DEVELOPING A NANOPORE SEQUENCING DATA PROCESSING PIPELINE FOR STRUCTURAL VARIATION IDENTIFICATION

By

TIMOTHY LIN

A thesis submitted to the

School of Graduate Studies

Rutgers, the State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Microbiology and Molecular Genetics

Written under the direction

Jinchuan Xing

And approved by

_____

_____

_____

New Brunswick, New Jersey

October, 2019

ABSTRACT OF THE THESIS

Developing a nanopore sequencing data processing pipeline for structural variation identification

By TIMOTHY KWANG LIN

Thesis Director:
Jinchuan Xing

Many genomic sequencing technologies have been developed since the Human Genome Project. These next-generation sequencing (NGS) technologies from various companies reshaped the genomics field and have improved rapidly. However, NGS has limitations for certain applications due to its short read length. The third generation of sequencing technology uses single molecule real-time sequencer that can generate long reads. Recently Oxford Nanopore entered the market with the release of its MinION sequencer. Oxford Nanopore's unique third generation sequencing technology allows for much longer read length than NGS technologies, potentially addressing some of the limitations of NGS. Due to the novelty of nanopore sequencing technology, the available tools for aligning long read data and detecting structural variants have not been thoroughly evaluated. Here we evaluate the performance of several alignment and structural variation detection tools on long read MinION data.

Acknowledgements

I would like to thank my advisor, Dr. Jinchuan Xing for his guidance and support as I learned more about human genetics and bioinformatics. He has always been a wealth of information and I have learned a great deal from him through our conversations.

I would like to thank Anbo Zhou for his advice and support, especially as I began learning to code, and for his work related to this project.

I would like to thank Dr. Premal Shah and Dr. Chris Ellison for their advice and support during our laboratory experiments.

Table of Contents

List of Tables

List of Figures

1. Introduction

Many genomic sequencing technologies have been developed since the Human Genome Project. Limitations on the Sanger sequencing technology lead to development of a more cost effective, high-throughput second generation of sequencing technology, termed next-generation sequencing (NGS). NGS technology reshaped the genomics field and has improved rapidly, dramatically driving the cost of whole genome sequencing down to the threshold of one thousand dollars [1]. However, NGS has limitations for structural variant (SV) calling due to its short read length. The arrival of a third generation of sequencing technology occurred as Pacific Biosciences launched its sequencer in 2011, the first available single molecule real-time sequencer that generated long reads [2]. Recently Oxford Nanopore entered the market with the release of its MinION sequencer. Nanopore sequencing technology is capable of obtaining much longer read lengths, potentially addressing some of the limitations of NGS [3]. Here we will first provide an overview of sequencing techniques from each generation.

Sanger sequencing, also called chain termination method, involves using dideoxyribonucleotides (ddNTPs) that lack the 3-prime hydroxyl group required by DNA polymerase to elongate the DNA strand. Strand elongation stops when a ddNTP is incorporated, leading to a mixture of fragment sizes. In classical Sanger sequencing, ddNTPs at low concentration are included with normal dNTPs in a reaction and each of the four ddNTPs needs a separate reaction. The four reactions are then placed in four

separate lanes and the fragments are size separated by gel electrophoresis. Smaller fragments will migrate further towards the bottom of the gel, and the sequence can determined by reading the order of fragments from bottom to top based on the lane that the band appears in. The method was improved with the introduction of fluorescent ddNTPs, allowing all four ddNTPs to be incorporated into one reaction and analyzed through capillary electrophoresis, where data is output as chromatogram trace peaks. These advances helped to automate sample preparation and sequencing for higher throughput applications [4]. The first generation of automated sequencers, such as the ABI PRISM, could sequence up to 384 reactions in parallel and accelerated the pace of the Human Genome Project.

Sanger sequencing is widely used in low throughput application such as sequencing a single gene or a small number of amplicons. The low error rate of Sanger sequencing is still often considered the de facto "gold standard" to which other sequencing platforms are compared, and Sanger sequencing is sometimes used to confirm results from NGS. The average read length for Sanger sequencing is between 700-1000 base pairs (bp), which is longer than most NGS platforms. However, the high cost per base and relatively low throughput made development of more cost efficient, massively parallel sequencing platforms necessary for sequencing large and complex genomes.

NGS platforms were introduced with the release of the 454 pyrosequencing platform in 2005 [5] and the field has continuously improved and evolved. Multiple NGS platforms utilizing different strategies were released and these technologies were capable of carrying out millions of sequencing reactions in parallel and generating vast

amounts of data in a single run in comparison to Sanger sequencing [6]. The most successful line of NGS platforms in terms of market share and widespread adoption is from Illumina. Obtained from Illumina's acquisition of Solexa, the technology underlying the platforms is termed "sequencing-by-synthesis" and involves the use of fluorescent reversible terminator deoxyribonucleotide triphosphates (dNTPs) and clonal amplification on the surface of a flow cell where sequencing reactions occur. As a review, the following are the general library preparation and sequencing steps using an Illumina platform.

First, a genomic library is sheared into smaller fragments through an enzymatic or physical method, such as sonication. Adapter sequences are ligated onto both ends of the fragmented DNA. These adapters include the sequencing primer binding site and a barcode or index that is used to identify the sample that a read originates from, used when samples are multiplexed and run together on the sequencer. The ends of the adapters also contain sequences that are complementary to oligonucleotide primers on the surface of the flow cell. After adapters have been ligated, the DNA library is washed over the surface of the flow cell, where the adapters allow the DNA library to anneal to complementary oligonucleotides on the surface and attach to the flow cell. The attached DNA fragments are clonally amplified through a process called bridge PCR, where clusters of monoclonal DNA fragments are generated. This cluster generation step is necessary to intensify the fluorescent signal of the sequencing reaction to a level that is detectable by the sequencer [7].

During bridge PCR, an attached DNA fragment is used as the template and the surface-bound oligonucleotide is elongated to generate the complement strand to the fragment. The original DNA fragment is denatured and washed away, and the newly extended complement strand arches over and anneals to a surface-bound oligonucleotide through the adapter sequence on the other end of the molecule. This molecule is now used as the template as the second oligonucleotide is extended, forming a double-stranded "bridge". The DNA fragments are then denatured and this amplification process repeats until a cluster of identical single stranded DNA fragments identical to the original DNA fragment are generated. Hundreds of millions of these single stranded monoclonal clusters are generated on the surface of the flow cell [8].

After clusters are generated, sequencing primers are introduced along with DNA polymerase and fluorescently labelled, reversible terminator dNTPs. The 3'-OH ends of the dNTPs are blocked by the fluorescent moiety, preventing elongation when incorporated into a DNA strand. In one cycle, each DNA cluster is extended by one base and a fluorescence reading is taken, after which the fluorescent moiety blocking the 3'-OH is enzymatically cleaved and ready for further extension in the next cycle. Millions of clusters undergoing sequencing reactions at the same time allow the method to be massively parallel, generating large amounts of data at high coverage.

Illumina sequencing technology attains a high accuracy rate above 99.5% and has the lowest cost per base of the major NGS platforms [9]. A wide variety of Illumina sequencers are on the market with various output options from the MiniSeq to the NovaSeq and the technology is capable of various applications like DNA methylation

sequencing, RNA sequencing, and ChIP-seq. The community using Illumina platforms is large and entrenched, contributing to its maturity as a technology and to the development of a large number of computational tools optimized around data produced from Illumina platforms. However, the startup costs of procuring Illumina sequencers may be expensive and prohibitive for some labs.

The main limitation to the technology is the relatively short read length. Illumina sequencers rely on of clusters of identical DNA fragments to obtain a detectable signal, which deteriorates with each cycle in the sequencing run. An increasing number of strands within a cluster become out of phase each cycle because of random incorporation errors. The maximum read length for Illumina platforms is 600 bp as a 300x300 paired end read, lower than the read lengths obtained from Sanger sequencing as well as some other NGS platforms. While high coverage and accurate reads are easily attainable through Illumina sequencing, the short read lengths are not optimal for the sequencing of long repetitive regions and detection of structural variation.

Third generation single molecule sequencers are under active development and are capable of directly sequencing single DNA molecules in long reads [10]. These long reads would be advantageous towards the study of repetitive regions and structural variations that are relevant to evolution or the development of diseases. One sequencing technology that has generated a great deal of interest is nanopore sequencing under development by Oxford Nanopore Technologies. The company released the MinION in 2015, the first nanopore-based sequencer to the market [11]. The device is small enough to be hand held and is easily transportable into the field.

The startup costs of procuring the MinION are also small in comparison to other platforms.

Unlike other sequencing chemistries, nanopore sequencing does not require the detection of fluorescence or a product of chemical or enzymatic reactions. Instead, single-stranded DNA molecules are sequenced by measuring current disruptions as molecules pass through a nanopore. These protein nanopores are embedded within a non-conductive polymer and allow ions to flow through, creating a baseline current. When a single strand of DNA is ratcheted through the nanopore by a protein motor, the current is disrupted into patterns that are characteristic to the nucleotides passing through the nanopore. This measurement of current disruption is often called "squiggle space" and the MinION software interprets the current disruption into 5 or 6-mer sequences [11].

The MinION is capable of '1D' and '2D' reads. For 1D reads, a single stranded molecule is simply read from beginning to end. For 2D reads, a hairpin adapter is used to tether together both strands of a DNA molecule. After the first strand is sequenced and transported through the nanopore, the complement strand immediately follows, allowing a consensus sequence to be called based on both strands and increasing the accuracy of the read. The reads obtained from the MinION can exceed an average of 10 kb, and some labs report routinely achieving read lengths of more than 50 kb [12]. One of the main weaknesses of nanopore sequencing is the low accuracy rate. While the accuracy rate has improved to 92%, the accuracy rate is much lower than NGS and such a high error rate makes the data problematic for applications such as single nucleotide

variation (SNV) detection. However the sequencing chemistry is rapidly improving and the development of computational tools designed specifically for long-read data will likely improve the accuracy rate in the near future.

