Whole-Genome Sequencing of two Swedish Individuals on PromethION

Nazeefa Fatima

Master’s Degree Project in Bioinformatics, 30 credits
2019

Department of Biology
Lund University

Reset form

Whole-Genome Sequencing of two Swedish Individuals on PromethION

Nazeefa Fatima[1,2]

1 Department of Biology, Faculty of Science, Lund University, Sölvegatan 37, 223 62, Lund, Sweden
2 National Genomics Infrastructure, Science for Life Laboratory, Husargatan 3, 752 37, Uppsala, Sweden

na1640fa-s@student.lu.se

1 Abstract

Background: Chromosomes can undergo various changes such as deletions, inversions, insertions, and/or translocations resulting in structural variation differences between individuals. Structural variants are a common source of variability in the human genome and have been known to be associated with common diseases such as autism, cancer, and rare human diseases [1, 2]. However, they have not yet been extensively studied at the higher resolution. SVs are complex genomic components partially due to being known to emerge in repetitive regions [3]. Alignment of short reads to repetitive regions can cause ambiguity and has, therefore, posed challenges in the past to detect SVs. New approaches for SV detection have been enabled by the recent improvements in sequencing technologies. In particular, the new long-read single-molecule sequencing instruments provided by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) produce a high yield in a short period while keeping a low cost for a library preparation. These instruments make it possible to generate high quality representations of whole genomes and enable reliable structural variant calling in human individuals [4, 5].

Objectives: A recent study performed on PacBio's Single-Molecule Real-Time sequencing of two Swedish human genomes, Swe1 (male) and Swe2 (female), as part of the SweGen 1000 Genomes project (https://swefreq.nbis.se), uncovered over 17K SVs per individual as well as various other genomic components [6] that are otherwise not detectable in short reads. As a follow-up study, we have now generated data for the same two Swedish individuals on the ONT's PromethION system, a new nanopore based sequencing instrument, that is known for its higher throughput as compared to the PacBio.

Results and Conclusion: We present a pilot study that evaluates nanopore data derived from whole-genome sequencing (WGS) on PromethION in comparison to the Single-Molecule Real-Time (SMRT) reads obtained from the PacBio RS II platform. We performed comparative analyses of single- molecule long-read technologies in a context of mappability, and SV detection that resulted in an average of 17k and 24k variants across nanopore and SMRT datasets, respectively. The results will be useful for the large-scale SweGen project in a context of validation and comparison of SVs in Swedish individuals. In addition, the study serves as a bioinformatics pipeline for future long-read data analyses and sets a basis for what to consider when designing future PromethION experiments.

2 Introduction

The field of high-throughput sequencing has been rapidly developing with new methods and technologies. Long-read technology is now becoming the go-to particularly for large-scale research projects mainly because it employs single-molecule approach that allows amplification-free sequencing. Single-molecule approach can be implemented in two ways i.e. sequencing-by-sensing (e.g. nanopore sequencing (Figure 1) commercialised by Oxford Nanopore Technologies (ONT) [7] and sequencing-by-synthesis (adopted by Pacific Biosciences (PacBio)) [8, 9]. This study involves comparative analyses of data generated by platforms that employ each of the two aforementioned approaches. In contrast to short-read sequencing platforms, long-read technology is far more efficient in uncovering both short and long range patterns in complex genomic regions such as repeat regions, and structural variations (SVs) [6, 10]. The question, however, the technology always brings with it is how to improve current approaches for error rates to be able to analyse genomic structures at a better resolution? This has fueled interest in improving protocols and platforms in a way that results in requiring minimal effort from a user-end and producing high throughput.

The use of nanopore sequencing is becoming a practice, democratizing life sciences research, with ONT's MinION device being more commonly used for both model [11, 12] and non-model organisms [13, 14, 15]. In a context of human biology research, nanopore sequencing has mainly been performed (with MinION) for individual cell lines and genomes [16, 17, 18, 19]. With latest improvements in devices and protocols, the attention is being driven steadily towards population-level whole-genome sequencing (WGS) experiments to gain insight into diversity of genomes across and within different communities. With large-scale sequencing comes a demand of a high speed and capacity to deliver data in a short turnaround. In a world of nanopore sequencing, PromethION (officially introduced in 2016 [20]) is the largest and highest-throughput platform so far to meet these demands [21].

The PromethION device has a capacity of up to 48 flow cells where each flow cell has 3000 channels collecting measurements from over 1 million pores, and generating sequencing data of multiple whole genomes. For example, the study by Nicholls *et al* [22] involved WGS of ten microbial communities that generated up to 300 Gb worth of data with over 100x of an average coverage for PromethION libraries. In a context of WGS of human genomes, there have been only two published studies that involved the use of PromethION. Last year, Roeck *et al* performed sequencing of 11 human individuals with each run obtaining up to 30x coverage for a maximum yield of 98 Gigabases (Gb) [23]. Following in the footsteps, the study by De Coster *et al* on Yoruban NA19240 genome found that PromethION allowed for sequencing of 59x median coverage across five flow cells yielding a total of 208 Gb throughput [24] - note, the genomic data analysed in the study was derived from Lymphoblastic cell lines (LCL). We report here the first pilot study in Sweden that involved PromethION sequencing of two human (blood-derived) genomes; we found that up to 30x coverage can be achieved on one flow cell with a maximum yield of nearly 78 Gb. In addition, our sequencing results are in line with results from the study by De Coster *et al* for the fact that sheared libraries resulted in a higher yield. In terms of read length N50, our finding is that sheared libraries have a low N50 as compared to unsheared libraries (**Supplementary Table 7.1**) which is also in agreement with aforementioned PromethION studies on human genomes. In addition to sequencing performance, the study lays out a bioinformatics workflow for SV detection in Swedish genomes sequenced on PromethION (Figure 3).

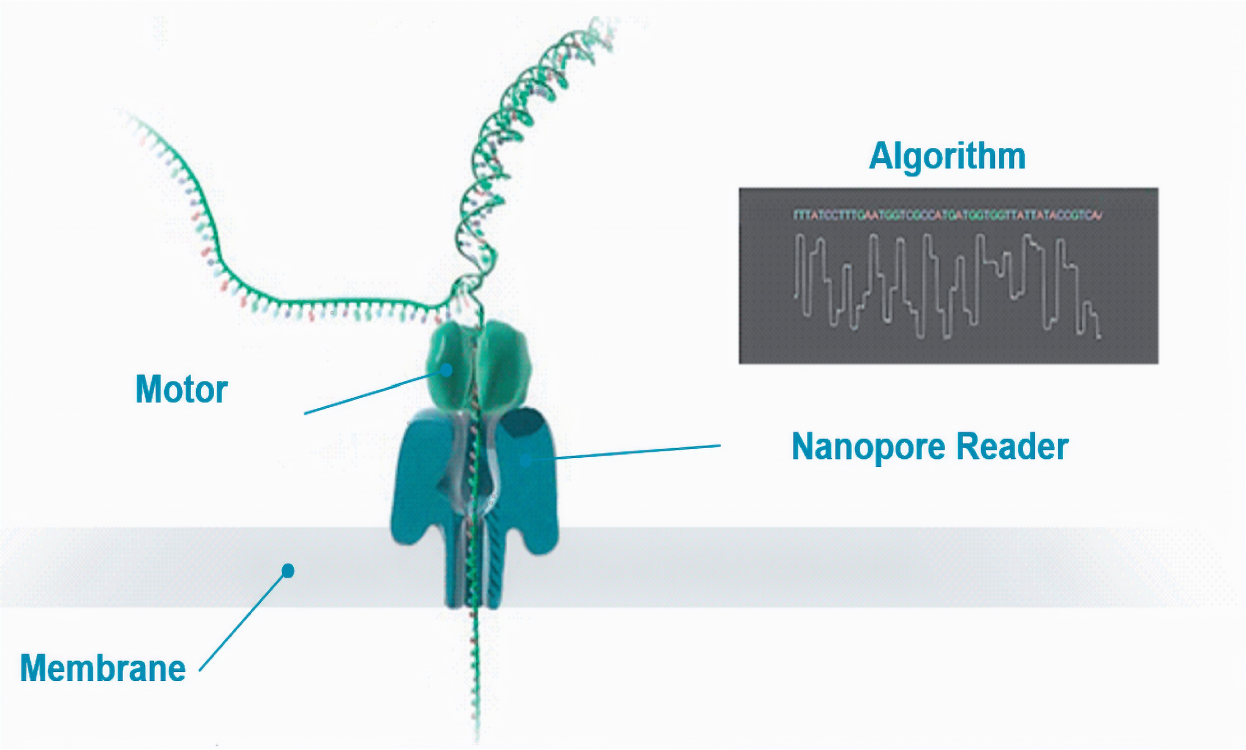

Figure 1. Sequencing-by-sensing: Through the nanopore. DNA helix is unzippped by a motor protein, and a molecule is passed through a pore ('Nanopore Reader') that acts as a hole for a membrane. Changes in current allow to identify a type of molecule. Real-time basecalling is performed as each molecule gets through resulting in a collection of sequenced bases represented by 'squiggles' (shown in grey box). Recurrent Neural Network implementation in basecalling algorithm enables raw signal-processing; assigning bases to data points. Image Source: Pollard *et al* [25].

Genomes are a mosaic of variants with rearrangements of DNA pieces making up large and small-scale SVs such as translocations, insertions and deletions (indels), copy number variants, and so on ranging from as little as 50 kilobases (kb) up to more than 1 Mb [26, 27]. SVs are essentially mutational events, known to contribute to a number of diseases including cancer [28, 29]. Despite several tools available, SV detection has always been a challenging step for two main reasons; a) complex structures of SVs such as ones embedded in repetitive regions and, b) incomplete genome assembly due to short-reads. Long reads promise to fill these gaps allowing to resolve and validate a repertoire of rare and novel SVs in human genomes. However, when it comes to variant detection pipelines for long-read data [30, 31, 32], there is no standard tool since identification and profiling are subjective and vary based on research questions. For clinical sequencing data [19, 34], for instance, algorithms that offer high precision and recall are often preferred [33, 35]. Given the length characteristic of sequencing reads in our data and the aims of our pilot study, we found Sniffles [36] to be the most appropriate caller. In this study, we report a catalogue of SVs identified in reads generated by two long-read technologies and perform comparative analysis of the two. In distinction between technologies for SV detection in human genomes, previous research has shown that SMRT sequencing appears more promising to identify a significant number of SVs [37, 38]. This was also the case in our study; a large proportion of SVs were identified in SMRT reads as compared to nanopore reads (**Table 2 and 3, and Supplementary Table 7.4.2)** and the evidence is rooted in the fact that PacBio RS II offers higher read accuracy and coverage. However, whether this suggests that PacBio RS II is a better platform in comparison to PromethION remains a question which is addressed in our study.

