky, D., Rhie, A., Eichler, E. E., Phillippy, A. M., & Koren, S. (2022). Verkko: telomere-to-telomere assembly of diploid chromosomes. In *bioRxiv* (p. 2022.06.24.497523). https://doi.org/10.1101/2022.06.24.497523

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*(2), 155–158. https://doi.org/10.1038/s41592-019-0669-3

Sanderson, N. D., Kapel, N., Rodger, G., Webster, H., Lipworth, S., Street, T. L., Peto, T., Crook, D., & Stoesser, N. (2023). Comparison of R9.4.1/Kit10 and R10/Kit12 Oxford Nanopore flowcells and chemistries in bacterial genome reconstruction. *Microbial Genomics*, *9*(1). https://doi.org/10.1099/mgen.0.000910

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-

molecule sequencing. *Nature Methods*, *15*(6), 461–468. https://doi.org/10.1038/s41592-018-0001-7

Shafin, K., Pesout, T., Chang, P.-C., Nattestad, M., Kolesnikov, A., Goel, S., Baid, G., Kolmogorov, M., Eizenga, J. M., Miga, K. H., Carnevali, P., Jain, M., Carroll, A., & Paten, B. (2021). Haplotype-aware variant calling with PEPPER-Margin-DeepVariant enables high accuracy in nanopore long-reads. *Nature Methods*, *18*(11), 1322–1332. https://doi.org/10.1038/s41592-021-01299-w

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., … Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, *38*(9), 1044–1053. https://doi.org/10.1038/s41587-020-0503-6

Smolka, M., Paulin, L. F., Grochowski, C. M., Mahmoud, M., Behera, S., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S. W., Carvalho, C. M. B., Proukakis, C., & Sedlazeck, F. J. (2022). Comprehensive Structural Variant Detection: From Mosaic to Population-Level. In *bioRxiv* (p. 2022.04.04.487055). https://doi.org/10.1101/2022.04.04.487055

Wagner, J., Olson, N. D., Harris, L., Khan, Z., Farek, J., Mahmoud, M., Stankovic, A., Kovacevic, V., Yoo, B., Miller, N., Rosenfeld, J. A., Ni, B., Zarate, S., Kirsche, M., Aganezov, S., Schatz, M. C., Narzisi, G., Byrska-Bishop, M., Clarke, W., … Zook, J. M. (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genomics*, *2*(5), 100128. https://doi.org/10.1016/j.xgen.2022.100128

Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.-C., Gupta, R., Wenger, A. M., Rowell, W. J., Khan, Z. M., Farek, J., Zhu, Y., Pisupati, A., Mahmoud, M., Xiao, C., Yoo, B., Sahraeian, S. M. E., Miller, D. E., … Sedlazeck, F. J. (2021). Towards a Comprehensive Variation Benchmark for Challenging Medically-

Relevant Autosomal Genes. In *bioRxiv* (p. 2021.06.07.444885). https://doi.org/10.1101/2021.06.07.444885

Wang, T., Antonacci-Fulton, L., Howe, K., Lawson, H. A., Lucas, J. K., Phillippy, A. M., Popejoy, A. B., Asri, M., Carson, C., Chaisson, M. J. P., Chang, X., Cook-Deegan, R., Felsenfeld, A. L., Fulton, R. S., Garrison, E. P., Garrison, N. A., Graves-Lindsay, T. A., Ji, H., Kenny, E. E., … Human Pangenome Reference Consortium. (2022). The Human Pangenome Project: a global resource to map genomic diversity. *Nature*, *604*(7906), 437–446. https://doi.org/10.1038/s41586-022-04601-8

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Willems, T., Gymrek, M., Highnam, G., 1000 Genomes Project Consortium, Mittelman, D., & Erlich, Y. (2014). The landscape of human STR variation. *Genome Research*, *24*(11), 1894–1904. https://doi.org/10.1101/gr.177774.114

Zheng, Z., Li, S., Su, J., Leung, A. W.-S., Lam, T.-W., & Luo, R. (2021). Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. In *bioRxiv* (p. 2021.12.29.474431). https://doi.org/10.1101/2021.12.29.474431

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials [Review of *Extensive sequencing of seven human genomes to characterize benchmark reference materials*]. *Scientific Data*, *3*, 160025. https://doi.org/10.1038/sdata.2016.25

Zook, J. M., McDaniel, J., Olson, N. D., Wagner, J., Parikh, H., Heaton, H., Irvine, S. A., Trigg, L., Truty, R., McLean, C. Y., De La Vega, F. M., Xiao, C., Sherry, S., & Salit, M. (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, *37*(5), 561–566. https://doi.org/10.1038/s41587-019-0074-6