While the high error rate of nanopore sequencing may be problematic for SNV detection, the error-prone long reads may be advantageous still be advantageous for detecting structural variations that can be problematic for NGS. Structural variants represent multiple types of genomic alterations larger than 50 bp and make up most of the heterogeneity between human genomes in terms of total number of bases [13]. SVs are associated with a variety of diseases, including severe neurological diseases and cancer. Studying SVs is therefore critical to understanding the underlying genetic conditions for these diseases. Long reads from nanopore data can span entire structural variations, enabling them to be detected in a single long read instead of inferred through strategies developed for short reads such as discordant read pairs. New computational tools have been developed specifically for long read data. However, since ONT's MinION was only recently launched, the available tools for aligning long read data and detecting structural variants (SVs) have not been thoroughly evaluated.

In this study, we evaluated several aligners, including BWA-MEM [14], GraphMap [15], LAST [16], ngmlr [17], and minimap2 [18], by aligning our MinION-generated data. GraphMap, ngmlr, and minimap2 were developed specifically for long read sequencing and have reported promising results. BWA-MEM and LAST are established mappers used for short-read data but have been tweaked for long-read data.

After evaluating the performance of the mappers, several tools for detecting structural variation were compared. We have chosen to evaluate Picky [19], Sniffles [17], and NanoSV [20]. These are all recently developed SV callers that were developed for long error-prone reads. Results from these structural variant callers were compared to publicly available high confidence structural variant calls to evaluate their performance.

2. Methods/Experimental Procedures

2.1. Library preparation and nanopore sequencing

DNA sample of individual NS12911 was purchased from Coriell (Camden, NJ, USA). Library preparation began with an input of 1.5 ug genomic DNA in 50 ul of water and followed the Oxford Nanopore (ONT) protocol for 1D Genomic DNA by ligation. The genomic DNA was sheared using a Covaris g-Tube (520079, Covaris, Woburn, MA, USA) spun at 6000 RPM for one minute to obtain 10 kb fragments. After recovering fragmented DNA from the g-Tube, fragment size was assessed using Agilent TapeStation with genomic DNA ScreenTape (5067-5365, Agilent, Santa Clara, CA, USA).

End repair and A-tailing were performed with NEBNext Ultra II End Repair/dA-Tailing module (E7546, New England Biosciences, Ipswich, MA, USA) and was followed by an AMPure XP bead (#A63880, Beckman Coulter, Indianapolis, IA, USA) cleanup following the manufacture protocol. Next, adapters were ligated using NEB Blunt/TA

Ligase master mix (#M0367, New England Biolabs, Woburn, MA, USA) with adapters from the 1D Genomic DNA by Ligation sequencing kit (SQK-LSK108, Oxford Nanopore Technologies, Oxford, UK). After adapter ligation the library was cleaned up by adding AMPure XP beads at a 0.4X ratio and aspirating the supernatant. The beads were washed with adapter bead binding buffer twice to remove free adapters and then eluted into 15 ul elution buffer from the 1D Genomic DNA by Ligation kit. One microliter was taken and assessed with Agilent genomic ScreenTape to assess library size.

The adapter-ligated library was mixed with library loading beads and running buffer provided in the ligation sequencing kit (SQK-LSK108, Oxford Nanopore Technologies, Oxford, UK) to prepare for flow cell loading. The library was then loaded into a flow cell (FLO-MIN106D, Oxford Nanopore Technologies, Oxford, UK) for sequencing. On the user interface, the kit SQK-LSK108, and flow cell FLO-MIN106 were selected and real-time basecalling was not used.

2.2. Data analysis methods

Basecalling was performed with ONT's Albacore 1.2.6 and Albacore 2.0.2 and the data quality and metrics were assessed with the nanopore-specific PoreTools and MinionQC. The data was aligned using the recently developed long-read mappers ngmlr, Graphmap, and minimap2, as well as the older mappers BWA-MEM and LAST, which have been optimized for long reads. The commands used for each run are listed

in Appendix C. Mappers were tested and benchmarked on data from our two Nanopore MinION sequencing runs.

Name	Type	Version	Description	Citation
Poretools	QC utility	0.6.0	QC toolkit for Oxford Nanopore sequencing data.	[21]
minionQC	QC Utility	1.3.5	QC toolkit for Oxford Nanopore sequencing data.	[22]
BWA-MEM	Aligner	0.7.15	Popular short read aligner	[14]
GraphMap	Aligner	0.5.2	Developed for long read sequencing.	[15]
LAST	Aligner	941	An older tool capable of handling long reads, similar to BLAST.	[23]
minimap2	Aligner	2.1	Developed for short and long reads, same lab as BWA-MEM	[18]
ngmlr	Aligner	0.2.6	An aligner working with Nanopore long reads to generate high quality SV calling	[17]
NanoSV	SV caller	1.2.0	Developed for long read sequencing. Identifies split and gap aligned reads, clusters based on orientation and genomic position to find breakpoint junctions	[20]
Sniffles	SV caller	1.0.11	Developed for long read sequencing. Identifies SVs using split-read alignments, high-mismatch regions, and coverage analysis	[17]
Picky	SV Caller	1.0	Developed for long read sequencing. Uses a greedy seed-and-extension algorithm to merge long read segments and detect breakpoints	[19]

Table 1: Tools used for analysis

The Linux command */usr/bin/time -v* was used to generate alignment time and maximum memory usage benchmarks, and Samtools' [24] flagstat option was used to generate the percentage of reads mapped. Samtools' depth was used to calculate average genomic coverage of the reads.

The structural variant callers Picky, Sniffles, and NanoSV were evaluated in combination with several of the best performing mappers using publicly available high coverage data from ONT (https://github.com/nanopore-wgs-consortium/NA12878). In order

to evaluate the SV calls generated from these mapper and caller combinations, a high coverage SV call set from Mt. Sinai was used as a benchmark (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai) [25]. This dataset was generated from 44X coverage PacBio reads from human reference NA12878 and the SV calls were evaluated using seven different methods. These methods incorporated different combinations of raw and error corrected reads, SV callers, BLASR versions, and assembly (Appendix A).

These SV calls were merged using bedtools [26] default merge function, which merges any calls with a single base overlap. Next, the calls were lifted over using liftOver [16] human reference from hg19 to hg38 because that the public data from ONT was aligned to hg38. SV calls from unincorporated contigs and the mitochondrial genome were filtered out to generate the final SV call set. The same merging and filtering steps were done with the passing calls in the dataset.

The filtered SV all set was used as a benchmark to compare the results of SV caller and long-read mapper combinations we tested. Minimap2 and ngmlr were used in conjunction with the SV callers Sniffles, NanoSV, and Picky. Picky was also run in combination with LAST because its authors developed it to run with LAST as its default mapper and incorporates LAST in the default command.

The deletion and insertion calls generated from these caller and mapper combinations were compared to the Mt. Sinai dataset using *bedtools intersect* and *bedtools window*. For deletions, b*edtools intersect* was set to report any calls with a one base overlap, which is its default setting. For insertions, *bedtools window* was used to

report any calls within a window of 100 bases from the reported insertion site in the genome. Both *bedtools intersect* and *bedtools window* were run twice for each caller and mapper combination, once to see which SV calls overlap the Mt. Sinai benchmark calls, and once to see which Mt. Sinai benchmark calls overlap the calls detected by the SV caller.

3. **Results**

3.1. **Library preparation**

The libraries prepared for the sequencing run fragmented uniformly according to our TapeStation results. The fragmentation gel image and electropherogram (Fig. 1a) from

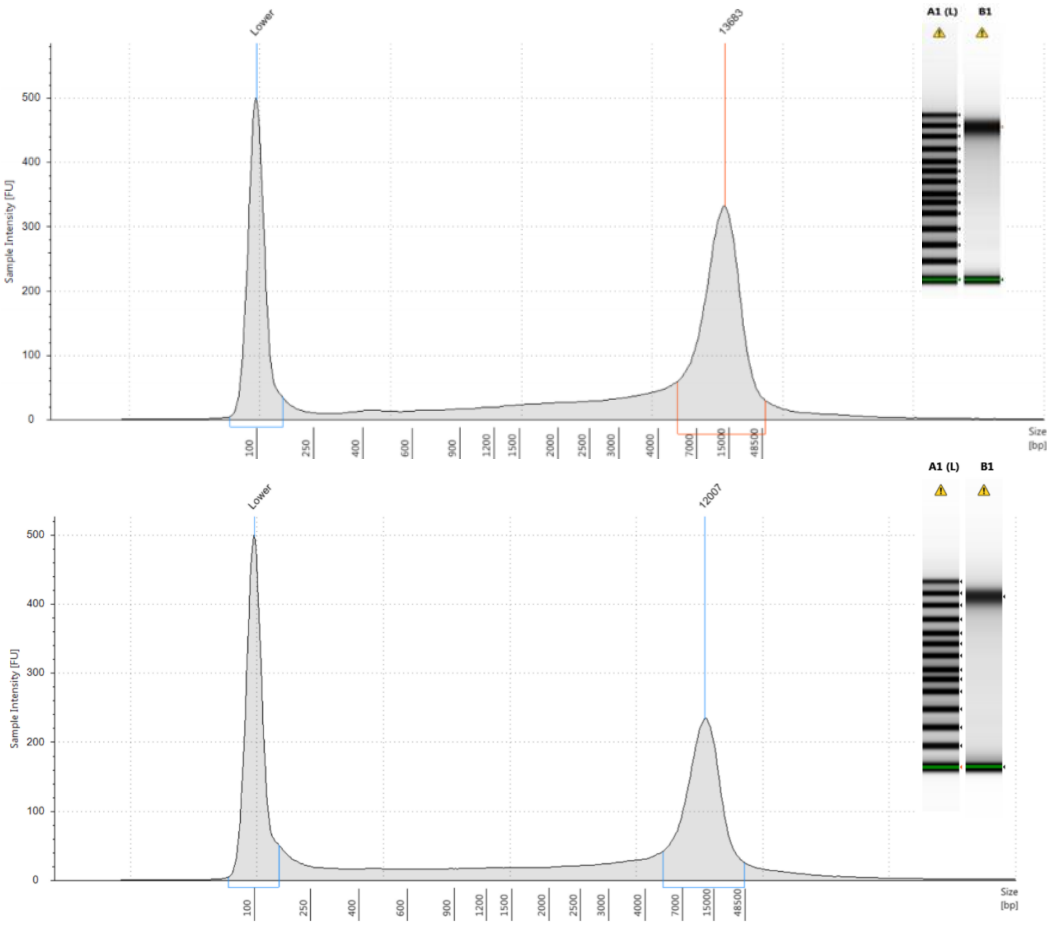


Figure 1: Library Preparation gel results after fragmentation (top) and adapter ligation (bottom)

the library preparation shows a large band around the 15,000 bp ladder band and a median fragment length of 13,483 bp. Similarly, the gel images from the post-adapter ligation results also show a single band around the 15,000 bp leader band and a median fragment length of 12,007 bp (Fig. 1b).

