Published online 26 April 2021

doi: 10.1093/nargab/lqab028

# The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools

Xueyi Dong[1,2,*], Luyi Tian[1,2], Quentin Gouil[1,2], Hasaru Kariyawasam[1], Shian Su[1,2], Ricardo De Paoli-Iseppi[3], Yair David Joseph Prawer[3], Michael B. Clark[3], Kelsey Breslin[1], Megan Iminitoff[1,2], Marnie E. Blewitt[1,2], Charity W. Law[1,2,*] and Matthew E. Ritchie[1,2,*]

[1]Epigenetics and Development Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia, [2]Department of Medical Biology, The University of Melbourne, Parkville, Victoria 3010, Australia and [3]Centre for Stem Cell Systems, Department of Anatomy and Neuroscience, The University of Melbourne, Parkville, Victoria 3010, Australia

Received June 22, 2020; Revised February 26, 2021; Editorial Decision March 23, 2021; Accepted March 30, 2021

## ABSTRACT

**Application of Oxford Nanopore Technologies' long-read sequencing platform to transcriptomic analysis is increasing in popularity. However, such analysis can be challenging due to the high sequence error and small library sizes, which decreases quantification accuracy and reduces power for statistical testing. Here, we report the analysis of two nanopore RNA-seq datasets with the goal of obtaining gene- and isoform-level differential expression information. A dataset of synthetic, spliced, spike-in RNAs ('sequins') as well as a mouse neural stem cell dataset from samples with a null mutation of the epigenetic regulator *Smchd1* was analysed using a mix of long-read specific tools for preprocessing together with established short-read RNA-seq methods for downstream analysis. We used *limma-voom* to perform differential gene expression analysis, and the novel *FLAMES* pipeline to perform isoform identification and quantification, followed by *DRIMSeq* and *limma-diffSplice* (with *stageR*) to perform differential transcript usage analysis. We compared results from the sequins dataset to the ground truth, and results of the mouse dataset to a previous short-read study on equivalent samples. Overall, our work shows that transcriptomic analysis of long-read nanopore data using long-read specific preprocessing methods together with short-read differential expression methods and software that are already in wide use can yield meaningful results.**

## INTRODUCTION

Short-read sequencing technology has underpinned transcriptomic profiling research over the past decade. The sequencing platforms offered by companies such as Illumina Inc. provide high read accuracy (>99.9%) and throughput which allows many samples to be profiled in parallel. One major limitation of short-read sequencing technology is the modest read lengths offered (currently up to 600 bases), which makes accurate isoform quantification and novel isoform discovery challenging. Long-read sequencing offers a distinct advantage in this regard, with the ability to generate reads that are typically in the 1–100 kilobase (kb) range (1), which spans the typical length distribution of spliced genes in human (for protein coding genes 1–3 kb is typical with outliers such as Titin at >80 kb) thereby allowing the sequencing of entire isoforms. This, however, comes at the expense of lower throughput and reduced accuracy compared to short-read sequencing. The two main technology platforms that dominate the field of long-read sequencing are Pacific Biosciences' (PacBio) Single-Molecule Real Time (SMRT) sequencing and Oxford Nanopore Technologies' (ONT) nanopore sequencing.

Previous work on long-read transcriptomic data focuses on transcript-level analysis, especially in the discovery of novel isoforms (2–4). Some long-read specific methods have been developed for this task. Reference-based methods, such as *TALON* (5), compares reads to existing gene and transcript models to create novel models. Reference-free methods, such as *FLAIR* (6), maps reads to the reference genome, clusters alignments into groups and collapses them into isoforms. Differential transcript usage (DTU) is another transcript-level analysis that is of great inter-

[*]To whom correspondence should be addressed. Tel: +61 9345 2555; Fax: +61 9347 0852; Email: mritchie@wehi.edu.au
Correspondence may also be addressed to Charity W. Law. Tel: +61 9345 2555; Fax: +61 9347 0852; Email: law@wehi.edu.au
Correspondence may also be addressed to Xueyi Dong. Tel: +61 9345 2555; Fax: +61 9347 0852; Email: dong.x@wehi.edu.au

est (6–8). DTU analyses examine differences in the relative proportions of expressed isoforms between two conditions. The *DRIMSeq* (9) method performs DTU analysis on transcript-level RNA-seq counts using a Dirichlet-multinomial model. Alternatively, tools developed for differential exon usage analysis, such as the *diffSplice* function (10) in the *limma* and *edgeR* packages, can also be used for DTU analyses by substituting transcript-level counts for exon-level counts (11). Both *DRIMSeq* and *diffSplice* methods were developed for short-read data. The *stageR* package (12) can be used to control the false discovery rate (FDR) of DTU analyses through its stage-wise method which screens potential DTU genes using gene-level *P*-values before selecting the transcripts with evidence of DTU.

In many analyses, a gene-level analysis is as far as researchers go, so understanding how well this type of analysis can be done for long-read data is of interest. Previous studies have looked at gene-level analysis of nanopore data but study design limited the methods used. Soneson *et al.* (13) concluded that read coverage in native RNA libraries (∼0.5 million aligned reads per flow cell) were too low for gene-level analyses, resulting in low power and high variability. Gleeson *et al.* (14) found differential expression analysis performed using *DESeq2* (15) was specific but not sensitive on their ONT MinION direct RNA neuroblastoma samples with synthetic spike-in controls, based on 6 samples and around 4 million reads in total. Li *et al.* (7) worked around this by simply using fold-changes to identify differentially expressed genes for three ONT MinION direct RNA *Caenorhabditis elegans* samples; however, the lack of statistical testing could lead to unreliable results. Jenjaroenpun *et al.* (16) used *DESeq2* (15) to perform differential expression analysis on direct RNA transcript-level counts, but gene-level expression was not studied. Cruz-Garcia *et al.* (17) used a Snakemake pipeline which included the *edgeR* (18) quasi-likelihood method to identify a radiation exposure signature consisting of 46 transcripts from a high coverage (40–75 million reads per sample) ONT PromethION PCR cDNA dataset of human blood cells.

In this study, we performed both gene- and isoform-level analyses of two nanopore long-read transcriptome sequencing datasets that follow a simple replicated experimental design: a synthetic 'sequins' (19) PCR-cDNA dataset, and a mouse neural stem cell direct-cDNA dataset. We obtained meaningful results using an analysis pipeline that mostly comprised of *'off-the-shelf'* methods developed for short-read data, despite our datasets having only a few million long reads per sample. We found our results for the common two-group experimental design to be reliable in that they are broadly consistent with the available ground-truth or findings from a previous short-read experiment. We found existing methods for isoform identification from long-read data to be unreliable, and introduce a novel method, *FLAMES*, as part of our isoform-level analysis pipeline.

## MATERIALS AND METHODS

### Study design

Mouse neural stem cells (NSCs) from 4 wild-type (WT) and 3 MommeD1 mutated (*Smchd1*-null) (20) samples were prepared and sequenced, together with 3 'other' samples from a different experiment. Samples were sequenced in two batches, each containing 6 samples. One WT and one 'other' sample were sequenced in both batches as technical replicates in order to obtain additional reads for these samples.

