Summer 2022

Comparison of an Oxford Nanopore Technologies Sequencing Platform to Existing Sequencing Methods for Differential Expression Studies

Nikola Klier
San Jose State University

Comparison of an Oxford Nanopore Technologies Sequencing Platform to Existing Sequencing Methods for Differential Expression Studies

A Research Project

Submitted in Partial Fulfillment of the Requirements for the

Master's Degree in

BIOINFORMATICS

Presented to

The Faculty of the Department of Computer Science and Department of Biological Sciences

San José State University

By

Nikola Klier

July 2022

ABSTRACT

As the genomics revolution continues, there is constant pressure to make sequencing technology more accessible and practical for a growing series of applications. Existing sequencing technologies are often prohibitively expensive, limiting their use for novel diagnostic and research applications. Additionally, existing technologies are often limited by short read lengths, which may present problems to certain quantitative sequencing applications. One such application is Differential Expression Analysis, in which RNA-Seq is performed in paired samples under different experimental conditions to identify differences in gene expression. In this study, an Oxford Nanopore Technologies sequencing platform was used to conduct a differential expression study to identify Notch targets. Notch is a transcription factor that regulates numerous functions related to cellular growth and development, and misregulations in the Notch pathway can lead to developmental disorders and cancer. Nanopore Sequencing offers a cheaper and potentially more effective way to conduct research on Notch-mediated expression. It was found that while nanopore sequencing offers a cheaper alternative to existing methods, additional development of the technology is required to perform at the same level as current research standard platforms in differential expression.

ACKNOWLEDGEMENTS

Thanks go out to my mentors and supporters over the course of this project:

- Dr. Brandon White, for the support, advising, and guidance as my research advisor.
- Dr. Wendy Lee, for comprehensive guidance on the computational aspects of bioinformatics.
- Dr. Philip Heller, for providing computational advice and for serving on my committee.
- The Student Research Grant Initiative from the Division of Research and Innovation at SJSU, for generously providing financial support for this project.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Modern sequencing technologies

1.1.1 Sequencing by Synthesis

The past few decades have seen a shift in how biology is studied. New, high-throughput nucleic acid sequencing technologies are able to sequence entire genomes and transcriptomes for routine scientific experiments. This has revolutionized cancer research [1][2], disease research, and other basic science applications [3][4], such as the complete human genome project [5]. Refining and improving existing sequencing methods remains an active area of research.

Second generation high-throughput sequencing methods predominantly work via sequencing by synthesis (SBS) mechanisms. In SBS, the nucleic acids to be sequenced are fragmented. These fragments are replicated, and reads are obtained as new nucleotides are integrated into the new fragments. In ion torrent sequencing [6], four nucleotides are cycled through the sequencing chamber. If the current nucleotide is integrated into the growing nucleic acid chain, it will produce hydrogen ions, producing a shift in the pH of the surrounding environment. This shift is detected and recorded as the next nucleotide in the chain, and the free nucleotides are flushed out to prepare for the introduction of the next nucleotide to be tested.

Similarly, Illumina sequencing uses chain-terminating, nucleotide analogs with an attached fluorescent group known as a fluorophore. Sequencing reads are generated as replication occurs. Polymerase replication of the template strand integrates these nucleotides into the copy, which is then imaged. Each nucleotide is represented by a unique color, so this image will represent the exact nucleotide integrated into the strand. After the image is taken, the fluorophore is removed, and replication proceeds to sequence the next nucleotide. Both technologies have become a research standard due to their high accuracy and throughput. The high redundancy of information contributes to this accuracy. In both cases, sequencing data is

taken on groups of identical strands that have been amplified immediately prior to sequencing. Each group represents a small fragment of the original sample, but both technologies allow all of these groups to be sequenced in parallel, dramatically increasing the throughput of these sequencing platforms [7][8].

Some differences exist between the two platforms. Most notably, Illumina sequencing allows for paired-end reads. By sequencing the complement of a fragment from the reverse end, accuracy in Illumina sequencing is increased. Ion torrent generally has more inconsistent read lengths than Illumina, whereas Illumina maintains a consistent read length. Practically, however, these technologies perform similarly for quantitative applications [9].

Widespread adoption of these platforms has permitted the development of many sequencing techniques, such as whole genome sequencing, metagenomic analysis, and quantitative resequencing techniques such as differential expression (DE) analysis. In DE analysis, RNA Sequencing (RNA-Seq) is performed on multiple experimental samples that differ. In RNA-Seq, RNA is captured from a biological sample, converted to cDNA, and sequenced. In differential expression, these may be a drug treated set of culture cells vs a control set of cultured cells, a genetically distinct population of organisms vs a control population, or any other situation in which the effects of a particular biological variable's effect on gene expression is being studied [9]. For every gene in the RNA-Seq dataset, genes that are expressed at statistically significantly different levels can be identified. DE studies strongly benefit from the high precision and throughput of SBS platforms. High throughput ensures that a significant portion of the sample is sequenced, reducing experimental variation, and high accuracy ensures that all of the sequenced reads can be accurately mapped to their corresponding transcripts for quantitation [10].

SBS, however, has several key drawbacks. SBS is dependent on DNA polymerases to synthesize new strands as part of the sequencing process. Most polymerases, however, have a limit in the length of new strand that is produced. When this limit is reached, the read cannot continue, limiting the size of each fragment being sequenced. With modern technologies, this limit is typically between 50bp and 300bp per fragment [11]. Most applications require sequencing of fragments much larger than this. Human mRNA typically averages 3kb in length, and genomic sequencing requires reads that cover large lengths of a chromosome. Several analysis techniques have been developed to address these issues. In de novo sequencing, where no reference sequence is available, overlaps between reads must be detected computationally. These reads are then assembled into a complete sequence [12]. The process of sequence assembly is computationally intensive, and often requires context specific changes to existing methods to be as accurate as possible [13]. Even after optimization, assembly still introduces potential error to sequencing experiments in particular situations, such as high error rate experiments or repetitive regions [14][15]. In resequencing applications such as DE, assembly is not necessary. Instead, reads that map to a particular reference are quantitated, but must then be normalized to allow comparison between genes and reference regions of different lengths [10]. Both methods are imperfect, and improving the computational basis of assembly and read quantitation is an area of active research. Additionally, the chemical environment and sensors required to facilitate SBS require large, expensive instrumentation, which may be prohibitive to use in fieldwork environments or any time that cost is prohibitive. Finally, since SBS does not directly sequence the original strand, epigenetic modifications cannot be directly detected, necessitating additional sample preparation steps [16].

1.1.2 Nanopore Sequencing

To address these shortcomings, novel ultralong sequencing techniques have been developed. PacBio sequencing has found numerous applications and offers long read length, but still suffers from high instrument cost and low accuracy [17]. To address the issues of traditional second-generation sequencing platforms, and provide an alternative to PacBio sequencing, Oxford Nanopore Technologies (ONT) developed a nanopore sequencing platform.

In ONT sequencing, nucleic acid molecules are drawn through a protein nanopore embedded in a membrane. An electric current is run across the membrane and through the pore. As different individual nucleotides are drawn through the pore, the electrical resistance of the pore changes dependent on the individual nucleotide currently in the pore, as well as surrounding nucleotides and epigenetic modifications of the nucleotides [18]. To generate sequence data from this electrical trace, machine learning algorithms such as guppy are used [19]. These algorithms are trained on known sequences, and correlate noisy electrical signals to a series of nucleotides [19].

ONT sequencing offers numerous benefits when compared to SBS. Most notably, there are no theoretical limits to the length of a single nucleic acid read on a Nanopore platform, and reads over 2MB in length have been recorded. In applications such as DE analysis, this allows for full transcripts to be sequenced, alleviating the need for assembly and providing more direct quantitation. Long reads also allow regions of the genome that were previously unsequenced to be sequenced and assembled, such as repetitive regions in centromeres and telomeres [20].

ONT devices have their own set of drawbacks, however. ONT sequencing platforms offer a 90% to 95% single nucleotide read accuracy, compared to Illumina’s 99.9% accuracy rate [21]. This problem can be alleviated in multiple ways. In resequencing applications, the length of the

overall read provides enough matches to accurately map a sequence to reference sequences [22]. In de novo sequencing and single nucleotide polymorphism identification, additional, redundant data must be used to form a consensus sequence [23]. This data may be supplied by additional nanopore runs, or from targeted sequencing runs performed on SBS devices in a process known as hybrid assembly [24].

The throughput of ONT devices is also limited by the lifespan and fuel availability of the pores themselves. Translocation of nucleic acids through a pore is dependent on ATP provided by added buffers, as well as the maintenance of the pore structure itself. As the sequencing run continues, ATP is consumed and pores lose their stability. Adding more ATP during sequencing is possible, however, limitations in the fluid capacity of the flow cell and structural changes in the protein pores eventually result in pore inactivation. As such, the number of available pores for sequencing decreases over time. As the number of pores decreases, the rate of data collection does as well. When the number of available pores decreases to zero, no more data can be collected from the sample. Pores can remain active for up to 72 hours, however, there is much variation in pore lifespan, often resulting in sequencing runs ending before this [21].

ONT seeks to alleviate this problem by offering platforms with additional pores. Flongle and MinION devices offer cheap upfront costs for low volumes of sequencing data, however, their flow cells only contain 126 and 512 pore channels respectively[25]. The PromethION sequencer offers 2975 pore channels per flow cell and can run up to 48 flow cells in parallel, however, the device itself is much more expensive than other ONT sequencers (Table 1). The total throughput of the Promethion, however, can reach a scale comparable to Illumina sequencing. While variable, 50-150GB of total PromethION reads have been recorded from a single flow cell when sufficient material is sequenced [26][18].

Platform	Instrument Cost	Cost per GB (Approximate)	Theoretical Maximum Number of Available Pores per Flow Cell (ONT Devices)
SBS Platforms			-
Illumina MiSeq	$128,000	$502	-
Illumina HiSeq 2000	$654,000	$40	-
Ion Torrent PGM	$80,000	$1000	-
ONT Devices			
Flongle	$2,500	$50-$1000	126
MinION	$1,000	$50-$1000	2048 (512 channels)
PromethION	$10,000	$20-$40	12000 (2675 channels)
Other long-read devices			
PacBio RS	$695,000	$40-$80	-

Table 1. Cost comparison of sequencing platforms. Initial instrument cost may vary dependent on manufacturer availability. Cost per GB varies dependent on technical variation of the experiment and total data to be collected per sample. In particular, Flongle and MinION sequencing cost is highly dependent on total data acquisition, which in turn is dependent on technical factors such as reagent quality and library concentration [27][28].