A second library was prepared using the identical protocol with similar TapeStation results (data not shown).

3.2. Sequencing and Base calling

The first sequencing run crashed after about 14 hours due to unconfirmed reasons, but we suspect it was a data transfer issue between the sequencer and computer due to the large number of temporary files being generated by the sequencer.

Basecalling was first performed with Oxford Nanopore's default basecaller Albacore

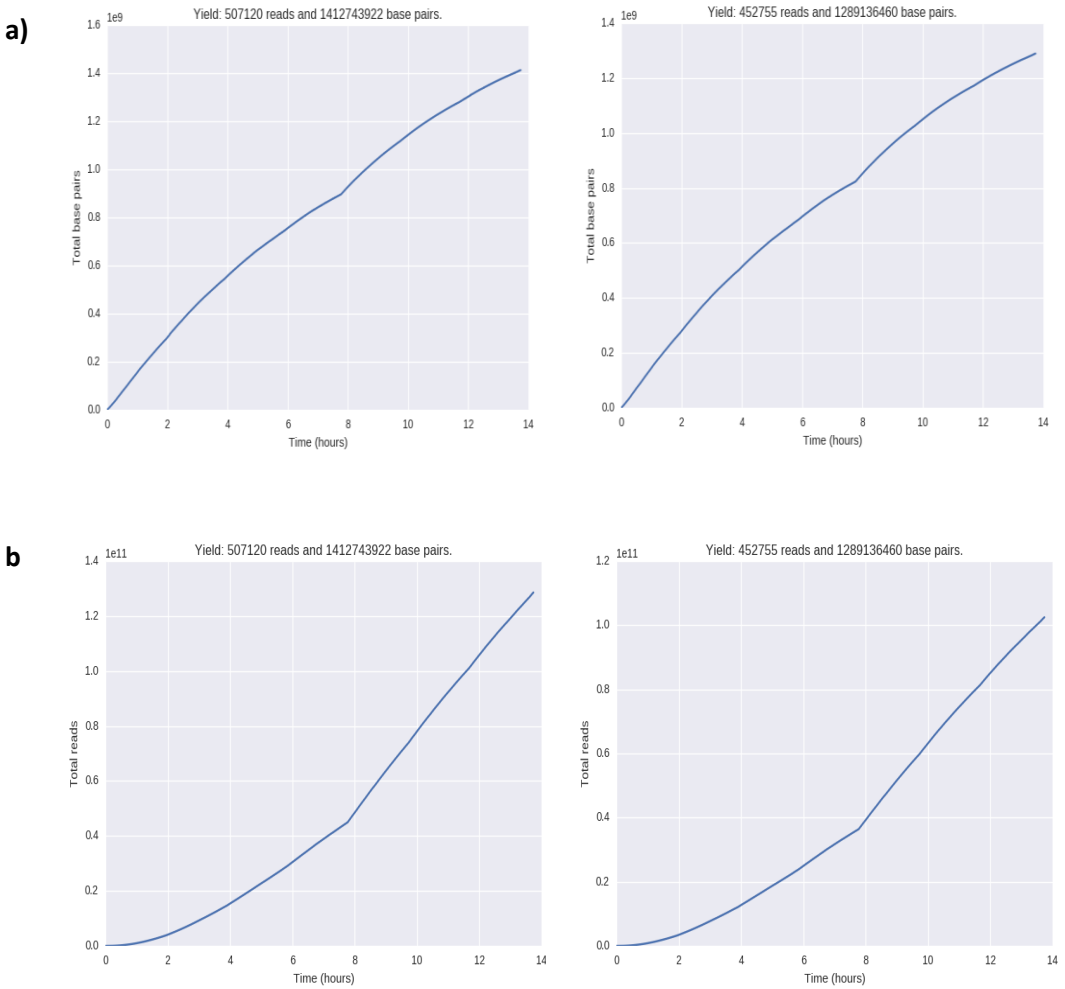


Figure 2: Poretools yield plot for (a) base pairs over time and (b) reads over time for Albacore 1.2.6 (left) and Albacore 2.0.2 (right)

ver. 1.2.6. While conducting our experiment, the updated Albacore ver. 2.0.2 was released and was subsequently used for basecalling. In comparison, 504,407 reads and 1,412,743,922 bp were produced from the Albacore ver. 1.2.6 while 452,755 reads and

1,289,136,460 bp from Albacore ver. 2.0.2 (Fig.2). The differences between the basecallers are small and likely due to Albacore 2.0.2 automatically filtering out reads with a Q score less than 7, while Albacore 1.2.6 did not have a quality filter by default. Since the differences were small and Albacore 2.0.2 was the updated version, we continued downstream analysis with results from this basecaller.

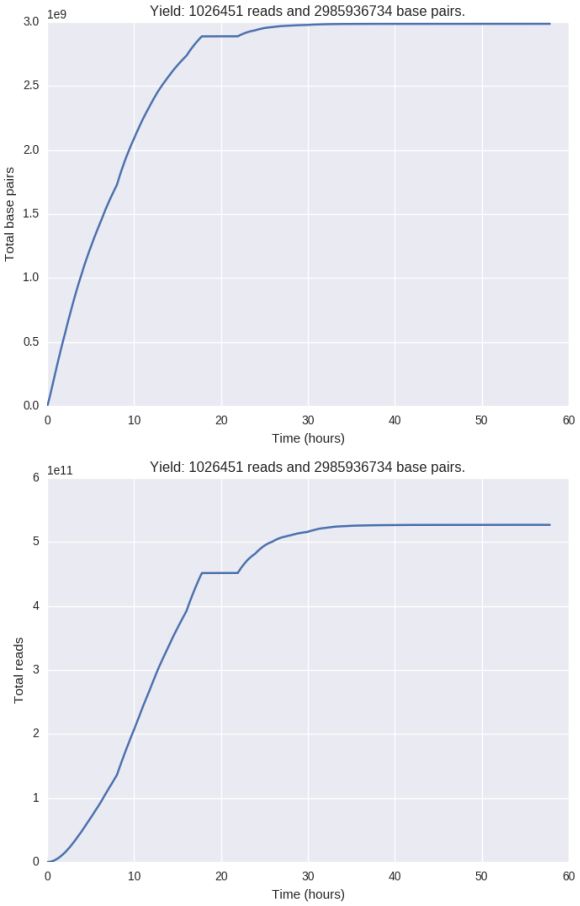


Figure 3: Yield plot for total base pairs over time (top) and total reads over time (bottom) generated by poretools

Because the first run failed at 14 hours, we constructed a new library following the same protocol as the previous library and repeated the sequencing run using a new flow cell. The second run also encountered a crash, but we were able to resume and finish a full 48 hour sequencing run.

Data from the second sequencing run was base called by Albacore 2.0.2. The sequencing run generated 1,026,451 reads and 2,985,936,734 bps (Fig. 3).

The poretools yield plots (Fig. 4) show that after about 18 hours the yield gain decreased dramatically. We encountered a crash at about the 18-hour mark but were able to continue the run. However, the sequencer produced only a small amount of sequence (<15%) after restarting the run. This reduction in throughput could be due to either the exhaustion of DNA templates, or the crash affected the sequencing performance.

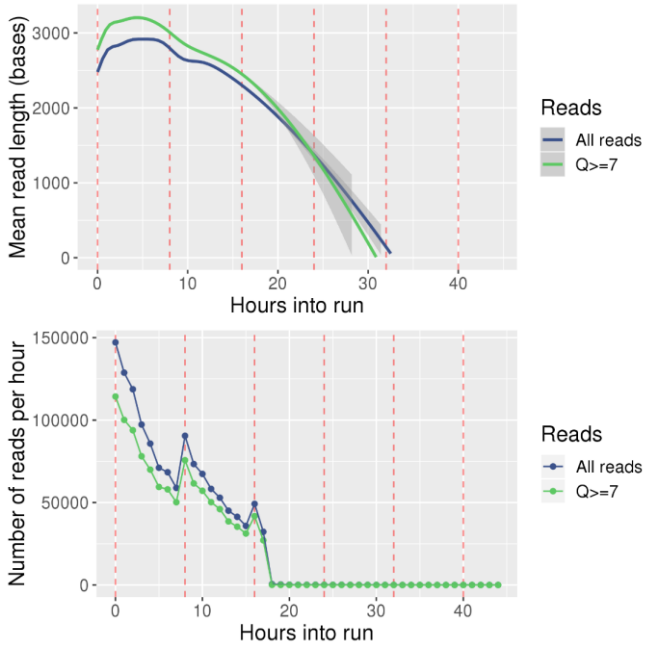


Figure 4: Mean read length per hour (top) and number of reads per hour (bottom) plots generated from MinionQC

Most of the longer fragments were sequenced towards the beginning of the sequencing run and the mean fragment size decreased over the length of the run. The number of reads per hour also decreased steadily but spiked immediately after each mux switch (Fig. 4). At about 18 hours the number of reads per hour drops to near zero due to the crash. When the run was resumed, more reads were generated but the number of reads being generated tapered off quickly at about 30 hours.