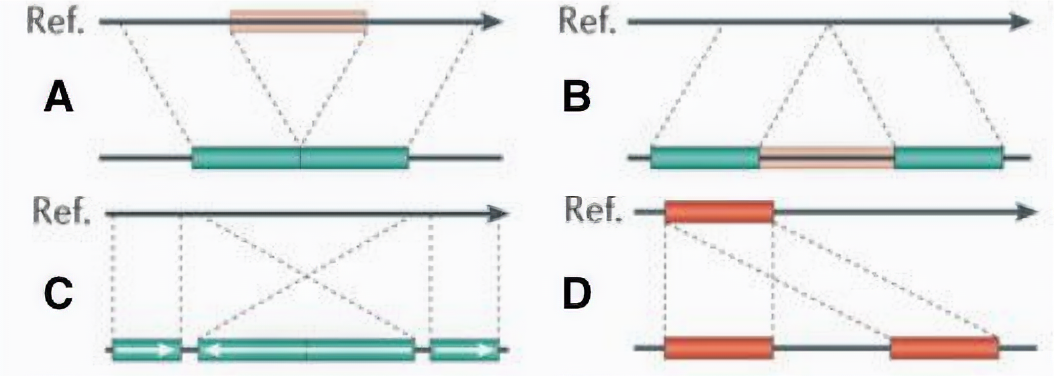

Figure 2. Schematic illustration of SVs. The figure shows the SV types, detected and analysed in this study, defined with regards to a reference (top part labelled Ref). (A) Deletion; removal of chromosomal segment in a genome relative to a reference (B) Insertion; addition of sequence between two adjacent sequences (C) Inversion; rearrangement of a chromosomal segment, in an inverted (180 degree) orientation, relative to flanking regions in a reference (D) Duplication; copy of a chromosomal segment in a non-reference genome. Image Source: Alkan *et al* [39]

3 Methods

3.1 Sample Collection

The procedure for genomic DNA extraction (from whole-blood samples) is described in the study by Ameur *et al* [6] (referred to as **Swe-2018 study** from here on). It is worth mentioning that the collection was done over a decade ago and, the two individuals were selected from a group of participants involved in the SweGen project [40].

3.2 PromethION Library Preparation and Nanopore Sequencing

The DNA ligation sequencing protocol SQK-LSK109 (based on a sequencing pore R9.4.1 chemistry) was used for four PromethION flowcells. Lambda-phage (Accession J02459.1) was used as a control DNA. The sequencing for both individuals was conducted on beta release of PromethION device. Real-time base-calling was one-directional (1D) since sequencing information from one strand was incorporated. Libraries were prepared using both native and sheared DNA; DNA for two libraries were sheared to 20 kilobases (kb) with the MegaRuptor system. Collectively, flowcells generated over 28 million reads with a total yield of 209 Gb. Experiment metrics and statistics on quality scores can be found in **Supplementary Table 7.1 and Section 7.2**.

3.3 Alignment

Genome indexing and alignments were performed with Minimap2 version 2.14-r883 [42]. The algorithmic approach Minimap2 employs is a standard seed-chain-align method [43] i.e. minimisers and matches are treated as seeds and chains, respectively, which is executed in an alignment-free and storage-saving manner. From here onward, **NanoSwe** is used to define PromethION (nanopore) data where NanoSwe1 and NanoSwe2 are used to refer to male and female genomes, respectively. Similarly, **Swe 1 and 2** is used for a reference to PacBio (SMRT) data.

3.3.1 The GRCh38 Reference Genome

The human reference genome assembly used for this study is the GRCh38 release [44] (GCA_000001405.15) [45] that does not contain alternative contigs. This choice is based on the fact that the assembly serves as a sufficiently reliable model for variant calling analyses as shown in previous studies [16, 24]. The release set represents a non-redundant haploid genome containing a total of 195 sequences; primary sequences of assembled chromosomes i.e. autosomes, chromosomes X and Y, and mitochondrial genome (chrM), and (unlocalized) scaffolds with unidentified location in a chromosome, unplaced scaffolds i.e. sequences with unknown chromosome assignment, and a decoy chromosome of 1718 bp for the Epstein-Barr virus (AJ507799.2). This patch of assembly is different from the release used for the Swe-2018 study, where the full set of the GRCh38 was used that includes its decoy version GCA_000786075.2. From here forward, the GRCh38 release used for this study will be referred to as **hg38** and the full set of it used in the Swe-2018 study will be referred to as **hg38-alt**.

3.3.2 Alignment of NanoSwe data

Reads were aligned to the hg38 using the command flags –ax; a serves as a preset and x allows to enable preset option, –map-ont; a type of preset option that sets a mapping mode suitable for nanopore data, and –MD that allows indel calling which is required for subsequent variant calling analyses. In a subsequent round of alignment, NanoSwe data was mapped to the extended version of the reference that includes the the SMRT novel sequences (detected in the Swe-2018 study) added to the assembly of the hg38. The alignment runs were performed with same aforementioned parameters; the indexing of the extended reference was built for a total of 5393 sequences.

3.3.3 Alignment of SMRT data

SMRT reads were aligned to the hg38 with –MD -ax map-pb flags. The map-pb sets a k-mer value of 19, for a reference, which allows indexing of homopolymer compressed minimsers (k-mers) that essentially means compression of homopolymers to a single base. This helps with finding more overlaps when mapping SMRT reads to a reference. The purpose of this particular alignment process was to find differences with previous data in addition to assess performance of Minimap2 for SMRT reads.

3.4 Homology Inference

BLAST version 2.7.1+ [54] was used to find whether a set of novel sequences are significantly related to sequences from other species. The searches were performed against the nucleotide database (blastn) with an e-value threshold set to 1e-10.

3.5 Structural Variant Calling

Alignments were investigated for SVs using Sniffles version 1.0.10. The algorithm adopts split-read approach and assigns scores to potential SVs based on key factors such as read support, SV type and length [36]. Parameter adjustment is important to ensure optimal calling, therefore testing was performed first for a critical parameter i.e a minimum read support (**Supplementary section 7.4.1**), defined as –min_support flag in Sniffles. It was concluded that, for the 30X coverage data generated by PromethION sequencing runs, it is appropriate and sufficient for a SV to be reported if a minimum read support is 10. We found that a read support below 10 reduces sensitivity and, therefore, is less likely to produce robust variant calls resulting in false-positives. For a minimum length (-l flag in Sniffles), the selection was based on the standard definition of a SV [55] that considers 50 bp as a lower limit. Runs were performed with the following parameters –report_seq, –report_BND, and –genotype; where report_seq retains sequences for SVs detected as insertions and deletions, report_BND enables detection of breaking-end events such as inversions, and genotype estimation is enabled with genotype flag.

For an intersection between the SVs detected in NanoSwe and SMRT alignments, the callsets were split into separate VCF files for each SV type. Bedtools version 2.27.1 [56] was used to identify a total amount of overlaps; -f and -r flags were used to define a fraction for overlap and that overlap must be reciprocal, respectively. In other words, count for an interval is reported if a SV position in the NanoSwe callsets overlaps at least n% of variant position in the SMRT callsets and vice versa.

3.6 Handling Sensitive Data

Since the study involves human subjects, the genomic data used for the analyses comes under a category of personal data. Handling of human genomic data, therefore, was carried out in a secure manner following the ethical conditions of respect for persons provided by the SciLifeLab. The analyses were performed on multi-processor cluster called Bianca which is a research system (designed only for sensitive data) without internet access and direct transfer of files.

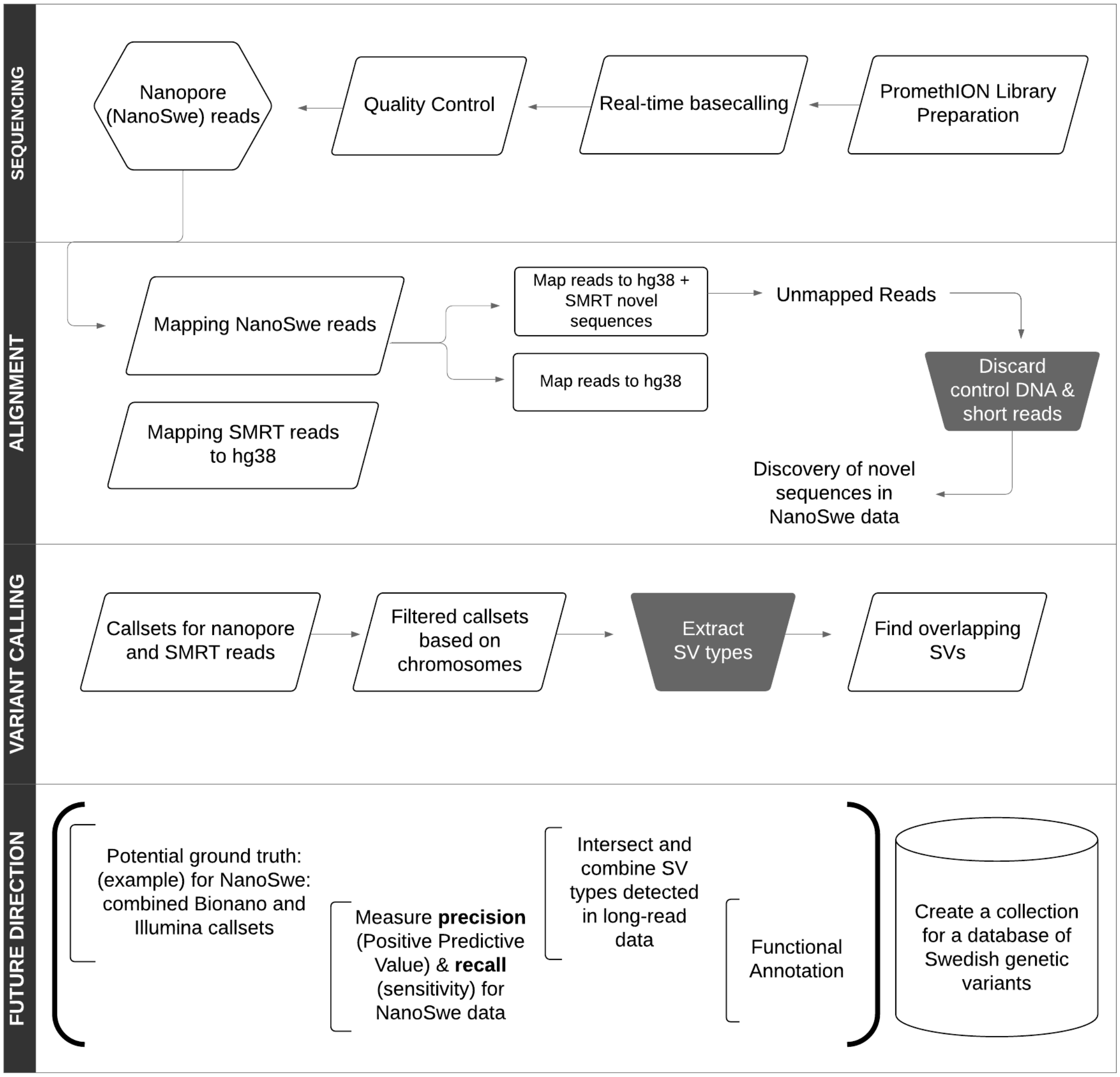
SEQUENCING
Nanopore (NanoSwe) reads
Quality Control
Real-time basecalling
PromethION Library Preparation
ALIGNMENT
Mapping NanoSwe reads
Map reads to hg38 + SMRT novel sequences
Map reads to hg38
Unmapped Reads
Discard control DNA & short reads
Mapping SMRT reads to hg38
Discovery of novel sequences in NanoSwe data
VARIANT CALLING
Callsets for nanopore and SMRT reads
Filtered callsets based on chromosomes
Extract SV types
Find overlapping SVs
FUTURE DIRECTION
Potential ground truth: (example) for NanoSwe: combined Bionano and Illumina callsets
Measure precision (Positive Predictive Value) & recall (sensitivity) for NanoSwe data
Intersect and combine SV types detected in long-read data
Functional Annotation
Create a collection for a database of Swedish genetic variants

Figure 3. NanoSwe Workflow. A process flowchart displaying steps addressed in current study, and potential future work.

