TITLE

Benchmarking long-read genome sequence alignment tools for human genomics applications

AUTHORS

Jonathan LoTempio Jr, BS; jlotempio@gwu.edu ; 1, 2, 3

Emmanuèle Délot, PhD; edelot@gwu.edu ; 1, 2, 4

Eric Vilain, MD, PhD; evilain@gwu.edu ; 1, 2, 3 *

1. Center for Genetic Medicine Research, Children's National Research Institute, Children's National Hospital, 111 Michigan Avenue NW, Washington DC, 20010

2. Department of Genomics and Precision Medicine, George Washington University, 111 Michigan Avenue NW, Washington DC, 20010

3. International Research Laboratory (IRL2006) "Epigenetics, Data, Politics (EpiDaPo)", Centre National de la Recherche Scientifique (CNRS), 2300 Eye St NW, Washington DC, 20052

4. Center for Translational Research, Children's National Research Institute, Children's National Hospital, 111 Michigan Avenue NW, Washington DC, 20010

* Corresponding author, evilain@gwu.edu

KEYWORDS

Long-read genome sequencing; nanopore; single-molecule real-time; benchmarking; reference-guided genome alignment; human whole genome sequencing

ABSTRACT

Background

The utility of long-read genome sequencing platforms has been shown in many fields including whole genome assembly, metagenomics, and amplicon sequencing. Less clear is the applicability of long reads to reference-guided human genomics, the foundation of genomic medicine. Here, we benchmark available platform-agnostic alignment tools on datasets from nanopore and single-molecule real-time platforms to understand their suitability in producing a genome representation.

Results

For this study, we leveraged publicly-available data from sample NA12878 generated on Oxford Nanopore and Pacific Biosciences platforms. Each tool that was benchmarked, including GraphMap, minimap2, and NGMLR, produced the same alignment file each time. However, the different tools widely disagreed on which reads to leave unaligned, affecting the end genome coverage and the number and locations of discoverable breakpoints. Only minimap2 was computationally lightweight enough for use at scale. No alignment from one tool independently resolved all large structural variants (10,000-100,000 basepairs) present in the Database of Genome Variants (DGV) for sample NA12878. For variants larger than 1,000,000 basepairs, nanopore sequence aligned with minimap2 and NGMLR, and single-molecule real-time sequence aligned with NGMLR contained more breakpoints than are present in DGV.

Conclusions

When computational resources are not a limiting factor, it should be best practice to use an analysis pipeline that generates alignments with both minimap2 and NGMLR, as neither results in a comprehensive genome representation. When computational resources are limited, use of minimap2 for human genome alignment produces files sufficient to answer hypotheses and generate new questions.

BACKGROUND

The diverse ecosystem of DNA preparation, sequencing, and mapping technologies is capable of generating computational representations of genome biology through different chemistries and processes (1–5). Sequencing-by-synthesis platforms produce highly accurate sequence data in the form of short reads (<300 basepairs, bp) (6) from high molecular weight DNA inputs. These molecular input libraries can be prepared for short-read sequencing (SRS) that are oriented in context long-range genome position through high-throughput chromatin confirmation capture (Hi-C) methods (7). In contrast, single-molecule real-time platforms can produce increasingly accurate reads on molecules >10 kilobasepairs (kbp) (8) while nanopore-based sequencing (9) and nanochannel-based mapping platforms (10) can sequence or visualize megabase-length DNA molecules (11). Each of these platforms has different abilities and available tools contributing to the diversity of projects enabled by these technologies (12,13).

Optical genome mapping (developed by Bionano Genomics) excels at detecting complex structural variants (SV), such as balanced translocations and deletions/insertions in the 1 kb to 1 Mb range, and its clinical utility was demonstrated in Duchenne (DMD) or facioscapulohumeral (FSHD) muscular dystrophies (14,15). But long-read sequence (LRS), developed by Pacific Biosciences and Oxford Nanopore Technologies, is the most appropriate technology to detect smaller variants in the 50 bp-1 kb range (16), especially in repetitive regions, and should therefore bring new diagnosis potential for diseases of trinucleotide repeat expansion and genome instability such as Huntington's Disease and Myotonic Dystrophies (17–21), as well as for discovery of SVs affecting regulatory regions of known genes. Fulfilling the promise of LRS for medical genetics requires understanding the tools that power the technology and their respective strengths.

LRS platforms have been shown to be capable of generating single, phaseable reads spanning repetitive or complex genomic regions that remain unresolved in current human genome assemblies (8,9,22,23). This has implication for resolution of highly homologous pseudogenes and large SV, even in low-complexity genomic regions. Read lengths in the tens of kbp have further allowed for end-to-end viral genome sequencing (24), while read lengths in the millions of bp have the potential to span whole mammalian chromosomes (9,25). This potential for more contiguous *de novo* human assembly has led to studies to specifically improve and benchmark basecalling and polishing tools (26), as well as assembly tools (27–30) that can be applied to human genomics.

For reference genome-guided experiments, LRS has proven useful for amplicon sequencing in cancer detection (31) and for metagenomics (32–34), but the field is still in early days for assessment of variation in high-coverage human whole genome sequence. Indeed, high-coverage human whole genome sequence on LRS platforms has yet to be undertaken on the scale of projects based on Illumina-developed SRS platform (35,36). This difference in data availability is reflected in the number of tools for SV calling - including at least 21 for SRS and five for LRS (13). It is certain that, due to the richness of the sequence data and diminishing cost of generating those data, these LRS platforms will be used in large-scale reference-guided whole human genome projects that are presently dominated by SRS data production (37).

The value of LRS in resolving SV has been demonstrated by efforts aggregating callsets across platforms and technologies to deeply characterize a genome (16). These experiments are concerned mainly with understanding the full scope of the architecture of a genome, rather than contrasting differences between results of one alignment tool or another. Each of these tools showcase their utility and strengths in their initial publications, especially in terms of the mapping quality of reads to the reference genome and precision and sensitivity to preserve

known variants in synthetic or downsampled genomic data, however, there have been no studies that specifically benchmark LRS platforms and tools for reference-guided experiments.

To address this gap, we benchmarked LRS alignment tools with the datasets generated from the Joint Initiative for Metrology in Biology's Genome in a Bottle Initiative (GIAB) sample, NA12878, which was sequenced on Oxford Nanopore's MinION-platform by the Whole Genome Sequencing Consortium (9) and on the Pacific Biosciences SMRT-sequence data by the National Institute of Standards and Technology (39). We compared computational performance (peak memory utilization, central processing unit (CPU) time, file size/storage requirements), genome coverage depth (x-times), and quantified the reads left unaligned in any given experiment. We have limited this manuscript to tools that are 1) widely adopted so that analysts may have something for comparison, 2) platform-agnostic so that LRS data of any sort may be used in pipeline, and 3) easy to install so as to be equitable in their utility to researchers and future clinical laboratories. Since the resolution of large SVs is a key application of this technology, and also allows comparison of the differences in genome alignments in an aggregate way, we ran the SV-calling tool Sniffles to highlight differences in breakpoint location in each binary alignment map (BAM) file (40). Taken together, these experiments present a comprehensive view of differences in the products of LRS whole-genome alignment pipelines.