Most of the reads were clustered around 1,000 bp (Fig. 5). The median read length for the sequencing run was 1,418 bp overall and 1,583 bp for reads with qscore >= 7. The longest read was 67,123 bp for both overall reads and reads with qscore >= 7. The

mean qscore for the reads was 8.6 and 9.2 for overall reads and read with qscore >= 7, respectively.

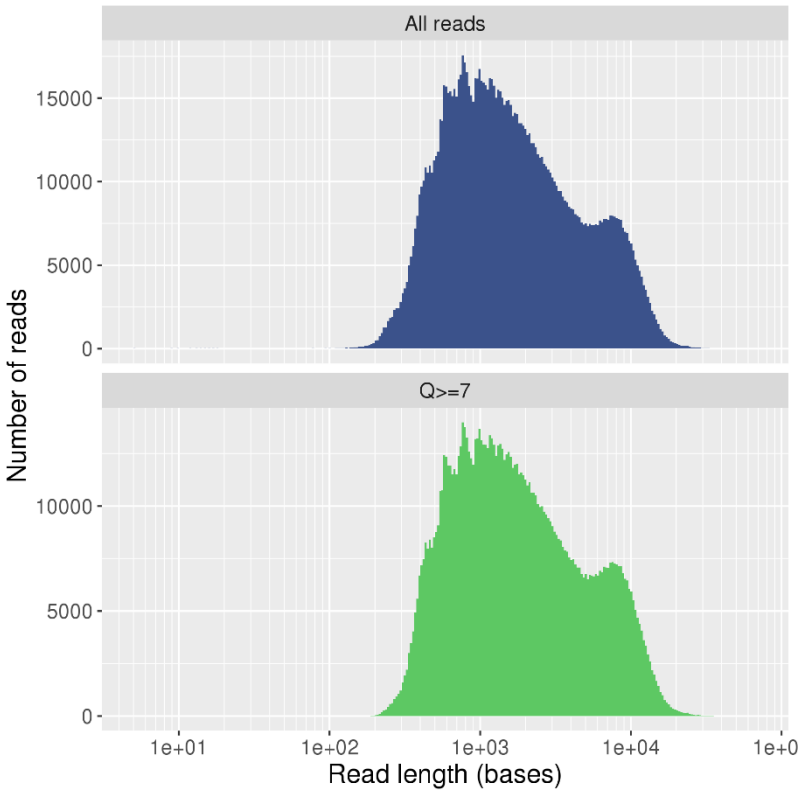


Figure 5: Fragment length histogram generated from minionQC

3.3. Read mapping results

Data from both sequencing runs were aligned with the mappers BWA-MEM, graphmap, LAST, minimap2, and ngmlr. These were evaluated with samtools and *time* Linux command. For the first sequencing run, minimap2 had the best performance, completing the alignment fastest in 0.19 hours, followed by ngmlr which finished in 1.05

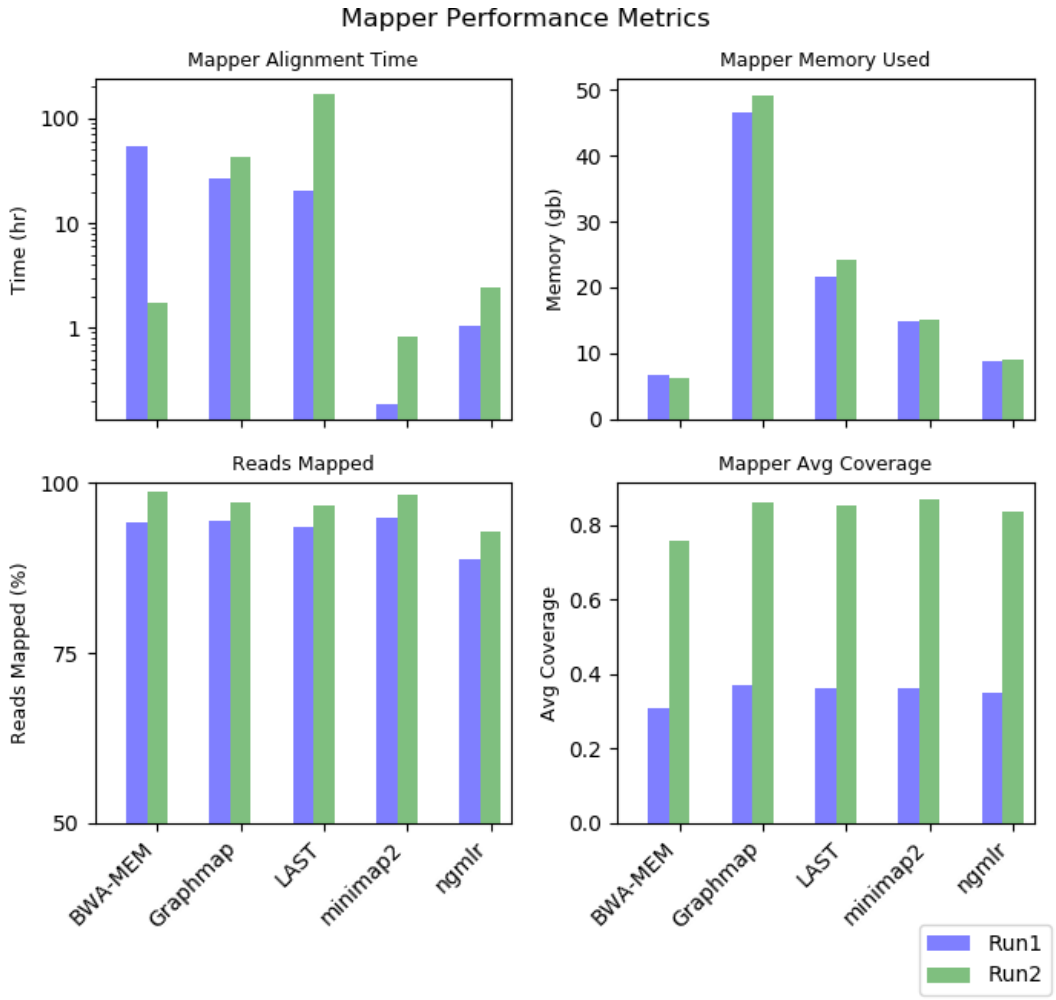


Figure 6: Mapper performance metrics

hours. BWA-MEM, LAST, and Graphmap completed alignment in 53.46, 20.53, and 26.55 hours respectively. For memory usage, BWA-MEM had the best performance, using 6.74 gb of memory, followed by ngmlr with 8.67 gb. Minimap2, LAST, and

graphmap used 14.95, 21.57, and 46.56 gb of memory respectively. Minimap2, graphmap, BWA-MEM and LAST performed similarly at 94.87%, 94.43%, 94.27%, and 93.61% reads mapped respectively. ngmlr had the worst performance of the mappers at 88.78% reads mapped. Graphmap had the best overall coverage from the at 0.37X, followed by minmap2, LAST, and ngmlr at 0.36X, 0.36X, and 0.35X coverage respectively. BWA-MEM had the worst coverage at 0.31X coverage.

For the second sequencing run, the pattern is largely consistent with the first run while mapping roughly double the amount of reads. Minimap2 completed alignment the fastest at 0.82 hours, followed by BWA-MEM at 1.71 hours and ngmlr at 2.47 hours. Graphmap and LAST finished in 43.27 and 168.30 hours respectively. BWA-MEM used the least amount of memory with 6.21 gb followed by ngmlr with 8.98 gb. Minimap2, LAST, and graphmap used 15.03, 24.22, and 49.20 gb memory respectively. BWA-MEM, minimap2, and graphmap had similar amount of mapped reads with 98.64%, 98.37%, and 97.18% reads mapped respectively, followed by LAST with 96.70% mapped reads. ngmlr had the worst mapping performance with 92.96% of reads mapped. Minimap2 had the best coverage at 0.87X coverage, followed by Graphamp, LAST, and ngmlr at 0.86X, 0.85X, and 0.84X coverage respectively. BWA-MEM had the worst coverage with 0.76X.

The alignment time increased in our second run due to the dataset being about twice as large as the first run. However, LAST alignment time took more than eight times as long as the first run which was much more proportionally than alignment time increases in the other mappers. Memory usage increased slightly for all mappers

between the two runs and mapped reads improved for all the mappers as well, except for LAST which was able to map all the reads for both runs.