4 Results

The research objectives of this study were to assess (i) alignment accuracy of long read data from different platforms, and (ii) SV detection in human genomes. Genomes of two Swedish individuals were sequenced across PromethION flowcells generating average read lengths of 6-8 kb for both female and male genomes. Nanopore reads that passed quality control filtering were used for downstream processing; a nanopore read is considered "passed" if it has a quality score above 7 [57]. Among passed reads, the longest read is derived from a sheared library of a male human genome spanning >1Mb (1,002,249 bp). Among all four libraries (regardless of quality score), the longest read is 1,021,893 bp which again happens to be derived from aforementioned library. In comparison, the study by De Coster *et al* [24] identified the longest read of 177 kb. Details of the individual NanoSwe sequencing runs can be found in **Supplementary section 7.1**.

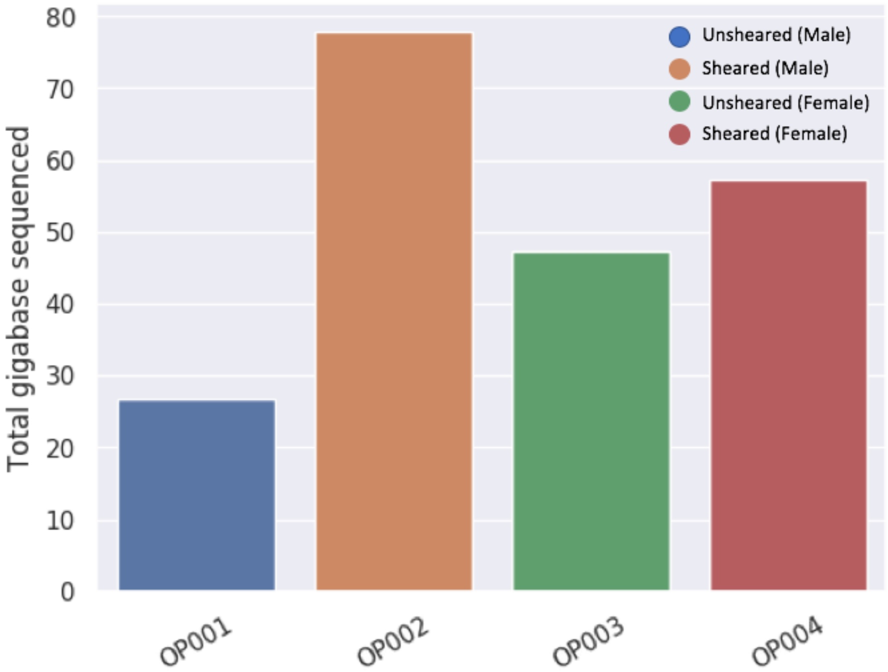

Figure 4. Sequencing Yield: A bar plot displaying PromethION sequencing yield in Gigabases. OP001 and OP002 libraries represent NanoSwe1 (male) whereas OP003 and OP004 libraries represent NanoSwe2 (female). Sheared libraries (OP002 and OP004) have a higher yield as compared to native DNA libraries (OP001 and OP003) and same pattern was observed in terms of coverage across all libraries.

4.1 Alignment Comparisons

In a context of NanoSwe libraries, we found that sheared libraries have a high sequence identity (above 80%) to the reference which reflects the successful sequencing run resulting in a higher coverage. In comparison between native libraries, reads for female human genome show a high alignment rate (76.38%) as compared to the male human genome (70.12%) (**Supplementary Table 7.1 and Table 7.3.1**). The quantity bias could be due to the fact that there are two copies of X-chromosome expressed in female and that there is a high abundance of reads in the X-chromosome annotation of the hg38. Conversely, chromosome Y is highly-repetitive [58, 59] yet there is an unequal representation of repeats in the reference genome particularly in the hg38-alt [60]. The insufficient repeat annotation in the hg38 can lead to a mapping bias. It is also worth mentioning here that nanopore sequencing is prone to error rate due to systematic bias in homopolymeric regions i.e. base-calling signal remains unchanged, resulting in a 10-30%

of genomic region being unmapped [61, 62] which could be another contributing factor here hence the distinction across alignment rate.

We pooled NanoSwe samples (based on sex), merged the data for another round of alignment, and found that the female human genome has a mappability along the same (but slightly higher) rate as the male human genome (Table 1). In contrast, SMRT reads have a higher number of regions aligning to the reference at a rate of 98.33% and 97.88% for Swe1 and Swe2, respectively. This is in line with the results from the Swe-2018 study where Swe1 and Swe2, respectively, displayed a mapping rate of 99.14% and 99.24% (*Section 3.2*, [6]). However, it is worth recalling that, SMRT reads were then aligned to the hg38-alt which suggests that the presence of alternative contigs/haplotypes bring a very little meaningful contribution. Recall that alternative contigs serve as alternative representation for highly variable parts (loci) of the hg38 and therefore generally contribute as a supplementary material. Inclusion of alternative contigs is a subjective matter and should be treated as one, reference genome does not always necessarily have to include alternative contigs for alignment. It is true that addition of decoy sequence can help to 1) reduce multimapped reads (assigning reads to a mapping quality below zero) due regions aligning to alternative contigs and 2) increases specificity (low false positive rate) for SV calling. However, this approach does not necessarily give high-quality output and holds relevance in subjective cases for example if scientific question requires parsing alignment regions for a detection of single nucleotide polymorphisms. To add more weight to this, the Swe-2018 study performed alignment both with and without alternative contigs and found that alternative-contig aware alignment overestimates gain (duplications) and loss (deletions) of SVs. The analyses were performed with BWA-MEM (https://github.com/lh3/bwa) which is a suitable tool for either primary and top-level reference assemblies. This brings the next point that, in the case of our study, Minimap2 was used which is more suitable for primary reference assemblies (Issues 58 and 72, https://github.com/lh3/minimap2).

Table 1. Summary of Alignment Statistics.

Genome	Aligned Reads, %	Forward Strand, %	Reverse Strand, %
[a]NanoSwe1	78.54	60.49	39.50
[a]NanoSwe2	78.79	60.23	39.76
[b]NanoSwe1 > hg38+SMRT	78.65	61.11	38.88
[b]NanoSwe2 > hg38+SMRT	78.89	60.91	39.08
[c]Swe1	98.33	50.95	49.04
[c]Swe2	97.88	51.26	48.73

[a]Pooled data. For alignment results of all samples, see **Supplementary Table 7.3.1**
[b]Statistics for NanoSwe reads mapped to the assembly of hg38 and novel SMRT sequences
[c]SMRT reads

Similar mappability rates were observed for alignment of NanoSwe data to the extended reference assembly (made up of the hg38 and the novel SMRT sequences detected in the Swe-2018 study). Nanopore reads were further investigated for the evaluation of NanoSwe sequences that did not map to the extended version of the reference. A total amount of 6.3 and 7.5 Gb of unmapped reads were found in NanoSwe1 and NanoSwe2, respectively. Majority of the unmapped data from both individuals is made up of short reads (Table 2). In a context of nanopore sequencing, a sequence is placed in a category of long read if it is above and/or equal to 1kb [63, 64] whereas a read is considered ultra-long if it is above 800kb [18]. In the author's view, 1kb is an appropriate minimum threshold for a read to be considered long and, therefore, reads below <1kb should be discarded as to circumvent bias in further analyses.

First, we take into account the phage Lambda that was used as a control DNA in NanoSwe libraries (Section 3.1), and which is expected to remain unaligned to the hg38. Therefore, BLAST analysis was first performed for all unmapped reads to confirm presence of control DNA sequences. Top hits were indeed retained from Lambda phage genome suggesting the unmapped data contains regions highly homologous to it. Subsequently, Lambda sequences were removed with NanoLyse [41]; 601246 reads from NanoSwe1 and 780193 reads from NanoSwe2. Secondly, short reads were then discarded from unmapped reads using a threshold of <1kb. For a final set of unmapped reads ("NanoSwe novel sequences"), a BLAST search against all databases was performed that retained majority of the hits from humans with 5% of the hits from non-human primate species. In addition, as found in the Swe-2018 study, our analysis also retained hits homologous to the flatworm species among which majority (50%) of hits are from species called *Spirometra erinaceieuropaei*.

Table 2. Short-Read Filtration: Extracting short reads from NanoSwe data that unmapped to the assembly of the hg38 and SMRT novel sequences. Values in brackets are written in percentage format.

Genome	Unmapped Reads	Short Reads	Long Reads
NanoSwe1	3,138,656	2,349,872 (74.87 %)	788,784 (25.13 %)
NanoSwe2	3,234,834	2,391,656 (73.93 %)	843,178 (26.07 %)

4.2 SV Detection in Swedish Human Genomes

We called an average of 17,584 and 24,046 variants across NanoSwe and SMRT datasets, respectively. The figures reported here are based on the SVs that passed the quality control filters (**Supplementary section 7.4.2**) i.e. SV has a minimum 10 reads support and has passed a minimum threshold of a mapping quality which is 20. In comparison in a context of SV calling tool performance, previously published study by De Coster *et al* reported 28,305 SVs for a genome sequenced on PromethION at the 59x median coverage. Narrowing down callsets based on each individual genome gives a total of 17,835 SVs for NanoSwe1 and 17,333 for NanoSwe2 whereas 23,837 and 24,256 events were called for SMRT reads in Swe1 and Swe2, respectively (**Supplementary Table 7.4.2**). Callset for each individual was subsampled based on autosomes and sex chromosomes, and the following subsections are mainly focused towards four SV types (Figure 2) present in subsampled data.

4.2.1 Structural Variation in NanoSwe Alignments

Sniffles detected a total of 17,157 and 16,715 SVs, respectively, for NanoSwe1 and NanoSwe2 across a set of chromosomes; this includes a count for translocations and nested events such as inversions flanked by deletions, and inverted duplications that are combinations of main SVs. In a context of type and size, we found that the most abundant type is insertions followed by deletions and majority of the SVs are less than 2kb long. We further evaluated callsets for breakpoint accuracy of all SV types including translocations and nested events; Sniffles determines alignment confidence of a breakpoint position categorising SVs into "precise" and "imprecise" events (*Supplementary Section 2.2.3*, [36]). For NanoSwe1, analysis revealed only 7658 SVs to have a high confidence breakpoint with remaining 9499 events with poor breakpoint accuracy. On the other hand, Sniffles determined 7515 precise and 9200 imprecise breakpoints in NanoSwe2 callset. Among SVs with a precise breakpoint in both individuals, the largest SV is an inversion located at chromosome 10 spanning 9.8 Mb followed by largest duplication of 1 Mb (located at chromosome 1 and Y in female and male genomes, respectively). However, insertions and deletions span more bases with the largest identified insertion being 3014 bp (chromosome 11) in NanoSwe1 and 2423 bp long (chromosome 17) in NanoSwe2. The callsets, subsampled for chromosomes, were further analysed for genotype estimation performed by Sniffles (*Supplementary Section 2.3*, [36]). In both individuals, a large portion of the SVs are heterozygous alleles i.e. events carrying a single copy of each of the reference and alternative alleles whereas less than 30% of SVs are homozygous alternate alleles (where a read is expected to be different from the hg38 read) with a ratio of allele frequency on each chromosome being close to the expected value of 1.

