

UC Irvine

UC Irvine Previously Published Works

Title

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

Permalink

<https://escholarship.org/uc/item/0951f0c3>

Authors

LoTempio, Jonathan

Delot, Emmanuele

Vilain, Eric

Publication Date

2023

DOI

10.7717/peerj.16515

Peer reviewed



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

Jonathan LoTempio^{1,2}, Emmanuele Delot^{3,4} and Eric Vilain^{1,2}

¹Institute for Clinical and Translational Science, University of California, Irvine, CA, United States of America

²International Research Laboratory (IRL2006) “Epigenetics, Data, Politics (EpiDaPo)”, Centre National de la Recherche Scientifique, Washington, DC, United States of America

³Center for Genetic Medicine Research, Children’s National Hospital, Washington, DC, United States of America

⁴Department of Genomics and Precision Medicine, George Washington University, Washington, DC, United States of America

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, which is 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 sample NA24385 on Pacific Biosciences platforms. We employed state of the art sequence alignment tools including GraphMap2, long-read aligner (LRA), Minimap2, CoNvex Gap-cost alignMents for Long Reads (NGMLR), and Winnowmap2. Minimap2 and Winnowmap2 were computationally lightweight enough for use at scale, while GraphMap2 was not. NGMLR took a long time and required many resources, but produced alignments each time. LRA was fast, but only worked on Pacific Biosciences data. Each tool widely disagreed on which reads to leave unaligned, affecting the end genome coverage and the number of discoverable breakpoints. No alignment tool independently resolved all large structural variants (1,001–100,000 base pairs) present in the Database of Genome Variants (DGV) for sample NA12878 or the truthset for NA24385.

Conclusions. These results suggest a combined approach is needed for LRS alignments for human genomics. Specifically, leveraging alignments from three tools will be more effective in generating a complete picture of genomic variability. It should be best practice to use an analysis pipeline that generates alignments with both Minimap2 and Winnowmap2 as they are lightweight and yield different views of the genome. Depending on the question at hand, the data available, and the time constraints, NGMLR and LRA are good options for a third tool. If computational resources and time are not a factor for a given case or experiment, NGMLR will provide another view, and another chance to resolve a case. LRA, while fast, did not work on the nanopore data for our cluster, but PacBio results were promising in that those computations completed faster than Minimap2. Due to its significant burden on computational resources and slow run time, Graphmap2 is not an ideal tool for exploration of a whole human genome generated on a long-read sequencing platform.

Submitted 25 July 2023
Accepted 2 November 2023
Published 18 December 2023

Corresponding author
Eric Vilain, evilain@hs.uci.edu

Academic editor
Xin Zhang

Additional Information and
Declarations can be found on
page 19

DOI [10.7717/peerj.16515](https://doi.org/10.7717/peerj.16515)

© Copyright
2023 LoTempio et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Bioinformatics, Computational Biology, Genomics, Translational Medicine

Keywords Long-read sequence, Genome alignment, Benchmark, Genomic medicine, Short-read sequence

INTRODUCTION

The diverse ecosystem of DNA preparation, sequencing, and mapping technologies is capable of generating computational representations of genome biology through different chemistries and processes on different sequencing and mapping platforms (*Sanger, Nicklen & Coulson, 1977; Fuller et al., 2009; Branton et al., 2008; Ardui et al., 2018; Levy-Sakin et al., 2019*). This has resulted in many different approaches for making use of the data of different sources, with algorithms and the tools built upon them used across platforms to differing success. Here we briefly outline the differences between these platforms and analysis strategies so that we may consider an important gap in applying the newest technologies to the problem of human genomics.

Sequencing-by-synthesis platforms produce highly accurate sequence data in the form of short-reads (<300 base pairs, bp) (*Modi et al., 2021*) from high molecular weight DNA inputs. In contrast, single-molecule real-time platforms can produce highly accurate reads through circular consensus sequencing (CCS) on molecules >10 kilobase pairs (kbp) (*Wenger et al., 2019*) while nanopore-based sequencing (*Jain et al., 2018a; Jain et al., 2018b*) and nanochannel-based mapping platforms (*Formenti et al., 2019*) can sequence or visualize megabase-length DNA molecules (*Pollard et al., 2018*). Each of these platforms has different utilities and available tools, which contributes to the diversity of projects enabled by these technologies (*Ho, Urban & Mills, 2020; Mahmoud et al., 2019*).

Optical genome mapping (developed by Bionano Genomics, San Diego, CA, USA) excels at detecting 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 (*Barseghyan et al., 2017; Sharim et al., 2019*), and cancer (*Neveling et al., 2021; Talsania et al., 2022; Bornhorst et al., 2023*). But long-read sequence (LRS), developed by Pacific Biosciences (Menlo Park, CA, USA) and Oxford Nanopore Technologies (Oxford, UK) among others, has been shown to be the most appropriate technology to detect smaller variants in the 50 bp-1 kb range (*Chaisson et al., 2019*). This is especially true in repetitive regions of the genome and should therefore bring new diagnosis potential for diseases of trinucleotide repeat expansion and genome instability such as Huntington's disease and myotonic dystrophies (*Kantartzis et al., 2012; Santoro et al., 2017; Liu, Gao & Wang, 2017; Mitsuhashi et al., 2019; Mantere, Kersten & Hoischen, 2019*), 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 further shown to be capable of generating single, phaseable reads spanning repetitive or complex genomic regions that remain unresolved in all (*Wenger et al., 2019; Jain et al., 2018a; Jain et al., 2018b; Huddleston et al., 2017; Ebbert et al., 2019*) but one current human genome assembly (*Nurk et al., 2022*). This has implications for

resolution of highly similar pseudogenes and large SV, even in low-complexity genomic regions, or from otherwise healthy diploid samples. Read lengths in the tens of kbp have further allowed for end-to-end viral genome sequencing ([Walker et al., 2020](#)), while read lengths in the millions of bp have the potential to span whole mammalian chromosomes ([Jain et al., 2018a](#); [Jain et al., 2018b](#); [Miga et al., 2020](#)). This potential for more contiguous *de novo* human assembly has led to studies to specifically improve and benchmark basecalling and polishing tools ([Wick, Judd & Holt, 2019](#)), as well as assembly tools ([Koren et al., 2018](#); [Cheng et al., 2021](#); [Kolmogorov et al., 2019](#)) that can be applied to human genomics.

For reference genome-guided experiments, LRS has proven useful for amplicon sequencing in cancer detection ([Norris et al., 2016](#)) and for metagenomics ([Deshpande et al., 2019](#); [Cuscó et al., 2019](#); [Sanderson et al., 2018](#)), but the field is still in early days for assessment of variation in high-coverage human whole genome sequence. There has been a demonstration of the added value of LRS in resolving SV has been demonstrated by efforts aggregating callsets across platforms and technologies to deeply characterize a genome ([Chaisson et al., 2019](#)). 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 published tool showcases its own 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 alignment platforms and tools for reference-guided experiments.

To address this gap, we benchmarked the most recent LRS alignment tools with the datasets generated from the Joint Initiative for Metrology in Biology's Genome in a Bottle Initiative (GIAB), specifically samples, NA12878 sequenced with nanopore technology and NA24385 sequenced with CCS (branded HiFi) technology. These were generated with Oxford 9.4 pore chemistry and with Guppy5 basecalling performed by the Whole Genome Sequencing Consortium ([Jain et al., 2018a](#); [Jain et al., 2018b](#)) and with Pacific Biosciences RSII SMRT CCS technology by PacBio and the National Institute of Standards and Technology ([Zook et al., 2019](#)). All alignments were performed on GRCh38, rather than the more complete CHM13 or more diverse pangenome graphs because there are not yet high-quality GIAB-like reference materials or variant truthsets that allow for this sort of benchmarking against these cutting-edge reference materials.

We compared computational performance (peak memory utilization, central processing unit (CPU) time, file size/storage requirements), genome depth and basepair coverage, and quantified the reads left unaligned in any given experiment. We have limited this study to tools that are platform agnostic and function on our cluster. 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 ([Sedlazeck et al., 2018](#)). Taken together, these experiments present a comprehensive view of differences in the products of LRS whole-genome alignment pipelines.

MATERIALS & METHODS

Portions of this text were previously published as part of a preprint (<https://www.biorxiv.org/content/10.1101/2021.07.09.451840v3.full>) (LoTempio, Délot & Vilain, 2023).

Tool selection criteria

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. We also used the website Long-Read Tools (<https://long-read-tools.org/>), searching their database with the filters “nanopore”, “pacbio”, and “alignment”. Since this yielded many software tools, we dove deeper and excluded any tools not designed for whole genome experiments. Tools which passed this test were then assessed for their suitability for use with nanopore and SMRT data and whether they produced a SAM/BAM file for analysis. 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.

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 most descriptive. Specifically, Sniffles outputs VCFs which call SVs as indels, translocations, duplications, inversions, and more classifications. This increases its utility for comparison to reference truthsets.

Data

Main experiments

Reference genome: GRCh38 was accessed on March 22, 2022 from NCBI (GRCh38_no_alt_analysis_set).

Sequence data

No new or original sequence data from new or existing samples were generated here. Instead, we leveraged data generated by third party consortia on well-characterized samples. These data include nanopore sequence data generated on NA12878 from the WGS Consortium, rel7 guppy 5 basecalls was accessed on March 22, 2022 (NA12878 rel7).