Technical replicates of synthetic 'sequin' RNA standards (19) from two mixes (A and B) were prepared and sequenced. These samples contain the same transcripts but at variable molar ratios to simulate biological differences in gene expression and alternative splicing. Among the 76 synthetic genes, 21 were up-regulated and 23 were down-regulated in mix B compared to mix A. The corresponding transcripts of 28 genes were expressed at different proportions between the two mixes, resulting in DTU for 62 out of 160 transcripts. A further two sequin mix A and two mix B samples were sequenced using Illumina short-read sequencing technology together with RNA from human lung adenocarcinoma cell lines.

### Biological materials

Synthetic 'sequin' RNA standards were obtained from the Garvan Institute of Medical Research.

NSCs were derived as described in Chen *et al.* (21). Cells were grown in NeuroCult Stem Cell medium (StemCell Technologies #05702) with cytokines: NeuroCult NSC Basal Medium (Mouse) (StemCell Technologies #05700) supplemented with NeuroCult Proliferation Supplement (Mouse) (StemCell Technologies #05701), 0.25 mg/mL rh EGF (StemCell Technologies #02633) and 0.25 mg/ml rh bFGF (StemCell Technologies #02634). We extracted total RNA with Trizol and purified polyA RNA with the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490).

### Nanopore sequencing and data preprocessing

Sequin cDNA libraries were constructed with SQK-PCS109 cDNA-PCR sequencing and SQK-PBK004 PCR Barcoding kits using the supplied protocol. Briefly, duplicate libraries of each mix (A1, A2, B1 and B2) were constructed using 15 ng as input for cDNA synthesis. Samples were barcoded 1 to 4 using the supplied PCR barcodes. Transcripts were amplified by 14 cycles of PCR with a 6-min extension time.

Sequencing libraries were individually purified using Beckman Coulter 0.8x AMPure XP beads and quantified using an Invitrogen Qubit 4.0 Fluorometer (ThermoFisher Scientific). Equimolar amounts of each sample were pooled to a total of approximately 160 fmol (assuming median transcript size is 1 kb), and quality control of the pooled library was performed using Agilent Technologies TapeStation 4200. The final library was loaded onto an R9.4.1 MinION flow cell and sequenced for 65 h with a buffer refuel at 24 h (using 250 ml buffer FB) using the ONT GridION platform. The *fast5* files were base-called by *Guppy* version 4.0.11 using configuration file *dna_r9.4.1_450bps_hac.cfg* to obtain `fastq` files, trim adaptor sequences and demultiplex barcoded reads. *Guppy* is only available to ONT customers via the community site (https://community.nanoporetech.com/).

For the NSC dataset we prepared direct-cDNA libraries from 40 ng purified polyadenylated RNA. We combined the ONT direct-cDNA sequencing (SQK-DCS108) protocol (version DCB_9036_v108_revG_30Jun2017) with the one-pot native barcoding protocol (http://lab.loman.net/protocols/) with extended incubation times (using SQK-LSK109 and EXP-NBD103 kits) for library preparation of the first batch, and used the updated kits SQK-DCS109 and EXP-NBD114 for the second batch (protocol PDCB_9093_v109_revA_04Feb2019). We loaded 100 ng of the final libraries on one PromethION flow cell (FLO-PRO002) per batch. The `fast5` files were basecalled by *Guppy* version 4.0.11 using configuration file *dna_r9.4.1_450bps_hac_prom.cfg* to yield `fastq` files. *Guppy* was also used to trim adaptor sequences from reads and demultiplex barcoded reads. The 'other' samples were removed in downstream analysis. For an overview of our analysis pipeline see Figure 1A.

## Illumina sequencing

About 6 ng of sequins mix A or B were added to 1 μg of total RNA from human lung adenocarcinoma cell line NCI-H1975 (mix A) and HCC827 (mix B). The pooled RNA was subjected to the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490) and NEBNext Ultra II Directional RNA Library Prep (E7760) for Illumina following manufacturer's instructions and default fragmentation conditions (15 min at 94°C), 5-fold adaptor dilution and a final amplification of nine PCR cycles. Indexed and pooled libraries were sequenced on a NextSeq500 (Illumina) High Output, 81 cycles paired-end.

## Genomic alignment

The NSC reads were aligned to mouse *mm10* genome using *minimap2* version 2.17-r943-dirty (22) to get `bam` files. The genome alignments were performed with the arguments

*-ax splice -uf -k14 –junc-bed*, allowing spliced alignments on the forward transcript strand to map with higher sensitivity. It also uses annotated splice junctions to improve the accuracy of mapping at junctions. Gencode release M23 (GRCm38.p6) annotation (23) was used to provide information on known splicing junctions. The sequins ONT reads were mapped to the artificial chromosome *chr*IS_R using *minimap2* with the arguments *-ax splice –MD*. The `bam` files were sorted and indexed using *samtools* version 1.6 (24). The sequin Illumina reads were mapped to *chr*IS_R using *Subread* version 1.6.3 (25).

## Gene abundance estimation

Mapped reads were assigned to individual genes and counted by the *featureCounts* (26) function in the R (https://www.R-project.org/) Bioconductor (27) package *Rsubread* version 1.34.4 (25,28). We used the Gencode release M23 (GRCm38.p6) annotation for the NSC data, and sequins annotation GTF file version 2.4 for the sequins data. Arguments *isLongRead=TRUE* (as recommended in the *featureCounts* help page when dealing with Nanopore data) and *primaryOnly=TRUE* to count primary alignments only were specified.

## Differential gene expression analysis

Genes in the NSC dataset were annotated using the R/Bioconductor package *Mus.musculus* (https://bioconductor.org/packages/Mus.musculus/ version 1.3.1) and read counts from technical replicates of the same sample run across different batches were combined (i.e. summed together). For all datasets, we organized and preprocessed the count data using the R/Bioconductor package *edgeR* version 3.26.8 (18,29). Lowly expressed genes were removed using the *filterByExpr* function with default arguments. Normalization factors were calculated using the trimmed mean of *M*-values method (30). Differential gene expression (DGE) analysis was performed using the *limma-voom* pipeline version 3.40.6 (10,31,32), with sample-specific quality weights (33). Linear models were fitted with either genotype or sequin mix information to create the design matrix, followed by empirical Bayes moderation of *t*-statistics (34). Raw *P*-values were adjusted for multiple testing (35).

## Mouse NSC short-read data

We obtained DGE results from a previous Illumina short-read RNA-seq study comparing mouse *Smchd1*-null and WT NSC samples (21,36) available at http://bioinf.wehi.edu.au/folders/smchd1/ and from GEO (accession number GSE65747). Using a *limma-voom* pipeline, the study reported 1197 differentially expressed (DE) genes (adjusted *P*-value cutoff of 0.01). We further restricted this list (adjusted *P*-value cutoff of 0.0001) to give us 218 up- and 54 down-regulated genes in *Smchd1*-null samples when compared to WT samples. This cutoff resulted in similar numbers of significant genes between the short- and long-read datasets. The DE genes were compared to that of the NSC long-read data using ROAST gene set testing (37) with 9999 rotations.