3.4. Mt. Sinai dataset

To evaluate the calling performance of the SV callers, we compared the SV calls from different callers with an SV call set produced by Mt. Sinai School of Medicine. In this

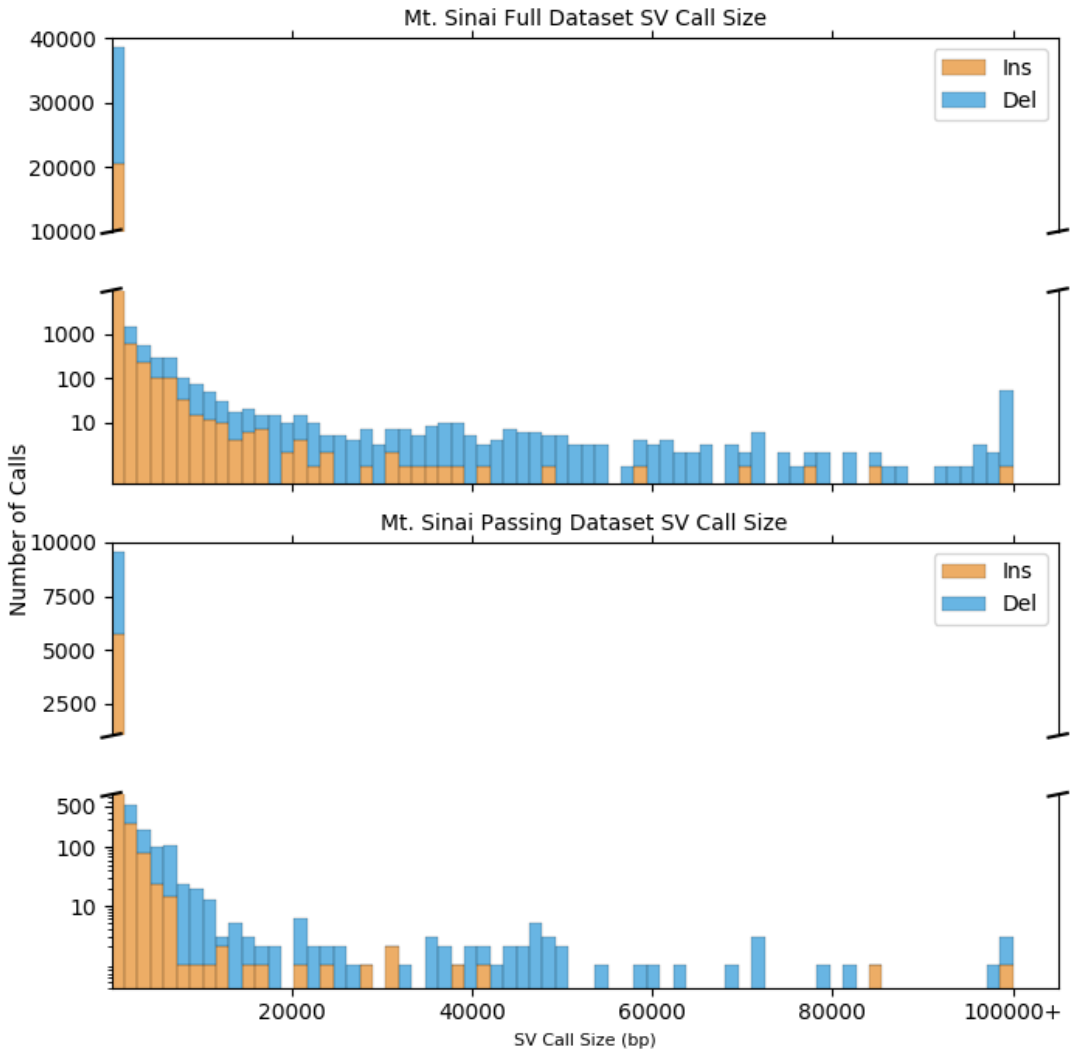


Figure 7: SV call size Mt. Sinai full dataset (top) and passing calls dataset (bottom)

dataset, SV calls were considered passing if called by at least three of the seven pipelines (Appendix A) and the dataset contains only deletion and insertion calls. The full dataset contains 20957 deletions and 22199 insertions totaling 43156 SV calls while the passing subset contains 4495 deletions and 6099 insertions, totaling 10594 calls. The SV calls in the dataset were merged, lifted over to hg38, and then filtered to remove SV calls from unincorporated contigs or from the mitochondrial genome, leaving the dataset with 16171 deletions and 21589 insertions, totaling 35120 total SV calls (Appendix B). The same merging and filtering steps were done with the passing calls in the dataset. Following these steps, 4354 deletions and 6066 insertions were present in the passing subset.

The SV calls in the Mt. Sinai dataset skewed towards smaller SVs (Fig. 7). While large calls over 100,000 bp existed in the dataset, the median SV call was 107 bp and 120 bp for deletions and insertions respectively. For passing calls, the median SV call was 312 bp and 299 bp for deletions and insertions respectively.

3.5. SV calling results

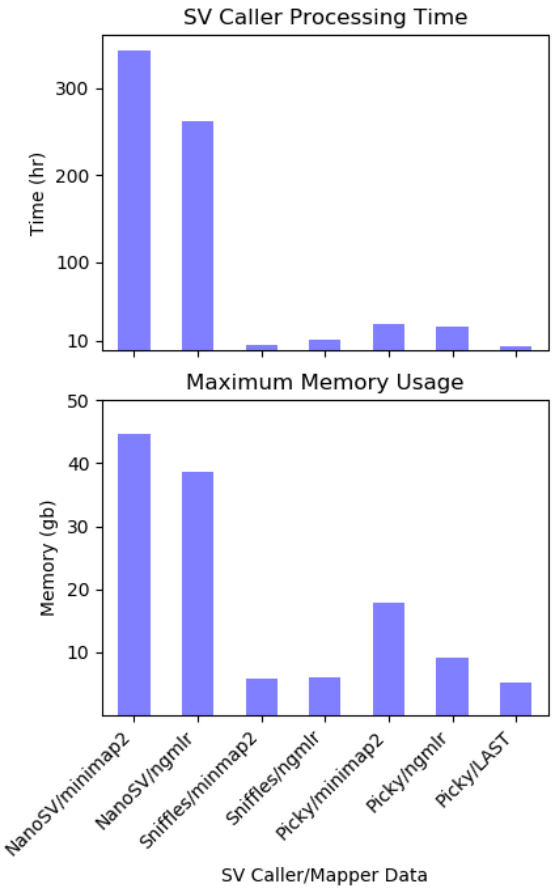


Figure 8: SV Caller processing time and memory usage

We evaluated the performance of the SV call software on ONT's sequencing data using a high-coverage (30x) ONT whole-genome sequencing dataset of individual NA12878. The SV callers Sniffles, NanoSV, and Picky were tested in combination with minimap2 and ngmlr. Minimap2 and ngmlr were chosen for use in this analysis because of their development specifically for long-read data, as well as their relatively short alignment time and comparable performance. Since minimap2 was developed by the same lab as BWA and designed to be its replacement for long-read data, we excluded BWA for the SV call analysis. We also included LAST in combination with Picky because the caller was initially designed to run with LAST, despite the long running time.

Picky calling LAST-mapped data took the least processing time at 4.63 hours, followed by Sniffles paired with minimap2 and ngmlr data, taking 5.58 and 11.19 hours respectively. Picky paired with ngmlr and minimap2 data took 29.19 and 25.88 hours to

process. NanoSV had the longest processing time overall, taking 343.24 and 261.71 hours to align minimap2 and ngmlr data respectively.

Picky paired with LAST data and Sniffles paired with minimap2 and ngmlr had similar memory usage with 5.12, 5.83, and 5.87 gb memory usage. Picky paired with ngmlr and minimap2 data used 9.13 and 17.88 gb respectively. NanoSV used the most memory with 38.59 and 44.72 gb when paired with ngmlr and minimap2 data, respectively.

Multiple SV types were detected by each SV caller. Deletions were the most frequent SV detected for all SV callers and insertions were the second most frequent call for Sniffles and NanoSV. Both SV callers detect duplications, and Sniffles also detects translocations and had a few ambiguous deletions/inversion calls. In contrast, Picky detected more duplications and inversions than insertions, especially when paired with ngmlr and minimap2. We have chosen only to evaluate deletions and insertions because the benchmark dataset only contains deletions and insertions (Table 2).

SV Caller	Mapper	Total	DEL	INS
Sniffles	minimap2	45273	37919	7354
Sniffles	ngmlr	40109	34769	5340
NanoSV	minimap2	122115	82840	39275
NanoSV	ngmlr	103950	82077	21873
Picky	minimap2	37178	35931	1247
Picky	ngmlr	20682	19407	1275
Picky	LAST	44793	42690	2103

Table 2: SV calls from caller/mapper combination

The distribution of the SV calls skews towards shorter calls for both deletions and insertions (Fig. 8). The median deletion call was under 100 bp for all SV callers, though each caller had a