We also used the 15kb insert size SMRT CCS data from sample NA24385 was accessed from the human pangenomics consortium GitHub on March 22, 2022 (NA24385 15kb insert).

While no DOI exist for these files, the citations provide the most stable URLs available ([PacBio, 2021](#)).

SV truthsets

Structural variant truthsets for NA12878, updated February 25, 2020, from DGV were accessed on March 31, 2022 (NA12878 DGV truthset).

Structural variant truthsets from Pacific Biosciences and the GIAB consortium were accessed on May 20, 2022 (PacBio GIAB SV Truthset).

Absolute coordinate systems within a reference genome have allowed for very accurate genome mapping and analysis within that reference. However, reference sets are tied to these absolute coordinate systems and require liftover and curation. Our study is subject

to the limitation that these variants were called with alignments to GRCh37, but we used the more advanced GRCh38. Accordingly, the absolute coordinates of the location of breakpoints differs across builds. By focusing our study on the size of variants within 1,000 and 10,000 bp windows, rather than their relative position within a set of reference genome coordinates, we are confident that this is a minor limitation.

Pilot experiments

Initial pilot experiments were performed on older reference builds and datasets. These can be found as [Supplement S4](#) and may be useful to investigators who have LRS data from the previous decade.

Reference genome

GRCh38.p12 was accessed and downloaded on April 8, 2019 (GRCh38.p12).

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 (NA12878 rel 5; NA12878 SMRT Mt Sinai).

Database of genome variants

DGV data were accessed on January 30, 2020 (Database of Genomic Variants, DGV). The full set of variants was reduced to 11,042 variants confirmed to be present in NA12878. A total of 14 of these variants were excluded as they were called on contigs present out of the main reference assembly contigs.

Hardware configuration

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

Whole genome alignment tool benchmarking and analysis

The following tool versions were used in the work presented in the main tables and figures.

1. Long-read aligner (LRA) v1.3.3. ([Ren & Chaisson, 2021](#))
2. GraphMap2 v0.6.3 ([Marić et al., 2019](#))
3. Minimap2 v2.24 ([Li, 2018](#))
4. CoNvex Gap-cost alignMents for Long Reads (NGMLR) v0.2.7 ([Sedlazeck et al., 2018](#))
5. Winnowmap v2.0.3 ([Jain et al., 2022](#))

The versions of alignment tools used in the preliminary experiments of this project available in [S4](#) included:

1. GraphMap v0.5.2 ([Sović et al., 2016](#))

2. Minimap2 v2.16 ([Li, 2018](#))
3. NGMLR v0.2.7 ([Sedlazeck et al., 2018](#))

All tools were run with recommended parameters and were flagged for nanopore or SMRT data based on the requirements of that experiment. Where multithreading was available, `-t`, or equivalent was set to 40, the max for our cluster. LRA, Minimap2, NGMLR, and Winnowmap2 all have specific presets for SMRT and nanopore-based data while Graphmap2 uses the same parameterization for all LRS platforms, with an option for alternative defaults to handle Illumina data. Its default parameters are optimized for accuracy and through graph-based handling of reads can address complex genomic regions and users have the option of modifying parameters to decrease accuracy.

LRA uses the same minimizers as seeds for both sequencing platforms, but decreases the density of minimizers in the genome by nearly 10-fold, which are then compared to minimizers from the reference, resulting in a set of query-reference anchors. Anchors are clustered based on read accuracy, with lower accuracy nanopore reads given a minimum of three clusters per read, and higher accuracy SMRT-CCS reads given a minimum of 10 clusters per read.

Minimap2, upon which Winnowmap2 builds, uses homopolymer-compressed minimizers in the seeding step of its algorithm. This is to account for errors in SMRT movie data which can struggle with homopolymer runs. For example, GATTAACCA becomes GATACA through homopolymer minimization. The nanopore default parameterization uses ordinary minimizers as seeds because in development, they saw no increased benefit to homopolymer compression, even though the technology has similar limitations to SMRT. Winnowmap2 builds on this by down-weighting frequently occurring minimizers to minimize masking seeds in complex genomic regions including tandem repeats.

For NGMLR generally, reads are handled in a way which is SV-aware and accounts for small (10 bp) indels or split reads over larger indels. Reads which can be mapped linearly are, and the remaining reads are handled through splitting them over SV. The nanopore preset further customizes parameters to address the characteristic high error rates, including substitutions, insertions, and deletions, inherent to nanopore sequencing. It achieves this by lowering match/mismatch scores, increasing gap penalties, and emphasizing sensitivity, making it well-suited for applications like structural variant analysis. Conversely, the SMRT preset retains higher match/mismatch scores to prioritize base accuracy, suitable for PacBio data known for lower substitution errors but higher indel error rates.

Computational metrics were printed from SLURM job records with `seff` and `sacct` with flags `format JobID, JobName, Elapsed, NCPUs, TotalCPU, CPUTime, ReqMem, MaxRSS, MaxDiskRead, MaxDiskWrite, State, ExitCode`. Samtools 1.15.1 was used for all alignment manipulations, alignment read depth coverage calculations, and to extract unmapped reads. Samtools `view -f 4` was used to generate bamstats files and Python Venn was used to compare the readnames across unmapped read files to assess the degree of overlap of unmapped reads across these subsetted alignment files ([Li et al., 2009](#); [Danecek et al., 2021](#)). R 3.5.2 version Eggshell Igloo with the tidyverse packages were used for preliminary experiments on deprecated data shown in [S4](#). Genome coverage was calculated with

samtools coverage. Binary conversion values were used for bytes (1,073,741,824) and kilobytes (1,048,576) to gigabytes.

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 semicolons) are not directly usable in python pandas. Relevant dataframes from genome alignment files were reshaped with shell scripts and Python scripts whose methodology and key intermediate files are available on our GitHub.

Comparison of breakpoints

Structural variants (SV) were called with sniffles/1.0.11 with default parameters (*Sedlazeck et al., 2018*). Only variants called on the main GRCh38 assembly were included. Python was used to bin and graph structural variants by SV Type. Intermediary files and scripts are available at our GitHub, archived stably here: <https://zenodo.org/badge/latestdoi/429362081>.

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 seaborn and matplotlib as well as Microsoft Excel (*Waskom, 2021; Hunter, 2007*).

RESULTS

Rigorously annotated variants were drawn from the Database of Genomic Variants (DGV) (*MacDonald et al., 2014*) for NA12878 and from a curated public repository for NA24385 (*PacBio, 2021*).

We present our results in four sections:

- (1) the tools that were included,
- (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.

Tools that passed the inclusion/exclusion criteria

Following a literature review of available alignment tools and search of Long-Read-Tools (*Amarasinghe, Ritchie & Gouil, 2021*), 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). All relevant tools and their reason for exclusion are outlined in [Table 1](#), with all tools annotated as platform agnostic and relevant to genome alignment included in [Supplemental S1](#). Five alignment tools, GraphMap2, LRA, minimap2, NGMLR, and Winnowmap2 were included in this study (*Marić et al., 2019; Li, 2018; Sedlazeck et al., 2018; Jain et al., 2022; Ren & Chaisson, 2021*).

Sixteen tools were excluded because they were superseded by other tools, produced an alignment file in a non-SAM format, were designed for only one type of input data, or did not work on our cluster (*Kielbasa et al., 2011; Li, 2013; Chaisson & Tesler, 2012; Faust &*

Table 1 Long-read genome sequence alignment tools. The tools included in the study are highlighted in this study. Further tools from Long-read-tools.org are available in [S1](#).

Tool	Year	PMID	DOI	Inclusion	If no, why?
BLASR	2012	22988817	10.1186/1471-2105-13-238	No	Designed only for SMRT seq
BWA-MEM	2011	NA	arXiv:1303.3997	No	Superseded by Minimap2
GraphMap	2016	27079541	10.1038/ncomms11307	No	Superseded by GraphMap2
GraphMap2	2019	NA	10.1101/720458	Yes	
Kart	2019	31057068	10.1142/S0219720019500082	No	Designed only for SMRT, SBS
LAMSA	2016	27667793	10.1093/bioinformatics/btw594	No	Deprecated dependencies resulting in seg fault
LAST	2011	21209072	10.1101/gr.113985.110	No	Not parameterized for the nuances of current platforms
lordFAST	2018	30561550	10.1093/bioinformatics/bty544	No	Runs resulted in partial completion and a seg fault
LRA	2021	34153026	10.1371/journal.pcbi.1009078	Yes	
Mashmap2	2017	30423094	10.1093/bioinformatics/bty597	No	Does not use SAM format needed for variant caller
MECAT	2017	28945707	10.1038/nmeth.4432	No	Superseded by Mecat2
MECAT2	2017	28945707	10.1038/nmeth.4432	No	Designed only for SMRT seq
Meta-aligner	2017	28231760	10.1186/s12859-017-1518-y	No	Designed only for SMRT seq
Minimap2	2018	29750242	10.1093/bioinformatics/bty191	Yes	
NGMLR	2018	29713083	10.1038/s41592-018-0001-7	Yes	
rHAT	2015	26568628	10.1093/bioinformatics/btv662	No	Designed only for SMRT seq
S-ConLSH	2021	33573603	10.1186/s12859-020-03918-3	No	Designed only for SMRT seq
smsMap	2020	32753028	10.1186/s12859-020-03698-w	No	Does not use SAM format needed for variant caller
Winnowmap	2020	32657365	10.1093/bioinformatics/btaa435	No	Superseded by Winnowmap2
Winnowmap2	2022	35365778	10.1101/2020.11.01.363887	Yes	
YAHA	2012	22829624	10.1093/bioinformatics/bts456	No	Not parameterized for the nuances of current platforms

Hall, 2012; Liu et al., 2016; Sović et al., 2016; Liu, Gao & Wang, 2017; Jain et al., 2018a; Jain et al., 2018b; Nashta-Ali et al., 2017; Xiao et al., 2017; Haghshenas, Sahinalp & Hach, 2019; Li, 2018; Sedlazeck et al., 2018; Marić et al., 2019; Kumar, Agarwal & Ranvijay, 2019; Wei, Zhang & Liu, 2020; Jain et al., 2020; Jain et al., 2022; Ren & Chaisson, 2021; Chakraborty, Morgenstern & Bandyopadhyay, 2021).