## Transcript-level analysis

We used three different tools to perform isoform detection and quantification: *FLAIR* version 1.5 (commit 2f8df51) (6) and *TALON* version 5.0 (commit c41b9dc) (5) for the sequins data, and *FLAMES* version 0.1.0 (available at https://github.com/XueyiDong/FLAMES) (38) for both sequins and NSC datasets. Default parameters were used to run *FLAIR*. *TranscriptClean* (39) version 1.02 which performs reference-based error correction was applied prior to running *TALON* version 5.0. Transcripts identified by *TALON* were filtered using default setting.

*FLAMES* (short for '**F**ull-**L**ength tr**A**nscript quantification, **M**utation and **S**plicing analysis') is a novel method and software tool developed for long-read RNA-seq data (38), available at https://github.com/LuyiTian/FLAMES. It requires sorted `bam` files with reads aligned to the genome as input. *FLAMES* first summarizes the alignment results by grouping reads with similar splice junctions to get a raw isoform annotation. The raw isoform annotation is compared against the reference annotation to correct potential splice site and transcript start/end errors. Transcripts that have similar splice junctions (<5 bp by default) and transcript start/end (<100 bp by default) to the reference tran-

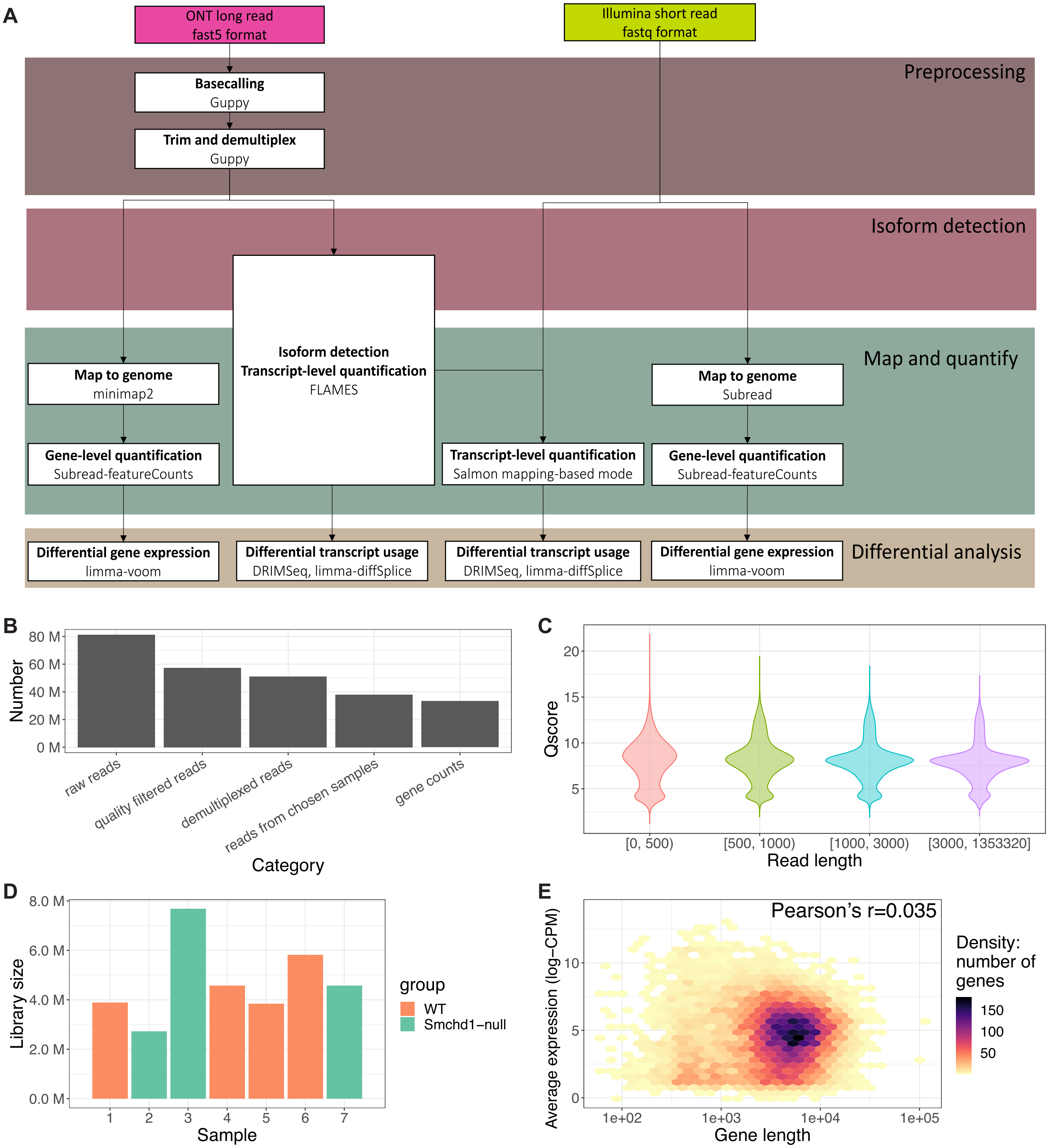

**Figure 1.** Analysis workflow and quality metrics. (**A**) Overview of the analysis workflow used to process the mouse NSC direct-cDNA long-read and short-read RNA-seq data. (**B**) The number of raw reads, quality filtered reads, trimmed and demultiplexed reads, reads from chosen samples and gene-level counts in the NSC dataset. (**C**) Distribution of read quality in the NSC dataset, stratified by read length. Read quality is defined by the average base quality score of a read. (**D**) The total number of reads assigned to each sample in the NSC dataset (green: *Smchd1*-null samples; orange: WT samples). (**E**) A hexagonal 2D density plot showing the correlation between gene length and average gene expression (log-CPM) in the NSC dataset.

script are merged with the reference. High-confidence isoforms are those identified with at least 10 supporting reads. This process will also collapse isoforms that are likely to be truncated transcripts, which are identified as having incomplete 3' regions compared to the full-length forms. A draft transcript assembly is generated at this stage, and all reads are re-aligned to both the known and assembled transcripts and quantified. For reads that align to multiple transcripts, both the proportion of transcript covered by the read alignment and the percentage of read sequence that aligns to each transcript are considered, with the read assigned to the transcript with highest coverage and >80% of its sequence aligned. A configuration file (JSON format) is used to set all parameters mentioned above so that different parameters can used to adapt to different situations. For example, one could set more stringent parameters for merging splice junctions (i.e. decreasing the default from 5 to 1 bp) to better suit long-read data generated by the PacBio platform, which generally has a lower error rate.

