RESEARCH ARTICLE

REVISED Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; peer review: 2 approved]

Jason L Weirather[1*], Mariateresa de Cesare[2*], Yunhao Wang[1,3,4*], Paolo Piazza[2], Vittorio Sebastiano[5,6], Xiu-Jie Wang[4], David Buck[2], Kin Fai Au[1,7]

[1]Department of Internal Medicine, University of Iowa, Iowa City, IA, USA
[2]Oxford Genomics Centre, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Key laboratory of Genetics Network Biology, Collaborative Innovation Center of Genetics and Development, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China
[5]Department of Obstetrics and Gynecology, Stanford University, Stanford, CA, USA
[6]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA, USA
[7]Department of Biostatistics, University of Iowa, Iowa City, USA

* Equal contributors

v2 **First published:** 03 Feb 2017, **6**:100
https://doi.org/10.12688/f1000research.10571.1
Latest published: 19 Jun 2017, **6**:100
https://doi.org/10.12688/f1000research.10571.2

Abstract

Background: Given the demonstrated utility of Third Generation Sequencing [Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT)] long reads in many studies, a comprehensive analysis and comparison of their data quality and applications is in high demand. *Methods:* Based on the transcriptome sequencing data from human embryonic stem cells, we analyzed multiple data features of PacBio and ONT, including error pattern, length, mappability and technical improvements over previous platforms. We also evaluated their application to transcriptome analyses, such as isoform identification and quantification and characterization of transcriptome complexity, by comparing the performance of size-selected PacBio, non-size-selected ONT and their corresponding Hybrid-Seq strategies (PacBio+Illumina and ONT+Illumina). *Results:* PacBio shows overall better data quality, while ONT provides a higher yield. As with data quality, PacBio performs marginally better than ONT in most aspects for both long reads only and Hybrid-Seq strategies in transcriptome analysis. In addition, Hybrid-Seq shows superior performance over long reads only in most transcriptome analyses. *Conclusions:* Both PacBio and ONT sequencing are suitable for full-length single-molecule transcriptome analysis. As this first use of ONT reads in a Hybrid-Seq analysis has shown, both PacBio and ONT can benefit from a combined Illumina strategy. The tools and analytical methods

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 19 Jun 2017	✓ view	✓ view
version 1 03 Feb 2017	? view	? view

1. **Jingyi Jessica Li**, University of California, Los Angeles, Los Angeles, USA

2. **Hagen Tilgner**, Weill Cornell Medical College , New York City, USA

Any reports and responses or comments on the article can be found at the end of the article.

developed here provide a resource for future applications and evaluations of these rapidly-changing technologies.

Keywords
Third Generation Sequencing, PacBio, Oxford Nanopore Technologies, Transcriptome

This article is included in the Nanopore Analysis gateway.

Corresponding authors: David Buck (dbuck@well.ox.ac.uk), Kin Fai Au (kinfai-au@uiowa.edu)

Author roles: Weirather JL: Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **de Cesare M**: Data Curation, Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Wang Y**: Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Piazza P**: Investigation, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Sebastiano V**: Investigation, Methodology, Resources, Validation; **Wang XJ**: Funding Acquisition, Project Administration, Resources, Supervision; **Buck D**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Au KF**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the National Human Genome Research Institute [R01HG008759 to KFA, YW and JLW]; the institutional fund of Department of Internal Medicine, University of Iowa [to KFA YW and JLW]. The Multidisciplinary Lung Research Career Development Program [T32HL007638 to JLW]; and the National Natural Science Foundation of China [91540204 to XW].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Weirather JL, de Cesare M, Wang Y *et al.* **Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis [version 2; peer review: 2 approved]** F1000Research 2017, **6**:100 https://doi.org/10.12688/f1000research.10571.2

First published: 03 Feb 2017, **6**:100 https://doi.org/10.12688/f1000research.10571.1

REVISED **Amendments from Version 1**

Notable changes to this version of the manuscript include clearly pointing out the differences in PacBio and ONT library preparations. The PacBio libraries were size-selected and ONT libraries were not. At the time of sequencing, size-selection was recommended for PacBio, but a comparable protocol was not yet established for ONT. Since this difference can have a profound influence on the distribution of read lengths, we have made prominent mention of it in this current version of the manuscript in the abstract, introduction and first two figures. Other notable changes include a more clear Table 2 to better display the evaluation of isoform identification by Second Generation Sequencing (SGS), Third Generation Sequencing (TGS), and Hybrid-seq (SGS+TGS) technologies using spike-in control transcripts. Additionally, the seven novel isoforms identified in Figure 7d (and Supplementary Table 3) were limited to five novel isoforms for which we have support from both ONT and PacBio reads. Furthermore, we made a minor edit to Figure 1 to make the axis label easier to read. Figure 2 has been modified to have the total number of aligned reads displayed above each panel. Figure 3 has been modified to have to have all the labels filled-in rather than implied. Figure 6 has been modified with updated results according to the requirement of a minimum 10% frequency of each alternative splicing event. Figure 7 and Supplementary Table 3 have been updated to show the five novel isoforms present. Previously, there were seven, but these five have support from the two different sequencing technologies. A number of other clarifications and corrections were made and we thank the referees for their recommendations; we have included detailed notes responding to the referees.