We excluded LAST and YAHA as they were not parameterized for the nuances of current sequencing platforms, *i.e.*, they did not have default settings that took into account long reads and low error rates because their success would require more careful consideration of research questions, rather than exploration. smsMap produced an output that would require post-processing to put it into a SAM compatible format, and since we did not aim to create new software here, we excluded it. LAMSA, lordFAST and were run, but resulted in segmentation faults (*Haghshenas, Sahinalp & Hach, 2019; Liu, Gao & Wang, 2017*). LAMSA was built to call precompiled GEM3 libraries, with which our current compute architecture was incompatible, while lordFast produced an undiagnosable error log (*Marco-Sola et al., 2020*). The other 11 tools, outlined in [Table 1](#), were excluded because they were only designed to handle data from one LRS platform.

Table 2 Benchmarking metrics. Each of the three runs for each tool is shown in this table. All values are taken directly from slurm's sacct and seff features. All metrics are intuitive except for CPU efficiency, which is a measure of idle CPUs to CPU time over wall clock time.

Platform	Tool	Wall clock time (D-H:M:S)	CPU time (D-H:M:S)	CPU efficiency	Memory utilized (Gb)
Nanopore	Graphmap2	14-00:00:04	560-00:02:40	65.25%	90.19
		14-00:00:17	560-00:11:20	63.34%	94.58
		14-00:00:15	560-00:10:00	64.23%	89.68
	Minimap2	1-11:00:08	58-08:05:20	7.51%	12.43
		1-10:48:18	58-00:12:00	7.51%	12.80
		1-11:07:00	58-12:40:00	7.51%	12.78
	NGMLR	10-00:00:03	400-00:02:00	97.70%	35.86
		10-00:00:10	400-00:06:40	99.56%	34.38
		9-14:55:56	384-21:17:20	93.43%	36.34
	Winnowmap2	15:25:58	25-17:18:40	83.82%	29.48
		15:19:14	25-12:49:20	83.61%	30.02
		15:15:24	25-10:16:00	83.06%	28.74
	Graphmap2	5-01:33:22	202-14:14:40	98.88%	45.07
		5-02:54:53	204-20:35:20	98.81%	44.90
		5-06:49:25	211-08:56:40	98.52%	48.96
LRA	6:53:00	11-11:20:00	33.20%	20.63	
	6:53:04	11-11:22:40	33.11%	21.10	
	6:32:33	10-21:42:00	33.22%	27.91	
SMRT-CCS	Minimap2	11:02:04	18-09:22:40	7.51%	9.10
		11:06:03	18-12:02:00	7.51%	9.13
		10:55:09	18-04:46:00	7.52%	9.09
	NGMLR	21:50:29	36-09:39:20	99.20%	25.53
		22:24:38	37-08:25:20	99.19%	25.69
		21:58:53	36-15:15:20	99.53%	25.82
Winnowmap2	3:39:42	6-02:28:00	88.10%	18.57	
	3:43:42	6-05:08:00	88.32%	18.71	
		3:40:32	6-03:01:20	87.91%	19.14

Computational performance and benchmarking

Computations were run three times on a single node allocated fully on our university's high-performance compute cluster (HPC), which includes 40 CPUs with a configuration described in the methods. All tools were run with default parameters to reflect typical use in exploratory studies. High-level computational benchmarks can be found in [Table 2](#). Full reports on each run can be examined in [Supplemental S2](#), while a composite table of the output of samtools coverage is available as [Supplemental S3](#).

Globally successful tools

Minimap2 was the least memory-demanding tool

Minimap2 successfully aligned nanopore data every time with unrestricted and restricted resources. Unrestricted runs used around 12.6 gigabytes (Gb) and the jobs took ~34 wall clock hours to complete.

Minimap2 also successfully aligned SMRT data every time with unrestricted or restricted resources. The runs used around 9 Gb of memory and the jobs took ~11 wall clock hours to complete, considerably faster than the (admittedly larger) nanopore dataset.

Memory usage 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.

Winnomap2 was the fastest tool

Winnomap2 was the fastest to run on any dataset. Specifically, it aligned the entire nanopore dataset in around 15 wall clock hours, using only a little more than two times more memory than minimap2 (~29 Gb). For SMRT, it was even faster at just over 3.5 wall clock hours and almost exactly twice as much memory (~18 Gb). These numbers were consistent over all runs.

Additionally, the high CPU efficiency of the jobs on our cluster were positive to note: over 80% efficiency, relative to Minimap2's ~7% efficiency. This is a measure of the ratio of CPU time used to the wall clock time times the number of CPUs. In the case of our cluster, CPUs are fully allocated to a node and not shared once allocated, but it does point the way towards either choosing more efficient tools, or optimizing jobs.

NGMLR completed 4/6 tasks and performance demanded great resources

NGMLR successfully aligned SMRT data every time but was the second slowest. The runs used around 25 Gb and the jobs took around 22 wall clock hours to complete, nearly twice as long as the next tool, Minimap2.

NGMLR was much more inconsistent on nanopore data. It successfully aligned the data on our first range finding experiment in approximately ~230 wall clock hours. Following this run, two jobs were set with a time of 10 days, and both resulted in timeouts. The single successful run used 36.34 Gb of memory, the most of all tools which produced an alignment.

Tools that presented a challenge

LRA did not work on nanopore-generated data

While LRA did show promise on SMRT data, aligning a genome in ~11 CPU days or ~6 wall clock hours, we could not run it successfully with nanopore. Each run resulted in unresolvable (undescriptive) segmentation faults that produced no partial alignment. In the case of SMRT data, it did fall in the front of the pack, aligning genomes quickly with typical memory usage relative to its peers (~20–27 GB).

GraphMap2 was the most resource-intensive tool

GraphMap2 took the longest, used the most memory, and had the most failures of all of the tools we examined. It did not produce alignment files on nanopore data in the time limit of 14 days imposed by our HPC. On the SMRT dataset which it was successful at aligning, it took more than 5 times longer to produce an alignment file than the next slowest tool, NGMLR. It had the highest memory usage, in excess of 89 GB nanopore on 44 GB on SMRT.

Whole genome alignment

Aligned reads show disagreement in coverage of each chromosome

In preliminary experiments that were completed successfully, alignments of the same genome with the same tool had the same genome coverage, whether resource-restricted or not (S4, column O). For this reason, one file generated from the first run was used for subsequent analysis, shown in Table 3, a summary of the coverage depth of each chromosome alignment and the percentage of the basepairs of GRCh38 covered. Global genome coverage of $\sim 30x$ was reported for nanopore and $\sim 34x$, 20x of which is present in the 15 kbp insert set, was reported for SMRT data sets. Breaking down the summary into chromosomes allows for more precision than referring to a genome by its coverage as a uniform metric, since there is discrepancy across the chromosomes.

For nanopore data, three alignment files were measured, while for SMRT data, four files could be included. Table 3 shows cover of each chromosome with high values represented in bold text, and low values represented in italic text. It becomes immediately apparent that Minimap2 and Winnowmap2 both retain the most reads (highest x coverage) and cover the most basepairs of the reference. NGMLR excludes the most reads and covering the fewest basepairs of the reference genome, which may impact downstream analysis. The final tool, LRA, which only worked on SMRT data, is intermediate with some lower values of coverage depth, and some higher values of coverage of basepairs.

Unaligned reads reveal differential exclusion of reads

The readname assigned to each read in a fastq retained in the BAM allowed us to directly compare the lists of reads that were not included in the alignment by each alignment tool with Python Venn (Fig. 1) (Grigorev, 2018).

All tools agreed to leave ~ 1 million of the same nanopore reads unaligned, but differed in their overall totals of unaligned reads. NGMLR left the highest number of reads unaligned, ~ 3.1 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 Winnowmap2 agreed on a separate ~ 0.2 million unaligned reads and Winnowmap2 and Minimap2 on ~ 0.03 million further unaligned reads.