Table 3. SVs Detected in NanoSwe Values represent data filtered for chromosomes.

SV Type	**NanoSwe1** (Nucleotides affected by SV, Mb)	**NanoSwe2** (Nucleotides affected by SV, Mb)
Deletions	7769 (482.2)	7820 (460.7)
Insertions	8746 (2.8)	8369 (2.54)
Duplications	185 (202.0)	147 (168.5)
Inversions	133 (1036.7)	128 (821.2)

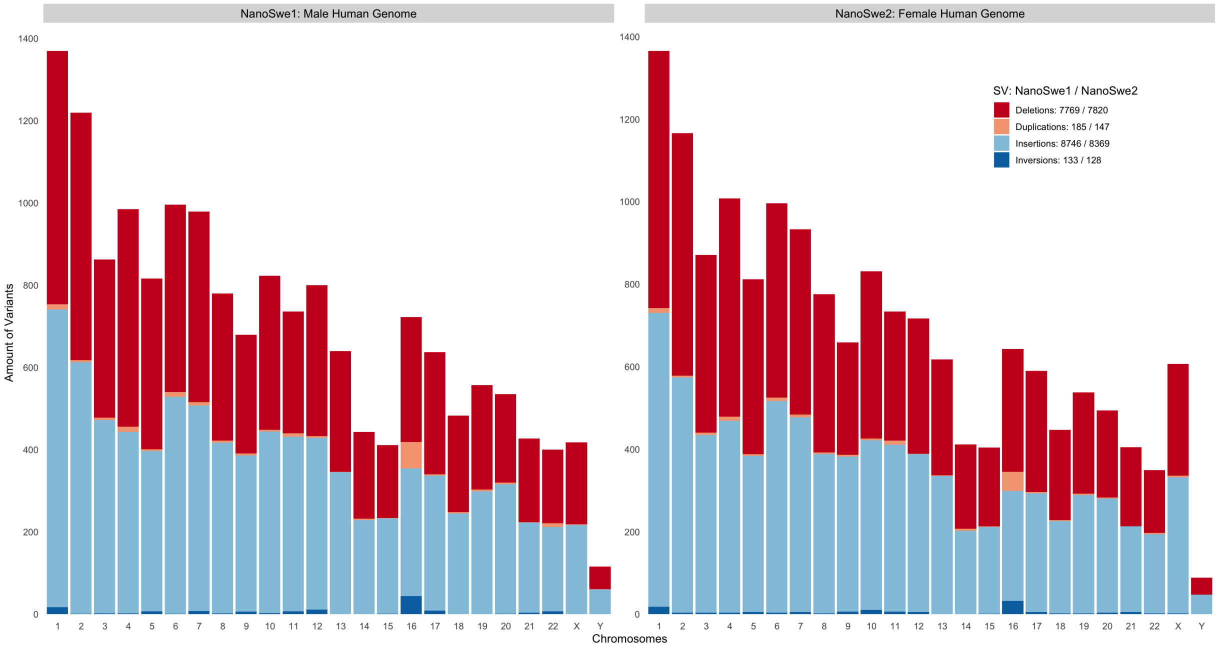

Figure 5. NanoSwe: SVs across Chromosomes Distribution of four SV types across autosomes (chromosomes 1-22) and sex chromosomes.

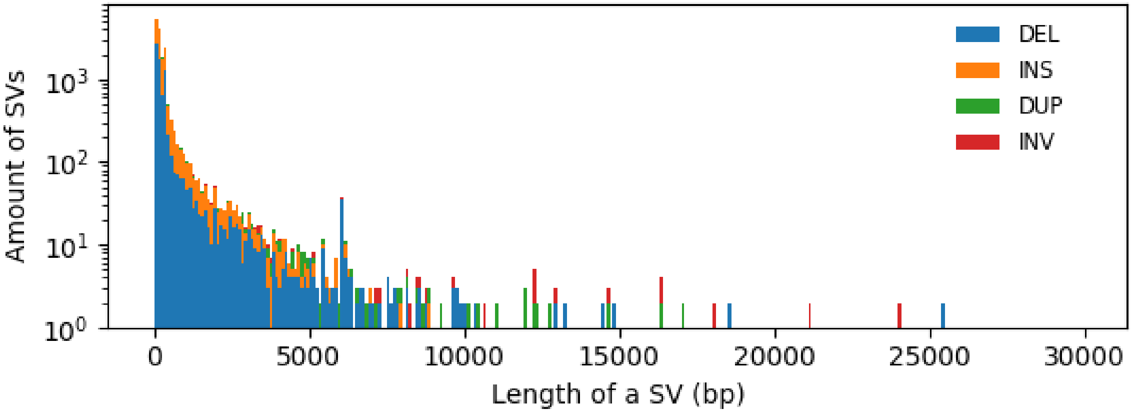

Figure 6. NanoSwe Callset: NanoSwe1. Length profile of SVs identified in NanoSwe1 (male human genome). The x-axis displays length distribution of SVs up to 30 kb with bins of a width of 100 and log-transformed count of variants is displayed on y-axis. The layout of the plot is the same in subsequent figures 7-9 (page 11-12).

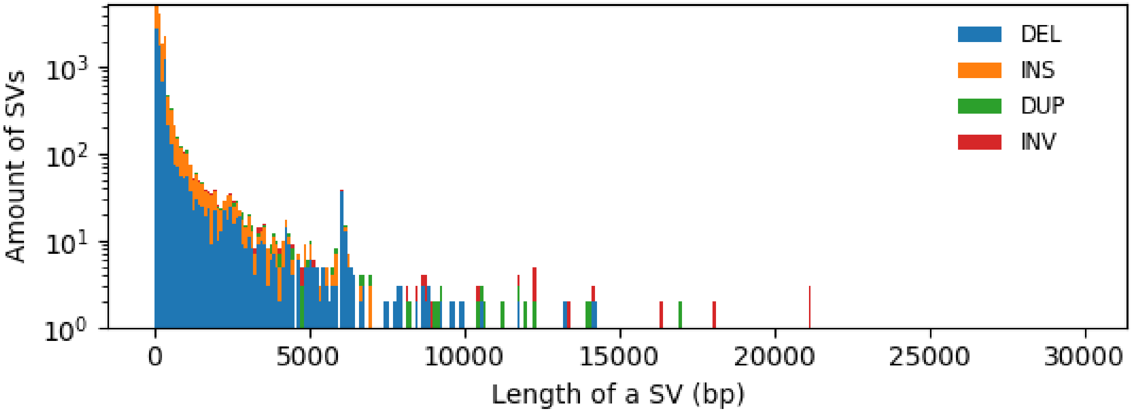

Figure 7. NanoSwe Callset: NanoSwe2. Length profile of SVs identified in NanoSwe2 (female human genome).

4.2.2 Structural Variation in SMRT Alignments

Chromosomes (both autosomes and sex chromosomes) in SMRT reads were found to contain a total of 22,829 and 23,284 SVs, respectively, for Swe1 and Swe2. The count reported here is higher as compared to (not only NanoSwe data but also) the Swe-2018 study where detection resulted in 17,936 SVs for Swe1 and 17,687 SVs for Swe2; the compared data represent four SV types (*Supplementary Table S5*, [6]). Although the SMRT reads analysed for SVs are obtained from the Swe-2018 study, the difference between findings is due to the choice of reference assembly release and the alignment tool used in our study. Inclusion of alternative contigs in the reference, as it is the case in the Swe-2018 study, results in lower count of SVs being called. In other words, representation of alternative contigs in the hg38-alt used in the Swe-2018 study resulted in regions in a subject genome finding alignment on decoy sequences. In terms of alignment tool, the Swe-2018 study performed SV calling following alignment using NGMLR [36]. Sniffles has been tested to perform slightly better (high recall), for example, for insertions after Minimap2 alignment [30]. The study by De Coster *et al* [24], although based on nanopore data, also reported that Sniffles performs better at higher precision and slightly higher recall rate after Minimap2 alignment.

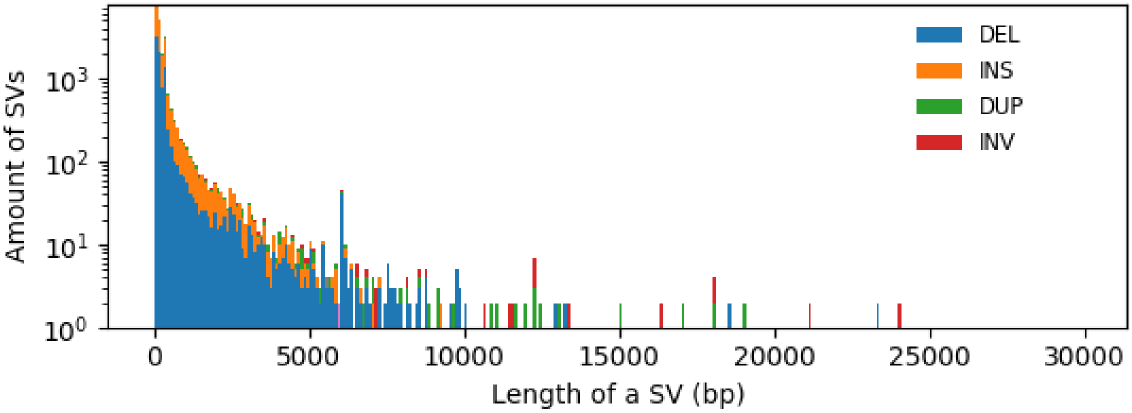

Figure 8. SMRT Callset: Swe1. Length profile of SVs identified in Swe1 (male human genome).

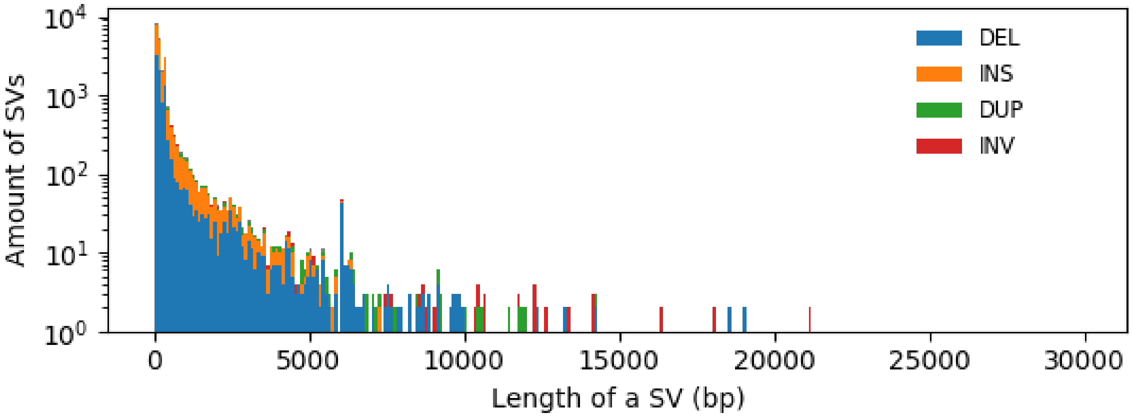

Figure 9. SMRT Callset: Swe2. Length profile of SVs identified in Swe2 (female human genome).