See referee reports

Introduction

Third Generation Sequencing (TGS) emerged more than 5 years ago when Pacific Biosciences (PacBio) commercialized Single Molecule Real Time (SMRT) sequencing technologies in 2011[1]. Although TGS platforms have significant technical differences, they all generate very long reads (1–100kb)[2–5], which is distinct from Second Generation Sequencing (SGS). Considering the paired-end information, the main SGS platform Illumina provides 50–600bp information from each DNA fragment; no SGS platforms provide >1000bp, including 454 sequencing, which generates the longest SGS reads (~700bp)[6,7]. Therefore, the short sequencing length limits the applications of SGS to large or complex genomic events, such as gene isoform reconstruction. TGS overcomes these challenging problems via long read lengths.

The most widely used TGS platforms [PacBio and Oxford Nanopore Technologies (ONT)] developed new biochemistry/biophysics methods to directly capture the very long nucleotide sequences from single DNA molecules. Other emerging TGS platforms (Moleculo[8] and 10X Genomics[9]) are based on the assembly of short reads from the same DNA molecules to generate synthetic long reads (SLR). Herein, we focus on data features of PacBio and ONT and their applications to transcriptome analysis.

PacBio adopts a similar sequencing-by-synthesis strategy as Illumina sequencing, except PacBio captures a single DNA molecule and Illumina detects augmented signals from a clonal *population* of amplified DNA fragments. The error rate of raw PacBio data is 13–15%, as the signal-to-noise ratio from single DNA molecules is not high[3]. To increase accuracy, the PacBio platform uses a circular DNA template by ligating hairpin adaptors to both ends of target double-stranded DNA. As the polymerase repeatedly traverses and replicates the circular molecule, the DNA template is sequenced multiple times to generate a continuous long read (CLR). The CLR can be split into multiple reads ("subreads") by removing adapter sequences, and multiple subreads generate circular consensus sequence ("CCS") reads with higher accuracy. The average length of a CLR is >10kb and up to 60kb, which depends on the polymerase lifetime[3]. Thus, the length and accuracy of CCS reads depends on the fragment sizes. PacBio sequencing has been utilized for genome (e.g., *de novo* assembly, detection of structural variants and haplotyping)[10] and transcriptome (e.g., gene isoform reconstruction and novel gene/isoform discovery)[11–13] studies.

ONT is a nanopore-based single molecule sequencing technology, and the first prototype MinION was released in 2014[14]. As compared to other sequencing technologies utilizing nucleotide incorporation or hybridization, ONT directly sequences a native single-stranded DNA (ssDNA) molecule by measuring characteristic current changes as the bases are threaded through the nanopore by a molecular motor protein. ONT MinION uses a hairpin library structure similar to the PacBio circular DNA template: the DNA template and its complement are bound by a hairpin adaptor. Therefore, the DNA template passes through the nanopore, followed by a hairpin and finally the complement. The raw read can be split into two "1D" reads ("template" and "complement") by removing the adaptor. The consensus sequence of two "1D" reads is a "2D" read with a higher accuracy[2]. Due to similar data features with PacBio, many researchers have utilized or are testing ONT in applications where PacBio has been applied.

PacBio and ONT platforms share the advantage of long read lengths, yet they also have the same drawback: higher sequencing error rate and lower throughput compared to SGS[3,14–16]. High sequencing error rates pose challenges for single-nucleotide-resolution analyses, such as accurate sequencing of transcripts, identification of splice sites and SNP calling. Low throughput is an obstacle for quantitative analysis, such as gene/isoform abundance estimation. Although PacBio CCS and ONT 2D consensus strategies can reduce error rates, the corresponding read lengths become shorter and throughput becomes lower. Therefore, hybrid sequencing ("Hybrid-Seq"), which integrates TGS and SGS data, has emerged as an approach to address the limitations associated with analysis of TGS data with assistance of SGS data. For example, error correction of PacBio or ONT long reads by SGS short reads improves the accuracy and mappability of long reads[17–19]. Hybrid-Seq can be applied to genome assembly and transcriptome characterization and improve the overall performance and resolution[11–13,17].