All tools agreed to leave 682 of the same SMRT reads unaligned. This is because Winnowmap2 excludes less than 1,000 reads in total, where the other tools excluded more reads over roughly three orders of magnitude. NGMLR left the highest number of reads unaligned, ~ 200 thousand, which explains why it produced the lowest coverage genome. LRA excluded around 90 thousand reads in all, of which around 70 thousand are common

Table 3 Sequencing depth and reference coverage. Chromosome level descriptions of the average number of reads covering each basepair, as well as the percentage of basepairs of the reference covered. The highest value for a chromosome is accented with bold text, while the lowest value is accented with italic text. Winnowmap2 abbreviated as WM2.

	SMRT-CCS								Nanopore						
	Mean depth (x-depth)				Coverage of basepairs (%)				Mean depth (x depth)			Coverage of basepairs (%)			
	LRA	Minimap2	NGMLR	WM2	LRA	Minimap2	NGMLR	WM2	Minimap2	NGMLR	WM2	Minimap2	NGMLR	WM2	
chr1	19.0627	20.0049	<i>17.9711</i>	20.002	91.7893	92.3244	<i>91.1358</i>	92.3622	chr1	37.7867	<i>33.3715</i>	37.4783	92.4802	<i>91.4793</i>	92.4861
chr2	20.0613	20.1343	<i>19.9393</i>	20.1463	98.7397	98.8041	98.707	<i>98.6286</i>	chr2	40.1079	36.3557	39.497	99.0601	<i>98.8241</i>	98.8272
chr3	20.1303	20.5557	<i>20.0446</i>	20.3575	98.984	99.5552	<i>98.7438</i>	99.5645	chr3	37.7894	<i>36.5499</i>	39.9339	99.773	<i>98.8066</i>	99.767
chr4	20.4284	20.7247	<i>20.1254</i>	21.8975	99.1528	99.0963	<i>98.6843</i>	99.2809	chr4	39.8297	<i>36.1401</i>	39.6729	99.4841	<i>98.7817</i>	99.5176
chr5	20.0292	20.3439	<i>19.7821</i>	20.0655	98.2323	98.1787	<i>97.6187</i>	98.2961	chr5	37.1564	<i>35.9384</i>	37.17	98.5218	<i>98.0472</i>	98.5096
chr6	19.9858	20.4916	<i>19.9745</i>	20.4963	98.795	99.0781	<i>98.674</i>	99.2314	chr6	37.4983	<i>36.4317</i>	38.3408	99.4746	<i>98.9018</i>	99.4822
chr7	19.5124	19.7661	<i>19.3123</i>	19.7686	98.1498	99.2092	<i>97.7858</i>	98.9308	chr7	37.5093	<i>35.7748</i>	37.3072	99.6968	<i>98.4965</i>	99.6698
chr8	20.0067	20.279	<i>19.7645</i>	20.272	98.5302	99.2526	<i>98.1774</i>	99.0982	chr8	37.2117	<i>35.8164</i>	37.4528	99.6024	<i>98.4431</i>	99.6027
chr9	17.085	17.2704	<i>16.5087</i>	17.2736	86.2352	86.2731	<i>85.4122</i>	86.3347	chr9	33.7857	<i>31.5634</i>	33.62	87.3625	<i>86.6206</i>	87.4715
chr10	19.8087	20.0826	<i>19.4999</i>	20.5247	98.2398	98.8789	<i>98.1124</i>	98.8153	chr10	37.9547	<i>35.509</i>	37.9365	99.3546	<i>98.2053</i>	99.3873
chr11	19.5361	19.8058	<i>19.4349</i>	19.8105	<i>98.0161</i>	99.1255	98.1379	98.264	chr11	37.5146	<i>35.9223</i>	37.282	99.562	<i>98.5893</i>	99.0318
chr12	19.6294	19.9493	<i>19.6137</i>	19.9618	98.3953	99.548	<i>98.151</i>	99.3563	chr12	37.9345	<i>36.4544</i>	37.6064	99.8008	<i>98.397</i>	99.7298
chr13	17.5257	18.1764	<i>16.9585</i>	17.6178	85.2628	85.3114	<i>84.1626</i>	85.3793	chr13	32.6205	<i>30.5633</i>	32.35	85.6486	<i>84.3648</i>	85.6546
chr14	16.815	16.8162	<i>16.6225</i>	16.9593	82.5962	82.5668	82.5949	82.5858	chr14	31.2147	<i>30.4015</i>	31.9599	82.5971	<i>82.5928</i>	82.5975
chr15	16.0919	16.3256	<i>15.7467</i>	16.3303	80.7849	81.7561	<i>80.0048</i>	81.9674	chr15	32.9232	<i>30.434</i>	32.7613	82.7815	<i>80.8381</i>	82.8274
chr16	18.4585	19.5281	<i>18.1335</i>	19.4345	89.1223	90.0939	<i>87.9862</i>	90.1285	chr16	36.9247	<i>34.5339</i>	37.2781	90.5049	<i>88.4413</i>	90.5183
chr17	18.0451	18.8278	<i>18.0519</i>	18.7767	95.1484	97.5541	<i>94.7803</i>	95.9119	chr17	37.2793	<i>35.2194</i>	36.9045	97.3704	<i>95.1519</i>	98.046
chr18	19.4328	19.7389	<i>18.8198</i>	19.7815	96.004	98.3045	<i>94.3649</i>	97.8677	chr18	36.8051	<i>34.2153</i>	36.5148	99.3802	<i>94.1824</i>	99.0731
chr19	16.8558	16.8674	<i>16.5895</i>	16.8926	95.0649	95.0588	<i>94.9148</i>	95.0692	chr19	36.3404	<i>33.9213</i>	39.4614	95.1037	<i>95.0137</i>	95.1095
chr20	19.34	20.0554	<i>19.0343</i>	20.1369	96.0126	98.1195	95.5606	98.0231	chr20	39.3238	<i>35.3079</i>	39.1318	99.1143	<i>96.3336</i>	99.1551
chr21	18.0549	18.4984	<i>16.8995</i>	18.2805	80.4792	80.4144	<i>78.9093</i>	80.16	chr21	35.4168	<i>30.0951</i>	34.298	81.3939	<i>80.763</i>	81.2955
chr22	13.9854	14.0173	<i>13.3346</i>	17.4726	72.7139	72.6378	<i>72.5114</i>	72.7444	chr22	32.7008	<i>25.9485</i>	31.1753	72.9147	<i>72.9092</i>	<i>72.9018</i>
chrX	9.8603	9.97983	<i>9.73401</i>	9.98727	97.639	97.9196	<i>96.4952</i>	98.0195	chrX	36.1372	<i>33.9833</i>	35.6273	99.1734	<i>97.0137</i>	98.9417
chrY	7.94848	8.21933	<i>3.85465</i>	8.37289	40.9697	40.9213	<i>37.9947</i>	40.8261	chrY	1.8305	<i>0.501775</i>	1.3745	3.1548	4.9006	<i>2.37513</i>
chrM	<i>1010.03</i>	1203.88	1203.67	1202.19	100	100	100	100	chrM	11954.1	<i>11413.1</i>	11685.6	100	100	100

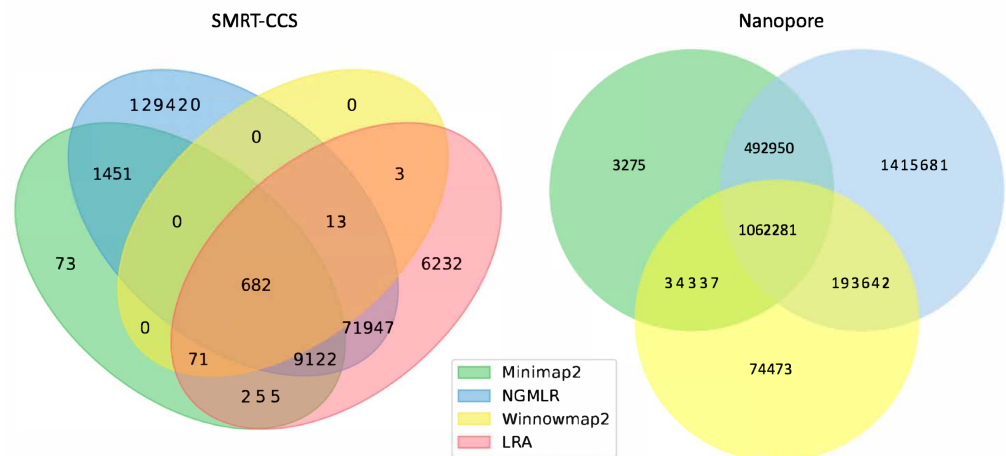


Figure 1 Unmapped reads. Venn diagrams show the total number of reads flagged as unmapped in the sorted BAM files. The left panel presents unmapped reads from the SMRT dataset. The right panel presents unmapped reads from the nanopore dataset.

Full-size DOI: 10.7717/peerj.16515/fig-1

to NGMLR. This also likely contributes to its lower-coverage status shown in Table 3. Minimap2 leaves out just shy of 12 thousand reads in all, most of which are shared with NGMLR and LRA.

In all, not enough information is available at this level of granularity to pass judgment on whether or not tools include or exclude the right reads. For this, an examination of breakpoints is required.

Breakpoints reveal differences between competing alignments of the same genome

Based on the above results, it became clear that we needed to examine the differences in alignment. To compare the alignments to each other, and assess their usability for variant calling across the whole genome, we looked to the SVs known to be present in the NA12878 genome as curated by the DGV resource. We ran a pilot study on now deprecated datasets for linear SMRT data before undertaking a study with current technology (NIST, 2015). The results of these pilot experiments are available as S4 and point to a gap between the largest SV curated in DGV and what is resolvable by long-read sequencing.