RESULTS

All experiments were performed on the NA12878 sample, for which publicly available datasets exist for both nanopore and SMRT sequencing, and rigorously annotated variants are available in the Database of Genomic Variants (DGV) (41). We present our results in four sections: 1) the tools that passed inclusion/exclusion criteria, 2) computational performance and benchmarking, 3) an analysis of aligned and unaligned reads, 4) an analysis of structural variation present in each alignment compared to a baseline.

1. Tools that passed the inlcusion/exclusion criteria

Following a literature review of available alignment tools and search of long-read-tools.org (38), we established a set of inclusion criteria (see Material and Methods), which accounted for both types of LRS data (a tool must be able to handle both nanopore and SMRT reads), as well as the state of the field in terms of software updates (must not have been superseded by another tool) and adoption of the tool, as measured by citations (must be cited by manuscripts written by outside parties). All surveyed tools and criteria are outlined in Table 1.

Three alignment tools, GraphMap, minimap2, and NGMLR were included in this study (40,42,43)**.** Kart, Mashmap2, LAMSA, MECAT2, lordFAST, smsMAP, and Winnowmap were excluded for lack of adoption based on citations (43–49). LAST, BLASR, rHAT, and Meta-aligner were excluded because they did not have modes to specifically account for both types of data (50–53). BWA-MEM and MECAT were excluded because they were superseded by newer tools (46,54).

2. Computational performance and benchmarking

Computations were run three times on full node capacity on our university's high-performance compute cluster (HPC), including 40 CPUs with a configuration described in the methods. Further one-off runs were performed with restricted node capacity of 30, 20, and 10 CPUs. All tools were run with default parameters to reflect typical use in exploratory studies. Summaries of findings can be examined in Table 2. Full reports on each run can be examined in Supplemental Table S1, while a composite table of the output of samtools bamstats is available as Supplemental Table S2.

Minimap2 was the least resource-demanding tool.

Minimap2 successfully aligned nanopore data every time with unrestricted and restricted resources. Unrestricted runs used an average of 14.2 gigabytes (Gb) and the jobs took ~540 CPU hours (~13 wall clock hours) to complete.

Minimap2 also successfully aligned SMRT data every time with unrestricted or restricted resources. The runs used an average of 16.9 Gb and the jobs took ~641 CPU hours (~16 wall clock hours) to complete, a little longer than for the nanopore dataset.

Memory usage (see peak_rss columns in Supplemental Table S1) and runtime were consistent across triplicate runs with unrestricted resources and did not change with restriction of resources when the tool was used with either dataset. The consistency of results, as well as the speed and relatively low computational demands of minimap2 make it a strong candidate for inclusion in clinical analysis pipelines.

GraphMap was the most resource-intensive tool, resulting in 3/12 run failures.

GraphMap successfully aligned nanopore data two out of three times in both restricted and unrestricted resources runs. For the two successful unrestricted runs, an average of ~96 Gb memory was used and the jobs took an average of ~8,328 CPU hours (~208 wall clock hours) to complete. This was 6.6 times more memory and 13 times longer than Minimap2. Memory usage was unchanged with resource restriction, but runtime did increase by approximately 150,000 CPU seconds (~7.5 hours).

GraphMap successfully aligned SMRT data every time with unrestricted resources and two out of three times with restricted resources. These runs used an average of ~50 Gb memory and the jobs took an average of ~10,642 CPU hours (~266 wall clock hours, 11

days) to complete. All unrestricted resource jobs used ~50 Gb of memory. One restricted run (30 threads) used ~49 Gb of memory while one (20) used ~46 Gb. The restricted run limited to 10 threads failed due to time out at 14 wall clock days.

Runtime across all runs did not directly relate to memory usage or resource restriction (Table 2, Supplemental Table S1). The resource-intensive requirements of GraphMap for whole human genome alignment combined with failures to run in both datasets likely precludes including it in the design of most analysis pipelines, however alignment files from successful runs were included in further analyses.

NGMLR completed all tasks, but performance was very resource-dependent. NGMLR successfully aligned nanopore data every time with unrestricted and restricted resources. The unrestricted runs used an average of 33.8 Gb and the jobs took ~2,539 CPU hours (~63.5 wall clock hours) to complete, intermediary between Minimap2 and GraphMap.

NGMLR was much faster on SMRT data. It successfully aligned every time even with restricted resources. The runs used an average of 29.7 Gb the jobs took ~658 CPU hours (~16.5 wall clock hours, almost 4 times faster than on the nanopore data) to complete, on par with the time taken by Minimap2.

Memory usage and runtime were consistent with the average across triplicate runs, but, unlike the other tools, NGMLR's performance was resource-dependent: runtime increased and memory usage decreased with resource restriction. While more computationally intensive than minimap2, there are no glaring concerns for NGMLR's

suitability for whole human genome alignment-based pipelines, as there is with GraphMap.

3. Whole genome alignment

Aligned reads. In every experiment that completed successfully, alignments of the same genome with the same tool had the same genome coverage, whether resource-restricted or not (Supplemental Table S1, column O). For this reason, one file generated from the first unrestricted run was used for subsequent analysis (shown in Table 2). Genome coverages of ~30x and ~44x were reported for nanopore (9) and SMRT (39,55) analyses respectively. We calculated the actual coverage after read alignment to the hg38 reference by each tool.

GraphMap produced higher coverage genomes than expected on both nanopore (32.8 x) and SMRT (45.4x) datasets.

Minimap2 produced higher coverage nanopore genomes (33.4x), and expected (within 1x) coverage SMRT genomes (43.6x).

NGMLR produced higher coverage nanopore genomes (31.1x) and lower coverage (39.7x) SMRT genomes.

For nanopore data, coverage was higher than expected, by 3.7%, 9.3% and 11.3% for NGMLR, GraphMap, and Minimap2 respectively. For SMRT data, GraphMap also produced slightly higher than expected coverage (by 3.2%), while NGMLR came in at 9.8% lower than expected. Minimap2 showed a small difference (-0.9%). These results are not sufficient evidence to pass judgment on the utility of any tool, but prompted us to examine where those discrepancies come from. We analyzed the unmapped reads to find the overlap in consensus-excluded reads.

Unaligned reads. The readname assigned to each read in a fastq retained in the BAM allowed us to write an R script to directly compare the lists of reads that were not included in the alignment by each alignment tool (Fig. 1).

All tools agreed to leave ~1.6 million of the same nanopore reads unaligned, but differed in their overall totals of unaligned reads. NGMLR left the highest number of reads unaligned, ~3.3 million, which explains why it produced the lowest coverage genome. It agreed with the other tools on approximately half of its discarded reads. Minimap2 and NGMLR agreed to leave a further ~0.5 million reads unaligned, while NGMLR and GraphMap agreed on a separate ~0.48 million unaligned reads and GraphMap and Minimap2 on ~0.4 million further unaligned reads.

All tools agreed to leave ~1.2 million of the same SMRT reads unaligned, but differed in their overall totals of unaligned reads. Similarly, for the ~44x SMRT dataset, all three tools agreed to exclude the same ~1.2 million reads. NGMLR left the highest number of reads unaligned, ~6 million, which explains why it produced the lowest coverage genome. Minimap2 and NGMLR agree to leave a further ~1.6 million reads unaligned, while NGMLR and GraphMap agree on a separate ~0.78 million unaligned reads and GraphMap and Minimap2 agree on ~0.01 million further unaligned reads. Perhaps the most interesting feature in this set of experiments is that NGMLR excludes ~6.5 times more reads than GraphMap or Minimap2.