1.2 Application of Nanopore Sequencing to Differential Expression

1.2.1 Adapting SBS Computational Tools to Nanopore Sequencing

As discussed previously, differential expression studies are well established on SBS platforms, however, they remain an emerging technology on newer platforms such as nanopore sequencing. In theory, ONT sequencers offer full transcript coverage and direct RNA sequencing, both of which reduce the complexity of analysis in DE studies. Current differential expression methodologies for ONT devices closely parallel those used for SBS platforms [29][30][31]. The differences between the traditional SBS reads and ONT reads, however, mean that some key factors must be considered.

One such factor is mapping quality of nanopore reads to a reference genome or transcriptome. Current ONT settings for minimap2 consider both the long-read length and low

single nucleotide accuracy of nanopore reads [29]. Single nucleotide mismatches between the read and a reference are penalized less, however, longer reads mean that more consensus of the overall sequence is required to match a read to a reference. The accuracy that long read mapping can provide despite low single nucleotide accuracy extends the utility of nanopore sequencing in resequencing applications such as differential expression [32].

Aside from scoring during alignment, long reads present additional challenges for mapping software. Current read quantitation software, such as RsubRead, expects an alignment file generated using a reference genome as opposed to a transcriptome. Genes are then quantitated based on alignment position using a separate gene annotation file. This allows certain information, such as chromosomal location of mapped genes, to be preserved [32]. Currently, recommended DE pipelines for ONT devices use a genomic reference set for alignment, as a genomic alignment paired with an annotation file describing gene locations should also properly match genes. These pipelines use Salmon to quantify based on Transcripts per Million (TPM), a normalization method used to account for short read data that does not cover a whole transcript [29]. Some studies have instead used transcriptomic reference data during mapping for nanopore sequencing of RNA in eukaryotes [31]. Longer nanopore reads are more likely to span multiple introns, and also more completely represent a splice variant, indicating that genomic mapping techniques used in SBS may not be entirely applicable [31][33]. This disagreement indicates that refining mapping strategies in nanopore sequencing pipelines remains an open area of research.

1.2.2 Application to Notch Signaling

The Notch family of transcription factors is responsible for regulating expression of a variety of genes involved in cellular growth and development. As such, misregulations in Notch

signaling are implicated in a variety of developmental disorders and cancers, making it a critical active area of research [34].

Normally, Notch is a transmembrane protein. When it binds to an extracellular ligand, however, the gamma-secretase enzyme complex cleaves the intracellular domain from the rest of the protein. From here, Notch translocates to the nucleus, where it forms a complex with other transcription factors such as MAML1 (Mastermind-like 1) and CSL (CBF1, suppressor of hairless, Lag-1). This complex drives transcription of downstream genes by binding to their promoters or enhancers at a consensus sequence known as a Notch Response Element (NRE). NREs may appear individually, or may also have a second, reverse complement NRE 12-17 nucleotides downstream of first NRE. This head-head configuration is known as a Sequence paired Site (SPS), which recruits a similarly dimeric form of the Notch Transcriptional complex [35].

This pathway is well characterized in canonical Notch targets such as Hes family genes, however, there are still many questions in Notch research. Notably, the exact role of dimeric Notch binding in driving transcription remains unclear. Impaired Notch dimerization is associated with a loss of expression in downstream Notch targets, suggesting that dimeric Notch in either the enhancer or the promoter of a Notch target gene is required for robust activation of transcription. These results, however, are largely based on canonical Notch targets, such as HES1 and HES5, and gaining a whole genome perspective on Notch-mediated gene expression remains an active area of research [36].

Since it is a newer technology than other sequencing platforms, nanopore sequencing has yet to be used for various applications that have already been studied on an Illumina platform. The benefits of an ONT platform discussed above mean that new insights may be gained from

applying nanopore sequencing to these scientific questions. Comparisons between platforms are necessary to determine the effectiveness of nanopore sequencing towards these scientific questions, however. In this study, we used an ONT MinION to identify Notch target genes with a differential expression workflow. These data were compared to data produced using an Illumina platform to assess the effectiveness of nanopore sequencing in this context.

2 Methods

2.1 cDNA library preparation from cultured T-ALL cells

Library preparation is summarized in Figure 1, with detailed methodology following.

2.1.1 Cell Culture, Treatment, and double-stranded complementary DNA (dscDNA) preparation

SUPT1 T-cell acute lymphoblastic leukemia (T-ALL) cells were used due to the constitutively active Notch signaling in T-ALL cell lines [37]. Cells were grown in Roswell Park Memorial Institute (RPMI) media supplied from Sigma-Aldrich with 10% Fetal Bovine Serum (FBS) and 1% antibiotic/antimycotic solution in a 37ºC incubator with a 5% CO_2 atmosphere. To study Notch signaling, the Gamma-secretase inhibitor DAPT was used to prevent NOTCH1 cleavage at the cell membrane, preventing downstream Notch-mediated transcription [38]. Genes found to be downregulated by DAPT during analysis are therefore identified as Notch targets. Experimental SUPT1 cells were treated with 5µM of DAPT for 24 hours. Control SUPT1 cells were treated with 5µM of DMSO for 24 hours. 10^6 cells were treated per sample. Total RNA extraction was performed using the recommended protocol for the Invitrogen TRIzol Reagent (Cat# 15596026). Five paired treatments of DAPT and DMSO were performed.

Double-stranded blunt-end cDNA was generated from extracted mRNA using Thermo Scientific Maxima H Minus Double-Stranded cDNA Synthesis Kit (Cat# K2561). Poly-A Plus mRNA was purified from 1 microgram of total RNA. dscDNA generation followed the recommended manufacturer's protocols for synthesis using oligo-dt primers with the purified polyA+ mRNA. Purified polyA+ mRNA was not quantitated.

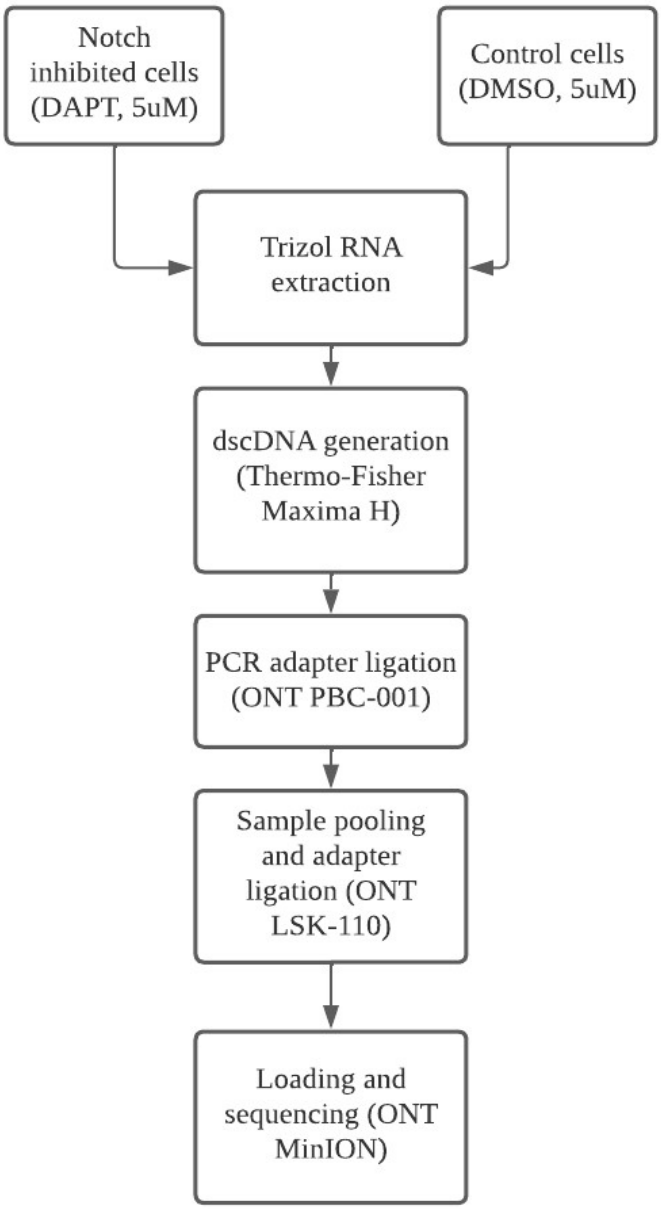


Figure 1. Wet lab workflow was consistent across 5 paired trials. SUPT1 T-ALL cells were treated with the Gamma Secretase Inhibitor DAPT to inhibit Notch mediated transcription. Genes that lost expression when compared to a DMSO control represent Notch target genes.

2.1.2 Oxford Nanopore Technologies Sequencing Library preparation

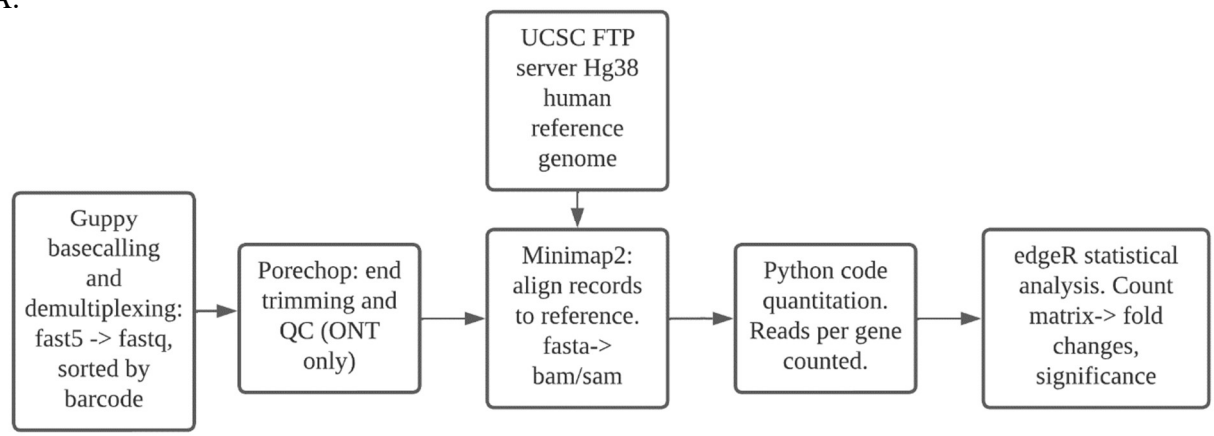
Blunt end double stranded cDNA is functionally similar to blunt end gDNA fragments, and as such, gDNA library preparation protocols may be used. Sequencing library preparation was performed using the ONT Ligation Sequencing Kit (Cat# SQK-LSK110). Additionally, samples were barcoded using the ONT PCR Barcoding Expansion 1-12 (Cat#EXP-PBC001). In barcoding, each sample is ligated with a unique sequence. This sequence is then used to PCR amplify each sample. Multiple samples can then be pooled together into a single library and demultiplexed during analysis using the known barcode sequences. Library preparation was performed according to recommended ONT protocols for the Ligation Sequencing Kit with the added PCR Barcoding Expansion[39].