After running each isoform detection tool, *SQANTI* (40) version 1.2 was used to classify identified isoforms into structural categories by comparing them to the annotation. An isoform matching a reference transcript can be categorized as 'full splice match' or 'incomplete splice match', based on whether all splice junctions are matched. 'Novel in catalog' is the category for novel isoforms of known genes containing new combinations of already annotated junctions or novel splice junctions composed of annotated donors and acceptors. 'Novel not in catalog' stands for the novel isoforms with novel donors and/or acceptors. Other isoforms can be categorized as either 'antisense', 'fusion' or 'genic/intergenic' isoforms. For both Illumina short-read datasets, *Salmon* version 1.3.0 (41) 'mapping-based mode' was used to obtain quantification of isoforms identified in corresponding ONT long-read data by *FLAMES*. Lowly expressed genes and transcripts were removed from downstream analysis using the *dmFilter* function from the *DRIMSeq* package. For all long- and short-read datasets, genes were required to have an associated gene count (obtained by summing counts across all transcripts for a given gene) of 10 or more in every sample. A second filter required transcripts to have 10 or more counts in at least 3 samples in the NSC data, and in at least 2 samples in the sequins data.

DTU analysis was performed using two methods: the R/Bioconductor package *DRIMSeq* version 1.12.0 (9), and *diffSplice* from the *limma* package version 3.40.6. Originally, *DRIMSeq* was designed for use with transcript-level counts in short-read data, giving adjusted *P*-values at both the gene-level and feature-level (transcripts). *DiffSplice* analyses exon-level counts in short-read data to indirectly call for differences in isoform proportions, and reports adjusted *P*-values at the gene-level (Simes adjustment and/or *F*-tests) and exon-level (*t*-tests). For long-read data, we applied the *diffSplice* to transcript-level counts rather than exon-level counts as carried out by Love *et al.* (11). Additionally, the stage-wise method from R/Bioconductor package *stageR* version 1.6.0 (12) was also applied to the raw *P*-values from *DRIMSeq* (gene- and transcript-level) and *diffSplice* (Simes and *t*-tests) methods for FDR control to give *stageR* gene- and transcript-level adjusted *P*-values.

### Data and code availability

RNA-seq data can be accessed from Gene Expression Omnibus (GEO) under accession numbers GSE151984 (sequins long-read), GSE151841 (NSC long-read data) and GSE164598 (sequins short-read). All code used to perform these analyses are available from https://github.com/XueyiDong/LongReadRNA.

## RESULTS

### Data quality

To assess the quality of our long-read datasets, raw long reads were pre-processed and assigned to gene-level counts using an appropriate reference genome. Figure 1B shows the number of reads (or amount of information) retained after some crucial steps in processing the NSC data. A total of ∼81 million reads were successfully sequenced and base-called. Approximately 70%, or ∼57 million reads had an average base quality score >7 and passed quality filtering. Around 51 million quality-passed reads were detected with adaptor and barcode sequences in the trimming and demultiplexing steps. The reads from our samples of interest were then mapped to the genome and assigned to genes, producing ∼33 million gene-level counts. For the sequins dataset, ∼11 million raw reads yielded ∼6.4 million gene-level counts (Supplementary Figure S1). In comparison, the sequins short-read dataset had ∼162 million gene-level counts.

In the NSC dataset, median read length is 752 bases. Figure 1C shows that shorter reads tend to have higher median read quality, but the difference is subtle. The quality of 'extra long' reads (≥3000 bases) were similar to reads of other length categories, indicating Nanopore's ability to detect transcripts in this size range. A small proportion of reads (∼1%) exceed 5 kilobases. Similar to the NSC dataset, the sequins dataset had a median read length of 716 bases, which is shorter than its expected value of 908 bases. The library size (sum of gene counts) of samples in the NSC dataset varied between 2.7 and 7.7 million reads (Figure 1D). In comparison, the library size of samples in the short-read NSC study (21) range between 18.6 and 23.2 million reads.

### Gene expression analysis

In short-read RNA-seq, transcripts (or genes) are fragmented for sequencing, such that longer transcripts can be over-represented relative to transcripts that are shorter. As a consequence, DGE analyses are biased towards the detection of genes (or transcripts) that are relatively long (42). Also, DGE analyses may be confounded by DTU such that gene-level counts are affected by the varying proportions of transcripts with varying lengths. One advantage to long-read RNA-seq protocols is that they do not include the fragmentation step, and should theoretically be unbiased to gene length. To examine this, we looked at the relationship between gene length and gene expression using $\log_2$-counts-per-million (log-CPM) values. Gene length is weakly associated with expression in both long-read datasets; Pearson correlation of 0.035 for the NSC dataset (Figure 1E) and

-0.056 for the sequins dataset (Supplementary Figure S2), whereas the equivalent correlation in the short-read NSC study (21) is greater at 0.20.

We applied the *limma*-*voom* workflow designed for short-read DGE analysis to the long-read data. We first analysed the sequins data to check whether the approach was appropriate using the ground-truth available. The analysis was carried out using *voomWithQualityWeights* to account for sample-level heterogeneity by estimating sample-specific quality weights based on similarity of gene expression within the same group. The sample weights were combined with *voom* precision weights that are based on the mean-variance relationship estimated from the data. Even though there were only 69 genes present in the dataset (as opposed to tens of thousands in a typical dataset), the mean-variance trend observed for the sequins data (Supplementary Figure S3) was similar to that of short-read RNA-seq data (31).

Linear modelling on the gene-level counts were carried to obtain estimated $\log_2$fold-change (logFC) values between mix A and B. Estimated values were highly correlated ($R^2 = 0.934$) with expected logFCs (Figure 2A). Using an adjusted *P*-value cutoff of 0.05, 21 down-regulated and 18 up-regulated genes were detected between mix B and A. There were no false discoveries, and only 2 of the true differentially expressed genes were not detected.

The same DGE analysis workflow was also applied to the short-read sequins data, with 2 false negative and 3 false positive discoveries made. Upon checking the annotation, we found that all of false positive genes in the short-read data were long (>2400 bp) DTU genes that were lowly expressed (abundance $\leq$ 0.24; median abundance among all genes = 7.55). The *t*-statistics from DGE analysis of short-read data were observed to be highly correlated ($R^2 = 0.838$) with the *t*-statistics from the long-read data (Figure 2B). Overall, results from the sequin synthetic control data indicate that the *limma*-*voom* pipeline is powerful and reliable when applied to long-read data, so we next applied it to the NSC dataset.

Unsupervised clustering by multidimensional scaling (MDS) was used to observe the relationships between NSC samples. Dimension 1 in the MDS plot roughly separates samples by genotype (Figure 2C), although a *Smchd1*-null sample (sample 7) is positioned more closely to WT samples. The mean-variance trend for this dataset is again typical of what is observed in short-read RNA-seq experiments (31) (Figure 2D). Estimated sample weights favoured samples that distinguished groups across dimension 2 of the MDS plot, giving samples 2 and 3 in the *Smchd1*-null group weights that are >1, as well as samples 1 and 5 in the WT group (Supplementary Figure S4). Using the default adjusted *P*-value cutoff of 0.05, only 12 genes were detected as DE. Using a more liberal adjusted *P*-value cutoff of 0.25 to account for the small library sizes, detected 81 down-regulated and 63 up-regulated genes between *Smchd1*-null and WT samples (Figure 2E). The *Smchd1* gene, which was depleted in *Smchd1*-null samples, was detected as the most significantly down-regulated gene in the comparison (highlighted in Figure 2E) and serves as a positive control for this analysis.