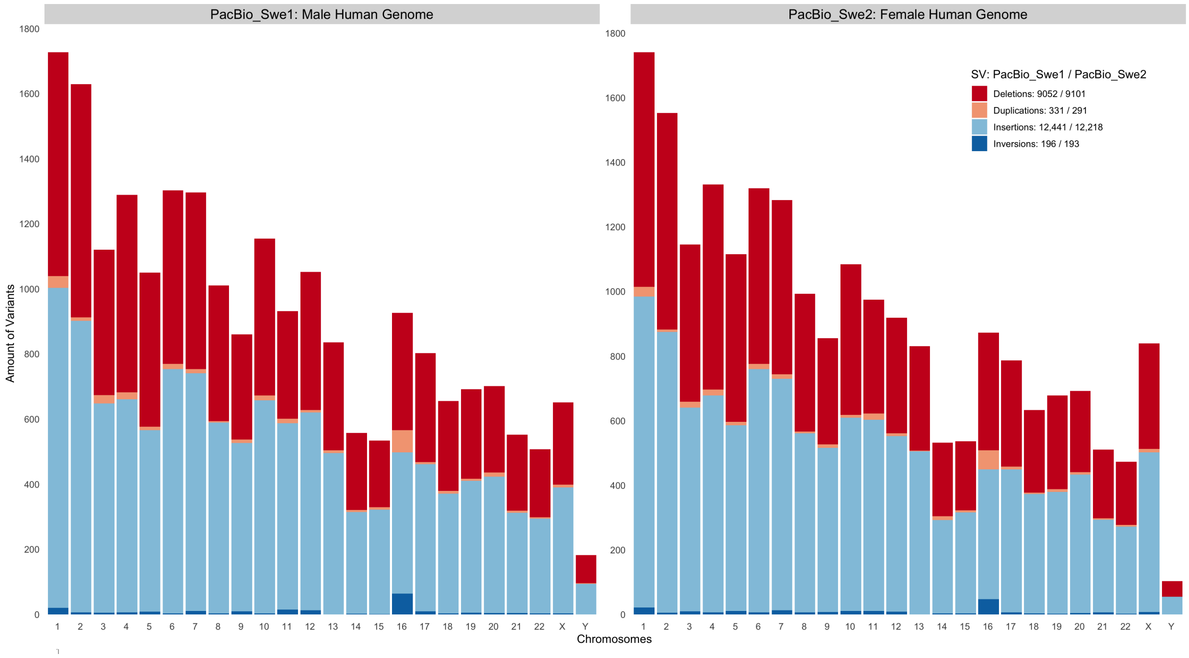

Figure 10. SMRT: SVs across Chromosomes. SVs detected in autosomes and sex chromosomes of two Swedish individuals sequenced on PacBio RS II.

Table 4. SVs Detected in SMRT Alignment: Values represent data filtered for chromosomes.

SV Type	**Swe1** (Nucleotides affected by SV, Mb)	**Swe2** (Nucleotides affected by SV, Mb)
Deletions	9052 (571.0)	9101 (635.2)
Insertions	12441 (3.8)	12218 (1003.6)
Duplications	331 (327.9)	291 (344.4)
Inversions	196 (2396.9)	193 (2405.0)

4.2.3 Comparison of NanoSwe and SMRT Callsets

Across all data, one common pattern is that a large fraction of SVs are located at chromosome 1 among autosomes - which is to be expected since majority of the reads mapped to chromosome 1 in the hg38 and it is the largest chromosome of all. Among sex chromosomes; collectively, over a thousand of SVs are located across X-chromosome in all individuals and again the proportion of SVs here correlate with chromosome size. Intriguingly, in contrast, there are SVs detected in Y-chromosome in female genomes *(NanoSw2: n= 94, and Swe2: n =114)* which could be due to mapping artefacts (Figure 5, page 10 and Figure 10, page 13). It is possible that the aligned regions for these SVs are either not from the Y-chromosome or that presence of repetitive sequences in the Y-chromosome in the hg38 could be resulting in reads being erroneously aligned. This is reflected in the finding that majority (over 60%) of the SVs detected in Y-chromosome in females have ambiguous breakpoint accuracy.

In terms of SV abundance, same pattern (as observed in NanoSwe callsets) was found for insertions being the most abundant type followed by deletions. SVs from both data types were compared for breakpoint accuracy and genotype estimation; similar to observations made for NanoSwe callsets, a large proportion of SVs have imprecise alignment breakpoints in both individuals with a majority of imprecise events being present in female genome. For genotype prediction, Sniffles reported majority of the SVs to be heterozygous alleles whereas 21% of the SVs account for homozygous alternate alleles.

Since a large proportion of SVs in both datasets are identified as indels (Figures 7-10), we investigated the callsets for the length distribution of the indels within the range of 50bp to 10kb in NanoSwe data. A consistent peak of 300bp was observed (Figure 11), and our finding is consonant with the observation made in the Swe-2018 study that found the peak to be denoting enrichment of highly abundant primate-specific *Alu* retrotransposons. NanoSwe callsets were further subsampled for a maximum of 10kb length that revealed a consistent peak made up of deletions around 6.2kb (Figure 12; the Swe-2018 study reported same observation attributing a peak to the presence of Long Interspersed Nuclear Elements (LINEs). These recurrent patterns have also been reported in the recent study on human genome sequencing on PromethION [24].

Given that both NanoSwe and SMRT data are from same individuals, we expect shared events across callsets; the assessment here is focused towards a unique count of a reciprocal overlap spanning 50% of length within SVs identified as indels. Callsets were analysed for intersecting SVs between same sex individuals sequenced from each platform (**Supplementary Figure 7.4.5**). Overlapping variants are referred to as two SVs of different lengths located within a position of a same chromosome. In comparison between insertions detected in female genomes, NanoSwe2 and Swe2 share 26.4% of events (n = *7372*) which is slightly lower as compared to shared insertions in male genomes (26.7%, n = *7720*) which reflects the fact that the mapping process of both SMRT and NanoSwe data to the hg38 is confounded by the repetitive characteristic of the Y-chromosome. In contrast, the amount of shared calls is higher for deletions with a unified SVs accounting for about 33% in female individuals and 30.1% in male individuals. The amount of intersecting SVs drop with a higher fraction set for reciprocal overlap; for example, for 90% overlap, we found that male genomes share 3480 insertions and 5829 deletions whereas female genomes have 3320 insertions and 5996 deletions shared between the callsets.

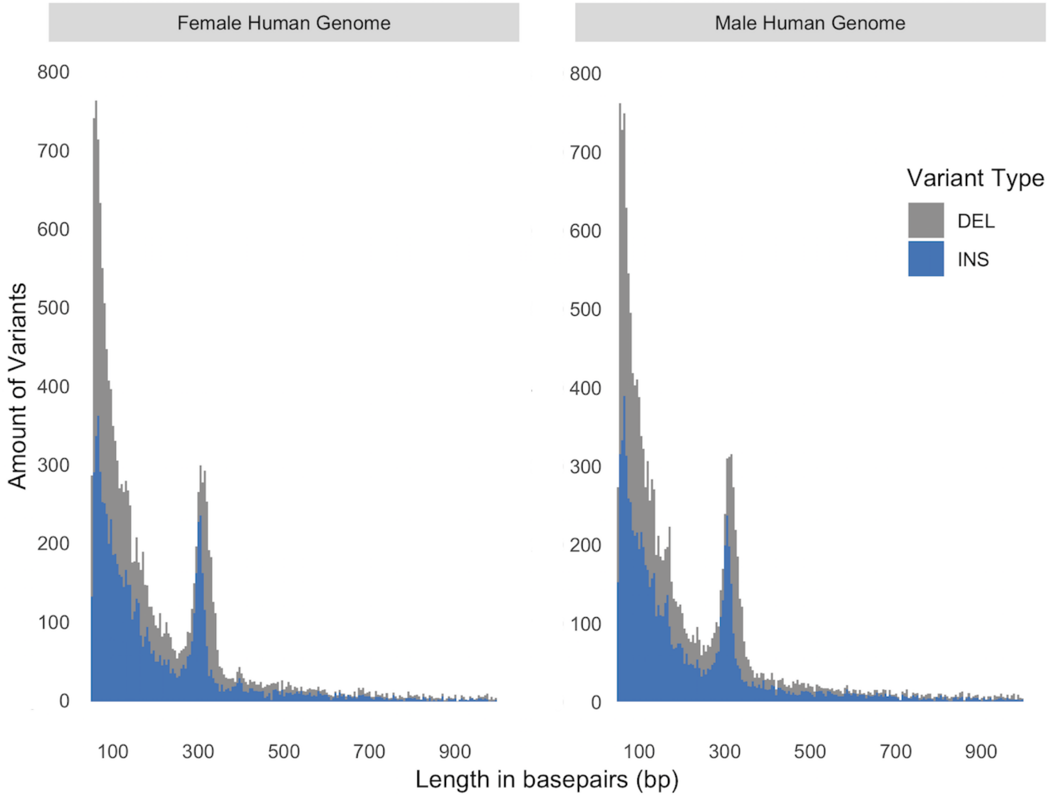

Figure 11. Length profile of indels below1kb. The plot displays a distribution of SVs identified as insertions and deletions in NanoSwe data. The y-axis displays the amount of SVs whereas the SV size is displayed on x-axis with bins of a width of 5.

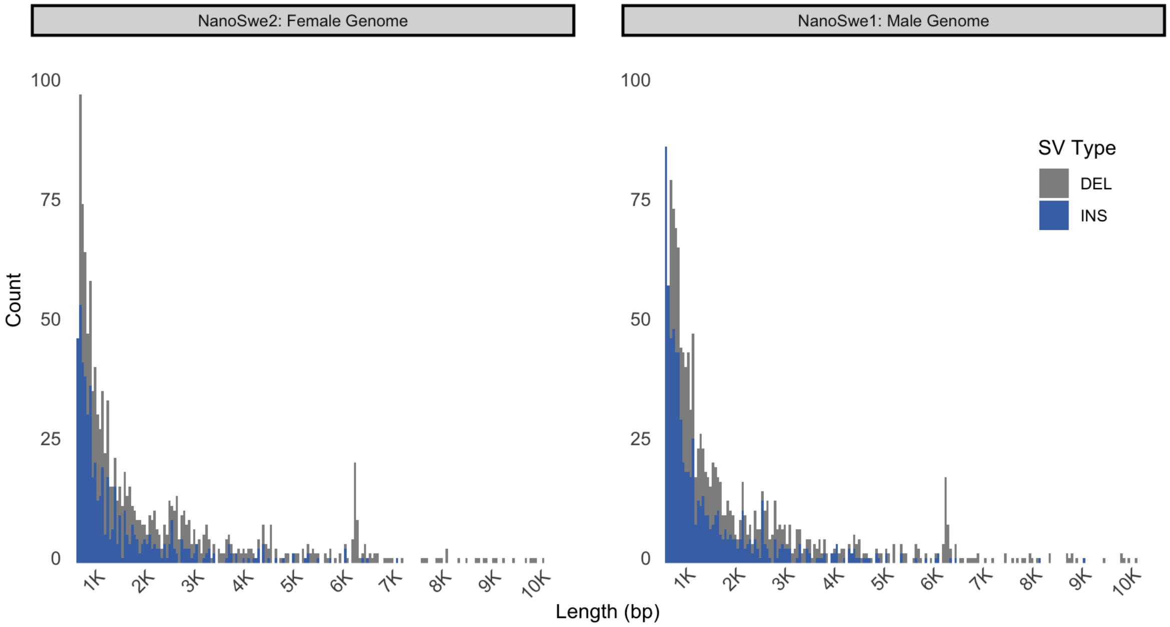

Figure 12. Length profile of indels: 1-10kb. The plot displays length of insertions and deletions detected in NanoSwe data. The x-axis displays the SV size with bins of a width of 50.