Samples were loaded onto an ONT R9.4.1 MinION flow cell (FLO-MIN106D) for sequencing on a MinION device. Assuming an average mRNA length of 3.4kb [36], 40 fmol of each sample was loaded onto a flow cell. cDNA libraries were quantified immediately prior to sequencing using the Promega QuantiFlour dsDNA System (Cat# E2671). Data was collected until the active pore count reached 0. 200µL of ATP containing ONT Flush Buffer (Cat#EXP-FLP002) was added to each flow cell 12 to 13 hours prior to the start of each sequencing run to extend active pore lifespan. Paired samples E1, E3, and E4 were sequenced simultaneously on the same flow cell. Samples E5 and E6 were sequenced together on a separate flow cell from the other three paired trials due to limitations in total load capacity of MinION flow cells.

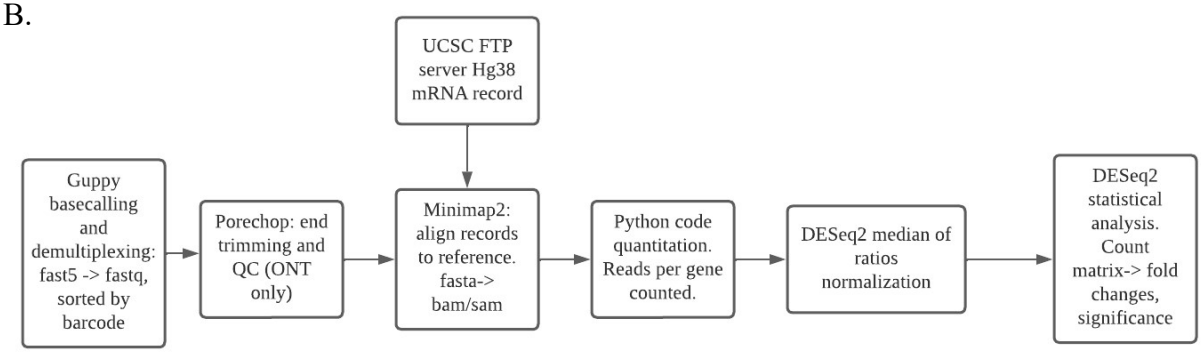
2.2 Differential Expression Analysis of RNA-Seq data

Computational pipelines used are summarized in Figure 2, with detailed methodology following.

A.



B.



C.

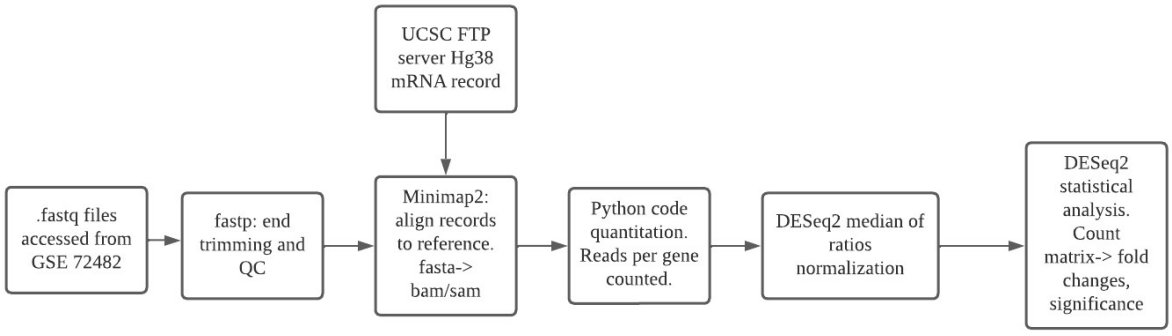


Figure 2. Comparison of different differential expression analysis pipelines for ONT data. 2A. DE expression analysis pipeline provided by ONT [29]. RNA-Seq reads are mapped to the human reference genome using Minimap2 ultralong read settings. 2B. Final analysis pipeline. Reads are mapped to the human reference transcriptome, and then normalized on a per-sample basis using DESeq2's median of ratios normalization method. 2C. Illumina Analysis pipeline. Data was accessed in .fastq format from Gene Expression Omnibus, and preprocessed using Fastp.

2.2.1 Computational Resources

Differential Expression analysis involves several computationally intensive steps. Most notably, mapping reads to a reference and quantitating mapped reads require large computational resources. The San José State University College of Science High Performance Cluster was used for mapping and quantitation steps. Basecalling and demultiplexing was performed using parallel CUDA processing with an NVIDA gtx 1660 ti.

2.2.2 Basecalling and Demultiplexing

The default output file of the ONT MinION is .fast5, an electrical trace file. To obtain sequencing reads, this electrical trace must be basecalled to translate electrical resistances to nucleotides. Additionally, barcodes must be identified within the pooled library to determine which sample each individual read originated from. Both of these steps were performed using the ONT Guppy basecaller [19]. ONT provides multiple preconfigured settings depending on library preparation kits and flow cell combinations. Settings were used for configuration dna_r9.4.1_450bps_hac, which is used for the LSK110 sequencing kit and PBC001 expansion. Reads were output in .fastq format [40].

2.2.3 Removal of Adapter and Barcode sequences using Porechop

During sequencing library preparation, two additional sequences are ligated to each read: a barcode to identify the original sample, and an adapter sequence that allows the read to bind to the nanopores themselves. Both of these sequences, however, are not part of the mRNA sequence itself, and are extraneous after demultiplexing has occurred. Additionally, reads are sometimes ligated to adapters in a read-adapter-read configuration, causing errors in quantitation. The

Porechop adapter trimmer was used to remove these nucleotides from the mRNA reads and split misligated reads from each other [41]. chopLoop.sh was written to run Porechop settings reproducibly, and is available at https://github.com/nklier38/ONT_MinION_Notch_DE [42].

2.2.4 Reference Genome and Transcriptome

cDNA reads were mapped to the Genome Reference Consortium Human Build 38 patch release 14 (GRCh38.p14) or the associated transcriptome in fasta format [43], obtained from UCSC genome browser ftp server. Initially, reads were aligned to the reference genome (Figure 2A), based on the recommended pipeline from ONT [29]. Custom developed pipelines for this project used the reference transcriptome (Figure 2B, 2C) due to concerns with gene discovery (Figure 4). All samples were run through both pipelines 2A and 2B for comparison. Since the SUPT1 cell line is not sequenced, alignment to the human reference genome is an accepted protocol for aligning sequencing reads [44].

2.2.5 Mapping Reads to Reference Sequences

cDNA reads were mapped to reference source using splice-aware ONT ultralong read settings for minimap2 (version 2.24) [44]. Reads were outputted in the .sam alignment file format. mapLoop.sh and mapLoopGenomic.sh were written to run the described settings reproducibly. Both are available at https://github.com/nklier38/ ONT_MinION_Notch_DE [42].For mapping to a reference transcriptome, standard long read ONT settings were used (-ax map-ont). For a genomic record, splice-aware settings were used (-ax splice). Data access instructions are available at https://github.com/nklier38/ONT_MinION_Notch_DE [42].

2.2.6 Quantitation of Transcripts Found for Each Gene

The number of reads aligned to each gene in the reference transcriptome were counted using the python script readCounter.py. The function of the python script is identical to the featureCounts function of RsubRead [32], however, RsubRead requires .sam files that have been aligned to a reference genome as opposed to the transcriptome used in Figure 2B and 2C. An additional gene annotation file is then required to show the locations of Transcription Start Sites, Exons, Introns, and Transcription Termination sites. This information is used to correlate sequencing data to known genes, allowing expression levels of these genes to be quantified. readCounter.py uses a more direct approach. In a reference transcriptome, each unique transcript is denoted by its accession number. This information is assigned to each read in the mapping file. The number of reads assigned to each accession number can then be counted by looping through each .sam file and counting how many times each transcript has reads mapped to it. readCounter.py is available at https://github.com/nklier38/ONT_MinION_Notch_DE [42].

To test that readCounter.py functioned similarly to RsubRead in this use case, an artificial annotation file was generated with entries for the five target genes in Table 2. The annotation file used for genomic data contains chromosome names to identify location. These were replaced with annotation numbers for the representative genes.

2.2.7 Statistical Analysis using EdgeR

Quantified reads were analyzed to calculate fold changes and significance between the DAPT and DMSO control groups. The EdgeR statistical analysis package was used in the pipeline shown in Figure 2A in accordance with ONT recommendations [29][45]. Fold-changes are calculated as the log base 2 of the change from the DMSO control to the experimental group. All p-values used are calculated assuming a negative binomial distribution. These are then

adjusted using the Benjamini-Hochberg (BH) post-hoc adjustment with an FDR cutoff of 0.05. The script edgerRunner.R was written to run edgeR analysis using these settings, and is available at https://github.com/nklier38/ONT_MinION_Notch_DE [42].

		Gene Name				
Quantitation Method	**Experiment**	HES4	HES5	HES1	HEY1	GAPDH
readCounter.py	e1 DMSO	416	0	132	266	6199
	e1 DAPT	17	0	6	24	1907
	e3 DMSO	110	0	34	120	2649
	e3 DAPT	31	0	7	3	1431
	e4 DMSO	87	0	15	149	1747
	e4 DAPT	6	0	4	23	3032
RSubRead	e1 DMSO	416	0	132	266	6199
	e1 DAPT	17	0	6	24	1907
	e3 DMSO	110	0	34	120	2649
	e3 DAPT	31	0	7	3	1431
	e4 DMSO	87	0	15	149	1747
	e4 DAPT	6	0	4	23	3032

Table 2. Comparison of Quantitation methods using methods described. Discovered counts were found to be consistent between readCounter.py and the RsubRead featureCounts function.

2.2.8 Normalization and Statistical Analysis using DESeq2

To account for variation between samples, DESeq2 median of ratios normalization was performed [46]. Median of ratios is an internal normalization performed independently on each sample to account for the individual variation between samples. While edgeR is capable of performing trimmed mean of M-value normalization, median of ratios normalization was selected based on its previous usage in similar differential expression studies [40]. In the computational pipelines shown in figure 2B, DESeq2 was used to perform a Wald test assuming a negative binomial distribution. BH adjustment was then performed with a p-value cutoff of

0.05. The script DESeqRun.R was written to run DESeq2 with these settings, and is available at https://github.com/nklier38/ONT_MinION_Notch_DE [42].