4. Structural variation differences across alignments

By examining the coverage of the genome and the overlap of the discarded reads, it became clear that more detail was needed to understand the differences in alignment. To globally

compare the alignments to each other, and assess their usability for variant calling, we looked to the SV known to be present in the NA12878 genome as curated by the DGV resource.

To compare the SV breakpoints preserved in the genome alignments created in this study, we leveraged the sniffles SV caller because of its highly-detailed output files. We used the curated SV present in DGV, annotated for presence in NA12878 as reference. Due to the different nomenclature for SV type in the sniffles output VCF specification document (56) and the DGV annotations, comparisons were limited to the four classes of variant that were most unambiguously defined in both: insertions, deletions, duplications, and inversions. We could not successfully install PBHoney on our cluster and NanoSV provided VCFs that did not have enough detail for comparison.

Sniffles variants were graphed by SV length on the X-axis in shades of blue, contrasted with DGV variants in red, organized by platform and SV type (Fig. 2). For both sequencing methods, deletions and insertions were overestimated in all data sets compared to the number curated in DGV. Conversely, the alignment files preserved fewer breakpoints recognized by sniffles as duplications and inversions than are present in DGV. Globally, there are more indels <10,000 bp present in VCFs than are present in DGV. This trend largely reverses above 10,000 bp, which is reflected in the literature (16). The summary of variants is presented in Table 3, with granular counts by SV length and type found in Supplemental Table S3.

a. Differences in variants among alignment files

Deletions (Fig. 2, A, B). The total number of identified deletions varied widely, from ~6,500 to ~14,300, across alignment files. For all three alignment tools, the nanopore data yielded the most sniffles-identified deletions. Minimap2 preserved the most deletions on both datasets. GraphMap-aligned genomes did not allow

detection of any deletions larger than 20 kbp in nanopore data or 6 kbp in SMRT data.

Insertions (Fig. 2, C, D)**.** The total number of identified insertions also varied (from ~6,000 to ~13,700). In contrast to deletions, the SMRT data yielded the most insertions for all three alignment tools. Minimap2 again allowed identification of the most insertions on both datasets. The three tools agreed for insertions in the 1,000-4,000 bp range. Strikingly, none of the tools allowed identification of insertions larger than 8,000 bp in nanopore sequence or 4,000 bp in SMRT sequence.

Duplications and inversions: GraphMap preserved no breakpoints that sniffles recognized as duplications or inversions. Both minimap2 and NGMLR datasets produced similar numbers of inversions (within 30-50 variants; Fig. 2, E, F) and inversions (within 4-7 variants; Fig. 2, G, H). NGMLR allowed calling of the highest number of duplications and inversions on both datasets, outperforming minimap2 in all but three SV length bins for duplications. This was particularly evident in the 6-10 kbp size range where no duplications were called from the minimap2 datasets (5-10 kbp in nanopore data).

b. Comparison to variants in DGV

Broad trends emerged with comparison to a truthset. Deletions and insertions were overestimated in all data sets compared to the number curated in DGV. Conversely, far fewer duplications and inversions than are present in DGV were called on the LRS data. While there were more indels <10,000 bp present in VCFs than are have been curated in DGV, this trend largely reversed above 10,000 bp, where most known variants were

missed. (DGV variants are shown in red in Fig. 2; details of counts are shown in Supplemental Table S3.) While these discrepancies may largely reflect the characteristics of LRS output since DGV variants are curated from several data sources, our analysis showed that the different alignment tools yielded different callsets from each other as well.

Deletions (Fig. 2A, B). The overestimation of deletions of all sizes called in the LRS datasets compared to DGV ranged from 4.8 times more in the SMRT/GraphMap data to 10.6 times more in the nanopore/Minimap2 data. For variants smaller than 10 kbp, the alignment files contained more SVs than are present in DGV. Indeed, for both datasets and all three alignment tools, 10 times more SVs between 51-1000 bp were present in VCFs than are present in DGV. Compared to the variants in DGV, minimap2 and NGMLR resolve tens of variants less in each bin greater than 20,000 bp but less than 100,000 bp.

Insertions (Fig. 2C, D). Sniffles identified more insertions ≤6,000 bp in nanopore alignments and ≤4,000 bp in SMRT alignments than are present in DGV. For both datasets and all three alignment tools, more than 100 times more insertions sized 51-1000 bp are present in VCFs compared to DGV. In contrast, none of the tools were able to identify large insertions beyond 8 kbp or 4 kbp on the nanopore or SMRT data, respectively. DGV contains 934 insertions between 9 and 80 kbp, yet none are resolved by Sniffles on any alignment file from the nanopore or SMRT set.

Duplications (Fig. 2E, F). None of the tools performed well for identification of duplications. NGMLR resolved at most half as many duplications present in

DGV while minimap2 resolved approximately 10%. GraphMap identified none. NGMLR detected ~50 variants larger than 100,000 bp, compared to the 19 variants present in DGV.

Inversions (Fig. 2G, H). GraphMap preserved no breakpoints Sniffles identified as inversions; minimap2 and NGMLR alignments led to overestimation of called small (51-1,000 bp) inversions. In total, DGV contains more inversions than are present in all VCFs. Less than a quarter of DGV inversions are greater than 100 kbp, compared to around a third of similarly sized variants in the VCFs, regardless of sequencing platform and alignment tool.

The fact that none of the callsets allowed identification of all large SVs of size greater than 10 kbp present in DGV is a significant limitation when trying to use this technology to identify causative SVs in patient samples. Since DGV variants have been curated by the community and have a high likelihood of being truly present in the studied genome, it means that, rather than a specific alignment tool underperforming, the long-read sequence platforms differ in their ability to resolve expected variants. This may yield false negative results for conditions where structural variation is a recognized etiology, such as DMD or FSHD, where variants range between the tens and hundreds of thousands of kbp (14,15), beyond the size of variants accurately detected in this study. It is also a critical issue when the technology is used for broad exploratory surveys seeking to identify new etiology in low-complexity regions of the genome where long-read sequencing should shine.

DISCUSSION

In this study we have highlighted the key differences between alignment files generated by three tools that have seen wide adoption for experiments involving the reference-guided alignment of large genomes. By using a well-characterized genome standard, NA12878, we were able to directly compare the performance of each tool on the sequencing datasets obtained on different platforms. We further analyzed the reads that were included or excluded from the alignment, and how these reads alignments revealed breakpoints that could be resolved as structural variants in genome architecture. We then compared those variants to a set of previously published variants discovered through multiple platforms.

Reassuringly, each alignment tool was internally consistent: when an alignment tool was given the same fastq and the same reference genome, it produced the same result as judged by bamstats and sniffles variant callsets. However, when looking across the alignment files produced by different tools on the same sequence data, the representations of the genome diverged in terms of which reads were included or excluded and the numbers and types of variants that were present in VCFs. This is impactful because the potential high value of LRS data in terms of potential for phasing and identification of epigenetic DNA modification (57). Since the majority of experiments that leverage large scale population surveys can be expected to rely on reference-guided alignment rather than *de novo* assembly because of both the cost and speed of analysis (58), it is key to understand the idiosyncrasies of each type of alignment files prior to generalizing clinical use for these technologies. Furthermore, clinical experiments in genomic medicine face human time constraints – speedier analyses will have higher appeal and adoption.