In a previous short-read study on the same mouse NSC groups (21,36), the imprinted genes *Ndn*, *Mkrn3* and *Peg12* were reported as up-regulated. These genes were also found to be DE in the long-read dataset (highlighted in Figure 2E). Further comparison between the short- and long-read datasets was carried out using a barcode plot (Figure 2F). The barcode plot shows that genes that were up-regulated in the short-read dataset (red vertical lines in the plot) also tend to be up-regulated in our long-read dataset (positioned towards the right of the plot). Specifically, the genes that were most highly up-regulated in the short-read dataset as ranked by logFC (long red vertical lines), are also highly up-regulated in the long-read data (further right in the plot). The same goes for down-regulated genes in the short-read dataset (blue vertical lines in the plot), which tend to be down-regulated in the long-read dataset (positioned towards the left of the plot). We tested concordance of the datasets formally by applying the ROAST gene set testing method (37) to our long-read data. Using both up- and down-regulated gene sets from the short-read dataset, weighted on their logFC values, ROAST returned an 'up' *P*-value of 0.086, which indicates that transcriptional changes for the comparison of *Smchd1*-null versus WT are somewhat consistent between the two datasets (up-regulated genes in the short-read data tend to be up-regulated in the long-read data, and down-regulated genes in the long-read data tend to be down-regulated in the long-read data). The relatively large ROAST *P*-value and overall lack of power to detect differentially expressed genes is likely due to relatively low sequencing levels per sample and within-genotype sample heterogeneity in the long-read experiment. Smoothed scatter plots of the logFCs and *t*-statistic for each gene between the short- and long-read datasets are presented in Supplementary Figure S5. Discordance between the long- and short-read data is likely caused by the fact that samples were not perfectly matched. The short-read experiment, which was performed a few years earlier, used samples derived from different animals which may have been at a slightly different developmental stage compared to the newer samples profiled using long-read RNA-seq. In addition, since these samples are from primary cells in culture, the number of passages post derivation can also influence gene expression and any subsequent differential expression results.

### Transcript-level analysis

Transcript-level analysis of nanopore RNA-seq data usually starts with isoform detection. To test which tool is best suited to nanopore data, we compared two popular tools, *FLAIR* and *TALON* with our novel *FLAMES* pipeline on the sequins dataset. Ideally, all transcripts that appear in the sequins annotation should be detected, and there should be no novel isoforms. Our results showed that *FLAMES* detected the most sequin transcripts (Figure 3A, 'full splice match' category) and fewer artefactual isoforms (Figure 3A, other categories). While most sequin transcripts were also detected by *FLAIR*, a large number of artefactual isoforms were also identified, especially those classified as 'novel in catalog' for which we know there should be none. *TALON* detected ~77% of the sequin transcripts, and a disproportionately large number of artefactual isoforms, especially in the 'antisense' and 'novel not in catalog' categories. When we looked into the number of reads assigned

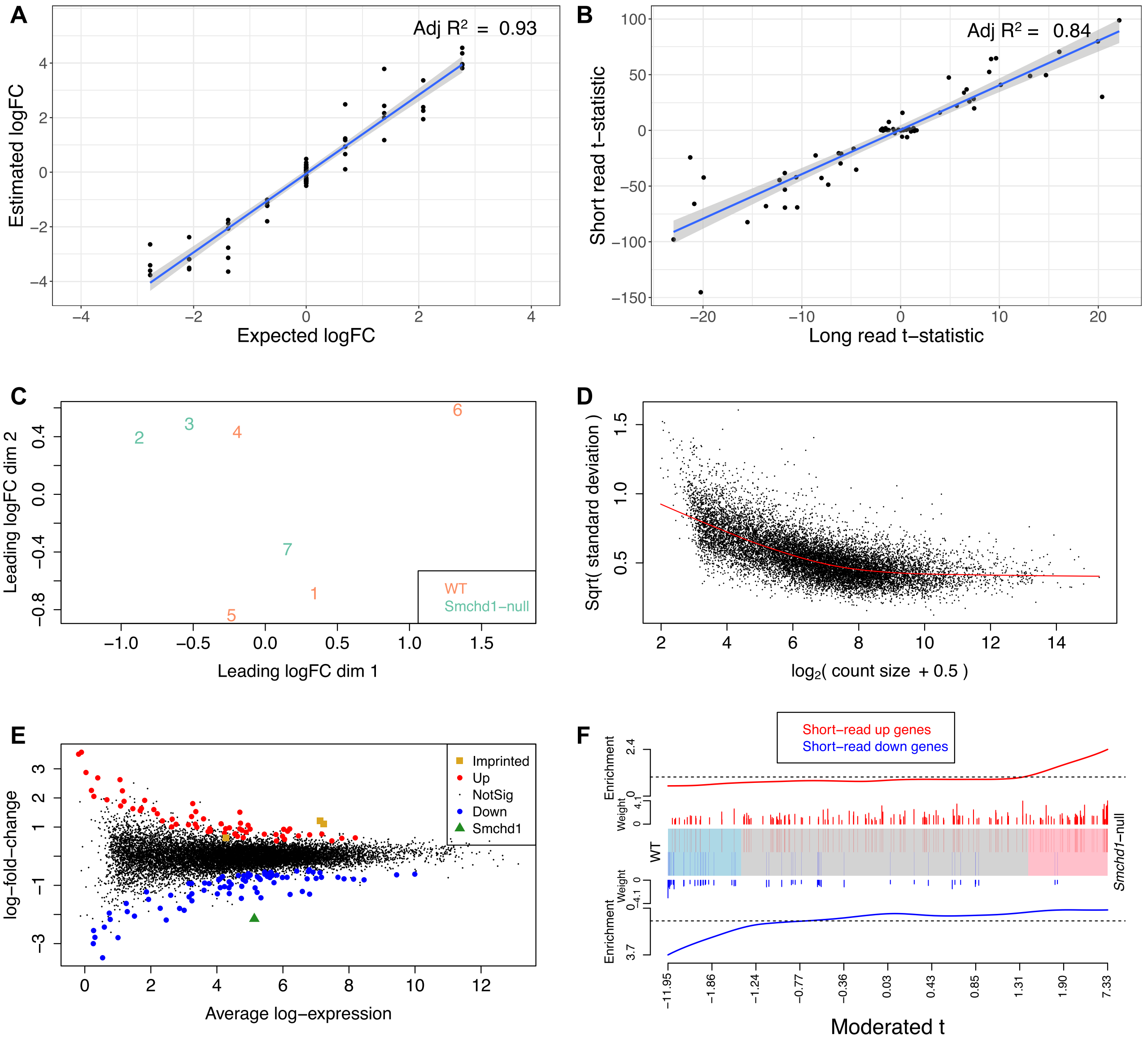

**Figure 2.** Results for differential gene expression analysis. (**A**) Scatter plot of the observed logFC between mix A and B versus the expected logFC in the sequins long-read data. The blue line is the linear regression line. (**B**) Scatter plot of the *t*-statistics calculated between mix A and B from the sequins short-read and long-read data. The blue line is the linear regression line. (**C**) MDS plot showing the relationship between NSC samples based on gene-level logCPM. (**D**) Voom mean-variance trend in NSC data where points represent genes. (**E**) Gene-level plot of logFC for *Smchd1*-null versus WT plotted against average $\log_2$-expression values. Differentially expressed genes are highlighted (red: up-regulated genes, blue: down-regulated genes). (**F**) The barcode plot shows the correlation between our long-read differential expression results and the results from a previous short-read dataset collected on the same NSC sample types. Each vertical bar represents a DE gene from the previous short-read study (red: up-regulated genes, blue: down-regulated genes), and the position of the bar on the *x*-axis represents the moderated *t*-statistic of the same gene in the long-read results. The length of the vertical lines represent the logFC of the gene in the short-read results. The red worm on the top and the blue worm at the bottom represent the relative enrichment of the vertical bars in each part of the plot with the smooth fit obtained using a moving average with tricube weights.