2.2.9 Procurement and Preprocessing of Illumina Data from CUTLL Ce1ls

The normalized DE analysis pipeline (Figure 2C) was also performed for a similar experiment performed on an Illumina platform. Data was obtained using a similar wet lab protocol to Figure 1, however, CUTLL cells were used, and sequencing was performed on an Illumina HiSeq 2000 [36]. The computational pipeline for Illumina analysis used the same software and paralleled the Nanopore pipeline when possible, with settings adjustments to account for short Illumina reads (Figure 2C). Data was sourced from NCBI Gene Expression Omnibus series GSE 72482. Measures of run performance, such as lane usage, were not provided. FastP was used for end trimming and quality control of Illumina data, using default single-end settings with no alteration [47]. CUTLL and SUPT1 cells are both T-ALL cell lines noted for their Notch activity and upregulation of key targets [48]. As such, this cell line was considered acceptable for comparison purposes; however, some variation in gene expression may be explained by these differing cell lines.

3 Results

3.1 Performance of a Nanopore Platform in Differential Expression Analysis

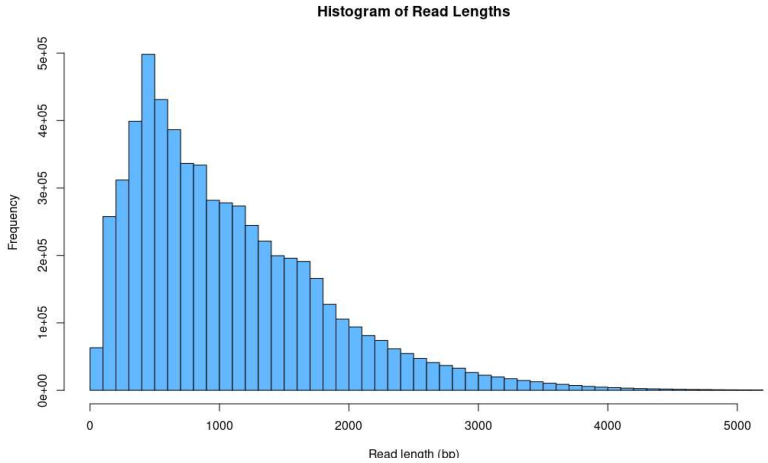
3.1.1 Read Length Performance

One of the primary advantages of nanopore sequencing is the potential to produce ultralong reads. The average length of mature human mRNA is 3.4 kb [36]. It is therefore well within the capabilities of nanopore sequencing to consistently sequence whole transcripts; however, this is not the case in these experiments. The modal fragment sequenced was 500-600 bp in length, and few fragments reached the expected size for mRNA (Figure 3A).

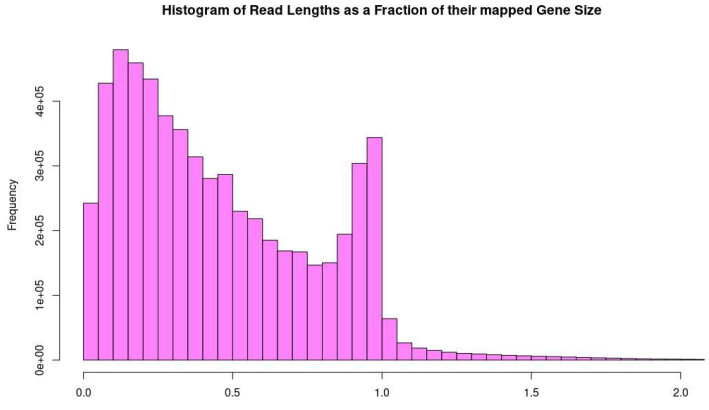
Read lengths were also represented as fractions of the size of the gene they mapped to (Figure 3B), calculated as the length of a read divided by the length of the gene that read mapped to. A value below 1.0 indicates a read smaller than the length of the gene it represents, whereas a value above 1.0 indicates a larger read than the gene. A small number of values above 1.0 are expected due to ligation products and alternative splicing, which is shown in the histogram. Porechop allows for ligation products in a read-adapter-read configuration to be properly split, however, ligation products that constitute reads directly ligated to other reads are outside of porechop's capabilities [41]. After Porechop processing, an average of 91.2% of read data was retained. The histogram peak indicates that many reads did in fact represent their whole target gene, however, a modal number of reads only exhibited 30%-50% coverage of a gene.

To examine whether reduced gene coverage was correlated with longer genes, gene length was plotted vs proportional gene coverage (Figure 3C). Each read in the scatter plot is represented as a single point. Due to the large number of reads in the dataset, this was further condensed into a heatmap of scatterplot points. Linear regression analysis shows a distinct negative correlation between gene length and coverage of that gene.

A.



B.



C.

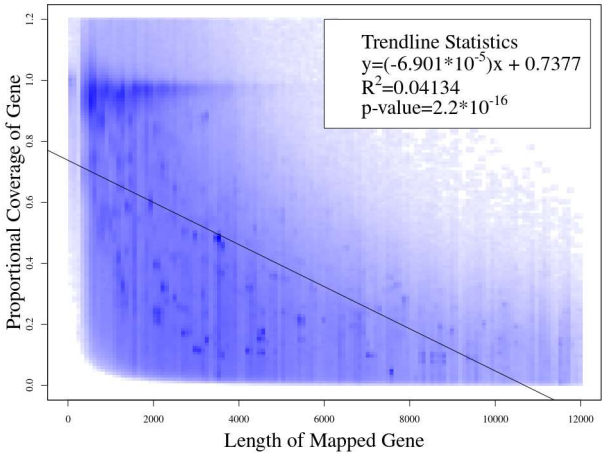


Figure 3. Assessment of read lengths of Nanopore reads. 3A. Histogram of lengths of mapped reads. The most reads were found to be less than 1kb in length. 3B. Histogram of read lengths represented as a fraction of the length of the gene they were mapped to. A value of 1.0 indicates full coverage of the gene. The peak at 1.0 indicates that full coverage did occur in some genes, however, most genes were represented by reads less than half their length. 3C. Smoothened scatterplot of the proportion described in 3B correlated to overall gene length. Scatterplot points are binned in square areas on the graph. More intense blue coloration represents more points in that bin. A weak but distinct negative correlation is observed, and a large number of fully mapped genes are seen as being less than 2000bp in length.

The negative correlation is weak, but highly significant($R^2 = 0.04124$, $p = 2.2 * 10^{-16}$), indicating a significantly negative but non-linear trend. Genes that are fully represented by their reads are more frequently shorter, as is shown by the heatmap intensity of shorter genes at 1.0 coverage (Figure 3C).

3.1.2 Total Acquired Nanopore Sequencing Reads Compared to MinION Flow Cell Loading Capacity

As discussed in 1.1.2, a MinION flow cell is limited by the lifespan of its sequencing pores [21]. Once all pores are inactive, no more sequencing data can be acquired. This is often well before all possible reads have been sequenced. Additionally, suboptimal ligation efficiency during library prep may result in some cDNA molecular being unreadable. To demonstrate the effects that these factors have on the data, the total mapped cDNA reads are shown for each sample, alongside the total unmapped reads and the total number of reads that were theoretically loaded onto the flow cell (Table 3). In all trials, the total number of acquired reads accounted for less than 0.01% of the total loaded DNA molecules actually producing a mapped read. While unmapped reads account for some of the loss in total data acquired, this accounts for far less than reads that were never initially acquired during the sequencing run.

Mapped Illumina reads generally made up a higher proportion of theoretical reads, compared to Nanopore reads, which made up small proportion of their theoretical library load. This is partially because total loaded library was not provided, and the total theoretical number of reads was computed based off of available data [36].

Sample	Mapped Transcripts	Unmapped Reads	Theoretical Transcripts	Percent of Theoretical Transcripts Mapped
E1-DMSO	1,400,381	309,949	$2.4*10^{10}$	0.005835
E1-DAPT	982,026	229,125	$2.4*10^{10}$	0.004092
E3-DMSO	802,217	242,784	$2.4*10^{10}$	0.003343
E3-DAPT	424,115	126,346	$2.4*10^{10}$	0.001767
E4-DMSO	650,376	74,299	$2.4*10^{10}$	0.00271
E4-DAPT	1,172,960	136,909	$2.4*10^{10}$	0.004887
E5-DMSO	88,071	21,434	$2.4*10^{10}$	0.000367
E5-DAPT	88,105	14,688	$2.4*10^{10}$	0.000367
E6-DMSO	29,665	4,882	$2.4*10^{10}$	0.000124
E6-DAPT	90,333	23,087	$2.4*10^{10}$	0.005835
Illumina Experiments	**Mapped Reads**	**Unmapped Reads**	**Total Reads**	
DMSO 1	33,337,673	11,269,514	44,607,187	
DMSO 2	39,269,378	9,762,649	49,032,027	
GSI 1	42,561,984	18,687,267	61,249,251	
GSI 2	34,622,606	13,089,058	47,711,664	

Table 3. Total number of transcripts read and mapped compared to unmapped reads and total number of loaded mRNA molecules. The theoretical number of transcripts in the sample represents the 40 loaded femtomoles of mRNA. Mapped reads exceeded unmapped reads, however, a low percentage of collected reads from the original sample indicates low throughput. Data is also provided from Illumina experiments [36]. Molar quantity of the library loaded was not given.

3.2 Comparison of Differential Expression Analysis Pipelines

3.2.1 Comparison of Transcriptomic and Genomic Alignments

To determine the most effective method of performing DE analysis, the two analysis pipelines shown in Figure 2 were compared. It was found that mapping nanopore reads to a genomic record caused notable problems during downstream analysis. Quantitation of many target genes varied depending on whether a genomic or transcriptomic record was used (Table 4).

Most notably, the well characterized Notch target Hes5 was found to be completely absent when quantifying alignments that had been mapped to a genomic record, disagreeing with previous data on the SUPT1 cell line [49][48]. Reads aligning to the HES5 record were recovered when aligning to a reference transcriptome in DMSO control cells, but not in DAPT treated cells, further matching known data about the cell line. A transcriptomic reference was therefore used for future analysis pipelines. Transcriptomic alignment additionally changed the number of identified reads in other genes compared to genomic alignments. Another key Notch target, HEY1, saw a reduced number of identified transcripts, whereas other targets such as HES1 remained mostly unchanged when compared to genomic alignments (Table 4).

Experiment		**Identified Transcripts per Gene**		
DMSO Controls	**Mapping reference type**	HES5	HEY1	HES1
E1-DMSO	Genomic	0	266	132
	Transcriptomic	17	84	133
E3-DMSO	Genomic	0	120	34
	Transcriptomic	12	38	33
E4-DMSO	Genomic	0	149	15
	Transcriptomic	12	46	14
DAPT Treatment				
E1-DAPT	Genomic	0	24	6
	Transcriptomic	0	8	5
E3-DAPT	Genomic	0	3	7
	Transcriptomic	2	0	6
E4-DAPT	Genomic	0	23	4
	Transcriptomic	0	5	4

Table 4. Total number of transcripts mapped to representative genes using a genomic reference (Figure 2A) compared to a transcriptomic reference (Figure 2B) for three representative genes. In HES5, no transcripts are found when mapping to the genome, but are present when mapping to the transcriptome. Less HEY1 transcripts mapped to the transcriptome than to the genome, whereas results for HES1 remained largely unchanged.