5 Discussion

The study reports first successful PromethION sequencing runs of Swedish human genomes. We leveraged long-read data with existing tools to curate a preliminary catalogue of SVs identified in two Swedish individuals. The complexity of genomic data derived from long read technologies indeed poses a number of interesting challenges; the key factors to take into account are error rate and accuracy when it comes to evaluation of platforms. The PacBio RS II gives a higher throughput with a rate of 10-15% accounting for errors [46, 47] that can be circumvented with repeated measurements and computational analyses because errors are 1) stochastic (correlation between depth and quality) and 2) mainly due to thermodynamic conditions [48]. The comparative analyses from this research concludes that SMRT data has a higher sequence identity to the reference owing to the reference-alignment bias, as well as the filtration step (performed in Swe-2018 study) where short reads below 500 bp length were discarded prior to further analyses. In addition, the results reflect the importance of trade-off between error rate and yield run; the key factors weighted less towards ONT platforms where unambiguous mapping cannot be easily achieved when there is a risk of 5-40% of sequencing errors [49]. It would be useful to report equivalent of precision-recall curve here to better understand the varied measurement of error rates across different platforms. Moving further, when it comes to downstream analyses such as alignment process, it is challenging and insufficient to accurately assess mappability by relying on coverage and yield only. The additional key contributors are the alignment-tool bias and reference-alignment bias. This study used Minimap2 and the hg38 for the reference whereas the Swe-2018 study used NGMLR and hg38-alt; the comparative analyses showed the mappability rate is along the same lines with a slightly higher number of aligned bases observed for male genome. To add weight to the alignment results, it would be useful to carry out an assessment of tools based on sensitivity of alignment in a context of errors such as false mappings and precisely aligned regions.

For SV detection, SMRT alignments displayed a higher representation of events. The high amount of SVs is congruent with a high percentage of sequence identity in alignments. Albeit, the results should be interpreted with a grain of salt since the callsets are yet to be evaluated for false positives. Also, since Sniffles searched for majority of the signatures by scanning alignments and the fact that long read alignments can incorrectly classify as SVs, it is very likely that SV detection is biased due to mapping artefacts. The remaining question, regarding how many of these events are true positives, must be addressed with the measurement of precision and recall which brings the need for a ground-truth that could serve as a baseline set. Previous studies have used the high-confidence set from the pilot NA12878 genome [18], from Genome in a Bottle, for evaluation of SV callsets [65, 50]. However, it is incomplete and biased towards non-Swedish human genome. The bottleneck can be circumvented with an ideal approach of building an ultimate biological truth set based on WGS data of Swedish genomes derived from different technologies as similar to the approach shown in previous studies [65, 51, 52, 53]; this is more suitable for our long-term work on Swedish genomes. Given the data availability, a relevant near-future solution for accurate precision and recall measurement could be to build a truth set from, for example, BioNano and Illumina callsets to use it as a benchmarking resource for nanopore data. In a context of reference genome, ideally more efforts should be pushed towards building a regional reference genome [66, 67] so it can be used as a benchmark ('regional gold standard') for sequencing analyses of other individuals from a selected population of said region. Building a local reference would allow to resolve (estimate and validate) population-level genetic variation therefore filling gaps in our current understanding of Swedish genomes. Besides SVs, it would be equally interesting to explore other regions in order to gain insight into genomic architecture of Swedish human genomes. Given that a large proportion of human genome is covered with repetitive elements including repeat arrays and satellites [68, 23], it would be worth looking into abundance of repeat elements across Swedish genomes.

As compared to the SMRT dat, the nanopore data analysed for this study appears to be of a reasonable coverage sufficient enough to detect SVs in human genomes. However, it can be improved in several ways such as by performing post-sequencing correction in a form of consensus calling or polishing of raw data [69]. In addition, the most appropriate way to improve output is to perform re-basecalling of existing raw data with latest algorithm Flappie (`https://github.com/nanoporetech/flappie`) which offers efficient acceleration in terms of read accuracy [70]. Besides improvement in error rate, the long-term goal of using new pore chemistry of R10 would be useful not only to obtain high-quality data but to assess the error rate as compared to the R9 chemistry used for current libraries.

Although PacBio RS II offers high-quality data and allow to detect large amount of SVs, it is costly and time-consuming (although a better choice to perform WGS for *de novo* assembly) as compared to the PromethION platform which is more affordable and time-saving particularly for multiple WGS experiments - the remaining question on how to reduce error-rate and increase read length N50 for nanopore reads is being addressed with improvements in basecalling algorithms and library protocols. The study serves as a stepping-stone for future research work on WGS of Swedish individuals including experimental design of future PromethION sequencing. In conclusion, the results demonstrate throughput variation each platform offers and how sequencing coverage and parameter selection impact downstream analyses. One interesting direction for future work is to compare sequencing data of same individuals obtained from other technologies to explore knowns and unknowns. We anticipate that this study will be useful to researchers for extracting information from their PromethION-based genome data.

6 Code and Data Availability

Sequencing data are available from the National Genomics Infrastructure for researchers who meet the criteria for access to confidential data. The Github repository for the project is available at https://github.com/Nazeeefa/NanoSwe.

7 Acknowledgments

The author would like to thank Adam Ameur for being one of the best mentors. The author would also like to thank Ulf Gyllensten for the research opportunity, Anna Petri for the wet-lab aspects of the study, Lars Feuk for being a great co-supervisor, Carla Giner for helping to appreciate data visualisation, and Christian Tellgren-Roth for thoughtful discussions.

Nanopore sequencing was performed at the Uppsala Genome Centre, part of the National Genomics Infrastructure hosted by the SciLifeLab. Computational work was performed on secured resources provided by the Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under projects b2015225 and sens-2016003.

8 References

1. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet. 2013;14(2):125-38. doi:10.1038/nrg3373

2. Sudmant, P., Rausch, T., Gardner, E., Handsaker, R., Abyzov, A., & Huddleston, J. et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature, 526(7571), 75-81. doi:10.1038/nature15394

3. Hastings, P. J., Lupski, J. R., Rosenberg, S. M. & Ira, G. Mechanisms of change in gene copy number. Nature Rev. Genet. 10, 551–564 (2009).

4. Shi, L., Guo, Y., Dong, C., Huddleston, J., Yang, H., Han, X., et al. (2016). Long-read sequencing and *de novo* assembly of a Chinese genome. Nature communications, 7, 12065. doi:10.1038/ncomms12065

5. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, et al. *De novo* assembly and phasing of a Korean human genome. Nature. 2016 Oct 13;538(7624):243-7.

6. Ameur A, Che H, Martin M, Bunikis I, Dahlberg J, Hoijer I, et al. *De novo* Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. Genes (Basel). 2018;9(10). doi:10.3390/genes9100486

7. Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. Nature Biotechnolgy. 2012 Apr 10;30(4):295-6.

8. Ameur A, Kloosterman WP, Hestand MS. Single-Molecule Sequencing: Towards Clinical Applications. Trends Biotechnol. 2019 Jan;37(1):72-85.

9. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol. 2016 Dec;56(4):394-404.

10. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, *et al.* Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015 Jan 29;517(7536):608-11.

11. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nature Methods. 2015 Aug;12(8):733-5.

12. Miller DE, Staber C, Zeitlinger J, Hawley RS. Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. G3 (Bethesda). 2018. doi:10.1534/g3.118.200160

13. Bainomugisa A, Duarte T, Lavu E, Pandey S, Coulter C, Marais BJ, et al. A complete high-quality MinION nanopore assembly of an extensively drug-resistant *Mycobacterium tuberculosis* Beijing lineage strain identifies novel variation in repetitive PE/PPE gene regions. Microb Genom. 2018;4(7).

14. Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien FA, Swinkels W, et al. Rapid *de novo* assembly of the European eel genome from nanopore sequencing reads. Sci Rep. 2017;7(1):7213.

15. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. Genome Biology. 2015 May 30;16:114.

16. Lee, Isac, *et al.* Simultaneous Profiling of Chromatin Accessibility and Methylation on Human Cell Lines with Nanopore Sequencing. 2018; doi:10.1101/504993.

17. Workman, Rachael E, *et al.* Nanopore Native RNA Sequencing of a Human Poly(A) Transcriptome. 2018. doi:10.1101/459529.

18. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nature Biotechnology. 2018 04;36(4):338-45.

19. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nature Communications. 2017;8(1):1326.

20. Highlights of Clive Brown's Technical Update. 2016. `https://nanoporetech.com/about-us/news/highlights-clive-g-browns-technical-update`

21. The highest throughput yet: PromethION breaks the 7 Terabase mark. 2019. https://nanoporetech.com/about-us/news/highest-throughput-yet-promethion-breaks-7-terabase-mark

22. Nicholls, S., Quick, J., Tang, S., & Loman, N. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. 2018; doi:10.1101/487033

23. Roeck, Arne De, et al. "Accurate Characterization of Expanded Tandem Repeat Length and Sequence through Whole Genome Long-Read Sequencing on PromethION." BioRxiv, 15 Nov. 2018, doi:10.1101/439026.

24. Coster WD, Roeck AD, Pooter TD, Dhert S, Rijk PD, Strazisar M, et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. BioRxiv. 2018; doi:10.1101/434118.

25. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Hum Mol Genet. 2018 08 1;27(R2):R234-R241.

26. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006 Feb;7(2):85-97.

27. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nat Genet. 2004 Sep;36(9):949-51.

28. Norris AL, Workman RE, Fan Y, Eshleman JR, Timp W. Nanopore sequencing detects structural variants in cancer. Cancer Biol & Therapy. 2016;17(3):246-53.

29. Vu, T., Davidson, S., Borgesi, J., Maksudul, M., Jeon, T., & Shim, J. (2017). Piecing together the puzzle: nanopore technology in detection and quantification of cancer biomarkers. RSC Advances, 7(68), 42653-42666. doi:10.1039/c7ra08063h

30. Heller D, Vingron M. SVIM: Structural Variant Identification using Mapped Long Reads. (2019) Bioinformatics. doi:10.1093/bioinformatics/btz041

31. English AC, Salerno WJ, Reid JG. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. BMC Bioinformatics. 2014 Jun 10;15:180.