to transcripts in each category (Figure 3B), the majority of counts in *FLAMES* were from known isoforms, while more than half of the counts in *FLAIR* and *TALON* were from artefactual isoforms. The total number of read counts from *FLAMES* (∼4.9 M) and *TALON* (∼4.2 M) are similar, while *FLAIR* recovered a lower number (∼2.7 M). Results from the sequins dataset indicated that the novel *FLAMES* pipeline outperformed the other two methods.

To further assess the performance of *FLAMES* and the quality of the dataset, we calculated the coverage fraction of transcripts by individual reads. Here, reads covering 95% or more of the bases of their corresponding transcript are defined as 'full-length'. In our sequins data, 47% of reads were found to be full-length, which is similar to Gleeson *et al* (14)'s ONT MinION direct RNA dataset. Reads assigned to longer transcripts are less likely to be full-length (Figure 3C), consistent with findings from Jenjaroenpun *et al.* (16). This suggests that some reads may be truncated in our sequins dataset, which presumably occurs during library preparation. The truncated reads may have resulted

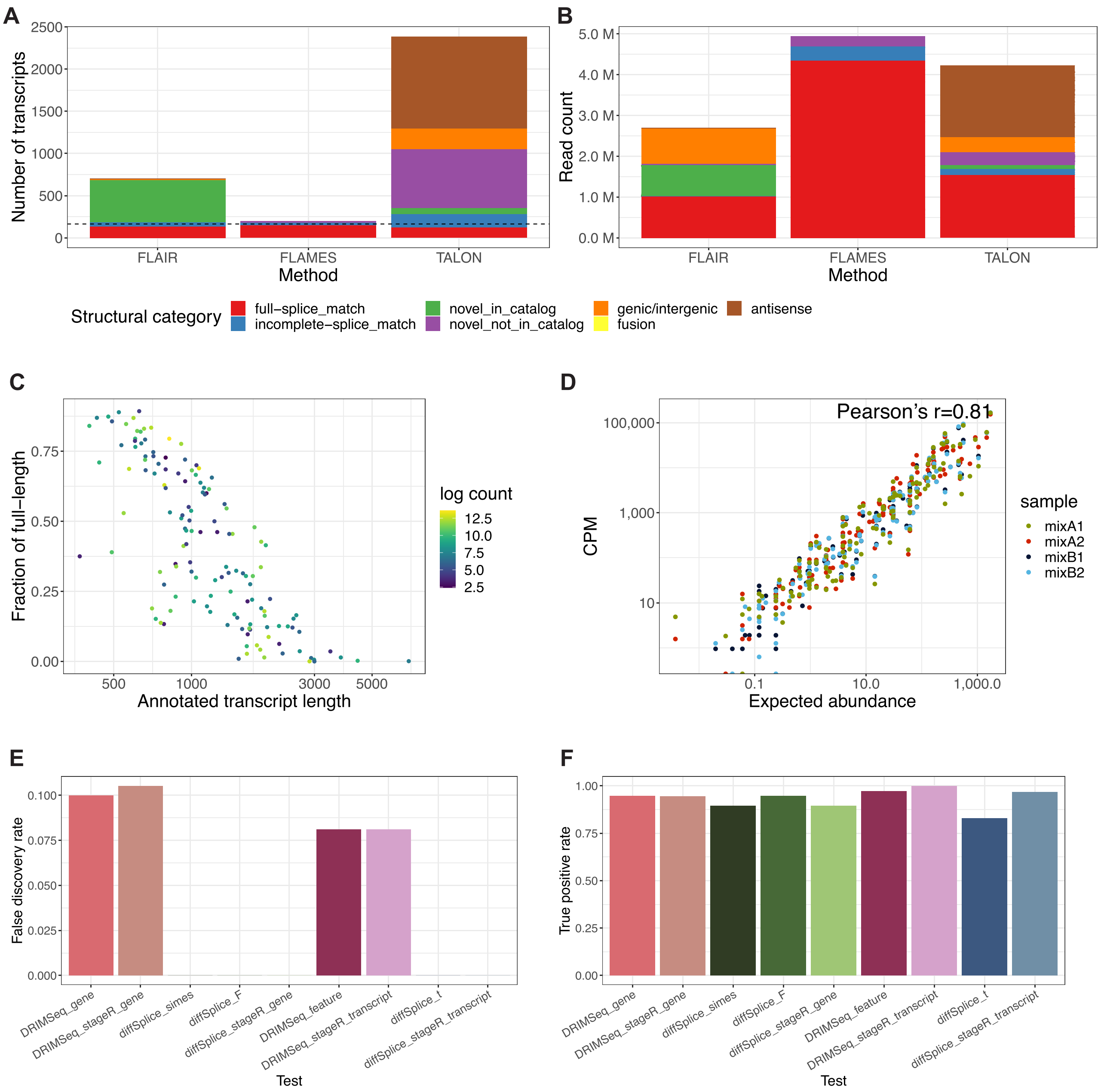

**Figure 3.** Isoform identification and differential transcript usage analysis. (**A**) A bar plot showing the number of discovered isoform types in the sequins long-read dataset. The bars are separated into isoform categories (by colour), and the dashed line represents the true number of isoform types. The red 'full-splice-match' category represents the known transcripts present in the sequin controls (i.e. true positive), while the other categories represent erroneous transcripts. (**B**) A bar plot showing the number of counts from isoforms in the sequins long-read dataset. The bars are separated into isoform categories (by colour) from which the counts are associated with. (**C**) A scatter plot showing the correlation between the fraction of full-length reads assigned to a transcript and the length of the annotated transcript. Dots are coloured by the transcript count (log$_2$-scale). (**D**) The correlation between observed transcript counts and expected transcript abundance of each gene from each sequins sample. **(E)** A bar plot showing the false discovery rate (FDR) from different tests of DTU in sequins long-read data. (**F**) A bar plot showing the true positive rate (FDR) from different tests of DTU in sequins long-read data.

in the detection of 'incomplete splice match' isoforms (blue in Figure 3A and 3B).