The long read length of PacBio and ONT is very informative for transcriptome research, especially for gene isoform identification. In addition to human transcriptomes[20–22], the PacBio transcript sequencing protocol, Iso-Seq, has been widely used to characterize transcriptome complexity in non-model organisms and particular genes/gene families[23–31]. In contrast, ONT has no standard transcript sequencing protocol and only a few pilot studies are publically available. Using MinION, Bolisetty *et al*. discovered very high isoform diversity of four genes in *Drosophila*, which

illustrates the utility of ONT in investigating complex transcriptional events[32]. Oikonomopoulos *et al*. also demonstrated the stability of ONT sequencing in quantifying transcriptome by analyzing an artificial mixture of 92 transcripts with Spike-In RNA[33]. Compared to these studies using PacBio or ONT alone, Hybrid-Seq can reduce the requirement of data size and improve the output, especially for transcriptome-wide studies. For example, a series of Hybrid-Seq methods (IDP, IDP-fusion, IDP-ASE) have been developed to improve the transcriptome studies to isoform levels (e.g., gene isoform reconstruction, fusion genes and allele phasing) with higher sensitivity and accuracy, and achieve a more accurate abundance estimation, which has been demonstrated in human embryonic stem cells (hESCs) and breast cancer[11–13].

Herein, we generated PacBio and ONT data from cDNA of hESCs. Using our tool AlignQC (http://www.healthcare.uiowa.edu/labs/au/AlignQC/), we performed a comprehensive analysis and comparison of PacBio and ONT data, including the raw data (subreads and 1D "template" reads) and their consensus (CCS and 2D reads). PacBio sequencing was performed on size-selected libraries, as size selection is the manufacturer recommendation. ONT libraries were not size selected, because size selection was not standard practice at the time of sequencing and was not performed for ONT. Since these technologies follow different library preparation protocols, it is important to consider these steps as potential sources of variability just as the sequencing platforms themselves can introduce variability. Comparisons analyzed included error rate and error pattern, read length, mappability and abnormal alignments, as well as technology improvements between the latest sequencing models (PacBio P6-C4 and ONT R9) and previous versions (C2 and R7). We also validated and compared the capability of PacBio and ONT alone to study a gold standard set of spike-in transcripts. Then, we applied long read only and the corresponding Hybrid-Seq approaches to human transcriptome analyses, including isoform identification, quantification and discovery of complex transcriptome events. In addition to a comprehensive evaluation of the characteristics of the two main TGS data platforms, this work serves as a guide for applications of PacBio and ONT and the corresponding Hybrid-Seq for transcriptome analysis.

Methods

Cell culture and RNA extraction

Human embryonic stem cells (H1 cell line; WiCell) were cultured as previously described[11]. In brief, cells were cultured in mTeSR1 (Stem Cell Technologies) on Matrigel matrix (BD). Cells were harvested between passages 50 and 55. Cells were fixed in 4% PFA for 10 minutes at room temperature and either incubated in blocking solution (2% FBS in PBS) or permeabilized in 0.2% Triton X-100 followed by incubation in blocking solution, where undifferentiated cells were verified by immunofluorescence (OCT4, NANOG, SSEA4, TRA-1-60, and TRA-1-81) as previous described[34]. Briefly, the primary antibodies used in the study were as follows: anti-OCT4 (mouse; Santa Cruz; sc-5279; 1:500), anti-h-Nanog (rabbit; Cosmo Bio; REC-RCAB0004P-F; 1:200), anti-SSEA-3 (rabbit; Millipore; MAB4303; 1:500), anti-TRA-1-60 (mouse; Millipore; MAB4360; 1:500), anti-CD31 (R&D Systems), and anti-desmin (Thermo Fisher Scientific). Primary antibodies were diluted 1:200 in blocking solution, unless otherwise stated, and incubated overnight at room temperature. Secondary antibodies (goat or donkey; Invitrogen; Alexa 488 and Alexa 594; 1:5000) were incubated for two hours at room temperature. Pluripotency was confirmed by teratoma assay where three germ layers formed *in vivo*[35].

Total RNA was extracted using RNeasy Plus Mini Kit (QIAGEN). Agilent RNA 6000 Pico Kit (Agilent) was used to assess the RNA quality, and Qubit RNA BR Assay Kit (ThermoFisher Scientific) was used to quantify the extracted RNA. SIRV (Spike-in RNA Variant) E0 mixture (Lexogen, Batch No. 216652830) was added to the extracted total RNA (about 2.83% SIRVs in the final mixture).