To expand upon these initial, promising findings, we accessed the most up-to-date data generated on nanopore and SMRT platforms. Presently, this includes the WGS consortium rel7 on NA12878 for nanopore and a new release of SMRT-CCS on NA24385. NA24385 is not present in DGV, so a high quality callset released by PacBio in collaboration with the Genome in a Bottle Consortium was used as a truthset. These differences are reflected in their labels within Figs. 2 and 3.

We leveraged the sniffles SV caller because of its highly-detailed output files. Our truthsets were the curated SV present in DGV and the set released by PacBio and GIAB for NA24385. Due to the different nomenclature for SV type in the sniffles output and the NA24385 calls, which follow VCF specification (GA4GH, 2023) but differ from DGV

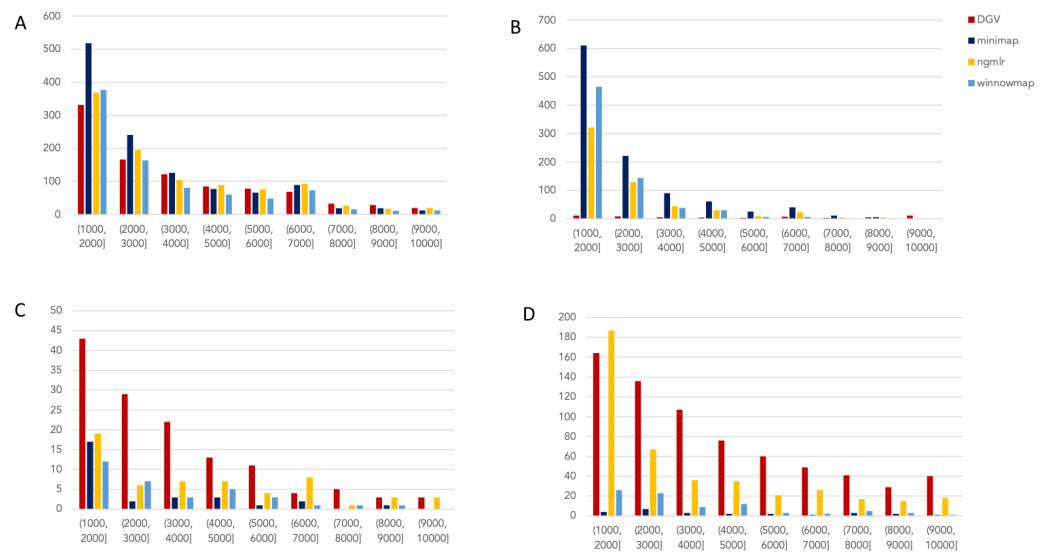


Figure 2 Nanopore SV 1,001–10,000 bp. Variants called by sniffles on each alignment file. (A) Deletions. (B) Insertions. (C) Inversions. (D) Duplications.

Full-size [DOI: 10.7717/peerj.16515/fig-2](https://doi.org/10.7717/peerj.16515/fig-2)

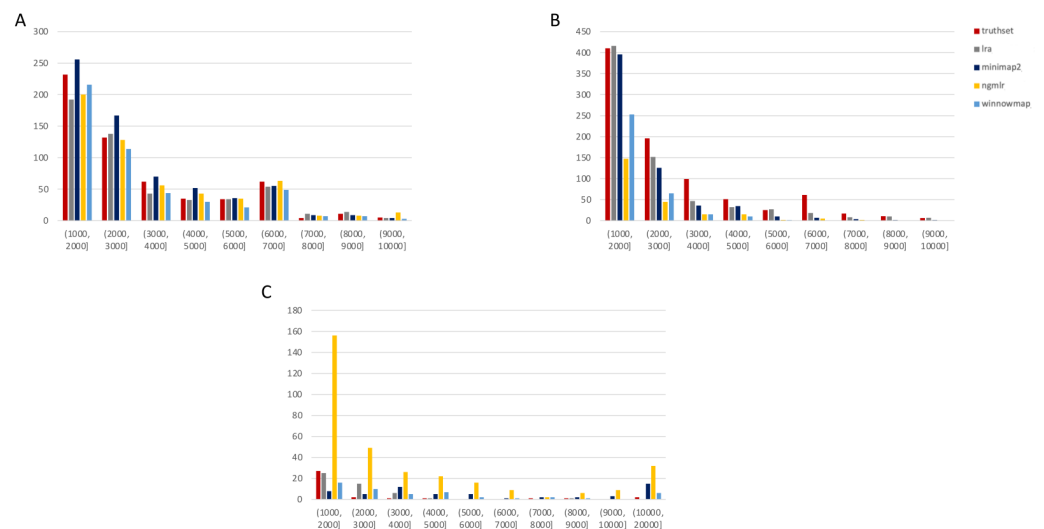


Figure 3 SMRT SVs 1001–10,000 bp. Variants called by sniffles on each alignment file. (A) Deletions. (B) Insertions. (C) Duplications.

Full-size [DOI: 10.7717/peerj.16515/fig-3](https://doi.org/10.7717/peerj.16515/fig-3)

annotations, comparisons were limited to the four classes of variant that were most unambiguously labeled in VCFs and the truth sets: insertions, deletions, duplications, and inversions.

The variants available for NA24385 were coded as all insertions or deletions, with subtypes within including duplications. Therefore, the callsets from SMRT do not include inversions as there was no baseline. Sniffles variants were graphed by SV length on the X-axis in shades of blue, grey, and yellow, contrasted with DGV variants in red, organized

by platform and SV type. Figures are organized by platform and size of SV, specifically between 1,001 and 100,000 bp, the range at which LRS platforms are known to excel.

Nanopore

For the variants between 1,001 and 10,000 bp (Fig. 2), alignment sets perform largely well on calling deletions that are curated in DGV while inversions were largely missed in all alignments. NGMLR alignments contained the highest number of breakpoints called as duplications. Above the 2001–3000 bp bin, the total duplications in DGV eclipses the number of breakpoints called in any of the alignment files. In contrast, there are very few insertions of this size range present in DGV, but hundreds in our callsets. This could be due to a differential interpretation of the breakpoints, where origin of inserted material has been ascertained (as a duplication) in DGV but remains unattributed (as an insertion) in call sets.

In the 10,001–100,000 bp size range (Supplement S5), DGV largely contains more variants, with the exception of inversions. Inversions present an interesting case here because DGV contains more very large inversions than are reported (>80,001), but various callsets do well at smaller ranges. Across deletions, insertions, and duplications, all alignment tools fall short of the curated reference. It is most interesting that DGV had few insertions smaller than 10,000 bp, but hundreds greater than that threshold. The trend of VCFs from our alignment files overperforming relative to DGV breaks down, and they highly underperform, calling only a few 10s of variants.

SMRT

The truthset for the SMRT dataset from NA24385 does not contain variants annotated as inversions, so Fig. 3 (SV 1001–10,000 bp) and S6 (SV 10,001–100,000 bp) only contain three panels each. We can see immediately in Fig. 3 that the bars generated from the reference set are much closer to the height of the bars from the VCFs, and this is not affected by the use of only the 15 kbp insert set. There is one notable exception with regard to duplications. As with the nanopore dataset, an NGMLR alignment contains the most duplications.

However, these duplications are far in excess of what is present in the reference set. This is notable, as the reference set has been carefully validated. What was missed, *versus* what is a false positive, is not resolvable in this experiment and likely not resolvable without wet lab bench work or orthogonal validation.

The largest SV show a complex picture (S6). NGMLR calls the most and the largest deletions and no tools do well for large insertions in absolute terms or relative to the truthset. However, Minimap2 and NGMLR preserve no breakpoints called as insertions. For duplications 10,001–20,000 bp, all tools except LRA resolve many more than are present in the reference. NGMLR continues to include many more duplications than the other alignments, and more than the reference.

The disparity between the number of large variants greater 10,001 bp in the NA24385 truthset *versus* the NA12878 is striking. This, along with the general lack of reference standards for all publicly-available genomes, highlights the need for better, more comprehensive reference sets for all publicly-available resources. Inability to resolve

large duplications 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 (*Barseghyan et al., 2017; Sharim et al., 2019*), 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 different tools. By using well-characterized genome standards, NA12878 and NA24385, 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 read alignments revealed breakpoints that could be resolved as structural variants in genome architecture. We then compared those variants to sets 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 of the potential high value of LRS data in terms of genome phasing and identification of epigenetic DNA modification (*Ni et al., 2019*). 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 (*Mardis, 2010*), 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, GraphMap2 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 was long and when it did work, Sniffles could not call variants on its product BAM. This is not unexpected, as the tool was designed in large part to increase single nucleotide variant sensitivity in noisy nanopore-sequenced reads.