Next, we used *DRIMSeq* and the *diffSplice* function in *limma* for DTU analysis of long-read data. We also combined the methods with the stage-wise analysis from the *stageR* (12) package since it was recommended in the *DRIMSeq* vignette for statistical improvement and enhanced biological interpretation of results. We expect good performance of DTU analyses comparing mixes A and B since the observed CPM values of sequin transcripts were highly correlated with their expected abundances (Pearson correlation = 0.81, Figure 3D). We expect up to 5% of false discoveries by applying at adjusted *P*-value cutoff of 0.05. *diffSplice* methods performed better than *DRIMSeq* overall with no false discoveries and high true positive rate (TPR > 0.82) (Figure 3E and 3F, Supplementary Figure S6). *DRIMSeq* had comparable TPR, but did not properly control the FDR, such that more false discoveries were found (FDR between 0.08 and 0.10) than expected. For transcript-level testing, *stageR* improved the TPR of both methods whilst maintaining the same FDR. However, *stageR* did not improve the performance in gene-level testing for either of the two methods (i.e. the TPRs were unchanged while the FDR increased for *DRIMSeq* by 0.005.

We also quantified the expression of the same transcripts and performed DTU analysis on short-read sequins data. The observed transcripts per kilobase million (TPM) values were also highly correlated with their expected abundances (Pearson correlation = 0.86, Supplementary Figure S7). However, the FDR of all DTU tests were higher than that of long-read data, and were not properly controlled (FDRs ranged between 0.10 and 0.17 using an adjusted *P*-value cutoff of 0.05, Supplementary Figure S8). The higher error rates in short-read data suggests that both methods performed better on long-read data than on the short-read data that they were originally designed for use on. We examined the concordance between the datasets by comparing the top *n* most significant DTU genes. For both *DRIMSeq* and *diffSplice*, in the top 19 genes, which is the number of true DTU genes in the datasets, there were 16 genes in common between the two methods (Supplementary Figure S9). We demonstrate that our pipeline and combination of methods for transcript-level analysis produces accurate transcript quantification and identification of DTU, and provides confidence for application to other long-read transcriptomic datasets.

We then applied our transcript-level analysis workflow to the NSC dataset. The *FLAMES* pipeline returned 38 857 unique isoforms from 9837 genes, of which 38% were classified as novel (Supplementary Figure S10), which is a lot more than what was observed in the sequins dataset. Since we observed a very low level of falsely discovered isoforms in the sequins data, we assume that that most of these novel isoforms are real, which suggests that the current mouse transcript annotation is incomplete. Of the mapped reads, 32.3% were assigned to novel isoforms, the majority of which were from the 'novel not in catalog' category (Supplementary Figure S11). Using an adjusted *P*-value cutoff of 0.05, *DRIMSeq* found one gene *Pisd* as having DTU in the *Smchd1*-null versus WT comparison, while *diffSplice* did not detect any DTU genes, although *Pisd* had the smallest *P*-value (raw *P*-value = 0.0002) among all the genes tested. Transcript *ENSMUST00000201980.4* ('Known5') and *ENSMUSG00000023452.19_32736305_32746312_1* ('Novel2') in *Pisd* was identified by *DRIMSeq* to have differential usage between the two groups (Supplementary Figures S12 and S13).

We also performed DTU analysis on the NSC short-read data. Based on the sequins results, DTU analysis on short reads may be less reliable, with more false discoveries than expected. Using an adjusted *P*-value cutoff of 0.05, *DRIMSeq* found 139 DTU genes, while *diffSplice* only found 9 DTU genes (Supplementary Figure S14). There were three DTU genes (*Pisd*, *Cyth2* and *Pabpn1*) found in common by both methods. Transcript *ENSMUST00000201980.4* in *Pisd* was identified by both methods to have a higher usage in *Smchd1*-null samples than in WT samples (Supplementary Figure S15, 'Known5'), which was concordant with what was observed in the long-read analysis. Since we observed high TPRs in our DTU analysis for the sequins short-read data and the library size of the NSC short-read dataset is relatively high, we expect our analysis to have sufficient power to detect DTU genes if they were present in this dataset.

## DISCUSSION

Our DGE analysis uses a *limma*-*voom* workflow and shows that results from PCR-cDNA and direct-cDNA long-reads are reliable, such that estimated results are comparable to the known truth in the sequins synthetic control dataset, and concordant with corresponding short-read studies. Although the total library size in the sequins dataset is lower than that of the NSC dataset, more reads were assigned per gene since the dataset contains a small set of genes, which improved power for DGE analysis. Overall, comparisons using long-read experiments suffer from a lack of statistical power due to low library sizes. It would be desirable for long-read transcriptomic studies to have total read numbers that are more comparable to what is routinely achieved in short-read experiments (20–50 million reads per sample is not unusual). We expect this to occur in the near future as throughput of long-read experiments increases.

We also looked into transcript-level analysis of long-read data and found our novel *FLAMES* pipeline to be reliable in both isoform detection and quantification. The high false positive rate of *FLAIR* and *TALON* for isoform detection suggests that these algorithms need further improvement to adapt to the high error rates in long-read sequencing. Despite methods being designed originally for short-read data, *diffSplice* (with or without *stageR*) performed very well in DTU analyses of the sequins long-read data using transcript-level counts, with better FDR control observed in long-read DTU analysis than in the short-read DTU analysis. We believe these methods could be applied to other datasets with confidence, but may lack power to detect DTU genes if transcript counts are very low. Notably, relative to DGE analyses, a DTU analysis further splits gene-level counts into associated isoforms which reduces power for statistical testing. For this reason, the power to detect DTU genes in the NSC long-read dataset is reduced relative to the sequins dataset since the latter contains far fewer expressed genes and transcripts to begin with, such that transcripts have higher counts on average. Another potential issue in the application of these methods to our NSC dataset is that altered expression of the gene *Smchd1* may not affect RNA splicing mechanisms, which could mean there is a biological explanation for why very little DTU was observed in these data across both the short and long-read analyses.

Our study is the first to test a pipeline for both gene-level DGE analysis and transcript-level DTU analysis of nanopore long-read RNA-seq data. Whilst the sequencing depth is relatively low, we are still able to obtain reasonable results using pre-existing methods designed for short reads, namely the *limma* software. We expect that other short-read tools such as *edgeR* and *DESeq2* may also be appropriate, as used in other studies (14,17) although further benchmarking efforts are required to confirm this. Exploring the strengths and weaknesses of different analysis methods on data arising from both the Nanopore and PacBio long-read platforms using a specially designed benchmarking dataset is planned as future work.

We hope that our analysis will encourage further research into the potential for long-read RNA-seq to be used in place of short-read RNA-seq, allowing for the simultaneous exploration of gene-level and isoform-level changes within the same experiment in a more comprehensive way.

## SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

## FUNDING

National Health and Medical Research Council (NHMRC) Project Grant [GNT1098290 to M.E.B., M.E.R.]; NHMRC Career Development Fellowship [GNT1104924 to M.E.R.]; NHMRC Early Career Fellowship [GNT1072662 to M.B.C.]; Bellberry-Viertel Senior Medical Research Fellowship (to M.E.B.); Melbourne Research Scholarship (to X.D., L.T.); Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS.