Library preparation and sequencing

For Illumina sequencing, TruSeq Stranded mRNA HT Sample Prep Kit (Illumina) was used to prepare the sequencing library by substituting the TruSeq barcoded adapter with Illumina Adapters (Multiplexing Sample Preparation Oligonucleotide Kit) and the PCR Primer Cocktail with Multiplex PCR primer 1.0 (5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′) and custom index primer (5′-CAAGCAGAAGACGGCATACGAGAT[index]CAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3') as described previously[36]. Sequencing was performed by Illumina HiSeq4000 with 150bp paired-end reads.

For PacBio sequencing, full-length cDNA and SMRTbell templates were prepared at the Centre of Genomic Research, University of Liverpool, following the Iso-Seq sample preparation protocol (Pacific Biosciences). For size selection, the full-length cDNA was fractioned into four contiguous size ranges (0–1kb, 1–2kb, 2–3kb, >3kb) on a Sage ELF (Sage Science) before constructing SMRTbell templates. Sequencing was performed by PacBio RS II using C4/P6 chemistry. The SMRT cell counts were 1, 4, 4 and 3 for 0–1kb, 1–2kb, 2–3kb and >3kb libraries, respectively.

For ONT sequencing, full-length cDNA was generated by the Smart-seq2 protocol, as described by Picelli *et al*.[37] using modified sequences for the TSO (5'-TTTCTGTTGGTGCTGATATTGCTGCCATTACGGCCrGrG+G-3') and the Oligo-dT$_{30}$VN (5'-ACTTGCCTGTCGCTCTATCTTCT$_{30}$VN-3') to allow amplification of the cDNA second strand with primers provided by ONT. The quality and size distribution of the cDNA was tested by a TapeStation Genomic DNA system (Agilent). For each ONT flowcell, 1 µg of double-stranded cDNA was converted in a Nanopore-compatible sequencing library using the Genomic DNA Sequencing Kit SQK– NSK007 (ONT), according to the manufacturer's protocol with minor modifications. In detail, the ds-cDNA was subjected to a combined end repairing and dA-tailing step using the NEBNext Ultra™ II End Repair/dA-Tailing Module (New England BioLabs) and incubated for 30 min at 20°C followed by 30 min at 65°C. The reactions were purified with 0.4x volume Agencourt AMPure XP beads (Beckman Coulter), according to manufacturer's instructions. The end-prepped cDNA was subsequently ligated to ONT leader- and HP-adapter using Blunt/TA Ligase Master Mix (New England BioLabs) with a 10 min incubation at room temperature. The ligated cDNA was annealed to a biotinylated tether oligo (ONT) that targets the hairpin-adapter (HP-adapter) by incubation for an additional 10 min at room temperature. The fragments with a HP-adapter ligated were selectively pulled down using Dynabeads MyOne Streptavidin C1 (Life Technologies). After

washing the DNA-bounded beads to remove unbounded DNA, the captured cDNA library was released from the streptavidin beads by incubating the beads re-suspended in ONT Elution Buffer for 10 min at 37°C. The beads were then pelleted using a magnetic rack and the supernatant containing the library was recovered. The full-length cDNA library was sequenced on a MinION Mk 1B using a 48h sequencing protocol on R7/R9 chemistry flowcells.

AlignQC software

Long reads require special considerations when accessing their quality; they have variable error rates and they are often size selected. These attributes make careful study of the alignments of long reads necessary to understand the quality and coverage of transcriptome sequencing.

Implementation: AlignQC (http://www.healthcare.uiowa.edu/labs/au/AlignQC/) is designed to provide comprehensive quality assessment for TGS long read sequencing alignment data by three layers: (1) basic statistics of the data, including read length, alignment and coverage across all chromosomes; (2) error pattern analysis if a reference genome is provided; (3) transcript-related statistics if a gene annotation is provided. AlignQC takes the standard BAM format file as the input, outputs XHTML format file for easy visualization, and provides links to access all analysis results.

For basic statistics of the data, AlignQC parses the CIGAR string and SEQ fields from the BAM file. Multiple alignment paths can be reported for each read, but only the longest aligned path is used in error rate calculations and annotation analyses. For alignment statistics, if two or more alignment paths are reasonably spaced across the read, and can together generate a longer alignment, they will be combined and classified as: (a) a gapped alignment of a gene if paths occur within close proximity to each other on the same strand; (b) a trans-chimeric alignment if paths occur on different loci; (c) a self-chimeric alignment if paths align to an overlapping genomic position; otherwise, the read is defined as (d) a single alignment.