At this time, GraphMap does not show utility for producing whole genome alignments that include structural variations when run with default parameters. The resource usage was large

and the time to complete the computation long with no added benefits to resolve SVs shown in this study. This is not unexpected, as the tool was designed in large part to increase single nucleotide variant sensitivity in noisy nanopore-sequenced reads (42); as such, the algorithm used smooths out some of the complexity of genome architecture. This reduction in complexity holds true for SMRT data as well, removing some of the key benefits of both LRS platforms.

Minimap2 used the least computational resources and ran successfully in the fastest time, an important point, should these platforms be tied closer to bedside applications. On data from both platforms, it allowed calling of the most insertions and deletions, but fell short on inversions and duplications.

NGMLR was the most discerning aligner, in that it left the highest number of reads unmapped. It used more compute power than Minimap2, but handled the SMRT-based data in nearly the same amount of time, while taking considerably more time when handling nanopore-based data. While it was designed specifically to resolve structural variation (40), it calls a high number of very large variants (>100,000 bp) that have not been validated with other methodologies and curated in DGV.

There is a great divergence in sniffles-called variants from alignment files generated by all tools from the variants present in DGV. This is a concerning expansion of seminal findings in a previous study (16) as none of the sniffles VCFs mirror the SVs present in the high-quality curated DGV database. Small variants are present in sniffles VCFs a much higher level than in DGV, and, conversely, between 10 and 100 kbp, each alignment tool produces an input for Sniffles that falls short of the number of variants in DGV.

The points above are critical in designing pipelines for genome analysis and structural variant discovery. In computationally unlimited research settings with onsite high-performance clusters, there is value added in the generation of alignment files from both minimap2 and NGMLR. These two perspectives on the same genome will account for some of the inherent differences of each tool and the algorithms they use to handle read alignment (59). If computational resources are limited, minimap2 is the best choice to move the greatest number of genomes through the pipeline quickly; however, the loss of comprehensiveness must be considered in cases where a suspected variant is not found.

This is impactful in genomic medicine. For example, as variants range from tens of kbp in FSHD and hundreds of kbp to mbp in DMD, diagnostic of these disorders will likely not benefit from data generated on LRS platforms at present, underscoring the need for optical mapping or array-based technologies. However, disorders resulting from smaller SVs such as Huntington's (~18-540 bp) (60), myotonic dystrophy 1 (~15-153 bp) and myotonic dystrophy 2 (~338-143,000 bp) (61) could be good candidates for deep study with LRS platforms based on the variants present in alignment files from minimap2 and NGMLR. Accordingly, LRS has been used to identify variants in many such disorders (21,62).

If pathogenic loci are known, a high diagnostic yield may be obtained by generating maps with each available alignment tool, and use a structural variant caller such as NanoSV (63). Unlike Sniffles, which provides a call of type of SV (deletion, insertion etc.), NanoSV only identifies breakpoints in the alignment, without assigning those breakpoints an SV type. A robust comparison of SV callers on nanopore datasets highlighted the relative strengths of variant calling pipelines and may help users determine the best caller for their experiments (64). NanoSV may be suitable for identifying breakpoints missed by Sniffles, but comes with the

further caveat that it is resource-intensive and may not scale in a clinical setting without vast computational resources (64).

Answers to biological questions can also leverage tools that work with data upstream and downstream of whole genome alignments; indeed, alternate solutions to discover differences in genome structure have been proposed that leverage new tools. Tools designed to specifically identify triplet repeat expansions in LRS data sets, such as Repeat Hidden Markov Model (Repeat HHM) or Tandem-Genotypes, have shown the superiority of the approach compared to other techniques (19,20). RepeatHMM uses raw, unaligned reads in alternate ways than whole genome alignment to a reference (19). Rather than aligning LRS, reads can be analyzed independently for microsatellites with specialized tools. The tandem-genotypes tool makes use of a LAST whole-genome sequence alignment to detect copy number changes in the genome (20). In line with this fact, LRS has been used to identify variants in many disorders (21).

The discrepancies between the VCFs generated from alignment files starkly show the need to design experiments with the appreciation that the genomics ecosystem is not dominated by one platform or pipeline and requires a multifaceted approach to discovery. We are at a position where simply because the breakpoint is missing from the VCF from a long-read genome, we cannot say that it is not present. We must therefore look across platforms and datatypes for comprehensive genome representations (16).

CONCLUSIONS

As the cost of long-read sequencing catches up to that of inexpensive short-read sequencing, the inevitable boom in data production will require well thought-out analysis pipelines. Pipeline design always involves a set of tradeoffs. To accurately assess these tradeoffs, we must have a rigorously benchmarked view into the tools available to create the analytic product. Here, we

looked to the differences in reference-guided human genome alignments to understand the difference in each tool's alignment of the same genome, and how it affects a structural variant callset. This informs our conclusion that, regardless of sequencing platform, when computational resources are not a limiting factor, it should be best practice to align an LRS human genome with both minimap2 and NGMLR to gain better insight into the architecture of a genome of interest.

MATERIALS AND METHODS

1. **Tool selection criteria:**
    a. Exclusion criteria were designed after a comprehensive literature review of alignment tools. Starting with tools that were recommended by the developers of each platform, we examined the tools that were cited or benchmarked against new, platform-specific tools. Since this yielded many software tools, we established exclusion criteria to limit our experiments to tools that had seen uptake by the community in terms of citations of that tool, while accounting for the idiosyncrasies of LRS sequencing platforms in their design and implementation. Since tools can be regularly updated or out-versioned, we wanted to use only the most up to date software at the time of analysis.

    Inclusion criteria:

        i. Must be cited in more than 100 original papers in the five year window between 2015-2020, other than the paper where the tool is initially published as determined by a Web of Science search query.
        ii. Must be designed with specific parameters for data produced on Oxford Nanopore Technologies and Pacific Biosciences platforms.
        iii. Must not be superseded by a new tool.

b. In our examination of structural variant calling pipelines, we elected to limit our study to the variant caller sniffles due to the fact that the output files were more descriptive than those produced by NanoSV. We were unable to install PBHoney on our cluster.

2. Data

Reference genome

GRCh38.p12 was accessed and downloaded on April 8, 2019 from https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/ (65).

Sequence data

Nanopore sequence data were accessed on April 8, 2019, version rel5-guppy-0.3.0-chunk10k.fastq, from AWS Open Data as made available by the Whole Genome Consortium. SMRT sequence data were accessed on May 28, 2019, version sorted_final_merged.bam from the NCBI FTP.

- https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md (66)
- ftp://ftp.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/ (67)

Database of genome variants

DGV data were accessed on January 30, 2020 from http://dgv.tcag.ca/dgv/docs/GRCh38_hg38_variants_2016-05-15.txt (68). The full set of variants was reduced to 11,042 variants confirmed to be present in NA12878. 14 of these variants were excluded as they were called on contigs present out of the main reference assembly contigs.

3. Hardware configuration

Computations were performed on the George Washington University High Performance Compute Center's Pegasus Cluster on SLURM-managed default queue compute nodes with the following configuration: Dell PowerEdge R740 server with Dual 20-Core 3.70GHz Intel Xeon Gold 6148 processors, 192GB of 2666MHz DDR4 ECC Register DRAM, 800 GB SSD onboard storage (used for boot and local scratch space), and Mellanox EDR InfiniBand controller to 100GB fabric.