32. Gong L, Wong CH, Cheng WC, Tjong H, Menghi F, Ngan CY, et al. Picky comprehensively detects high-resolution structural variants in nanopore long reads. Nat Methods. 2018 Jun;15(6):455-60.

33. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 04 15;32(8):1220-2.

34. Eisfeldt J, Pettersson M, Vezzi F, Wincent J, Käller M, Gruselius J, et al. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. PLoS Genet. 2019 02;15(2):e1007858.

35. Eisfeldt, J., Vezzi, F., Olason, P., Nilsson, D., & Lindstrand, A. (2017). TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000research, 6, 664. doi:10.12688/f1000research.11168.1

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

37. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2017 05;27(5):677-85.

38. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al. Assembly and diploid architecture of an individual human genome via single-molecule technologies. Nat Methods. 2015 Aug;12(8):780-6.

39. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nature Reviews Genetics. 2011 May;12(5):363-76.

40. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, et al. SweGen: A whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur J Hum Genet. 2017 11;25(11):1253-60.

41. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics. 2018 Aug 1;34(15):2666-9.

42. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094-100.

43. Roberts, M., Hayes, W., Hunt, B., Mount, S., & Yorke, J. Reducing storage requirements for biological sequence comparison. Bioinformatics, 2004;doi:10.1093/bioinformatics/bth408

44. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 2017 05;27(5):849-64.

45. GRCh38 assembly patch release - `ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/seqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz`

46. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. Nat Biotechnol. 2012 Jul 1;30(7):693-700.

47. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018 Mar 16;46(5):2159-68.

48. Potapov V, Ong JL, Langhorst BW, Bilotti K, Cahoon D, Canton B, et al. A single-molecule sequencing assay for the comprehensive profiling of T4 DNA ligase fidelity and bias during DNA end-joining. Nucleic Acids Res. 2018 Jul 27;46(13):e79.

49. Goldfeder RL, Wall DP, Khoury MJ, Ioannidis JPA, Ashley EA. Human Genome Sequencing at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. Am J Epidemiol. 2017 Oct 15;186(8):1000-9.

50. Bowden, R., Davies, R., Heger, A., Pagnamenta, A., de Cesare, M., & Oikkonen, L. et al. (2019). Sequencing of human genomes with nanopore technology. Nature Communications, 10(1). doi:10.1038/s41467-019-09637-5

51. Parikh H, Mohiyuddin M, Lam HY, Iyer H, Chen D, Pratt M, et al. svclassify: a method to establish benchmark structural variant calls. BMC Genomics. 2016 Jan 16;17:64.

52. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019 04 16;10(1):1784.

53. Fang, L. T., Zhu, B., Zhao, Y., Chen, W., Yang, Z., Kerrigan, L., et al. (2019). Establishing reference samples for detection of somatic mutations and germline variants with NGS technologies. doi:10.1101/625624

54. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, *et al.* BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10:421.

55. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biol. 2010;11(5):R52.

56. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014 Sep 8;47:11.12.1-34.

57. Lu H, Giordano F, Ning Z. Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics. 2016 Oct;14(5):265-79.

58. Kuderna, L., Lizano, E., Julià, E., Gomez-Garrido, J., Serres-Armero, A., & Kuhlwilm, M. et al. Selective single molecule sequencing and assembly of a human Y chromosome of African origin. Nature Communications, 10(1). 2019. doi:10.1038/s41467-018-07885-5

59. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, et al. Linear assembly of a human centromere on the Y chromosome. Nat Biotechnology. 2018 04;36(4):321-3.

60. Miga KH, Eisenhart C, Kent WJ. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. Nucleic Acids Res. 2015 Nov 16;43(20):e133.

61. Sarkozy, P., Jobbágy, Á., & Antal, P. Calling Homopolymer Stretches from Raw Nanopore Reads by Analyzing k-mer Dwell Times. EMBEC & NBC. 2017, 241-244. doi:10.1007/978-981-10-5122-7_61

62. Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, et al. Assessing the performance of the Oxford Nanopore Technologies MinION. Biomol Detect Quantif. 2015 Mar;3:1-8.

63. Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. MinION-based long-read sequencing and assembly extends the. Genome Res. 2018 02;28(2):266-74.

64. Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, et al. Rapid Low-Cost Assembly of the. G3 (Bethesda). 2018 10 3;8(10):3143-54.

65. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016 Jun 7;3:160025.

66. Besenbacher S, Liu S, Izarzugaza JM, Grove J, Belling K, Bork-Jensen J, et al. Novel variation and *de novo* mutation rates in population-wide *de novo* assembled Danish trios. Nat Commun. 2015 Jan 19;6:5969.

67. Fakhro KA, Staudt MR, Ramstetter MD, Robay A, Malek JA, Badii R, et al. The Qatar genome: a population-specific tool for precision medicine in the Middle East. Hum Genome Var. 2016;3:16016.

68. Mitsuhashi, S., Frith, M., Mizuguchi, T., Miyatake, S., Toyota, T., & Adachi, H. *et al.* (2018). Robust detection of tandem repeat expansions from long DNA reads. doi:10.1101/356931

69. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018 07 13;19(1):90.

70. Wick, R., Judd, L., & Holt, K. Performance of neural network basecalling tools for Oxford Nanopore sequencing. 2019;doi:10.1101/543439

7 Supplementary Information

Table 7.1 Read Length Metrics of PromethION sequencing. PromethION (beta) device was used for sequencing for all four flowcells. Yield and mean read length were calculated with NanoComp (https://github.com/wdecoster/nanocomp) after basecalling. Key: Gb = Gigabases, bp = basepairs.

Library	Yield (Gb)	Estimated Genome Coverage	Maximum Length (bp)	Total Number of Reads	Read N50**	Total Number of Bases
OP001 (NanoSwe1)	26.6	8.31x	838,077	4,208,279	26, 471	26,663,122,288
OP002* (NanoSwe1)	77.9	24.34x	1,021,893	9,359,793	14, 311	77,920,585,462
OP003 (NanoSwe2)	47.2	14.75x	679,796	6,805,305	20, 975	47,289,196,719
OP004* (NanoSwe2)	57.18	17.87x	858,408	8,120,176	11, 944	57,185,367,919

**Read length of N50 is half of the total sequenced bases of reads equal or larger than the values mentioned.
*Sheared libraries.

7.2 PromethION Basecalling Quality

All the metrics were obtained post-basecalling using NanoStat *(*https://github.com/wdecoster/nanostat*).*

Table 7.2.1 Quality Score Information.

General summary	OP001	OP002	OP003	OP004
Mean read length	6,335	8,325	6,949	7,042
Mean read quality	6.9	7.8	7.5	7.6
Median read length	1,910	7,201	2,932	5,332
Median read quality	8.7	9.0	9.1	8.8

Table 7.2.2 Quality Thresholds: Number, percentage, and megabases (Mb) of reads above quality cutoffs.

Libraries	>Q7	>Q10
OP001	2606397 (61.9%) 23352.6	1040598 (24.7%) 10101.7
OP002	6744676 (72.1%) 67280.4	2405780 (25.7%) 24647.5
OP003	4784449 (70.3%) 42364.9	1777056 (26.1%) 16687.7
OP004	5758625 (70.9%) 47807.8	1239302 (15.3%) 10242.9

Table 7.2.3. NanoSwe: Highest average quality scores. Top three reads (and their read lengths).

OP001	OP002	OP003	OP004
13.4 (1120)	13.4 (1089)	13.0 (1678)	13.7 (1637)
13.3 (318)	13.2 (588)	12.9 (3980)	12.9 (401)
13.2 (3552)	13.2 (536)	12.9 (450)	12.8 (394)

Table 7.2.4. NanoSwe: Longest reads. Top three long reads (and their mean basecall quality score).

OP001	OP002	OP003	OP004
838077 (4.2)	1021893 (7.0)	679796 (4.8)	858408 (7.6)
784429 (3.2)	1002249 (7.2)	640828 (4.1)	657574 (5.6)
661504 (3.1)	970761 (6.5)	637946 (4.7)	570944 (6.4)

7.3. Alignment Statistics

Table 7.3.1. **Mapping NanoSwe to the GRCh38 (hg38):** Alignment statistics are based on raw data mapped to the reference assembly. Statistics were obtained with samtools v1.9 after each alignment run. The resulting BAM files were later merged to a single file for each individual genome for variant calling.

Metrics	**Libraries**			
	OP001	**OP002**	**OP003**	**OP004**
Total Raw Sequences	4,208,279	9,359,793	6,797,511*	8,120,176
Mapped Reads (secondary alignments) [1]	2,644,132 (816, 743)	7,172,690 (2,199,882)	4,771,239 (1,425,707)	6,120,418 (1,869,187)
Mapped Reads ** (Forward & Reverse), Percentage	70.12 (64.9 & 35.1)	82.15 (58.7 & 41.3)	76.38 (61.6 & 38.4)	80.79 (59.0 & 41.0)
Bases Mapped [2]	23,963,442,045	70,960,805,580	43,007,519,860	50,321,571,655
Unmapped Reads	1,564,147	2,187,103	2,026,272	1,999,758
Runtime (CPU), hours	2.8 (18.6)	7 (48.1)	5 (31.6)	7 (6.5)

* 7794 sequences were filtered out due to being invalid i.e. fastq entries being concatenated together.

[1] In a context of Minimap2 aligner, secondary alignments are alternative aligned reads that are tagged by “272” and “256” flags in SAM format. Primary alignments are the reads with longest alignments.

** Percentage representing a count of reference locations for mapped reads.

[2] Accurate estimate of mapped bases (based on CIGAR format).

Table 7.3.2. **Mapping SMRT Reads to hg38**.

SMRT Libraries	**Total Sequences**	**Average Length (basepairs)**	**Maximum Length (basepairs)**	**Mapped Reads**	**Unmapped Reads**	**Runtime (CPU), hours**
Swe1	26,395,733	8930	55,282	25,729,221	666,512	16.5 (133)
Swe2	26,780,995	8698	73,235	25,912,828	868,167	16.8 (135)

7.3.3 NanoSwe Data and Reference Assembly of hg38 + SMRT novel sequences. FASTQ files for sheared and natives PromethION libraries of each individual were combined.

PromethION Libraries	Total Sequences	Average Length (basepairs)	Maximum Length (basepairs)	Mapped Reads	Unmapped Reads	Runtime (CPU), hours
NanoSwe1	13,568,072	7708	1,021,893	9,828,170	3,739,902	~10 (66.6)
NanoSwe2	14,917,687	6999	858, 408	10,902,660	4,015,027	11 (65.5)

7.4. Variant Calling

7.4.1 Parameter Testing

Table 7.4.1.1 Variant calling with a test for read support parameter variation. Three different values for supporting reads (2, 5, and 10) were tested, whereas the minimum length for a variant event to be reported was set to 30 bp (default). Only one sample (OP001) used for testing.