3.2.2 Effects of Normalization on Gene Quantitation

To evaluate technical variation between samples, the total number of successfully mapped transcripts found in each sample was calculated (Table 5). Each number represents the total number of transcripts that were successfully mapped to a gene. These counts should be approximately equal, as sequencing libraries were loaded in equal quantities. Despite this, there is notable technical variation between samples. This effect was most pronounced between different flow cells. Paired trials E5 and E6 both exhibited lower total transcript counts than any trial on the first flow cell.

DMSO Experiments	E1-DMSO	E3-DMSO	E4-DMSO	E5-DMSO	E6-DMSO
Total Number of Mapped Transcripts	1,400,381	802,217	650,376	88,071	29,665
DAPT Experiments	E1-DAPT	E3-DAPT	E4-DAPT	E5-DAPT	E6-DAPT
Total Number of Mapped Transcripts	982,026	424,115	1,172,960	88,105	90,333

Table 5. Total number of mapped transcripts found in each experimental sample. Variability in total number of mapped reads provided initial justification for using normalized counts for downstream statistical analysis.

Because of the variation in read counts between the different flow cells was seen between each sample, DESeq2 median of ratios normalization was performed. Normalization of read counts dramatically increased the number of significantly differentially expressed genes identified. Only 19 genes were found to be differentially expressed upon DAPT treatment when a genomic mapping pipeline was used. Mapping to a transcriptomic record and normalization increased the number of identified genes to 69 (Table 6). Total gene discovery was also assessed for both platforms. Of the 11,265 unique transcripts identified on the Nanopore platform, 9976 were also found on the Illumina platform. The remaining 11.4% of unique transcripts on the

Nanopore platform may be the result of differences between the CUTLL and SUPT1 cell lines; however, this high overlap indicates that the two largely share patterns in gene expression.

Platform	Total Up-regulated Genes with post-hoc adjustment	Total Downregulated Genes with post-hoc adjustment	Significant genes per GB	Total Unique Transcripts	Total throughput (GB)
Illumina- analysis from data source[36]	N.D.	388	N.D.	N.D.	N.D.
Illumina	572	1295	184.12	46,556	10.140
Nanopore- normalized data	13	56	9.994	11,265	6.903
Nanopore- pre-normalization	3	16	2.752	11,265	6.903

Table 6. Comparison of expression analysis to similar experiments performed in CUTLL cells on an Illumina platform. To account for the increased total throughput of the Illumina platform, total significant gene identification was divided by the total data output between all trials to normalize for total throughput. For both measurements, Illumina outperformed Nanopore sequencing in this differential expression application. Additionally, normalized vs raw data was analyzed for target gene discovery. Normalization dramatically aided differentially expressed gene discovery on a nanopore platform. Total unique transcripts, whether significant or not, is also provided, as well as total throughput. In both measures, Illumina outperformed Nanopore. The number of identified Notch targets presented by the authors in the source paper of the Illumina data is also included in the top row [36]. While the source paper provided all raw data, genes presented as Notch targets also had to meet additional experimental criteria. The total number of upregulated genes was not provided. In contrast, our analysis of the same data in the second row yielded more potential Notch target genes, as it was based on differential expression analysis alone.

3.3 Identification of Notch Target Genes using a Nanopore Platform

3.3.1 Comparison of Nanopore Sequencing with an Illumina Sequencing Platform in Identifying Notch Target Genes

The total number of significantly differentially expressed genes found on the ONT Minion was 69, 56 of which were downregulated during GSI treatment and 13 upregulated during GSI treatment. This set is dwarfed in comparison to the set of genes found on an Illumina platform, which identified 572 upregulated genes and 1295 downregulated genes (Table 6). In

both cases, significance was determined using Benjamini-Hochberg (BH) adjusted p-values. All genes identified as significant in the nanopore sequencing data were also found to be significant on the Illumina platform, suggesting that Illumina devices may be more applicable to DE analysis. As shown in the last column of Table 5, this is partially because Illumina platforms produce more data. Even when normalized for total throughput, however, nanopore sequencing only yields about 8% of the number of significant genes as an Illumina platform. Extending beyond significantly differentially expressed genes, the nanopore platform identified fewer unique transcripts overall as well.

The source paper for Illumina RNA-Seq data provides different experimental criteria to identify Notch target genes [36]. 388 Notch target genes were identified in total based on two criteria: genes that were able to have expression rescued post-GSI washout, and genes identified as Notch targets through Chromatin Immunoprecipitation sequencing. Due to these more stringent conditions to identify a Notch target, the 388 genes presented in the source paper is smaller than the 1295 downregulated genes identified using the analysis in Figure 2C.

3.3.2 Volcano Plot Representation of Nanopore and Illumina Data

For every gene found in the set of sequencing data, log-fold change in expression from DMSO to DAPT was plotted versus the adjusted p-value of the difference in means between the two groups (Figure 4A, 4B). Points further up and to the left indicate more significant and more extreme downregulation of that gene. Points further up and to the right indicate more extreme and significant upregulation. Key Notch targets are labeled, as well as the most significant and most extreme up and down regulated genes in the nanopore data (Figure 4A) and Illumina data (Figure 4B).

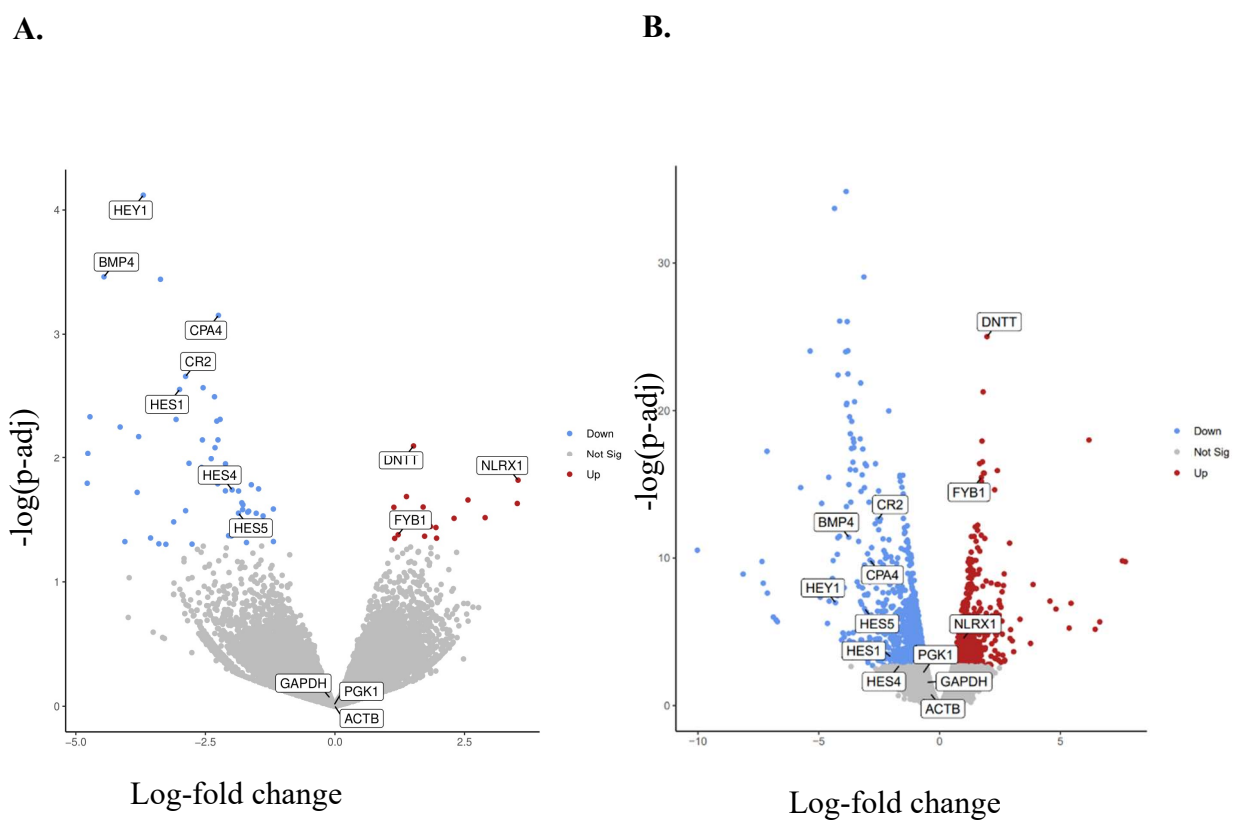


Figure 4. Volcano Plots of differential expression on both an ONT and Illumina platform. 4A. Volcano plot of GSI differential expression data collected in SUPT1 cells with an ONT MinION. Log-fold change is calculated as fold change from control to GSI inhibition trials. Blue points represented genes that are downregulated upon Notch inhibition. Labeled blue genes indicate known Notch targets that exhibited downregulation. Red points indicate upregulated genes. Labeled genes are the most extremely upregulated gene (NLRX1) and the most significantly upregulated gene (DNTT). Grey genes were not significantly differentially expressed between samples. Labeled grey genes represent housekeeping genes that are not expected to be differentially expressed between trials. 4B. Volcano plot of GSI differential expression data collected in CUTLL cells with an Illumina HiSeq2000. The same genes as the Nanopore Volcano plot are labeled.

3.3.3 Significance, Fold Change, and Transcript Counts for Representative Genes

Log2 fold-change and p-value was calculated between the DMSO controls and DAPT experimental groups (Table 7). Genes identified as downregulated upon DAPT treatment can be considered Notch targets. HES1, HES4, HES5 are canonical Notch targets that were correctly identified as being downregulated upon GSI treatment.