4. Whole genome alignment tool benchmarking and analysis

There versions of alignment tools used in this project included: GraphMap/0.5.2, minimap2/2.16, and ngmlr/0.2.7.

- https://github.com/isovic/GraphMapLink (42)
- https://github.com/lh3/minimap2 (43)
- https://github.com/philres/ngmlr (40)

Computational metrics were printed from SLURM job records. Samtools 1.9 was used for all alignment manipulations, alignment read depth coverage calculations, and to extract unmapped reads (69). Samtools 1.10 was used to generate bamstats files. R 3.5.2 version Eggshell Igloo with the tidyverse packages were used to compare the readnames across unmapped read files to assess the degree of overlap of unmapped reads across alignment files. Genome coverage was calculated with samtools 1.9 assuming a 3 billion basepair genome. Binary conversion values were used for bytes (1,073,741,824) and kilobytes (1,048,576) to gigabytes. CPU and wall clock hours were calculated from CPU and wall clock seconds.

5. **Data reshaping and visualization**

As data were integrated across multiple sources and formats, they needed to be reshaped for comparison and visualization. For example, the multiple separators found in VCFs (commas and semicolon) are not directly usable in R or python. Relevant dataframes from genome alignment files were reshaped with shell scripts and r scripts whose methodology and key intermediate files are available as supplemental files to this manuscript.

6. **Comparison of breakpoints**

Structural variants (SV) were called with sniffles/1.0.11 with default parameters. Only variants called on the main GRCh38 assembly were included. R 3.5.2 was used to bin and graph structural variants by SV Type with tidyverse packages and cutr (70,71). Intermediary files and scripts are available upon request. Data from sniffles VCFs and from DGV were subsetted by comparable fields including SV Type, SV Length, and Chromosome since a common format was unavailable for direct comparison. Figures were made with R packages tidyverse/ggplot and VennDiagram (72).

DECLARATIONS

- Ethics approval and consent to participate
  - N/A
- Consent for publication
  - N/A

AVAILABILITY OF DATA AND MATERIALS

All data and tools used for this study are publicly available at the URLs indicated above in Methods. The scripts we used to process those data are available in supplementary materials. Raw and post-processed VCFs, DGV source material, and other dataframes used to create the figures are available at:

https://github.com/jlotempiojr/longread_alignment_benchmarking (73)

Interested parties can further reach out to the authors for intermediate BAMs if needed.

COMPETING INTERESTS

EV is on the board of scientific advisors to Bionano Genomics.

FUNDING

EV is supported in part by A. James Clark Distinguished Professorship in Molecular Genetics at Children's National Hospital. EV, JL, and ED are supported by NIH Award Number U01HG011745 from the National Human Genome Research Institute. ED is supported in part by NIH Award Number UL1TR001876 from the National Center for Advancing Translational Sciences. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

AUTHORS CONTRIBUTIONS

JL conceived of the study, performed the experiments, and wrote the manuscript

ED gave comment and direction on experimental design, and edited the manuscript

EV gave comment and direction on experimental design, edited the manuscript, and allocated laboratory resources to the project

ACKNOWLEDGEMENTS

Adam Kai Leung Wong, PhD, High Performance Computing Specialist for Genomics at the GWU Computational Biology Institute provided critical support in preparing the HPC for this study.

REFERENCES

1. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7.
2. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, et al. The challenges of sequencing by synthesis. Vol. 27, Nature Biotechnology. Nat Biotechnol; 2009. p. 1013–23.
3. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Vol. 26, Nature Biotechnology. Nat Biotechnol; 2008. p. 1146–53.
4. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res [Internet]. 2018 [cited 2019 Jul 10];46(5):2159–68. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29401301
5. Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. Nat Commun [Internet]. 2019 [cited 2019 Jul 10];10(1):1025. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30833565
6. Modi A, Vai S, Caramelli D, Lari M. The Illumina Sequencing Protocol and the NovaSeq 6000 System. In: Methods in Molecular Biology [Internet]. Humana Press Inc.; 2021 [cited 2021 Jun 14]. p. 15–42. Available from: https://link.springer.com/protocol/10.1007/978-1-0716-1099-2_2
7. Kronenberg ZN, Rhie A, Koren S, Concepcion GT, Peluso P, Munson KM, et al. Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. Nat Commun 2021 121 [Internet]. 2021 Apr 28 [cited 2021 Jun 4];12(1):1–10. Available from: https://doi.org/10.1038/s41467-020-20536-y
8. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, et al. Accurate

circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019 Oct;37(10):1155–62.

9. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol [Internet]. 2018 [cited 2019 Jul 10];36(4):338–45. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29431738

10. Formenti G, Chiara M, Poveda L, Francoijs KJ, Bonisoli-Alquati A, Canova L, et al. SMRT long reads and direct label and stain optical maps allow the generation of a high-quality genome assembly for the european barn swallow (Hirundo rustica rustica). Gigascience [Internet]. 2018 Nov 29 [cited 2021 Jun 14];8(1):1–9. Available from: http://orcid.org/0000-0002-7554-5991.

11. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Vol. 27, Human molecular genetics. NLM (Medline); 2018. p. R234–41.

12. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era [Internet]. Vol. 21, Nature Reviews Genetics. Nature Research; 2020 [cited 2021 Jun 6]. p. 171–89. Available from: /pmc/articles/PMC7402362/

13. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: The long and the short of it. Vol. 20, Genome Biology. BioMed Central Ltd.; 2019. p. 246.

14. Barseghyan H, Tang W, Wang RT, Almalvez M, Segura E, Bramble MS, et al. Next-generation mapping: A novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. Genome Med. 2017 Oct;9(1).

15. Sharim H, Grunwald A, Gabrieli T, Michaeli Y, Margalit S, Torchinsky D, et al. Long-read single-molecule maps of the functional methylome. Genome Res. 2019 Apr;29(4):646–56.

16. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-

platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun [Internet]. 2019 [cited 2019 Jul 10];10(1):1784. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30992455

17. Kantartzis A, Williams GM, Balakrishnan L, Roberts RL, Surtees JA, Bambara RA. Msh2-Msh3 Interferes with Okazaki Fragment Processing to Promote Trinucleotide Repeat Expansions. Cell Rep. 2012 Aug;2(2):216–22.

18. Santoro M, Masciullo M, Silvestri G, Novelli G, Botta A. Myotonic dystrophy type 1: role of CCG, CTC and CGG interruptions within DMPK alleles in the pathogenesis and molecular diagnosis. Vol. 92, Clinical Genetics. Blackwell Publishing Ltd; 2017. p. 355–64.

19. Liu Q, Zhang P, Wang D, Gu W, Wang K. Interrogating the “unsequenceable” genomic trinucleotide repeat disorders by long-read sequencing. Genome Med. 2017 Jul;9(1):65.

20. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019 Mar;20(1):58.

21. Mantere T, Kersten S, Hoischen A. Long-read sequencing emerging in medical genetics [Internet]. Vol. 10, Frontiers in Genetics. Frontiers Media S.A.; 2019 [cited 2021 Jun 4]. p. 426. Available from: www.frontiersin.org

22. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017 May;27(5):677–85.

23. Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, et al. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biol 2019 201 [Internet]. 2019 May 20 [cited 2021 Jun 4];20(1):1–23. Available from: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1707-2

24. Walker A, Houwaart T, Wienemann T, Vasconcelos MK, Strelow D, Senff T, et al. Genetic structure of SARS-CoV-2 reflects clonal superspreading and multiple independent introduction events, North-Rhine Westphalia, Germany, February and March 2020. Eurosurveillance. 2020 Jun;25(22):2000746.

25. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, et al. Telomere-to-telomere assembly of a complete human X chromosome. Nature [Internet]. 2020 Sep 3 [cited 2021 Jun 4];585(7823):79–84. Available from: https://doi.org/10.1038/s41586-020-2547-7

26. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019 Jun;20(1).

27. Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, et al. De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol. 2018 Dec;36(12):1174–82.

28. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540–6.

29. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020 Feb;17(2):155–8.

30. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods [Internet]. 2021 Feb 1 [cited 2021 Jun 16];18(2):170–5. Available from: https://doi.org/10.1038/s41592-020-01056-5

31. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther. 2016 Mar;17(3):246–53.

32. Deshpande S V., Reed TM, Sullivan RF, Kerkhof LJ, Beigel KM, Wade MM. Offline next generation metagenomics sequence analysis using minION detection software (MINDS). Genes (Basel). 2019 Aug;10(8).

33. Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. Microbiota profiling with long

amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the rrn operon. F1000Research. 2019 Aug;7:1755.

34. Sanderson ND, Street TL, Foster D, Swann J, Atkins BL, Brent AJ, et al. Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices 06 Biological Sciences 0604 Genetics 06 Biological Sciences 0605 Microbiology. BMC Genomics. 2018 Sep;19(1).

35. Burgess DJ. The TOPMed genomic resource for human health. Nat Rev Genet [Internet]. 2021 Apr 1 [cited 2021 Jun 16];22(4):200. Available from: https://www.nature.com/articles/s41576-021-00343-x

36. NIH genome sequencing program targets the genomic bases of common, rare disease | National Institutes of Health (NIH) [Internet]. [cited 2021 Jun 16]. Available from: https://www.nih.gov/news-events/news-releases/nih-genome-sequencing-program-targets-genomic-bases-common-rare-disease

37. De Coster W, Weissensteiner MH, Sedlazeck FJ. Towards population-scale long-read sequencing. Nat Rev Genet [Internet]. 2021 May 28 [cited 2021 Jun 16];1–16. Available from: http://www.nature.com/articles/s41576-021-00367-3

38. Amarasinghe SL, Ritchie ME, Gouil Q. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. Gigascience. 2021 Feb 16;10(2).

39. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, et al. An open resource for accurately benchmarking small variant and reference calls. Nat Biotechnol. 2019 May;37(5):561–6.

40. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018 Jun;15(6):461–8.

41. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: A curated collection of structural variation in the human genome. Nucleic Acids

Res [Internet]. 2014 Jan 1 [cited 2021 Jun 16];42(D1):D986–92. Available from: http://www.ebi.ac.uk/dgva

42. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. 2016 Apr;7.

43. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics [Internet]. Volume 34(Issue 18):Pages 3094–3100. Available from: https://academic.oup.com/bioinformatics/article/34/18/3094/4994778

44. Lin H-N, Hsu W-L. Kart: a divide-and-conquer algorithm for NGS read alignment. Bioinformatics. 2017 Aug;33(15):2281–7.

45. Liu B, Gao Y, Wang Y. LAMSA: fast split read alignment with long approximate matches. Bioinformatics [Internet]. 2017 Jan 15 [cited 2021 Nov 17];33(2):192–201. Available from: https://academic.oup.com/bioinformatics/article/33/2/192/2525703

46. Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, et al. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods 2017 1411 [Internet]. 2017 Sep 18 [cited 2021 Nov 17];14(11):1072–4. Available from: https://www.nature.com/articles/nmeth.4432

47. Haghshenas E, Sahinalp SC, Hach F. lordFAST: sensitive and Fast Alignment Search Tool for LOng noisy Read sequencing Data. Bioinformatics [Internet]. 2019 Jan 1 [cited 2021 Nov 17];35(1):20–7. Available from: https://academic.oup.com/bioinformatics/article/35/1/20/5047762

48. Wei Z-G, Zhang S-W, Liu F. smsMap: mapping single molecule sequencing reads by locating the alignment starting positions. BMC Bioinforma 2020 211 [Internet]. 2020 Aug 4 [cited 2021 Nov 17];21(1):1–15. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-020-03698-w

49. Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, et al. Weighted minimizer sampling improves long read mapping. Bioinformatics [Internet]. 2020 Jul 1 [cited 2021

Nov 17];36(Supplement_1):i111–8. Available from: https://academic.oup.com/bioinformatics/article/36/Supplement_1/i111/5870473

50. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. BMC Bioinformatics. 2012 Sep;13(1).

51. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics [Internet]. 2010 Mar 1;26(5):589–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20080505

52. Liu B, Guan D, Teng M, Wang Y. rHAT: fast alignment of noisy long reads with regional hashing. Bioinformatics [Internet]. 2016 Jun 1 [cited 2021 Nov 17];32(11):1625–31. Available from: https://academic.oup.com/bioinformatics/article/32/11/1625/1742681

53. Nashta-ali D, Aliyari A, Ahmadian Moghadam A, Edrisi MA, Motahari SA, Hossein Khalaj B. Meta-aligner: long-read alignment based on genome statistics. BMC Bioinforma 2017 181 [Internet]. 2017 Feb 23 [cited 2021 Nov 17];18(1):1–9. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1518-y

54. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013 Mar 16; Available from: http://arxiv.org/abs/1303.3997

55. GIAB Consortium. Genome in a Bottle [Internet]. 2021. Available from: https://jimb.stanford.edu/giab

56. The Variant Call Format Specification [Internet]. 2021. Available from: https://github.com/samtools/hts-specs

57. Ni P, Huang N, Zhang Z, Wang D-P, Liang F, Miao Y, et al. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. Elofsson A, editor. Bioinformatics [Internet]. 2019 Nov 1;35(22):4586–95. Available from: https://academic.oup.com/bioinformatics/article/35/22/4586/5474907

58. Mardis ER. The $1,000 genome, the $100,000 analysis? Vol. 2, Genome Medicine.

BioMed Central; 2010. p. 84.

59. Jenko Bizjan B, Katsila T, Tesovnik T, Šket R, Debeljak M, Matsoukas MT, et al. Challenges in identifying large germline structural variants for clinical use by long read sequencing. Vol. 18, Computational and Structural Biotechnology Journal. Elsevier B.V.; 2020. p. 83–92.

60. Counting CAG repeats in the Huntington's disease gene by restriction endonuclease EcoP15I cleavage.

61. Yum K, Wang ET, Kalsotra A. Myotonic dystrophy: disease repeat range, penetrance, age of onset, and relationship between repeat size and phenotypes. Vol. 44, Current Opinion in Genetics and Development. Elsevier Ltd; 2017. p. 30–7.

62. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. Vol. 65, Journal of Human Genetics. Springer Nature; 2020. p. 11–9.

63. Cretu Stancu M, Van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, De Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017 Dec;8(1):1–13.