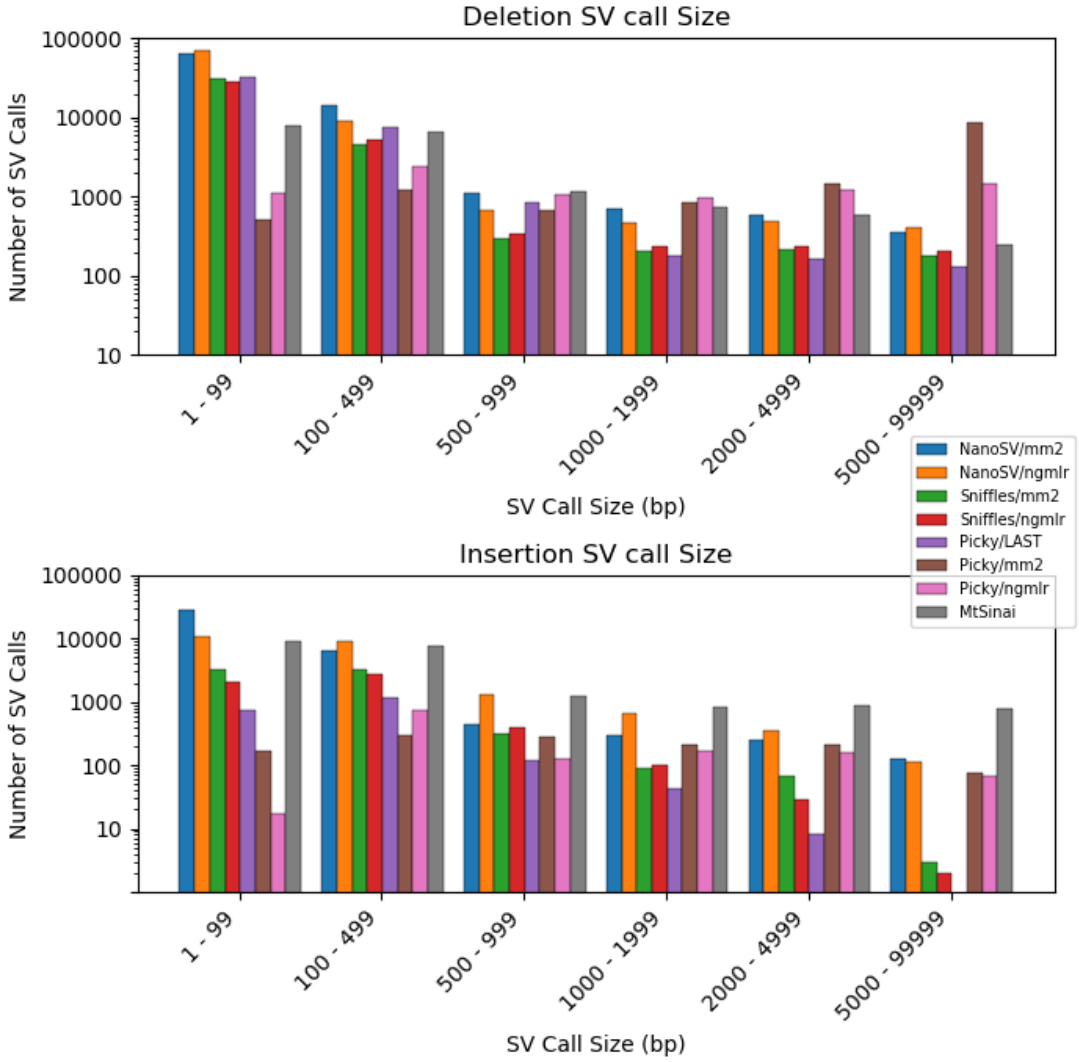


Figure 9: SV call size distribution for each SV type

few deletion calls over 100,000 bp. The median insertion size was under 250 bp for all SV callers except with while for Picky when paired with LAST, where the median size was higher at 750 bp. The other SV types detected vary widely depending on the SV caller and mapper combination. Since the SV calls in the Mt. Sinai dataset contained only deletions and insertions, we did not evaluate other SV types against the Mt. Sinai dataset. Within these calls, SV calls

over 100,000 bp were filtered out. We observed a few very large calls within the dataset but reasoned that these calls were likely inaccurate. The 100,000 bp threshold filtered out these calls but was loose enough to retain most of the SV calls and was close to the maximum call length in the Mt. Sinai dataset. The overlapping calls were then merged and the deletion and insertion calls were counted (Table 3).

Caller	Mapper	All SVs called	Deletions called	Insertions called
Sniffles	minimap2	44943	37681	7262
Sniffles	ngmlr	39913	34586	5327
NanoSV	minimap2	107945	68671	39274
NanoSV	ngmlr	88009	66137	21872
Picky	minimap2	5966	4768	1198
Picky	ngmlr	8448	7185	1263
Picky	LAST	44774	42672	2102

Table 3: SVs called from caller/mapper combination after filtering

3.6. Evaluating SV caller results using the Mt. Sinai dataset

Since we regarded the Mt. Sinai calls as the benchmark, we considered any Mt. Sinai call which overlapped with an SV call as a true positive. We considered any Mt. Sinai call that did not overlap with an SV call as a false negative and any SV call that did not overlap with a Mt. Sinai call a false positive. The number of true positive, false negative, and false positive calls from each call set is shown in Appendix D.

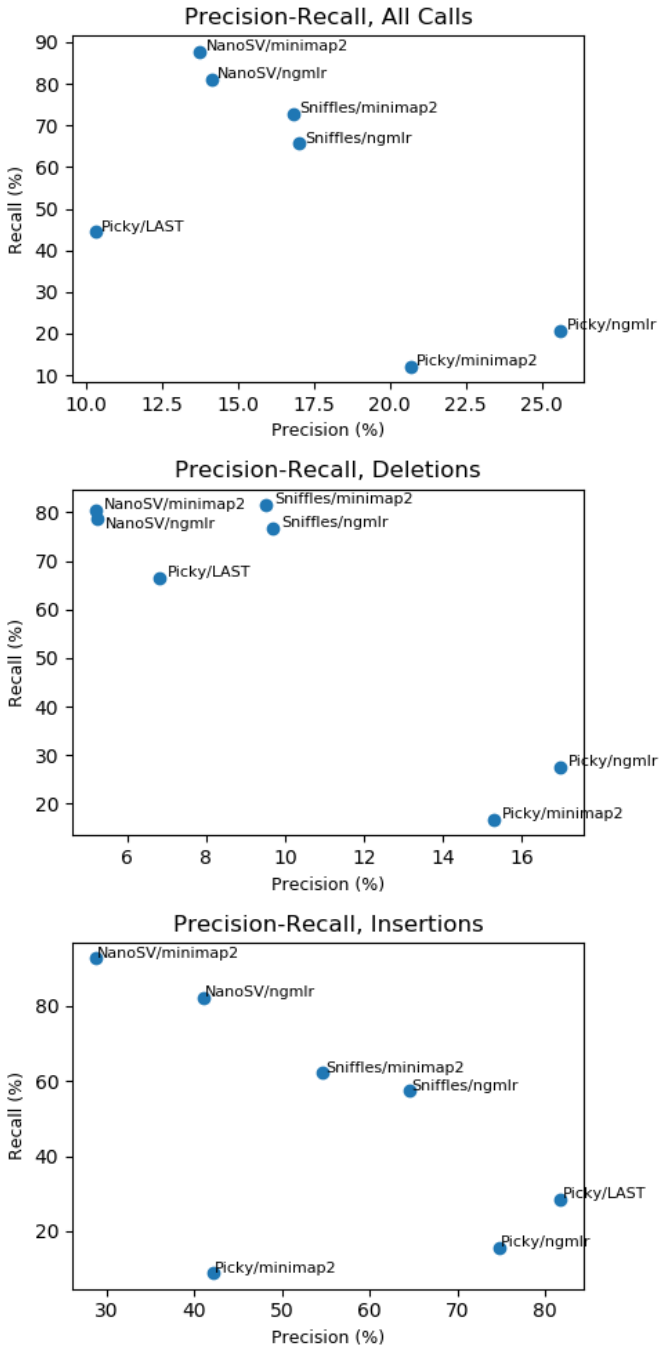


Figure 10: SV Caller Precision-Recall for combined calls (top), deletions (middle) and insertions (bottom)

Using these results, we calculated the recall, precision, and F1 score to evaluate the SV caller performance. Recall, precision, and F1 score of the SV calls from these mapper and caller combinations were benchmarked against the passing calls from Mt. Sinai dataset (Fig. 10, 11). For recall, NanoSV with minimap2 had the

best overall performance with 87.73% of calls detected and was also the most sensitive for insertions with 92.71% of insertions detected. Sniffles with minimap2 was the best performer for detecting deletions at 81.49%, slightly more sensitive than NanoSV, which detected 80.45%. Both Sniffles and NanoSV seem to be slightly more sensitive in combination with minimap2 than ngmlr. Picky had overall poor performance for recall, and in combination with minimap2 and ngmlr the results were especially poor.

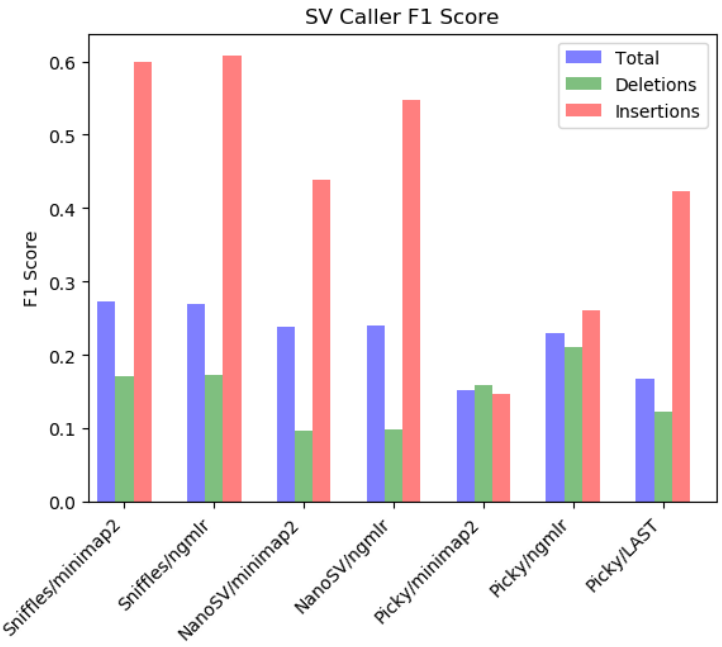


Figure 11: SV Caller F1 Score

In terms of precision, nearly all the caller and mapper combinations performed poorly, due to the large number of calls generated by the callers. However, Picky in combination with LAST had the highest precision in combination with LAST at 81.78%. Picky with ngmlr had the highest overall precision with 25.62% of calls. In general, Picky with minimap2 and ngmlr outperformed the other combinations in terms of precision, though both combinations also had very few insertion calls.

F1 score was highest for Sniffles overall, while also having the highest F1 score when evaluating insertions in particular (Fig. 11). F1 score was similar for Sniffles in

combination with both minimap2 and ngmlr. F1 score for deletions was lower across most SV caller and mapper combinations due to the high number of false positives called. Picky in combination with ngmlr had the highest F1 score for deletions but called very few deletions in comparison to other SV callers.