			Transcript Counts per Experiment									
			DMSO Trials					DAPT Trials				
Downregulated Genes	**log2 fold change**	**p-adj**	**E1**	**E3**	**E4**	**E5**	**E6**	**E1**	**E3**	**E4**	**E5**	**E6**
HEY1	-3.699	7.60E-05	55	30	47	2	0	5	0	4	0	0
BMP4	-4.457	3.45E-4	49	8	5	2	0	0	0	1	0	0
CPA4	-2.249	7.05E-4	207	186	49	6	0	28	10	19	4	1
CR2	-2.882	2.18E-3	121	14	48	1	0	13	2	3	0	0
HES1	-2.998	2.78E-3	140	33	12	0	0	3	6	3	0	0
HES4	-1.985	0.0182	209	54	44	1	1	8	17	2	0	5
HES5	-1.862	0.0280	225	66	48	1	0	0	2	0	2	5
Housekeeping genes	**log2 fold change**	**p-adj**	**E1**	**E3**	**E4**	**E5**	**E6**	**E1**	**E3**	**E4**	**E5**	**E6**
GAPDH	-0.0971	1	586	361	160	91	23	133	239	268	83	82
PGK1	-0.0165	1	1887	615	1419	55	23	777	344	2857	86	48
ACTB	-0.00659	1	13557	10416	2929	266	160	10325	4266	4883	468	406
Upregulated genes	**log2 fold change**	**p-adj**	**E1**	**E3**	**E4**	**E5**	**E6**	**E1**	**E3**	**E4**	**E5**	**E6**
DNTT	1.516	7.93E-3	75	22	51	2	1	119	38	202	7	20
FYB1	1.219	0.0418	42	21	24	6	0	95	19	128	3	8
NLRX1	3.533	0.0152	0	0	0	0	0	2	5	1	2	1

Table 7. Data table of differential expression in target genes. Log2 fold change represents fold change in the number of identified transcripts from the DMSO control to DAPT treated samples. Negative values indicate a decrease in expression, and represent identified Notch targets. P-value was calculated between the DMSO control and DAPT groups using a binomial exact test and Benjamini-Hochberg adjustment. Prior to calculation of fold change and significance, transcript counts were normalized using DESeq2 median of ratios normalization. Displayed transcript counts represent actual, non-normalized transcript counts identified in each sample.

HEY1 was found to have the most significant downregulation, followed by CPA4. BMP4 exhibited the most extreme downregulation, followed by HEY1, HES1, and CR2. Canonical housekeeping genes (HKG) GAPDH, PGK1, and ACTB were chosen based on their standard as RNA-Seq HKGs [50]. No significant difference was identified in these genes between experimental groups. DNTT was found to have the most significant increase in expression, while NLRX1 had the largest increase

in expression upon DAPT treatment.

3.3.4 Effects of Post-Hoc Adjustment on Notch Target Gene Discovery

Benjamini-Hochberg (BH) adjustment was used to investigate potential loss of Notch target identification due to the high stringency provided by this post-hoc test. 255 total significantly differentially expressed genes were identified before BH adjustment. Of these, only 135 were found to be significantly differentially expressed in the Illumina dataset before BH adjustment. Two of these genes, PCGF5 and APP, were not represented in the Illumina dataset at all, significant or not. In contrast, the 69 genes identified using Benjamini-Hochberg (BH) adjustment on Nanopore DE data were all found to be significant within the Illumina set (Table 8).

	raw p-value hits	**post-hoc hits**	**Percent retained after post-hoc adjustment**
Illumina	255	69	27.059%
Nanopore	6625	1866	28.166%

Table 8. Effects of BH FDR adjustment and comparison to Illumina. Retention of significant hits when a post-hoc test was applied was similar between platforms.

3.3.5 Comparison to previous Nanopore RNA-Seq studies

To investigate whether our nanopore results were similar in throughput to previous studies, RNA-Seq analysis was examined from Massui et al. (2021)[51]. This study is not a differential expression study, however, broad statistics about total throughput are provided for comparison. The question of whether nanopore results were comparable to Illumina data was assessed by the authors by total gene discovery and agreement of discovered genes between platforms. This study found results that largely agreed with Illumina data, however, they collected 23 GB of total reads to achieve this goal (Table 9).

	Total read data	**Number of MinION Flow Cells used**	**Total Nucleic Acid load**	**Comparable results to Illumina?**
SUPT1 Notch DE	6.9 Gigabase	2	400 fmol	No
Cardiac Fibroblast raw RNA-Seq[51]	23 Gigabase	5	1,000 fmol	Yes

Table 9. Comparison of throughput statistics from Massui et al. (bottom) to collected SUPT1 differential expression data (Top). While we did not achieve Illumina level results, data from cardiac fibroblasts suggests that more samples, spread across more flow cells, collected more data, may be necessary to achieve this.

4 Discussion

4.1 Optimizing Existing Computational Resources for Long Nanopore Reads

Differential Expression Analysis remains an application that is not deeply explored on Oxford Nanopore technologies platforms. The existing recommended pipeline is largely based upon applications designed for short, high-accuracy Illumina reads. Most notably, the pipeline fails to address the nature of long read data in two key ways.

Existing DE analysis pipelines map collected cDNA reads to genomic DNA. An additional gene annotation file is then required to show the locations of Transcription Start Sites, Exons, Introns, and Transcription Termination sites. This information is used to correlate sequencing data to known genes, allowing expression levels of these genes to be quantified. Genomic mapping is standard practice in RNA-Seq analysis [23][29][33]. RNA-Seq data is sometimes mapped to a reference transcriptome instead [52][53]. The reference is composed of known transcripts instead of a complete genomic record. Functionally, the major difference between mapping a read to a genome and mapping to a transcript is that the genome includes introns, whereas a transcriptome will include different splice variations. Separate annotation files are used to add splice variant and intron position information to genomic references. Modern mapping tools, such as Minimap2, can be run in a splice-aware fashion that ensures mapping can occur between an RNA read and a genomic reference [44]. Due to the information found in the annotation file, if a transcript is found in a reference transcriptome, there should therefore be no difference between genomic mapping and transcriptomic mapping. If the transcript is not in the reference transcriptome, however, then it will not be mapped when using a transcriptomic reference. Genomic mapping is therefore capable of discovering novel transcripts, such as new splice variants, while transcriptomic mapping is not [53]. Novel transcripts are one contributing

factor to the total number of unmapped reads, alongside potential contamination or fragments that are too small or inaccurate to map. Therefore, studies targeting unmapped reads often focus on novel gene or splice variant discovery [54][55]. Novel transcripts are therefore likely a component of the unmapped reads found in this study (Table 5).

When using genomic mapping, quantitation of known target genes changed in ways that disagreed with existing studies. Most notably, the extremely well characterized Notch target HES5 was completely absent from the genomic mapped dataset (Table 2). In addition to being a known Notch target [56], HES5 is also noted for having high expression in the SUPT1 cell line under normal conditions [49][48]. When mapped to a reference transcriptome, however, transcripts for HES5 are found in the sequencing dataset. suggesting that the issue with quantitation is purely computational. Additional Notch targets, such as HEY1, are also affected by these different mapping techniques, however, others still seem to be mostly unaffected, such as HES1 (Table 2).

Theoretically, these differences between mapping strategies should not be present. These differences do not appear to be associated with the total number of introns in a gene, however. HES1, which remained similar between mapping references, has 3 introns. HES5 and HEY1, which were both affected, have 2 and 5 respectively [57][58][59]. With the splice-aware settings used by minimap2 when aligning to the genome, introns should not affect overall alignment [44]. This discrepancy could have many underlying explanations. Other preliminary studies have noted small differences between minimap2 genomic alignment and transcriptome-specific alignment tools [60]. Algorithmic differences between splice-aware and map-ont settings of minimap2 may explain this difference, however, further experimentation in other DE contexts with known target genes would be necessary to verify this.

The second key way in which the ONT pipeline should be updates is in normalization. This pipeline uses the software tool Salmon for read quantitation and normalization [29]. Salmon normalizes based on Transcripts Per Million (TPM), a simple normalization method that only accounts for bulk quantities of RNA reads in each sample. This creates problems when the distribution of RNA is different between samples, which is expected in differential expression studies. Instead of TPM, modern RNA-Seq analysis uses direct counts of mapped transcripts, which are then normalized using downstream analysis [61]. Recently, TPM has been replaced by various other normalization methods, including DESeq2 median-of-ratios normalization [46].

As mentioned previously, nanopore sequencing allows for single reads to provide coverage on an entire transcript. Direct quantitation should theoretically be possible, without the need for any form of normalization. Additionally, barcoding samples and loading them on the same flow cell should theoretically reduce technical variation. Based on the total read counts from each sample, this is not the case (Table 5). Notably, E5 and E6 occurred on a different flow cell than the other three trials, and both exhibit greatly reduced total read counts in each treatment. Normalizing just on a flow cell basis does not account for all of the variation, however. E1-DMSO and E4-DAPT both exhibited higher read counts than the other samples on the flow cell, including the treatment they were paired with. In a situation with high variability such as this, per-sample normalization strategies, such as DESeq2 median of ratios, should be performed.

E5 and E6 were the only pairs of samples sequenced in their sequencing run. Since each sample was loaded at 40 fmol, this meant that the total loaded library during this run was 160 contrasting with the 240 fmol load during the sequencing run for E1, E3, and E4. Initially, it was hypothesized that the low overall loading capacity on a flow cell from running less samples at

once would mean less reads competing for pores, and therefore more data would be collected per sample. This proved to not be the case (Figure 4). This, however, is obscured by flow cell health and reuse. Oxford Nanopore provides reagents to allow reuse of MinION flow cells [62]. While this kit is capable of recovering some pores inactivated during previous sequencing runs, ONT notes that it is not intended to recover all of them. E5 and E6 were run on a flow cell that had previously been used for sequencing and subsequently washed, while other samples were run on a fresh flow cell, offering a potential explanation for the greatly reduced total volume of data in these two trials.

In this differential expression study, accounting for RNA composition bias is crucial. Notch is often described as a master regulator gene, and is therefore responsible for regulating a wide variety of genes [36][63]. Therefore, it is expected that when the Notch pathway is inhibited, reduced expression of Notch target genes will result in a decrease in the total mRNA produced. Since sequencing libraries are ultimately loaded in the same molar amounts, this discrepancy can cause abnormalities in the ratios between gene read counts. DESeq2 median of Ratios normalization was chosen to combat this and ensure that DAPT and DMSO samples were normalized to the same scale [46]. Since this is a function of the differential expression experimental setup and not the sequencing platform, DESeq2 median of ratios is generally favored over TPM for Illumina applications [30][61][64].

Median of ratios normalization through DESeq2 appeared to have a positive effect on the ability of this pipeline to identify differentially expressed genes. 69 differentially expressed genes were identified in the normalized set, a notable increase from the 31 found without normalization (Table 6). Normalization, however, cannot fully account for all of the discrepancies between samples. In some cases, Notch target genes were completely absent from

datasets with lower overall library sizes. Trials E5 and E6 often did not contain any transcripts that were found to be significantly up or down regulated. Notably, the E6 DMSO trial did not exhibit any transcripts in identified Notch targets except for HES4. In this case, as well as any other sample in which a total transcript count is 0, normalization cannot account for the variation between samples.

The final step of statistical analysis was a Benjamini-Hochberg p-value adjustment. When performing large numbers of statistical tests at once, such as during sequencing analysis, a certain number of significant hits are expected to yield false positives. Post-hoc adjustments are intended to correct for this false positive rate by making tests more stringent. To analyze the effectiveness of this strategy in nanopore sequencing, the total number of genes retained by post-hoc analysis was compared to the total number of pre-adjustment hits (Table 7). On both platforms, post-hoc testing appears to retain a similar percentage of the total genes, suggesting that the effectiveness of post-hoc testing is not platform specific.