Minimap2 used the least memory while Winnowmap2 ran successfully in the fastest time, an important point, should these platforms be tied closer to bedside applications. On data from both platforms, they both 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 and took much longer (3–10 times) than the next fastest tool. While it was designed specifically to resolve structural variation, it calls a high number of SV that have not been validated with other methodologies or 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 ([Chaisson et al., 2019](#)) as none of the Sniffles VCFs mirror the SVs present in the high-quality curated DGV database. Furthermore, there are many genomes released across multiple consortia including GIAB and the Human Pangenome Reference Consortium ([Miga & Wang, 2021](#)). But not all of these genomes are sequenced on multiple platforms, and comprehensive reference sets which are harmonious across samples, genomes, and datasets are still limited.

This resulted in us using two reference sets for this project to accommodate the latest releases of data. The callsets generated in our experiments used annotations which were different from the NA12878 variants in DGV and the curated variant set for NA24385, which was surprising. Finally, the most up to date variant benchmark for NA24385 is called on GRCh37-aligned genomes, which is from 2009 and not as comprehensive as current builds ([Nurk et al., 2022](#)). Taken together, this underscores the need for orthogonal approaches and collaboration between wet and dry labs to solve this problem.

The points above are critical in designing pipelines for genome analysis and structural variant discovery. In short, it is not a high burden to generate two alignment files per genome with each Minimap2 and Winnowmap2. In computationally unlimited research settings, there is value added in the generation of alignment files from NGMLR as well, although the time to complete these experiments makes them less appealing. These three 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 ([Bizjan et al., 2020](#)). If computational resources are limited, Minimap2 is the best choice to move the greatest number of genomes through the pipeline quickly with small memory needs; 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, diagnosis 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) ([Moncke-Buchner, 2002](#)), myotonic dystrophy 1 (~15–153 bp) and myotonic dystrophy 2 (~338–143,000 bp) ([Yum, Wang & Kalsotra, 2017](#)) could be good candidates for deep study with LRS platforms based on the variants present in alignment files from Minimap2, Winnowmap2 and NGMLR. Accordingly, LRS has been used to identify variants in many such disorders ([Mitsuhashi & Matsumoto, 2020](#)).

If pathogenic loci are known, a high diagnostic yield may be obtained by generating maps with each available alignment tool, and use of a structural variant caller such as NanoSV ([Cretu Stancu et al., 2017](#)). 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. 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 (*Zhou, Lin & Xing, 2019*).

The discrepancies between the VCFs generated from alignment files starkly show the need to design experiments with the appreciation that the genomics ecosystem cannot yet be 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 an LRS genome, we cannot say that it is not present. We must therefore look across platforms and data types for comprehensive genome representations (*Chaisson et al., 2019*). However, this requires more data from the community-generation of NA24385 data on new nanopore chemistry would allow for direct comparison to the current CCS standard, or alternatively new benchmark CCS data could be generated on NA12878. This would address a critical gap in our ability to benchmark new tools in a platform-agnostic manner.

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 at 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 the sequencing platform, when computational resources are not a limiting factor, it should be best practice to align an LRS human genome with three alignment tools. Minimapp2, Winnowmap2, and NGMLR will provide a strong foundation to gain better insight into the architecture of a genome of interest, but there are circumstances where use of LRA for SMRT data may make sense in lieu of NGMLR. When compute resources are limited, minimapp2 is a strong choice, and when time is a limiting factor, Winnowmap2 is the best choice.

As we enter the era of multiple references, pangenomics, and graph genomes, analytic substrate from the unmapped reads of a genome gains higher value. Through finding consensus reads unaligned by different aligners, teams interested in data excluded from the reference genome will yield more fertile ground for the discovery of novel genomic material.

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. 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.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Eric Vilain was supported by the A. James Clark Distinguished Professorship in Molecular Genetics at Children's National Hospital. Eric Vilain, Jonathan LoTempio, Emmanuele Delot are part of the GREGoR consortium, supported by NIH Award Number U01HG011745 from the National Human Genome Research Institute. The University of California, Irvine's Institute for Clinical and Translational Science, and the National Center for Advancing Translational Sciences (UL1TR001414) supported the publication fee for this article. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:

Molecular Genetics at Children's National Hospital.

The National Human Genome Research Institute: U01HG011745.

The University of California, Irvine's Institute for Clinical and Translational Science, and the National Center for Advancing Translational Sciences: UL1TR001414.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Jonathan LoTempio conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Emmanuele Delot conceived and designed the experiments, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Eric Vilain conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The code is available at GitHub and Zenodo:

- https://github.com/jlotempiojr/longread_alignment_benchmarking

- jlotempiojr. (2023). jlotempiojr/longread_alignment_benchmarking: Benchmarking long-read genome sequence alignment tools for human genomics applications (1.1).

Zenodo. <https://doi.org/10.5281/zenodo.8416228>

No new sequence data was generated in this study. The data used is available at GitHub:

- <https://github.com/nanopore-wgs-consortium/NA12878/blob/master/Genome.md>

- https://github.com/human-pangenomics/HG002_Data_Freeze_v1.0

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.16515#supplemental-information>.

REFERENCES

- Amarasinghe SL, Ritchie ME, Gouil Q. 2021. long-read-tools.org: an interactive catalogue of analysis methods for long-read sequencing data. *GigaScience* 10(2):giab003 DOI 10.1093/gigascience/giab003.
- Ardui S, Ameer A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research* 46:2159–2168 DOI 10.1093/nar/gky066.
- Barseghyan H, Tang W, Wang RT, Almalvez M, Segura E, Bramble MS, Lipson A, Douine ED, Lee H, Délot EC, Nelson SF, Vilain E. 2017. Next-generation mapping: a novel approach for detection of pathogenic structural variants with a potential utility in clinical diagnosis. *Genome Medicine* 9:90 DOI 10.1186/s13073-017-0479-0.
- Bizjan BJ, Katsila T, Tesovnik T, Šket R, Debeljak M, Matsoukas MT, Kovač J. 2020. Challenges in identifying large germline structural variants for clinical use by long read sequencing. *Computational and Structural Biotechnology Journal* 18:83–92 DOI 10.1016/j.csbj.2019.11.008.
- Bornhorst M, Eze Jr A, Bhattacharya S, Putnam E, Almira-Suarez MI, Rossi C, Kambhampati M, Almalvez M, Barseghyan M, Del Risco N, Dotson D, Turner J, Myseros JS, Vilain E, Packer RJ, Nazarian J, Rood B, Barseghyan H. 2023. Optical genome mapping identifies a novel pediatric embryonal tumor with a ZNF532::NUTM1 fusion. *Journal of Pathology* 260(3):329–338 DOI 10.1002/path.6085.
- Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA. 2008. The potential and challenges of nanopore sequencing. *Nature Biotechnology* 26:1146–1153 DOI 10.1038/nbt.1495.
- Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D, Anantharaman T, Audano PA, Brand H, Cantsilieris S, Cao H, Cerveira E, Chen C, Chen X, Chin C-S, Chong Z, Chuang NT, Lambert CC, Church DM, Clarke L, Farrell A, Flores J, Galeev T, Gorkin DU, Gujral M, Guryev V, Heaton WH, Korlach J, Kumar S, Kwon JY, Lam ET, Lee JE, Lee J, Lee W-P, Lee SP, Li S, Marks P, Viaud-Martinez K, Meiers S, Munson KM, Navarro FCP, Nelson BJ, Nodzak C, Noor A, Kyriazopoulou-Panagiotopoulou S, Pang AWC, Qiu Y, Rosanio G, Ryan M, Stütz A, Spierings DCJ, Ward A, Welch AE, Xiao M, Xu W, Zhang C, Zhu Q, Zheng-Bradley X, Lowy E, Yakneen S, McCarroll S, Jun G, Ding L, Koh CL, Ren B, Flicek P, Chen K, Gerstein MB, Kwok P-Y, Lansdorp PM, Marth GT, Sebat J, Shi X, Bashir A, Ye K, Devine SE, Talkowski ME, Mills RE, Marschall T, Korbelt JO, Eichler EE, Lee C. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications* 10:1784 DOI 10.1038/s41467-018-08148-z.