*Conflict of interest statement.* MBC and RDP have received support from Oxford Nanopore Technologies (ONT) to present their findings at scientific conferences. However, ONT played no role in study design, execution, analysis or publication.

## REFERENCES

1. Pollard,M.O., Gurdasani,D., Mentzer,A.J., Porter,T. and Sandhu,M.S. (2018) Long reads: their purpose and place. *Hum. Mol. Genet.*, **27**, R234–R241.
2. Gupta,I., Collier,P.G., Haase,B., Mahfouz,A., Joglekar,A., Floyd,T., Koopmans,F., Barres,B., Smit,A.B., Sloan,S.A. *et al.* (2018) Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat. Biotechnol.*, **36**, 1197–1202.
3. Sahlin,K. and Medvedev,P. (2020) De novo clustering of long-read transcriptome data using a greedy, quality value-based algorithm. *J. Comput. Biol.*, **27**, 472–484.
4. Au,K.F., Sebastiano,V., Afshar,P.T., Durruthy,J.D., Lee,L., Williams,B.A., van Bakel,H., Schadt,E.E., Reijo-Pera,R.A., Underwood,J.G. *et al.* (2013) Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl. Acad. Sci. USA*, **110**, E4821–E4830.
5. Wyman,D., Balderrama-Gutierrez,G., Reese,F., Jiang,S., Rahmanian,S., Zeng,W., Williams,B., Trout,D., England,W., Chu,S. *et al.* (2019) A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. bioRxiv doi: https://doi.org/10.1101/672931, 18 June 2019, pre-print: not peer reviewed.
6. Tang,A.D., Soulette,C.M., van Baren,M.J., Hart,K., Hrabeta-Robinson,E., Wu,C.J. and Brooks,A.N. (2020) Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. Commun.*, **11**, 1–12.
7. Li,R., Ren,X., Ding,Q., Bi,Y., Xie,D. and Zhao,Z. (2020) Direct full-length RNA sequencing reveals unexpected transcriptome complexity during Caenorhabditis elegans development. *Genome Res.*, **30**, 287–298.
8. Byrne,A., Beaudin,A.E., Olsen,H.E., Jain,M., Cole,C., Palmer,T., DuBois,R.M., Forsberg,E.C., Akeson,M. and Vollmers,C. (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.*, **8**, 1–11.
9. Nowicka,M. and Robinson,M.D. (2016) DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Res*, **5**, 1356.
10. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
11. Love,M.I., Soneson,C. and Patro,R. (2018) Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Res*, **7**, 952.
12. Van den Berge,K., Soneson,C., Robinson,M.D. and Clement,L. (2017) stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome Biol.*, **18**, 151.
13. Soneson,C., Yao,Y., Bratus-Neuenschwander,A., Patrignani,A., Robinson,M.D. and Hussain,S. (2019) A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.*, **10**, 3359.
14. Gleeson,J., Lane,T.A., Harrison,P.J., Haerty,W. and Clark,M.B. (2020) Nanopore direct RNA sequencing detects differential expression between human cell populations. bioRxiv doi: https://doi.org/10.1101/2020.08.02.232785, 02 August 2020, pre-print: not peer-reviewed.
15. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
16. Jenjaroenpun,P., Wongsurawat,T., Pereira,R., Patumcharoenpol,P., Ussery,D.W., Nielsen,J. and Nookaew,I. (2018) Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of Saccharomyces cerevisiae CEN.PK113-7D. *Nucleic Acids Res.*, **46**, e38.
17. Cruz-Garcia,L., O'Brien,G., Sipos,B., Mayes,S., Love,M.I., Turner,D.J. and Badie,C. (2019) Generation of a transcriptional radiation exposure signature in human blood using long-read nanopore sequencing. *Radiat. Res.*, **193**, 143.
18. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
19. Hardwick,S.A., Chen,W.Y., Wong,T., Deveson,I.W., Blackburn,J., Andersen,S.B., Nielsen,L.K., Mattick,J.S. and Mercer,T.R. (2016) Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat. Methods*, **13**, 792–798.
20. Blewitt,M.E., Gendrel,A.V., Pang,Z., Sparrow,D.B., Whitelaw,N., Craig,J.M., Apedaile,A., Hilton,D.J., Dunwoodie,S.L., Brockdorff,N. *et al.* (2008) SmcHD1, containing a structural-maintenance-of-chromosomes hinge domain, has a critical role in X inactivation. *Nat. Genet.*, **40**, 663–669.
21. Chen,K., Hu,J., Moore,D.L., Liu,R., Kessans,S.A., Breslin,K., Lucet,I.S., Keniry,A., Leong,H.S., Parish,C.L. *et al.* (2015) Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. *Proc. Natl. Acad. Sci. USA*, **112**, E3535–E3544.
22. Li,H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
23. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J. *et al.*

(2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
24. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
25. Liao,Y., Smyth,G.K. and Shi,W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, **41**, e108.
26. Liao,Y., Smyth,G.K. and Shi,W. (2014) FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
27. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L., Girke,T. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.
28. Liao,Y., Smyth,G.K. and Shi,W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47.
29. McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.*, **40**, 4288–4297.
30. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
31. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
32. Law,C.W., Alhamdoosh,M., Su,S., Smyth,G.K. and Ritchie,M.E. (2016) RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Res*, **5**, 1408.
33. Liu,R., Holik,A.Z., Su,S., Jansz,N., Chen,K., Leong,H.S., Blewitt,M.E., Asselin-Labat,M.-L., Smyth,G.K. and Ritchie,M.E. (2015) Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res.*, **43**, e97.
34. Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, doi:10.2202/1544-6115.1027.
35. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B*, **57**, 289–300.
36. Liu,R., Chen,K., Jansz,N., Blewitt,M.E. and Ritchie,M.E. (2016) Transcriptional profiling of the epigenetic regulator Smchd1. *Genom. Data.*, **7**, 144–147.
37. Wu,D., Lim,E., Vaillant,F., Asselin-Labat,M.L., Visvader,J.E. and Smyth,G.K. (2010) ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**, 2176–2182.
38. Tian,L., Jabbari,J.S., Thijssen,R., Gouil,Q., Amarasinghe,S.L., Kariyawasam,H., Su,S., Dong,X., Law,C.W., Lucattini,A. *et al.* (2020) Comprehensive characterization of single cell full-length isoforms in human and mouse with long-read sequencing. bioRxiv doi: https://doi.org/10.1101/2020.08.10.243543, 10 August 2020, pre-print: not peer-reviewed.
39. Wyman,D. and Mortazavi,A. (2019) TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics*, **35**, 340–342.
40. Tardaguila,M., De La Fuente,L., Marti,C., Pereira,C., Pardo-Palacios,F.J., Del Risco,H., Ferrell,M., Mellado,M., Macchietto,M., Verheggen,K. *et al.* (2018) SQANTI: Extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res.*, **28**, 396–411.
41. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
42. Oshlack,A. and Wakefield,M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, **4**, 14.