SV Type	Read support: 2	Read support: 5	Read support: 10
Insertions (INS)	11, 641	5051	1045
Duplications (DUP)	398	97	27
Duplication Insertion (DUP/INS)	10	0	0
Inverted Duplication (INVDUP)	63	1	0
Deletion (DEL)	22, 528	5835	1255
Inversion (INV)	283	77	18
Deletion/Inversion (DEL/INV)	12	6	3
Translocation (TRA)	0	0	69
Total	34, 935	11, 067	2, 417

Table 7.4.1.2. Assessment of SV calls across sheared and native libraries. Test was performed with a minimum length of 50 bp and a minimum supporting read of 10. Testing was performed for native/unsheared (OP001) and sheared (OP002) libraries of NanoSwe1 male human genome.

SV Type	OP001	OP002
Insertions (INS)	794	6955
Duplications (DUP)	29	152
Duplication Insertion (DUP/INS)	0	2
Inverted Duplication (INVDUP)	0	0
Deletion (DEL)	807	6472
Inversion (INV)	19	97
Deletion/Inversion (DEL/INV)	1	5
Translocation (TRA)	69	284
Total	1,719	13, 967

Table 7.4.2 SV Calling Results: NanoSwe and SMRT Data. The events were called with Sniffles version 1.0.10 (https://github.com/fritzsedlazeck/Sniffles).

SV Type	NanoSwe1	NanoSwe2	Swe1	Swe2
Insertions (INS)	9027	8608	12, 834	12, 615
Deletion (DEL)	8078	8133	9495	9531
Duplications (DUP)	209	163	374	330
Inversion (INV)	138	132	211	202
Translocation (TRA)	377	292	897	801
Duplication Insertion (DUP/INS)	1	0	4	6
Inverted Duplication (INVDUP)	1	1	13	761
Deletion/Inversion (DEL/INV)	4	4	9	10
Total	17,835	17,333	23, 837	24, 256
Run Time (CPU), hours	2.4 (6.1)	2.3 (6)	6.5 (15.9)	6 (15.5)

7.4.3 Length profiles of SVs: Raw NanoSwe Data

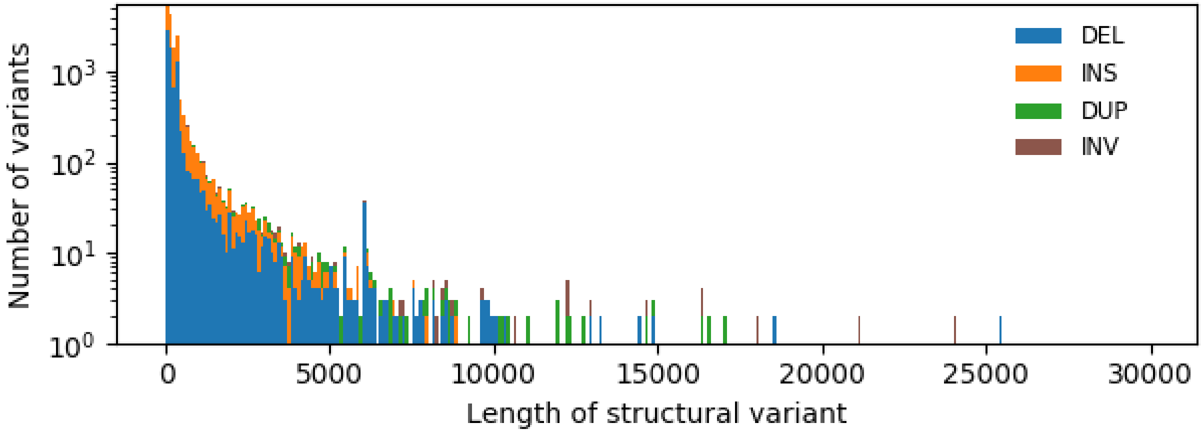

Figure 7.4.3.1. Length Distribution of SVs in NanoSwe1 (male individual). The x-axis displays length distribution of SVs in basepairs up to 30 kilobases, and the log-transformed count of variants is displayed on y-axis. The layout of the plot is the same in subsequent figures (page 28-29).

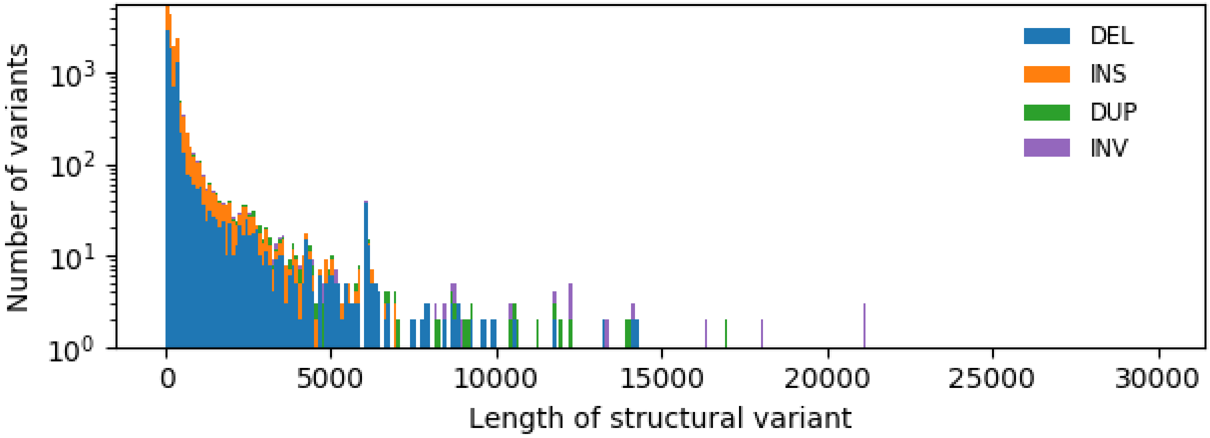

Figure 7.4.3.2. Length Distribution of SVs in NanoSwe2 (female individual).

7.4.4 Length profiles of SVs: Raw SMRT Data

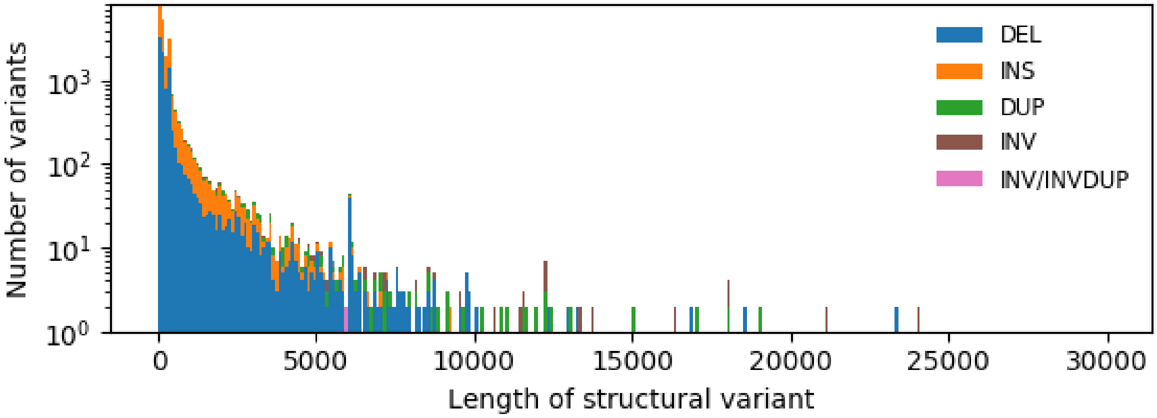

Figure 7.4.4.1. Length Distribution of SVs in Swe1 (male individual).

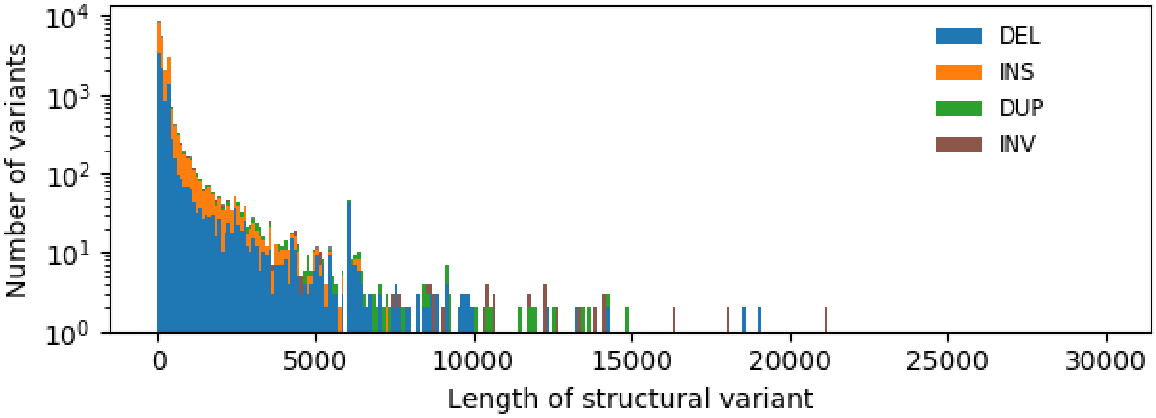

Figure 7.4.4.2. Length Distribution of SVs in Swe2 (female individual).

7.4.5 NanoSwe and SMRT: Variant Overlaps

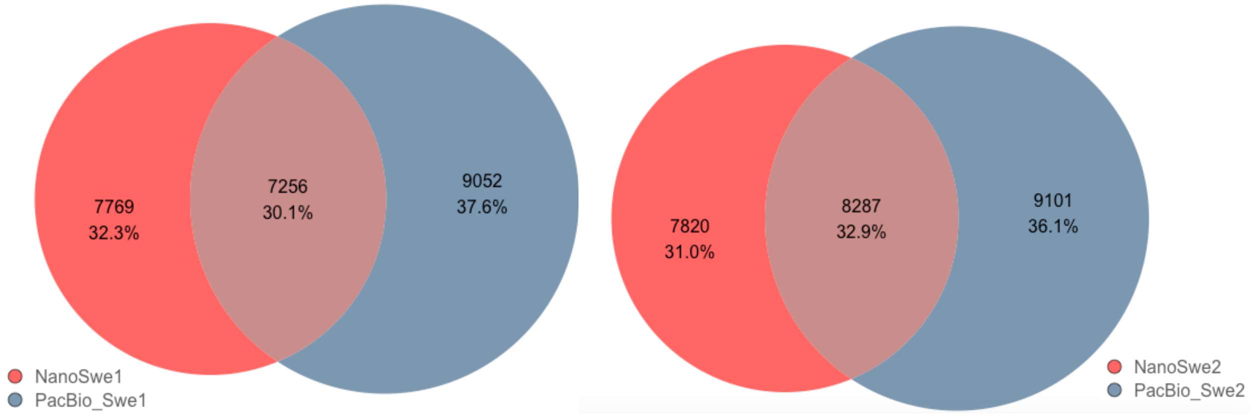

Figure 7.4.5 Unified Callset of Deletions.

Reciprocal overlap of (>=50bp) deletions detected in male and female individuals across NanoSwe and PacBio callsets. The figure is shown for visualization purpose only, it was produced in R version 3.5.1 with ggplot2 [https://github.com/tidyverse/ggplot2] and eulerr [https://github.com/jolars/eulerr] packages. The script is available at https://github.com/Nazeeefa/NanoSwe.