Appropriate post-hoc testing remains a persistent problem in differential expression analysis. While correcting for false positives in necessary, many tests may be too stringent for certain DE applications. In particular, most post-hoc tests assume that an extremely small proportion of tests performed will actually be true positives [65]. The calculation of a post-hoc test is partially dependent on the significance “rank” of the test, which brings this assumption into the calculation. Genes higher in the significance rank order are therefore more likely to be retained. In the case of DE analysis performed on master regulator transcription factors, however, this may not always be the case. In this situation, Notch is responsible for regulating a wide variety of downstream target genes, creating a situation where the number of true positives could potentially be very high.

Despite these limitations, the addition of both transcriptomic mapping and normalization to a DE computational pipeline improved results for this Notch target gene identification.

4.2 Technical Limitations of Nanopore Sequencing for DE Analysis

One of the primary advantages of nanopore sequencing is its long read length. In a differential expression study, this means that whole transcripts are theoretically able to be sequenced with a single read. In practice, however, we found that this was not the case (Figure 3B). Modal read length was 400-500bp, with 90% of reads being under 2kb, well under the average length of human mRNA (Figure 3A). There are two possible sources of this variation based on read lengths alone. Either sequencing is biased towards shorter genes, or genes are only being represented by a small fraction of their overall length. Plotting reads as a fraction of their corresponding gene reveals that both are the case. While there is a distinct peak representing reads that exhibit full coverage of their corresponding genes, over half of all reads represent less than half of a complete gene transcript. Additionally, shorter genes appear to be more likely to have reads that fully represent the gene. This is demonstrated by the dense line of genes less than 2kb long that achieve a coverage of 1.0 in Figure 3C, as well as the negative correlation between gene length and proportional coverage of a gene by a read. This is potentially caused by fragmentation, but it could be a shortcoming of the platform itself. Fast5 files are generated in real time, meaning that they do not check for a complete read before writing data [66]. Future library preparation should utilize size distribution analysis immediately prior to sequencing to determine which is the case.

It was also found that genes can sometimes be represented by very few reads (Table 7). For example, BMP4 was represented by less than 10 reads in all DMSO samples except for E1.

Additionally, it was completely absent in E6. HES1, a gene that is notably active in SUPT1 cells[48], was completely absent from the DMSO controls of E5 and E6. In addition from genes being absent from specific samples, genes are likely absent from the dataset as a whole. In total, nanopore sequencing identified 11265 unique transcripts in the SUPT1 cell line, whereas Illumina identified 46556 in the closely related CUTLL cell line (Table 5). This indicates that many genes may simply be missing from the dataset entirely, being represented by zero transcripts in any experiment.

4.3 Identification of Notch Targets

Nanopore DE analysis identified 69 differentially expressed genes between the DMSO and DAPT treatments. Of these, 56 were downregulated upon Notch inhibition, and can therefore be identified as potential Notch target genes (Table 5). Among the downregulated genes are many canonical Notch targets, such as the HES family genes HES1, HES4, and HES5. All three of these genes are well known downstream elements of Notch signaling, with roles in cellular development [56][57][67]. Identification of these genes as Notch targets therefore serves as an important confirmation of the effectiveness of the technology. HEY family genes also serve important roles in downstream Notch signaling [68]. HEY1 was the most significantly downregulated gene upon Notch inhibition, providing another canonical Notch target to support the validity of the platform.

DE analysis also identified less canonical Notch targets. BMP4 and CR2 have both been shown to be transcriptionally activated by Notch in low throughput, targeted studies [69][70]. The most significantly downregulated gene after HEY1 was CPA4, a carboxypeptidase and known oncogene [71]. While no direct link between Notch and CPA4 has been established

through targeted techniques, other high throughput studies have identified CPA4 as a potential Notch target through both DE analysis and examination of Notch binding sites in chromatin [72][36].

None of the genes found to be upregulated are known Notch targets. Some, such as DNTT, have related functions. DNTT plays a role in lymphoblast differentiation alongside Notch target genes, but Notch signaling does not induce its transcription [73]. Upregulated genes may be explained by continuations of other Notch-related pathways, however. The HES family of genes, including HES1 and HES5, function as transcriptional repressors[57][67]. Consequently, the observed downregulation in these genes due to a loss in Notch may cause transcription of other genes, or upregulation of other signaling pathways. While no direct molecular link has been observed, both HES1 and DNTT play a role in lymphoblast differentiation, suggesting they may share downstream pathways [74].

The higher throughput PromethION sequencing platform may help alleviate some of these problems. The underlying pore chemistry of a PromethION is largely shared with the MinION; however, it offers more pores per flow cell, as well as the ability to sequence on multiple flow cells in parallel [75][26]. The end result of this is larger total read acquisition, increasing the overall throughput of a sequencing run. Since important details of pore chemistry, such as translocation speed, remain similar between platforms, the PromethION is unlikely to solve the noted issues regarding read length (Figure 3)[75][26].

4.4 Comparison to Illumina DE Analysis

The power of this platform to identify both canonical Notch targets and less well characterized Notch targets through DE analysis demonstrates its utility, however, nanopore

sequencing still dramatically underperforms when compared to Illumina sequencing. All 69 significantly differentially expressed genes identified using nanopore sequencing were also identified using Illumina sequencing. Illumina sequencing also identified an additional 539 genes that were upregulated upon Notch inhibition, and an additional 627 that were downregulated (Table 6). This is likely due to the technical limitations of nanopore sequencing discussed previously, ultimately affecting the total number of significant hits found. It should be noted that Illumina data was collected in a CUTLL cell line, as opposed to the SUPT1 cells used for nanopore sequencing. However, this difference is likely not enough to account for the dramatically different numbers of significant genes identified, as both cell lines are T-ALL. Cell line differences are potentially seen in two genes, however. Amyloid beat precursor protein (APP) and Polycomb Group Ring Finger 5 (PCGF5) are significant by pre-adjustment p-value in the Nanopore dataset, but are completely absent from the Illumina set. BH adjustment removes this significance. Even if they are not truly significant between DAPT and DMSO trials, their presence in SUPT1 data and not CUTLL indicates that they may be cell line specific. APP is most notable for its role in Alzheimer's disease as a precursor to beta-amyloid [76], but expression has also been reported in lymph nodes and immune cells [77]. PCG5 promotes RNA polymerase II function, and has also been reported in immune cells [78][79]. Both have some implication in transcriptional regulation.

A likely factor in the MinION's reduced ability to identify differentially expressed genes is the low overall throughput of the platform when compared with Illumina, offering about 68% of total bp read. When normalizing the number of significant genes discovered to throughput, however, nanopore still underperforms when compared to Illumina (Table 6), weakening its applicability to DE analysis.

In addition to the low throughput, differences between the platform may also be explained by the different cell lines used between the two experiments. 88.6% of all genes with transcript counts above 0 in any experiment were also found in the Illumina experiments. As the results discussed above have shown, the Illumina platform generally outperforms the nanopore platform in throughput and gene discovery. As such, this difference is likely not entirely attributable to the nanopore system discovering transcripts that the Illumina experiment failed to. Rather, it possibly represents genes that are expressed in the SUPT1 cell line that are not present in CUTLL cells.

Previous uses of nanopore sequencing in RNA-Seq have generally had mixed results. In viral genomes, nanopore sequencing appears to provide better accuracy for dense, overlapping genes [80]. In Eukaryotic yeast, RNA-Seq was found to provide lower accuracy and inconsistent mapping when compared to Illumina. Further refinement to long-read mapping may help alleviate this problem. Despite this, overall coverage of RNA-Seq reads appears to be similar to Illumina due to the long length of individual reads [81]. In our study, 6.903 GB of data were collected over five paired trials, indicating a large number of experiments to achieve Illumina-level results (Table 5). Together with the data shown in this study, this shows a consistent trend. In metrics where coverage and precise assembly is required, long nanopore sequencing performs well, however, when precise quantitation is required, it falls behind existing sequencing methodologies.

A previous study by Massui et al. demonstrated similar results to Illumina in RNA-Seq trials [51]. In their experiments, RNA-Seq data was collected from cardiac fibroblasts and examined for total gene discovery. These genes were then assessed for agreement to Illumina datasets used in similar studies. Since this study is not a differential expression study and

addresses a different biological question, the results are not quantitatively comparable to our experiments. Despite this, the total throughput that the authors used to achieve Illumina level results is worth noting (Table 9). Compared to our 6.9 GB of total read data collected, this study collected 23 GB in total. They were able to achieve this many reads by spreading their trials across five separate flow cells, each loaded to a 200 fmol capacity. This means more transcripts were loaded to be sequenced in total. Possibly more importantly, however, is the number of flow cells used. As noted previously, the throughput of nanopore sequencing is primarily limited by the lifespan of the pores [21]. Using more flow cells, each containing their own set of pores, is one method of increasing the total number of pores utilized in each experimental run.

Oxford Nanopore sequencing still has other potential applications. The consistency at which it can identify a subset of significant genes identified by Illumina suggest utility in preliminary research and education, where Illumina sequencing would be prohibitively expensive. Nanopore sequencing can also be used alongside Illumina to produce accurate hybrid assemblies, allowing for both accurate *de novo* sequencing and more precise quantitation [5][82][24].

5 CONCLUSION

Existing analysis tools and pipelines [29][44] require careful consideration and updating when being applied to third-generation, ultralong read sequencing technologies. The pipeline used in this study was able to effectively identify Notch target genes through differential expression analysis. When compared to existing Illumina sequencing datasets, however, it was found that nanopore sequencing still cannot perform at the same level as second-generation sequencing techniques for this application, even when considering the reduced overall throughput of the platform. All genes identified on an ONT platform were also identified on an Illumina platform, and the Illumina platform was able to identify many additional Notch targets. Despite being a useful tool for applications where long reads are required, nanopore sequencing appears to consistently struggle with precise quantitation. Despite this, it may still be applied to differential expression problems in which cost would normally be prohibitive, such as preliminary studies, field studies, and education.