64. Zhou A, Lin T, Xing J. Evaluating nanopore sequencing data processing pipelines for structural variation identification. Genome Biol [Internet]. 2019 Nov 14 [cited 2021 Jun 4];20(1). Available from: /pmc/articles/PMC6857234/

65. Genome Reference Consortium. GRCh38.p12 [Internet]. [cited 2019 Apr 8]. Available from: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.38/

66. WGS Consortium. NA12878 Nanopore Data [Internet]. [cited 2019 Apr 8]. Available from: https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md

67. GIAB Consortium. NA12878 PacBio Data.

68. The Cancer Genome Atlas. Database of Genomic Variants [Internet]. [cited 2020 Jan 30]. Available from: http://dgv.tcag.ca/dgv/docs/GRCh38_hg38_variants_2016-05-15.txt.

69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence

Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug;25(16):2078–9.

70. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019 Nov;4(43):1686.

71. moodymudskipper/cutr: Enhanced cut And Useful Related Functions version 0.0.0.9000 from GitHub.

72. Chen H, Boutros PC. VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 2011 Jan;12(1):35.

73. LoTempio, JE. Github. Available from: https://github.com/jlotempiojr/longread_alignment_benchmarking

TABLES

Tool	Version	Year	PMID	DOI	Web of Knowledge Citations (2015-20)	Platform	Included (Exclusion criterion)
BWA-MEM	0.7	2011	NA	arXiv:1303.3997	NA	ILMN, ONT, PB	No (3)
LAST	V-984	2011	21209072	10.1101/gr.113985.110	437	ILMN, PB	No (2)
BLASR		2012	22988817	10.1186/1471-2105-13-238	532	PB	No (2)
rHAT	2.2.2	2015	26568628	10.1093/bioinformatics/btv662	13	PB	No (2)
GraphMap	0.5.2	2016	27079541	10.1038/ncomms11307	146	ONT, PB	Yes
LAMSA	NA	2016	27667793	10.1093/bioinformatics/btw594	8	ONT, PB	No (1)
Mashmap2	2.0	2017	30423094	10.1093/bioinformatics/bty597	17	ILMN, ONT, PB	No (1)
Meta-aligner	NA	2017	28231760	10.1186/s12859-017-1518-y	1	PB	No (2)
MECAT	V1.3	2017	28945707	10.1038/nmeth.4432	94	ONT, PB	No (3)
MECAT2	20193.14	2017	28945707	10.1038/nmeth.4432	94	ONT, PB	No (1)
lordFAST	v0.0.10	2018	30561550	10.1093/bioinformatics/bty544	5	ONT, PB	No (1)
Minimap2	2.17	2018	29750242	10.1093/bioinformatics/bty191	782	ILMN, ONT, PB	Yes
NGMLR	0.2.7	2018	29713083	10.1038/s41592-018-0001-7	227	ONT, PB	Yes
Kart	2.5.2	2019	31057068	10.1142/S0219720019500082	2	ONT, PB	No (1)
smsMap	ver1.0.0	2020	32753028	10.1186/s12859-020-03698-w	0	ONT, PB	No (1)
Winnowmap	V2.0	2020	32657365	10.1093/bioinformatics/btaa435	1	ILMN, ONT, PB	No (1)

Table 1. Long-read genome sequence alignment tools.

The tools included in the study are highlighted in green. Shown in the last column is the criterion failed by excluded tools (1= widely adopted, 2= platform agnostic, 3= not superseded by a newer version), as delineated in the Materials and Methods. For details on specificity of

methodological approaches for aligners see Bizjan 2020. ILMN = Illumina SRS, ONT = Oxford Nanopore Technologies, PB= Pacific Biosciences SMRT.

		Computational Benchmarking				Genome Benchmarking		
	Alignment Tool	Wall Clock Hours	CPU Hours	CPU Seconds	Peak Memory Usage (Gb)	x-coverage 3-billion bp	Mapped reads	Unmappe reads
Nanopore	GraphMap	208.2	8327.9	29,980,420	96.30	32.8	13,254,609	2,637,28
	Minmap2	13.5	540.5	1,945,707	14.16	33.5	13,217,207	2,674,69
	NGMLR	63.5	2538.7	9,139,347	33.81	31.1	12,602,742	3,288,41
SMRT	GraphMap	266.1	10642.4	38,312,720	50.06	45.4	38,471,526	2,349,17
	Minmap2	16.0	641.5	2,309,293	16.93	43.6	37,661,104	3,159,60
	NGMLR	16.5	658.4	2,370,240	29.69	39.7	34,765,559	6,041,05

Table 2. Benchmarking metrics summary. Average runtime, memory usage, coverage, and mapped and unmapped reads are shown for three identical runs of each tool, GraphMap, minimap2, and NGMLR on NA12878 genomes with unlimited computational resources, described in Materials and Methods. Data for each individual run and for the resource-restricted runs are show in S1 and a summary of bamstats is shown in S2. Standard error was not calculated as less than 3 measurements were available in some sets due to run failures.

Structural Variant Type	Platform	GraphMap	Minimap2	NGMLR	DGV Variants
DEL	Nanopore	8,666	14,380	11,819	1,353
	SMRT	6,567	10,945	8,561	
INS	Nanopore	6,092	11,469	9,302	4,698
	SMRT	8,200	13,770	10,377	
DUP	Nanopore	0	127	774	1,619
	SMRT	0	155	820	
INV	Nanopore	0	98	183	237
	SMRT	0	102	190	

Table 3. Summary of deletions, insertions, inversions, and duplications present in Sniffles variant call files and the Database of Genome Variants (DGV). No variants to alternate loci in the hg38 reference genome were included in this study.

FIGURE LEGENDS

Figure 1. Unmapped reads. Venn diagrams show the total number of reads (in millions) flagged as unmapped in the sorted BAM files produced by each of the three alignment tools, GraphMap, Minimap2, and NGMLR. Panel A presents unmapped reads from the nanopore dataset. Panel B presents unmapped reads from an SMRT dataset.