4. Discussion

Structural variants comprise about 1% of heterogeneity between human genomes and have a significant role in phenotypic variation and disease susceptibility. SVs have been implicated as driver alterations in a variety of disease, including severe neurological diseases and cancer. Thus, advances in our ability to detect and evaluate SVs in the genome is crucial to improving our understanding of the biological impact of SVs

While NGS has revolutionized genomics, its short read length is not optimal for resolving SVs. Recently, ONT released its nanopore-based sequencers capable of generating long reads, potentially improving our ability to detect SVs. We generated nanopore sequencing data from the MinION to evaluate current mappers and used publicly released high-coverage nanopore data to assess SV callers developed for long read sequencing data.

We prepared a library for sequencing following ONT's 1D genomic DNA by ligation protocol and sequenced on a single MinION flow cell. We experienced crashes on two different sequencing runs but successfully resumed the second crash for a full 48-hour run. While we have not pinpointed the reason for the crashes, we suspect the crashes

are related to insufficient memory or communications issue between the sequencer and the computer. Since the MinION is being actively developed, the software and system requirements have both been updated multiple times since our experiment and we expect these issues are likely resolved.

Two versions of ONT's Albacore base caller, Albacore ver 1.2.6 and Albacore ver 2.0.2, were tested. In terms of quality control, the differences between the two versions were small and mainly due to the incorporation of a qscore filter by default in Albacore ver 2.0.2 and we did downstream analysis with this version. Albacore ver 2.0.2 switched from event-based base calling to raw signal base calling, which improves single read accuracy [25]. While we tested the most recent version of Albacore available to us, base calling is an active and rapidly changing area of development within ONT and the nanopore community. Older base callers like Nanonet, DeepNano [27], and basecRAWller [28] are no longer maintained while ONT released six different versions of Albacore within the same year. Guppy, a new GPU-based caller produced by ONT that is not yet publicly available, recently replaced Albacore as the production basecaller for the MinION system [29].

After base calling, we observed that the read lengths were shorter than expected. However, the library QC showed that we had fragments over 10 kbp after both the fragmentation and adapter ligation steps. We suspect that the shorter read lengths may be due to nicks on the DNA fragments that occurred during library preparation. While we followed the standard protocol published by ONT the optional DNA repair step was omitted. This step may have been necessary if many of the fragments were damaged, in

which case part of the fragment may have disassociated from the adapter and would not have been sequenced in the nanopore. Several others have included the DNA repair step [12, 30] and attained fragment sizes above 10 kbp with an otherwise similar library preparation. However, one lab that included the DNA repair step also reports shorter fragment lengths for some of the libraries [20]. The chemistry is still under development by ONT and the newest version of the 1D sequencing by ligation kit, SQK-LSK109, removes the fragmentation step and requires DNA repair [31]. The new kit and protocol have already produced improved fragment lengths and we expect longer fragment lengths to be more reproducible as the chemistry improves and the nanopore sequencing community expands.

We benchmarked several mappers using our sequencing data including both older established mappers and recent long-read mappers. Alignment time and memory usage varied widely between mappers while differences in percent mapped reads were more moderate. All mappers improved percent reads mapped for the second run possibly because the number of reads to align was doubled. Minimap2, a recently developed long-read mapper was extremely fast with high percentage of reads mapped. BWA-MEM, developed by the same author as minimap2, had similar read mapping performance to minimap2 and used the least memory of all the mappers. We may need to map longer reads to see demonstrable benefits in read mapping that minmap2 is designed to have over BWA-MEM. Ngmlr had the worst reads mapping performance but had comparatively fast alignment time and low memory usage. Graphmap had the longest alignment time of the three new long-read mappers and high memory usage.

LAST was comparable to the other long read mappers but had a very long alignment time for the second sequencing run.

We chose minimap2 and ngmlr as mappers to test in combination with all the SV callers tested because of their development as long-read specific tools and relatively short alignment times. Additionally, ngmlr was developed alongside Sniffles and is the preferred mapper according to its authors. LAST was also chosen to run with Picky due to its claimed synergy and its incorporation into the default Picky workflow. We excluded graphmap from the evaluation because of its long run time and excluded BWA-MEM since it had the worst overall coverage and minimap2 was developed to be its replacement for long reads.

Since the fragment length and coverage of our data were not optimal for evaluating SV calls, we used publicly released NA12878 data from Nanopore Sequencing Consortium for the SV caller evaluation. We used an SV call set generated from PacBio data from Mt. Sinai School of Medicine as our benchmark, reasoning that SV calls in common between our nanopore-derived data and the PacBio generated data would more likely be true.

The SV call results differ significantly when evaluated for deletions and insertions separately. If the user does not have any specific requirements for SV calling, we would recommend Sniffles paired with minimap2. This combination of tools would have the fastest processing time and the overall performance is balanced for both deletions and insertions.

For deletion focused performance, we recommend Sniffles because of its relatively balanced performance. Sniffles and NanoSV had similar recall for deletions but Sniffles was more precise. Sniffles deletion performance did not change drastically between minimap2 and ngmlr, but Sniffles paired with minimap2 is recommended over ngmlr because of its recall and fast run time. Picky showed especially low recall with minimap2 and ngmlr. We suspect the Picky may require further development before we can use it in combination with these mappers. Although the F1 score for Picky paired with ngmlr is the highest of the F1 scores for deletions, we do not recommend this combination due to extremely poor recall.

We also recommend Sniffles for calling insertions due to its more balanced performance between recall and precision, though NanoSV may be considered if high insertion recall is desired and additional false positives can be tolerated. Sniffles overall performance paired with minimap2 and ngmlr is similar, but we recommend minimap2 for higher recall and much faster run time. Like with deletions, Picky had poor recall for deletions with minimap2 and ngmlr and further optimization is required. Picky run with LAST alignment had slightly better recall than paired with the other callers and high precision and may be useful for applications where high insertion precision is required and recall is not a priority (Table 4).

SV Caller/Mapper Performance	Optimal For
Sniffles/minimap2	Overall deletion calling
Sniffles/minimap2	Overall Insertion calling
NanoSV/minimap2	Insertion recall
Picky/LAST	High insertion precision

Table 4: SV Caller/Mapper recommendation by application

We evaluated our calls against a 44X coverage dataset generated through Pacific Biosciences sequencing technology, one of few long-read datasets that attempt to contain high confidence SV calls by employing several different pipelines. While we reason that SV calls in this benchmark dataset are more likely be true if also found in our own SV calls generated from nanopore sequencing chemistry, the benchmark is limited in several ways. The benchmark dataset is limited to deletions and insertions, which did not allow us to fully evaluate SV callers that also generate calls for other SV types. One of the SV callers used to generate the benchmark, PBHoney [32], is an older SV caller that is not actively maintained. Some of the pipelines used to generate the benchmark set were similar to each other, differing only in different versions of the software used. We predict that as long-read sequencing technology matures, more comprehensive SV call sets will become available. However, experimental validation of some SV calls is necessary to empirically assess the accuracy of the calls.

The SV calling results from the tools tested in this study show that there is significant room for improvement in both recall and precision. In order to improve performance of the current SV callers, we plan to follow an integrative approach and combine the calls from multiple callers. The consensus SV call set can be expected to increase the precision and recall of our SV calls.

5. Conclusion

Nanopore sequencing is a rapidly developing technology in both sequencing chemistry and data analysis. ONT continues to enhance nanopore sequencing

performance with regular chemistry, reagent kit, and analysis algorithm updates. For example, ONT released an updated ligation sequencing kit last year and recently launched its new R10 nanopore into an early access program. Recently, the MinION default basecaller was also changed to Guppy.

New mappers and SV callers have been developed to leverage long-read sequencing data. Though few resources are available to benchmark these tools, we have established a workflow for evaluating these mappers and SV callers. We found that SV caller performance diverges depending on the type of SV the user desires to evaluate. Therefore our recommendation may differ depending on the desired application. However, for an initial analysis without specific requirements, we recommend minimap2 and Sniffles due to its rapid speed and relatively balanced performance calling both insertions and deletions.

As can be expected with first generation tools, significant optimization to improve accuracy and recall is needed. In future evaluations we intend to combine results between SV callers to form a consensus set as well as evaluate SV callers with simulated nanopore data. We expect resources from ONT and the nanopore sequencing community to improve as the technology improves and adds to its user base.

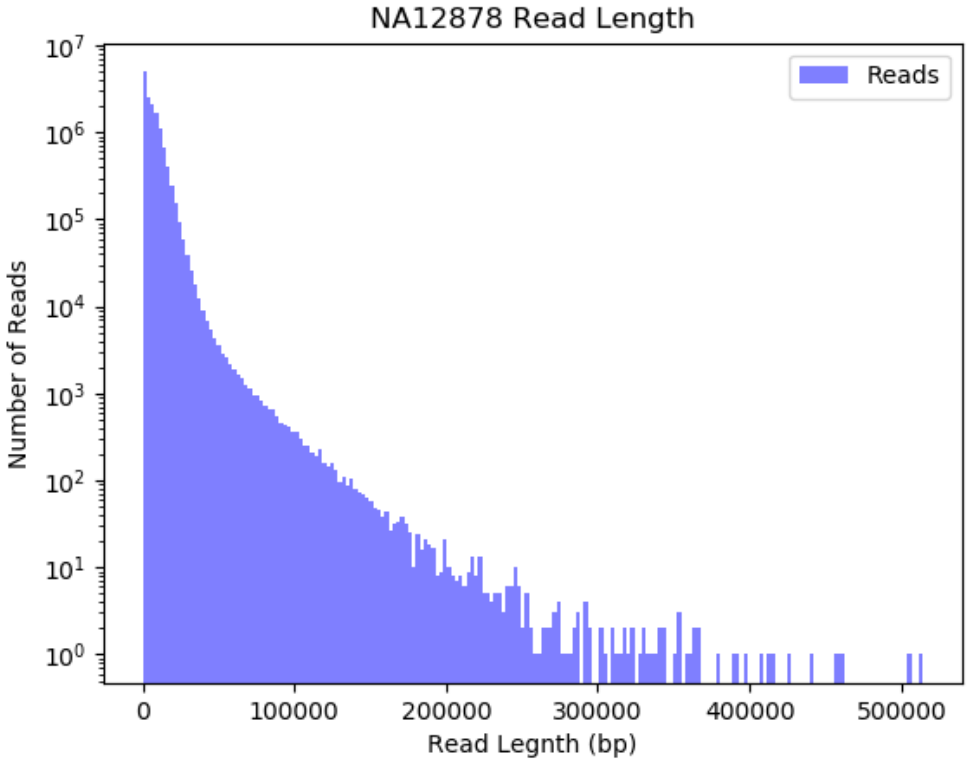
Appendix A. Methods used to generate Mt. Sinai Dataset