- Chaisson MJ, Tesler G. 2012.** Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**:238 DOI [10.1186/1471-2105-13-238](https://doi.org/10.1186/1471-2105-13-238).
- Chakraborty A, Morgenstern B, Bandyopadhyay S. 2021.** S-conLSH: alignment-free gapped mapping of noisy long reads. *BMC Bioinformatics* **22**:64 DOI [10.1186/s12859-020-03918-3](https://doi.org/10.1186/s12859-020-03918-3).
- Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021.** Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**:170–175 DOI [10.1038/s41592-020-01056-5](https://doi.org/10.1038/s41592-020-01056-5).
- Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, Pregno G, Giachino D, Mandrile G, Espejo Valle-Inclan J, Korzelius J, de Bruijn E, Cuppen E, Talkowski ME, Marschall T, de Ridder J, Kloosterman WP. 2017.** Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nature Communications* **8**:1326 DOI [10.1038/s41467-017-01343-4](https://doi.org/10.1038/s41467-017-01343-4).
- Cuscó A, Catozzi C, Viñes J, Sanchez A, Francino O. 2019.** Microbiota profiling with long amplicons using Nanopore sequencing: full-length 16S rRNA gene and the 16S-ITS-23S of the *rrn* operon. *F1000Research* **7**:1755 DOI [10.12688/f1000research.16817.2](https://doi.org/10.12688/f1000research.16817.2).
- Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021.** Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2):giab008 DOI [10.1093/gigascience/giab008](https://doi.org/10.1093/gigascience/giab008).
- Deshpande SV, Reed TM, Sullivan RF, Kerkhof LJ, Beigel KM, Wade MM. 2019.** Offline next generation metagenomics sequence analysis using MinION detection software (MINDS). *Gene* **10**(8):578 DOI [10.3390/genes10080578](https://doi.org/10.3390/genes10080578).
- Ebbert MTW, Jensen TD, Jansen-West K, Sens JP, Reddy JS, Ridge PG, Kauwe JSK, Belzil V, Pregent L, Carrasquillo MM, Keene D, Larson E, Crane P, Asmann YW, Ertekin-Taner N, Younkin SG, Ross OA, Rademakers R, Petrucelli L, Fryer JD. 2019.** Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology* **20**:97 DOI [10.1186/s13059-019-1707-2](https://doi.org/10.1186/s13059-019-1707-2).
- Faust GG, Hall IM. 2012.** YAHA: fast and flexible long-read alignment with optimal breakpoint detection. *Bioinformatics* **28**:2417–2424 DOI [10.1093/bioinformatics/bts456](https://doi.org/10.1093/bioinformatics/bts456).
- Formenti G, Chiara M, Poveda L, Francoijs K-J, Bonisoli-Alquati A, Canova L, Gianfranceschi L, Horner DS, Saino N. 2019.** 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* **8**(1):giy142 DOI [10.1093/gigascience/giy142](https://doi.org/10.1093/gigascience/giy142).
- Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV. 2009.** The challenges of sequencing by synthesis. *Nature Biotechnology* **27**:1013–1023 DOI [10.1038/nbt.1585](https://doi.org/10.1038/nbt.1585).
- GA4GH. 2023.** hts-specs: specifications of SAM/BAM and related high-throughput sequencing file formats. *GitHub*.
- Grigorev K. 2018.** Venn. Available at <https://pypi.org/project/venn/> (accessed on 11 July 2023).

- Haghshenas E, Sahinalp SC, Hach F. 2019. lordFAST: sensitive and fast alignment search tool for long noisy read sequencing data. *Bioinformatics* 35:20–27 DOI [10.1093/bioinformatics/bty544](https://doi.org/10.1093/bioinformatics/bty544).
- Ho SS, Urban AE, Mills RE. 2020. Structural variation in the sequencing era. *Nature Reviews. Genetics* 21:171–189 DOI [10.1038/s41576-019-0180-9](https://doi.org/10.1038/s41576-019-0180-9).
- Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, Peluso P, Boitano M, Chin C-S, Korlach J, Wilson RK, Eichler EE. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research* 27:677–685 DOI [10.1101/gr.214007.116](https://doi.org/10.1101/gr.214007.116).
- Hunter. 2007. Matplotlib: a 2D graphics environment. 9:90–95.
- Jain C, Koren S, Dilthey A, Phillippy AM, Aluru S. 2018a. A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics* 34:i748–i756 DOI [10.1093/bioinformatics/bty597](https://doi.org/10.1093/bioinformatics/bty597).
- Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. 2022. Long-read mapping to repetitive reference sequences using Winnommap2. *Nature Methods* DOI [10.1038/s41592-022-01457-8](https://doi.org/10.1038/s41592-022-01457-8).
- Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM. 2020. Weighted minimizer sampling improves long read mapping. *Bioinformatics* 36:i111–i118 DOI [10.1093/bioinformatics/btaa435](https://doi.org/10.1093/bioinformatics/btaa435).
- Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, Malla S, Marriott H, Nieto T, O’Grady J, Olsen HE, Pedersen BS, Rhie A, Richardson H, Quinlan AR, Snutch TP, Tee L, Paten B, Phillippy AM, Simpson JT, Loman NJ, Loose M. 2018b. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* 36:338–345 DOI [10.1038/nbt.4060](https://doi.org/10.1038/nbt.4060).
- Kantartzis A, Williams GM, Balakrishnan L, Roberts RL, Surtees JA, Bambara RA. 2012. Msh2-Msh3 interferes with Okazaki fragment processing to promote trinucleotide repeat expansions. *Cell Reports* 2:216–222 DOI [10.1016/j.celrep.2012.06.020](https://doi.org/10.1016/j.celrep.2012.06.020).
- Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. *Genome Research* 21:487–493 DOI [10.1101/gr.113985.110](https://doi.org/10.1101/gr.113985.110).
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology* 37:540–546 DOI [10.1038/s41587-019-0072-8](https://doi.org/10.1038/s41587-019-0072-8).
- Koren S, Rhie A, Walenz BP, Dilthey AT, Bickhart DM, Kingan SB, Hiendleder S, Williams JL, Smith TPL, Phillippy AM. 2018. De novo assembly of haplotype-resolved genomes with trio binning. *Nature Biotechnology* 36:1174–1182 DOI [10.1038/nbt.4277](https://doi.org/10.1038/nbt.4277).
- Kumar S, Agarwal S, Ranvijay . 2019. Fast and memory efficient approach for mapping NGS reads to a reference genome. *Journal of Bioinformatics and Computational Biology* 17:1950008 DOI [10.1142/S0219720019500082](https://doi.org/10.1142/S0219720019500082).
- Levy-Sakin M, Pastor S, Mostovoy Y, Li L, Leung AKY, McCaffrey J, Young E, Lam ET, Hastie AR, Wong KHY, Chung CYL, Ma W, Sibert J, Rajagopalan R, Jin N,

- Chow EYC, Chu C, Poon A, Lin C, Naguib A, Wang W-P, Cao H, Chan T-F, Yip KY, Xiao M, Kwok P-Y. 2019.** Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications* **10**:1025 DOI [10.1038/s41467-019-08992-7](https://doi.org/10.1038/s41467-019-08992-7).
- Li H. 2013.** Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv [q-bio.GN].
- Li H. 2018.** Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**:3094–3100 DOI [10.1093/bioinformatics/bty191](https://doi.org/10.1093/bioinformatics/bty191).
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009.** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16):2078–2079 DOI [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).
- Liu B, Gao Y, Wang Y. 2017.** LAMSA: fast split read alignment with long approximate matches. *Bioinformatics* **33**:192–201 DOI [10.1093/bioinformatics/btw594](https://doi.org/10.1093/bioinformatics/btw594).
- Liu B, Guan D, Teng M, Wang Y. 2016.** rHAT: fast alignment of noisy long reads with regional hashing. *Bioinformatics* **32**:1625–1631 DOI [10.1093/bioinformatics/btv662](https://doi.org/10.1093/bioinformatics/btv662).
- Liu Q, Zhang P, Wang D, Gu W, Wang K. 2017.** Interrogating the unsequenceable genomic trinucleotide repeat disorders by long-read sequencing. *Genome Medicine* **9**:65 DOI [10.1186/s13073-017-0456-7](https://doi.org/10.1186/s13073-017-0456-7).
- LoTempio J, Délot E, Vilain E. 2023.** Benchmarking long-read genome sequence alignment tools for human genomics applications. *bioRxiv* DOI [10.1101/2021.07.09.451840](https://doi.org/10.1101/2021.07.09.451840).
- MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. 2014.** The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research* **42**:D986–D992 DOI [10.1093/nar/gkt958](https://doi.org/10.1093/nar/gkt958).
- Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. 2019.** Structural variant calling: the long and the short of it. *Genome Biology* **20**:246 DOI [10.1186/s13059-019-1828-7](https://doi.org/10.1186/s13059-019-1828-7).
- Mantere T, Kersten S, Hoischen A. 2019.** Long-read sequencing emerging in medical genetics. *Frontiers in Genetics* **10**:426 DOI [10.3389/fgene.2019.00426](https://doi.org/10.3389/fgene.2019.00426).
- Marco-Sola S, Moure JC, Moreto M, Espinosa A. 2020.** Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics* **37**(4):456–463.
- Mardis ER. 2010.** The 1,000 genome, the 100, 000 analysis? *Genome Medicine* **2**:84 DOI [10.1186/gm205](https://doi.org/10.1186/gm205).
- Marić J, Sović I, Križanović K, Nagarajan N, Šikić M. 2019.** Graphmap2—splice-aware RNA-seq mapper for long reads. *BioRxiv* 720458 DOI [10.1101/720458](https://doi.org/10.1101/720458).
- Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, Schneider VA, Potapova T, Wood J, Chow W, Armstrong J, Fredrickson J, Pak E, Tigyi K, Kremitzki M, Markovic C, Maduro V, Dutra A, Bouffard GG, Chang AM, Hansen NF, Wilfert AB, Thibaud-Nissen F, Schmitt AD, Belton J-M, Selvaraj S, Dennis MY, Soto DC, Sahasrabudhe R, Kaya G, Quick J, Loman NJ, Holmes N, Loose M, Surti U, Risques RA, Graves Lindsay TA, Fulton R, Hall I, Paten B, Howe K, Timp W, Young A, Mullikin JC,**