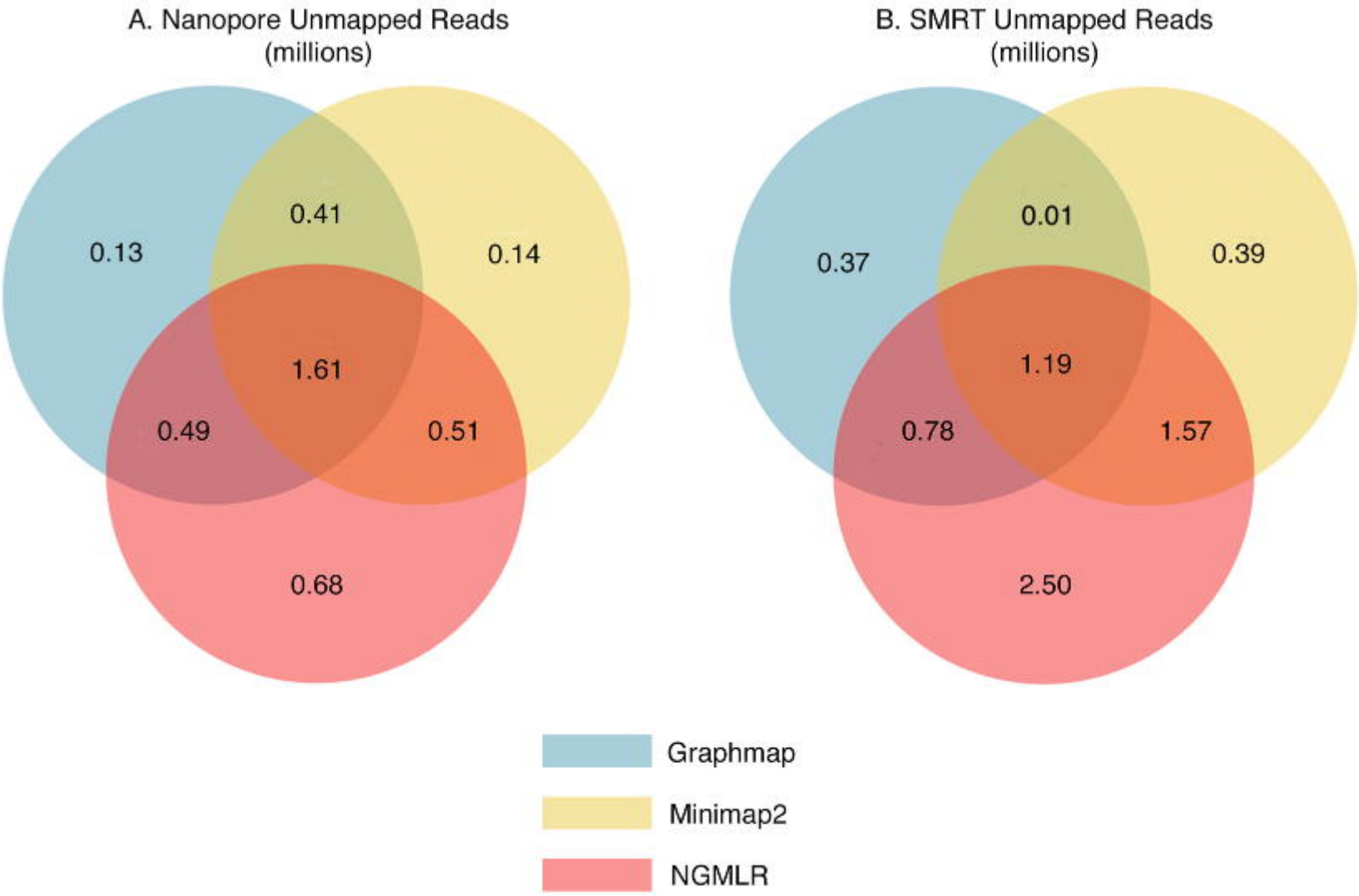
Figure 2. Structural variants present in alignment VCFs. Structural variants called by sniffles on NA12878 BAMs. Calls in the three alignment datasets (shades of blue) compared to variants present in the Database of Genomic Variants (red), organized by SV type, size, and platform. X-axes represent size bins of structural variants, while y-axes represent the count of structural variants present in that bin. GraphMap called only insertions and deletions on these data sets. Please note that the y-axis represents 0-10,000 for insertions and deletion, 1-1,000 for duplications and 1-100 for inversions.

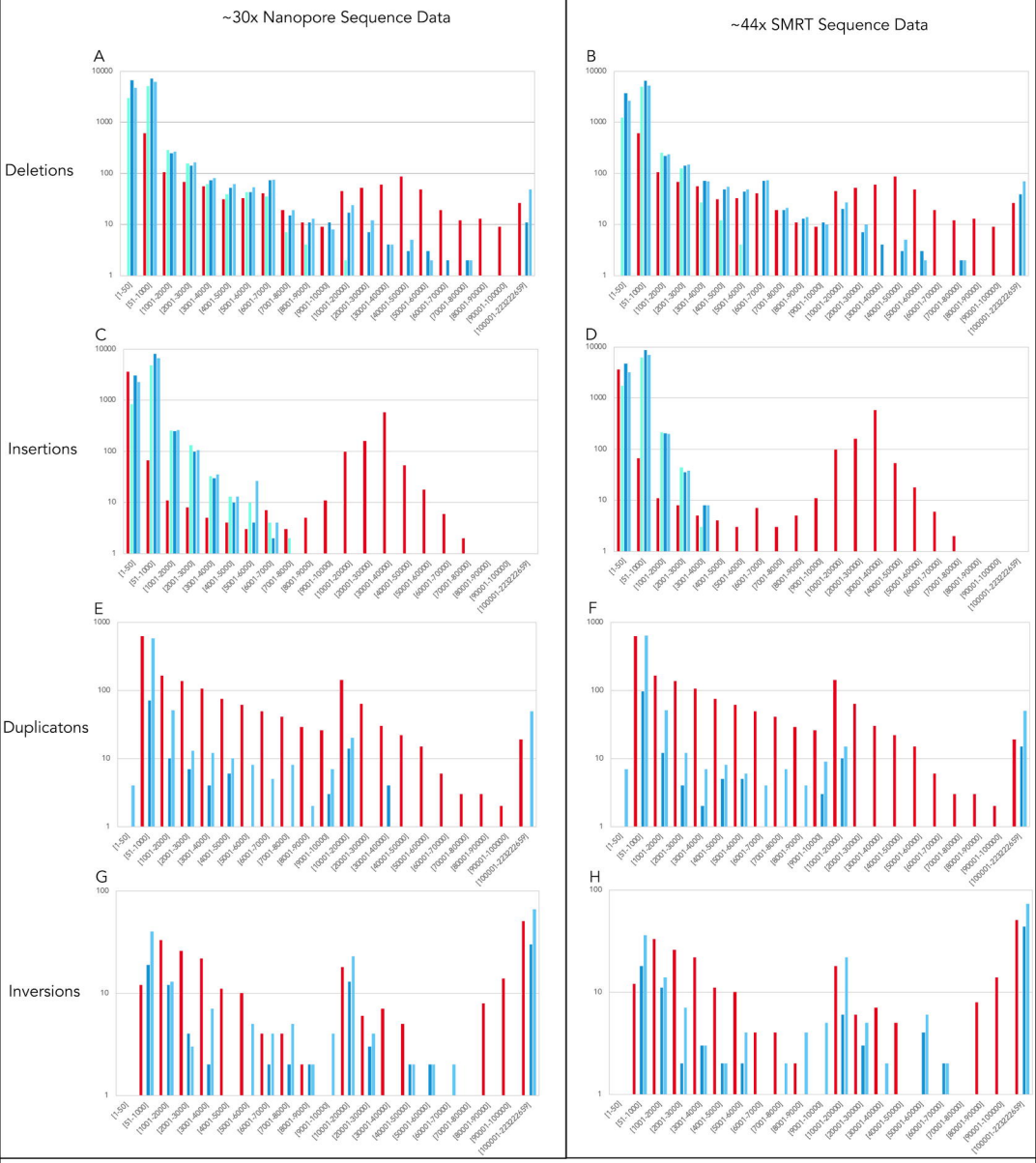
SUPPLEMENTAL TABLE OF CONTENTS

S1: Table of computational metrics

S2: Table of samtools bamstats outputs

S3: Table of structural variant calls by size


A. Nanopore Unmapped Reads
(millions)
0.13
0.41
0.14
1.61
0.49
0.51
0.68
B. SMRT Unmapped Reads
(millions)
0.37
0.01
0.39
1.19
0.78
1.57
2.50
Graphmap
Minimap2
NGMLR


~30x Nanopore Sequence Data
~44x SMRT Sequence Data
Deletions
Insertions
Duplicatons
Inversions
A
B
C
D
E
F
G
H
DGV
GRAPHMAP
MINIMAP2
NGMLR