Methods used to generate Mt. Sinai Dataset:
PBHoney, raw reads, blasr1.3.1
Custom pipeline, raw reads, blasr1.3.1
PBHoney, error-corrected reads, blasr1.3.1
Custom pipeline, error-corrected reads, blasr1.3.1
Assembly
Custom pipeline, error-corrected reads, blasr1.3.2
Custom pipeline, raw reads, blasr1.3.2

Appendix B. Mt. Sinai Dataset filtering

	Full dataset:	After merging:	Liftover to hg38:	Filtering:
Total Calls	43156	35836	35496	35120
Deletions	20957	16649	16375	16171
Insertions	22199	22174	22046	21589
	Passing dataset:	**After merging:**	**Liftover to hg38:**	**Filtering:**
Passing Calls	10594	10503	10425	10404
Deletions	4495	4430	4363	4354
Insertions	6099	6089	6078	6066

Appendix C. NA12878 read length size distribution



Appendix D. SV caller/Mapper performance

True Positives			
Caller/Mapper combination	**Total:**	**Deletions:**	**Insertions:**
Sniffles/minimap2	7568	3548	4020
Picky/minimap2	1264	727	537
NanoSV/minimap2	9127	3503	5624
Sniffles/ngmlr	6825	3344	3481
Picky/ngmlr	2161	1202	959
NanoSV/ngmlr	8426	3431	4995
Picky/LAST	4628	2896	1732

False Negatives			
Caller/Mapper combination	**Total:**	**Deletions:**	**Insertions:**

Sniffles/minimap2	2836	806	2046
Picky/minimap2	9140	3627	5529
NanoSV/minimap2	1277	851	442
Sniffles/ngmlr	3579	1010	2585
Picky/ngmlr	8243	3152	5107
NanoSV/ngmlr	1978	923	1071
Last/picky	5776	1458	4334

False Positives			
Caller/Mapper combination	**Total:**	**Deletions:**	**Insertions:**
Sniffles/minimap2	37382	34090	3292
Picky/minimap2	4732	4039	693
NanoSV/minimap2	93092	65094	27998
Sniffles/ngmlr	33121	31231	1890
Picky/ngmlr	6284	5965	319
NanoSV/ngmlr	75561	62658	12903
Last/picky	40151	39768	383

Recall			
Caller/Mapper combination	**Total:**	**Deletions:**	**Insertions:**
Sniffles/minimap2	72.43%	81.49%	65.74%
Picky/minimap2	10.26%	16.70%	5.61%
NanoSV/minimap2	87.73%	80.45%	92.71%
Sniffles/ngmlr	64.04%	76.76%	54.75%
Picky/ngmlr	20.51%	27.61%	15.36%
NanoSV/ngmlr	80.99%	78.80%	82.34%
Last/picky	44.26%	66.51%	28.17%

Precision			
Caller/Mapper combination	**Total:**	**Deletions:**	**Insertions:**
Sniffles/minimap2	16.75%	9.53%	53.79%
Picky/minimap2	17.50%	15.29%	25.96%
NanoSV/minimap2	13.76%	5.21%	28.71%
Sniffles/ngmlr	16.63%	9.70%	61.41%
Picky/ngmlr	25.29%	16.98%	72.14%
NanoSV/ngmlr	14.14%	5.26%	41.01%
Last/picky	10.29%	6.81%	80.88%

F1 Score			
Caller/Mapper combination	**Total:**	**Deletions:**	**Insertions:**
Sniffles/minimap2	0.27	0.17	0.60
Picky/minimap2	0.15	0.16	0.15
NanoSV/minimap2	0.24	0.10	0.44
Sniffles/ngmlr	0.27	0.17	0.61
Picky/ngmlr	0.23	0.21	0.26
NanoSV/ngmlr	0.24	0.10	0.55
Last/picky	0.17	0.12	0.42

Appendix E. – Mapper and SV caller commands used

BWA-MEM:

/bwa mem -t 8 hg19.fa nanopore_data.fastq > nanopore_bwamem.sam

Graphmap:

/graphmap align -r hg19.fa -t 8 -d nanopore_data.fastq -o nanopore_graphmap.sam

LAST:

/ lastal -Q1 -P 8 -p last_nanopore.param referencedb nanopore_data.fastq | \last-split >

nanopore_reads.maf

/maf-convert sam nanopore_reads.maf > nanopore_last.sam

Minimap2:

/minimap2 -t 8 -ax map-ont hg19.fa nanopore_data.fastq > nanopore_minimap.sam

ngmlr:

/ngmlr -r hg19.fa -t 8 -q nanopore_data.fastq -o nanopore_ngmlr.sam -x ont

NanoSV:

NanoSV -t 16 -s samtools -b human_b38.bed -o nanosv.vcf mapper_ONT.sort.bam

Sniffles:

sniffles -t 16 -m mapper_ONT.sort.bam -v sniffles.vcf

Picky:

```
./picky.pl script --fastq LongRead.fastq --thread 4 > LongRead.sh

./LongRead.sh
```

Bibliography

1. Kircher, M. and J. Kelso, *High-throughput DNA sequencing--concepts and limitations.* Bioessays, 2010. **32**(6): p. 524-36.
2. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
3. Bayley, H., *Nanopore sequencing: from imagination to reality.* Clin Chem, 2015. **61**(1): p. 25-31.
4. Franca, L.T., E. Carrilho, and T.B. Kist, *A review of DNA sequencing techniques.* Q Rev Biophys, 2002. **35**(2): p. 169-200.
5. Ahmadian, A., M. Ehn, and S. Hober, *Pyrosequencing: history, biochemistry and future.* Clin Chim Acta, 2006. **363**(1-2): p. 83-94.
6. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
7. Ju, J., et al., *Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators.* Proc Natl Acad Sci U S A, 2006. **103**(52): p. 19635-40.
8. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.
9. Liu, L., et al., *Comparison of next-generation sequencing systems.* J Biomed Biotechnol, 2012. **2012**: p. 251364.
10. Niedringhaus, T.P., et al., *Landscape of next-generation sequencing technologies.* Anal Chem, 2011. **83**(12): p. 4327-41.
11. Jain, M., et al., *The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.* Genome Biol, 2016. **17**(1): p. 239.
12. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads.* Nat Biotechnol, 2018. **36**(4): p. 338-345.
13. Pang, A.W., et al., *Towards a comprehensive structural variation map of an individual human genome.* Genome Biol, 2010. **11**(5): p. R52.
14. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv:1303.3997, 2013.
15. Sovic, I., et al., *Fast and sensitive mapping of nanopore sequencing reads with GraphMap.* Nat Commun, 2016. **7**: p. 11307.
16. Hinrichs, A.S., et al., *The UCSC Genome Browser Database: update 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D590-8.
17. Sedlazeck, F.J., et al., *Accurate detection of complex structural variations using single-molecule sequencing.* Nat Methods, 2018. **15**(6): p. 461-468.
18. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
19. Gong, L., et al., *Picky comprehensively detects high-resolution structural variants in nanopore long reads.* Nat Methods, 2018. **15**(6): p. 455-460.
20. Cretu Stancu, M., et al., *Mapping and phasing of structural variation in patient genomes using nanopore sequencing.* Nat Commun, 2017. **8**(1): p. 1326.
21. Loman, N.J. and A.R. Quinlan, *Poretools: a toolkit for analyzing nanopore sequence data.* Bioinformatics, 2014. **30**(23): p. 3399-401.
22. Lanfear, R., et al., *MinIONQC: fast and simple quality control for MinION sequencing data.* Bioinformatics, 2019. **35**(3): p. 523-525.

23. Kielbasa, S.M., et al., *Adaptive seeds tame genomic sequence comparison.* Genome Res, 2011. **21**(3): p. 487-93.
24. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
25. Dokos, R., R. Ronan, and J. Pugh. *Release of Albacore v2.0.1*. 2017 March 24th, 2019]; Available from: https://community.nanoporetech.com/posts/release-of-albacore-2-01.
26. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.
27. Boza, V., B. Brejova, and T. Vinar, *DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads.* PLoS One, 2017. **12**(6): p. e0178751.
28. Stoiber, M. and J. Brown, *BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal.* bioRxiv, 2017: p. 133058.
29. Wick, R.R., L.M. Judd, and K.E. Holt, *Performance of neural network basecalling tools for Oxford Nanopore sequencing.* bioRxiv, 2019: p. 543439.
30. De Coster, W., et al., *Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome.* bioRxiv, 2018: p. 434118.
31. Kuznetsova, O. *Ligation Sequencing Kit (SQK-LSK109) release*. 2018 March 24th, 2019]; Available from: https://community.nanoporetech.com/posts/ligation-sequencing-kit-s.
32. English, A.C., W.J. Salerno, and J.G. Reid, *PBHoney: identifying genomic variants via long-read discordance and interrupted mapping.* BMC Bioinformatics, 2014. **15**(1): p. 180.