- Pevzner PA, Gerton JL, Sullivan BA, Eichler EE, Phillippy AM. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585:79–84 DOI 10.1038/s41586-020-2547-7.
- Miga KH, Wang T. 2021. The need for a human pangenome reference sequence. *Annual Review of Genomics and Human Genetics* 22:81–102 DOI 10.1146/annurev-genom-120120-081921.
- Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, Oma Y, Kino Y, Mitsuhashi H, Matsumoto N. 2019. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biology* 20:58 DOI 10.1186/s13059-019-1667-6.
- Mitsuhashi S, Matsumoto N. 2020. Long-read sequencing for rare human genetic diseases. *Journal of Human Genetics* 65:11–19 DOI 10.1038/s10038-019-0671-8.
- Modi A, Vai S, Caramelli D, Lari M. 2021. The illumina sequencing protocol and the NovaSeq 6000 system. *Methods in Molecular Biology* 2242:15–42 DOI 10.1007/978-1-0716-1099-2_2.
- Moncke-Buchner E. 2002. Counting CAG repeats in the Huntington's disease gene by restriction endonuclease EcoP15I cleavage. *Nucleic Acids Research* 30:83e–83 DOI 10.1093/nar/gnf082.
- Nashta-Ali D, Aliyari A, Ahmadian Moghadam A, Edrisi MA, Motahari SA, Hossein Khalaj B. 2017. Meta-aligner: long-read alignment based on genome statistics. *BMC Bioinformatics* 18:126 DOI 10.1186/s12859-017-1518-y.
- Neveling K, Mantere T, Vermeulen S, Oorsprong M, van Beek R, Kater-Baats E, Pauper M, van der Zande G, Smeets D, Weghuis DO, Stevens-Kroef MJPL, Hoischen A. 2021. Next-generation cytogenetics: Comprehensive assessment of 52 hematological malignancy genomes by optical genome mapping. *American Journal of Human Genetics* 108(8):1423–1435 DOI 10.1016/j.ajhg.2021.06.001.
- Ni P, Huang N, Zhang Z, Wang D-P, Liang F, Miao Y, Xiao C-L, Luo F, Wang J. 2019. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics* 35:4586–4595 DOI 10.1093/bioinformatics/btz276.
- NIST. 2015. Index of /giab/ftp/data/NA12878/NA12878_PacBio_MtSinai. Available at https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai/ (accessed on 11 July 2023).
- Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. 2016. Nanopore sequencing detects structural variants in cancer. *Cancer Biology & Therapy* 17:246–253 DOI 10.1080/15384047.2016.1139236.
- Nurk S, Koren S, Rhie A, Rautiainen M, Bizikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, Aganezov S, Hoyt SJ, Diekhans M, Logsdon GA, Alonge M, Antonarakis SE, Borchers M, Bouffard GG, Brooks SY, Caldas GV, Chen N-C, Cheng H, Chin C-S, Chow W, de Lima LG, Dishuck PC, Durbin R, Dvorkina T, Fiddes IT, Formenti G, Fulton RS, Functammasan A, Garrison E, Grady PGS, Graves-Lindsay TA, Hall IM, Hansen NF, Hartley GA, Haukness M, Howe K, Hunkapiller MW, Jain C, Jain M, Jarvis ED, Kerpedjiev P, Kirsche

- M, Kolmogorov M, Korf J, Kremitzki M, Li H, Maduro VV, Marschall T, McCartney AM, McDaniel J, Miller DE, Mullikin JC, Myers EW, Olson ND, Paten B, Peluso P, Pevzner PA, Porubsky D, Potapova T, Rogaev EI, Rosenfeld JA, Salzberg SL, Schneider VA, Sedlazeck FJ, Shafin K, Shew CJ, Shumate A, Sims Y, Smit AFA, Soto DC, Sović I, Storer JM, Streets A, Sullivan BA, Thibaud-Nissen F, Torrance J, Wagner J, Walenz BP, Wenger A, Wood JMD, Xiao C, Yan SM, Young AC, Zarate S, Surti U, McCoy RC, Dennis MY, Alexandrov IA, Gerton JL, O'Neill RJ, Timp W, Zook JM, Schatz MC, Eichler EE, Miga KH, Phillippy AM. 2022. The complete sequence of a human genome. *Science* 376:44–53 DOI 10.1126/science.abj6987.
- PacBio NIST. 2021. sv-benchmark: public benchmark of long-read structural variant caller on PacBio CCS HG002 data. *Github*.
- Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their purpose and place. *Human Molecular Genetics* 27:R234–R241 DOI 10.1093/hmg/ddy177.
- Ren J, Chaisson MJP. 2021. Ira: a long read aligner for sequences and contigs. *PLOS Computational Biology* 17:e1009078 DOI 10.1371/journal.pcbi.1009078.
- Sanderson ND, Street TL, Foster D, Swann J, Atkins BL, Brent AJ, McNally MA, Oakley S, Taylor A, Peto TEA, Crook DW, Eyre DW. 2018. Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices. *BMC Genomics* 19:714 DOI 10.1186/s12864-018-5094-y.
- Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74:5463–5467.
- Santoro M, Masciullo M, Silvestri G, Novelli G, Botta A. 2017. Myotonic dystrophy type 1: role of CCG, CTC and CGG interruptions within DMPK alleles in the pathogenesis and molecular diagnosis. *Clinical Genetics* 92:355–364 DOI 10.1111/cge.12954.
- Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Haeseler Avon, Schatz MC. 2018. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* 15:461–468 DOI 10.1038/s41592-018-0001-7.
- Sharim H, Grunwald A, Gabrieli T, Michaeli Y, Margalit S, Torchinsky D, Arielly R, Nifker G, Juhasz M, Gularek F, Almalvez M, Dufault B, Chandra SS, Liu A, Bhattacharya S, Chen Y-W, Vilain E, Wagner KR, Pevsner J, Reifengerger J, Lam ET, Hastie AR, Cao H, Barseghyan H, Weinhold E, Ebenstein Y. 2019. Long-read single-molecule maps of the functional methylome. *Genome Research* 29:646–656 DOI 10.1101/gr.240739.118.
- Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. 2016. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature Communications* 7:11307 DOI 10.1038/ncomms11307.
- Talsania K, Shen TW, Chen X, Jaeger E, Li Z, Chen Z, Chen W, Tran B, Kusko R, Wang L, Pang AWC, Yang Z, Choudhari S, Colgan M, Fang LT, Carroll A, Shetty J, Kriga Y, German O, Smirnova T, Liu T, Li J, Kellman B, Hong K, Hastie AR, Natarajan A, Moshrefi A, Granat A, Truong T, Bombardi R, Mankinen V, Meerzaman D, Mason CE, Collins J, Stahlberg E, Xiao C, Wang C, Xiao W, Zhao Y. 2022. Structural

- variant analysis of a cancer reference cell line sample using multiple sequencing technologies. *Genome Biology* **23**(1):255 DOI [10.1186/s13059-022-02816-6](https://doi.org/10.1186/s13059-022-02816-6).
- Walker A, Houwaart T, Wienemann T, Vasconcelos MK, Strelow D, Senff T, Hülse L, Adams O, Andree M, Hauka S, Feldt T, Jensen B-E, Keitel V, Kindgen-Milles D, Timm J, Pfeffer K, Dilthey AT. 2020.** Genetic structure of SARS-CoV-2 reflects clonal superspreading and multiple independent introduction events, North-Rhine Westphalia, Germany, February and 2020. *Euro Surveillanc: Bulletin European sur Les Maladies Transmissibles = European Communicable Disease Bulletin* **25**(22):2000746 DOI [10.2807/1560-7917.ES.2020.25.22.2000746](https://doi.org/10.2807/1560-7917.ES.2020.25.22.2000746).
- Waskom M. 2021.** seaborn: statistical data visualization. *Journal of Open Source Software* **6**:3021 DOI [10.21105/joss.03021](https://doi.org/10.21105/joss.03021).
- Wei Z-G, Zhang S-W, Liu F. 2020.** smsMap: mapping single molecule sequencing reads by locating the alignment starting positions. *BMC Bioinformatics* **21**:341 DOI [10.1186/s12859-020-03698-w](https://doi.org/10.1186/s12859-020-03698-w).
- Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, De Pristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW. 2019.** Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**:1155–1162 DOI [10.1038/s41587-019-0217-9](https://doi.org/10.1038/s41587-019-0217-9).
- Wick RR, Judd LM, Holt KE. 2019.** Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* **20**:129 DOI [10.1186/s13059-019-1727-y](https://doi.org/10.1186/s13059-019-1727-y).
- Xiao C-L, Chen Y, Xie S-Q, Chen K-N, Wang Y, Han Y, Luo F, Xie Z. 2017.** MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature Methods* **14**:1072–1074 DOI [10.1038/nmeth.4432](https://doi.org/10.1038/nmeth.4432).
- Yum K, Wang ET, Kalsotra A. 2017.** Myotonic dystrophy: disease repeat range, penetrance, age of onset, and relationship between repeat size and phenotypes. *Current Opinion in Genetics & Development* **44**:30–37 DOI [10.1016/j.gde.2017.01.007](https://doi.org/10.1016/j.gde.2017.01.007).
- Zhou A, Lin T, Xing J. 2019.** Evaluating nanopore sequencing data processing pipelines for structural variation identification. *Genome Biology* **20**(1):237 DOI [10.1186/s13059-019-1858-1](https://doi.org/10.1186/s13059-019-1858-1).
- Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, De La Vega FM, Xiao C, Sherry S, Salit M. 2019.** An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology* **37**:561–566 DOI [10.1038/s41587-019-0074-6](https://doi.org/10.1038/s41587-019-0074-6).