REFERENCES

[1] Guan, Y.-F. et al. Application of next-generation sequencing in clinical oncology to advance personalized treatment of cancer. Chin. J. Cancer 31, 463–470 (2012)

[2] Hong, M. et al. RNA sequencing: new technologies and applications in cancer research. J. Hematol. Oncol. 13, 166 (2020)

[3] Maljkovic Berry & Melendrez. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. The Journal of Infectious Disease doi:10.1093/infdis/jiz286

[4] Fernandez-Marmiesse, A., Gouveia, S. & Couce, M. L. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. Curr. Med. Chem. 25, 404–432 (2018)

[5] Nurk, S. et al. The complete sequence of a human genome. Science 376, 44–53 (2022)

[6] Rothberg, J. M. et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 475, 348–352 (2011)

[7] Shen, R. et al. High-throughput SNP genotyping on universal bead arrays. Mutat. Res. 573, 70–82 (2005)

[8] Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next-Generation Sequencing Technologies. Curr. Protoc. Mol. Biol. 122, e59 (2018)

[9] Lahens, N. F. et al. A comparison of Illumina and Ion Torrent sequencing platforms in the context of differential gene expression. BMC Genomics 18, 602 (2017)

[10] Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14, 91 (2013)

[11] Maximum read length for Illumina sequencing platforms. https://support.illumina.com/bulletins/2020/04/maximum-read-length-for-illumina-sequencing-platforms.html

[12] Rhie, A. et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature 592, 737–746 (2021)

[13] Dominguez Del Angel, V. et al. Ten steps to get started in Genome Assembly and Annotation. F1000Res. 7, (2018)

[14] Tørresen, O. K. et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 47, 10994–11006 (2019)

[15] Heydari, M., Miclotte, G., Demeester, P., Van de Peer, Y. & Fostier, J. Evaluation of the impact of Illumina error correction tools on de novo genome assembly. BMC Bioinformatics 18, 374 (2017)

[16] Li, Y. & Tollefsbol, T. O. DNA methylation detection: bisulfite genomic sequencing analysis. Methods Mol. Biol. 791, 11–21 (2011)

[17] Rhoads, A. & Au, K. F. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics (2015)

[18] Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 21, 30 (2020)

[19] Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 20, 129 (2019)

[20] Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36, 338–345 (2018)

[21] Tyler, A. D. et al. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci. Rep. 8, 10931 (2018)

[22] Jain, M., Olsen, H. E., Akeson, M. & Abu-Shumays, R. Adaptation of Human Ribosomal RNA for Nanopore Sequencing of Canonical and Modified Nucleotides. Methods Mol. Biol. 2298, 53–74 (2021)

[23] Feng, Z., Clemente, J. C., Wong, B. & Schadt, E. E. Detecting and phasing minor single-nucleotide variants from long-read sequencing data. Nat. Commun. 12, 3032 (2021)

[24] Chen, Z., Erickson, D. L. & Meng, J. Benchmarking hybrid assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. BMC Genomics 21, 631 (2020)

[25] Jain, M., Olsen, H. E., Paten, B. & Akeson, M. Erratum to: The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 17, 256 (2016)

[26] Sara Goodwin, W. R. M. Sequencing Complex Genomes with PromethION Technology in a Core Setting. Journal of Biomolecular Technology S36–S37 (2019)
[27] Lin, B., Hui, J. & Mao, H. Nanopore Technology and Its Applications in Gene Sequencing. Biosensors 11, (2021)

[28] Quail, M. A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341 (2012)

[29] Love, M. I., Soneson, C. & Patro, R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Res. 7, 952 (2018)

[30] Chung, M. et al. Best practices on the differential expression analysis of multi-species RNA-seq. Genome Biol. 22, 121 (2021)

[31] Gleeson, J. et al. Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. Nucleic Acids Res. 50, e19 (2022)

[32] Yang Liao, Gordon K Smyth, Wei Shi. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 47, e47 (2019)

[33] Dong, X. et al. The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools. NAR Genomics and Bioinformatics vol. 3 (2021)

[34] Aster, J. C., Pear, W. S. & Blacklow, S. C. The Varied Roles of Notch in Cancer. Annu. Rev. Pathol. 12, 245–275 (2017)

[35] Kopan, R. & Ilagan, M. X. G. The canonical Notch signaling pathway: unfolding the activation mechanism. Cell 137, 216–233 (2009)

[36] Severson, E. et al. Genome-wide identification and characterization of Notch transcription complex-binding sequence-paired sites in leukemia cells. Sci. Signal. 10, (2017)

[37] Xiaoyu Li, H. von B. Notch Signaling in T-Cell Development and T-ALL. ISRN Hematol. (2011) doi:10.5402/2011/921706

[38] Weng AP, Nam Y, Wolfe MS, Pear WS, Griffin JD, Blacklow SC, Aster JC. Growth suppression of pre-T acute lymphoblastic leukemia cells by inhibition of notch signaling. Mol Cell Biol. 2003 Jan;23(2):655-64. doi: 10.1128/MCB.23.2.655-664.2003. PMID: 12509463; PMCID: PMC151540.

[39] Ligation sequencing amplicons - PCR barcoding (SQK-LSK110 with EXP-PBC001). https://community.nanoporetech.com/docs/prepare/library_prep_protocols/pcr-barcoding-amplicons-sqk-lsk110/v/pbac12_9112_v110_revg_10nov2020 (2021)

[40] Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC Bioinformatics 19, 274 (2018)

[41] Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. Microb Genom 3, e000132 (2017)

[42] Klier, N. ONT MinION Notch DE. GitHub https://github.com/nklier38/ONT_MinION_Notch_DE

[43] Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res. 27, 849–864 (2017)

[44] Li, H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018)

[45] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140 (2010)

[46] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014)

[47] Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890 (2018)

[48] Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607 (2012)

[49] Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database 2016, (2016)

[50] Chang, C.-W. et al. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS One 6, e22859 (2011)

[51] Massaiu, I. et al. Evaluation of Oxford Nanopore MinION RNA-Seq Performance for Human Primary Cells. Int. J. Mol. Sci. 22, (2021)

[52] Piovesan, A., Caracausi, M., Antonaros, F., Pelleri, M. C. & Vitale, L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database 2016, (2016)

[53] Conesa, A. et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13 (2016)

[54] Kazemian, M. et al. Comprehensive assembly of novel transcripts from unmapped human RNA-Seq data and their association with cancer. Mol. Syst. Biol. 11, 826 (2015)

[55] Dongwei Li, Qitong Huang, Lei Huang, Jikai Wen, Jing Luo, Qing Li, Yanling Peng, Yubo Zhang. Baiting out a full length sequence from unmapped RNA-seq data. BMC Genomics (2021) doi:10.1186/s12864-021-08146-4

[56] Ohtsuka, T. et al. Hes1 and Hes5 as notch effectors in mammalian neuronal differentiation. EMBO J. 18, 2196–2207 (1999)

[57] Takebayashi, K., Akazawa, C., Nakanishi, S. & Kageyama, R. Structure and promoter analysis of the gene encoding the mouse helix-loop-helix factor HES-5. Identification of the neural precursor cell-specific promoter element. J. Biol. Chem. 270, 1342–1349 (1995)

[58] C Leimester, A Externbrink, B Klamt, M Gessler. Hey genes: a novel subfamily of hairy- and Enhancer of split related genes specifically expressed during mouse embryogenesis. Mech. Dev. (1999) doi:10.1016/s0925-4773(99)00080-5

[59] J N Feder, L Li, L Y Jan, Y N Jan. Genomic cloning and chromosomal localization of HRY, the human homolog to the Drosophila segmentation gene, hairy. Genomics (1994) doi:10.1006/geno.1994.1126

[60] Mikheenko, A., Prjibelski, A. D., Joglekar, A. & Tilgner, H. U. Sequencing of individual barcoded cDNAs using Pacific Biosciences and Oxford Nanopore Technologies reveals platform-specific error patterns. Genome Res. 32, 726–737 (2022)

[61] Zhao, S., Ye, Z. & Stanton, R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. RNA 26, 903–909 (2020)

[62] Flow Cell Wash Kit. nanoporetech https://store.nanoporetech.com/flow-cell-wash-kit-r9.html

[63] Cai, W. et al. Master regulator genes and their impact on major diseases. PeerJ 8, e9952 (2020)

[64] Maza, E. In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. Front. Genet. 7, 164 (2016)

[65] Grinde, K. E. et al. Illustrating, Quantifying, and Correcting for Bias in Post-hoc Analysis of Gene-Based Rare Variant Tests of Association. Front. Genet. 8, 117 (2017)

[66] Zhang, H. et al. Real-time mapping of nanopore raw signals. Bioinformatics 37, i477–i483 (2021)

[67] Kageyama, R., Ohtsuka, T. & Kobayashi, T. The Hes gene family: repressors and oscillators that orchestrate embryogenesis. Development 134, 1243–1251 (2007)

[68] Fischer, A., Schumacher, N., Maier, M., Sendtner, M. & Gessler, M. The Notch target genes Hey1 and Hey2 are required for embryonic vascular development. Genes Dev. 18, 901–911 (2004)

[69] Ng, H. L. et al. Notch signaling induces a transcriptionally permissive state at the Complement C3d Receptor 2 (CR2) promoter in a pre-B cell model. Mol. Immunol. 128, 150–164 (2020)

[70] Dahlqvist, C. et al. Functional Notch signaling is required for BMP4-induced inhibition of myogenic differentiation. Development 130, 6089–6099 (2003)

[71] Sun, L. et al. CPA4 is a Novel Diagnostic and Prognostic Marker for Human Non-Small-Cell Lung Cancer. J. Cancer 7, 1197–1204 (2016)

[72] Wang, H. et al. NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. Proc. Natl. Acad. Sci. U. S. A. 111, 705–710 (2014)

[73] Weerkamp, F. et al. Identification of Notch target genes in uncommitted T-cell progenitors: No direct induction of a T-cell specific gene program. Leukemia 20, 1967–1977 (2006)

[74] Rothenberg, E. V. Transcriptional control of early T and B cell developmental choices. Annu. Rev. Immunol. 32, 283–321 (2014)

[75] De Coster, W. et al. Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. Genome Res. 29, 1178–1187 (2019)

[76] O'Brien RJ, Wong PC. Amyloid precursor protein processing and Alzheimer's disease. Annu Rev Neurosci. 2011;34:185-204.

[77] Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. NM_201414, Amyloid Beta Precursor Protein. Available from: https://www.ncbi.nlm.nih.gov/gene/?term=NM_201414

[78] Meng Y, Liu Y, Dakou E, Gutierrez GJ, Leyns L. Polycomb group RING finger protein 5 influences several developmental signaling pathways during the in vitro differentiation of mouse embryonic stem cells. Dev Growth Differ. 2020 May;62(4):232-242. doi: 10.1111/dgd.12659. Epub 2020 Apr 22. PMID: 32130724

[79] Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – . Accession No. NM_001257101, polycomb group ring finger 5. Available from: https://www.ncbi.nlm.nih.gov/gene/?term=NM_001257101

[80] Depledge, D. P. et al. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. Nat. Commun. 10, 754 (2019)

[81] Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods 15, 201–206 (2018)

[82] Lemay, M.-A. et al. Combined use of Oxford Nanopore and Illumina sequencing yields insights into soybean structural variation biology. BMC Biol. 20, 53 (2022)