For the error pattern analysis, AlignQC compares the aligned reads to the provided reference genome. Based on the difference between aligned reads and reference genome, it estimates the error rates of total and different error types, including substitutions, insertions, and deletions. The overall error rates are calculated by sampling alignments until at least 1 million aligned bases have been included. Context-specific error pattern is analyzed by randomly sampling the best alignments until each individual context has been observed at least 10,000 times.

For transcript related statistics, AlignQC firstly annotates the aligned reads according to their overlap with provided genes/transcripts. A read is assigned to a reference transcript if it can cover the first and last exons with any length, and the internal exons with ≥ 80% length. When multiple exons are present and both the read and the reference transcript have the same consecutive exons, the match is called as a "full-length" match, otherwise, it is referred to as a "partial" match.

Operation: AlignQC usage can be divided in to report generation, and report viewing. Report generation requires a Linux operating system with coreutils (version 8.6 or newer) and python (2.7 or newer); both are present in most current Linux releases. R must be installed (tested with version 3.3.0; https://www.r-project.org/). At least 16GB of RAM is recommended to run AlignQC. A full analysis of an alignment from a PacBio SMRT cell containing 107,960 molecules was processed by 4 threads in 32m21.307s. A full analysis of an alignment from an ONT R9 flow cell containing 387,810 molecules required 52m22.163s.

Report viewing can be done through any modern web browser and does not require any specific operating system. The primary output of AlignQC is an XHTML format report. Analysis files are embedded in the report; these include high quality plots and the long read mappings that are compatible with the UCSC genome browser[38]. These reports can serve as both an analysis archive and a convenient means to share results.

Short read and long read data processing and alignment

For Illumina short reads, the quality was assessed by FastQC. The sequencing adapters were trimmed by 9 bases on the 5' end and adapters were removed by cutadapt[39] with the parameter "-a AGATCGGAAGAG -A AGATCGGAAGAG -m 50". Short read alignment was performed by HISAT with default parameters. For SIRV, the reference genome (SIRV_151124a.fasta; https://www.lexogen.com/wp-content/uploads/2015/11/SIRV_Sequences_151124.zip; Supplementary Table 1) was provided by Lexogen. For the hESCs analysis, the reference genome was downloaded from UCSC (hg38 assembly; GCA_000001305.2; http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/).

For PacBio, the subreads and CCS reads were extracted using SMRT Analysis software (version 2.3.0; http://www.pacb.com/products-and-services/analytical-software/smrt-analysis/). For technical comparisons, CCS and subreads are used as referred. For the transcriptome analyses, PacBio data sets were comprised of "best reads". These were constructed with the goal of 1) having each molecule represented in the dataset once and only once and 2) choosing the best-quality read of each molecule for transcriptome analysis. Below is the priority order of reads to be selected as the "best read" for each molecule in different analysis strategies:

For "PacBio only" strategy, "best reads" were selected by using 1) the best aligned CCS reads (determined by the number of mapped bases in the read), or if no CCS read or alignment was available, 2) the best aligned subread. For the "PacBio+Illumina" Hybrid-Seq analysis the "best read" is 1) the best aligned CCS reads with >2 passes and an accuracy greater than 95 (estimated by SMRT Analysis software); otherwise 2) the best aligned CCS reads corrected by short reads; otherwise 3) the best aligned subread.

For ONT, the template, complement and 2D reads were extracted by poretools software (version 0.5.1; https://poretools.readthedocs.io/en/latest/). For technical comparisons 2D reads and 1D template strand reads are referred as 2D and 1D, respectively. For the transcriptome analyses, a "best reads" set analogous to PacBio was

used. For the "ONT only" strategy, "best reads" were selected by 1) the best aligned 2D reads, otherwise 2) the best aligned 1D template strand reads, otherwise 3) the best aligned 1D complement strand reads. For "ONT+Illumina" Hybrid-Seq strategy, the "best read" was selected by the same order after error correction by short reads.

For PacBio and ONT long read alignment, GMAP[40] (version 2016-06-30) was used with the parameter "-n 10".

Isoform identification in SIRVs by Illumina, PacBio, ONT

The SIRV (Lexogen) transcriptome, which consists of 69 transcripts, mimics 7 human model genes and includes all kinds of complex alternative splicing events. SIRV is useful to assess the performance of sequencing technology applied to studying human transcriptome. This study used the SIRV E0 mix (Batch No. 216652830, in which isoform SIRV502 is missing) with 68 RNA variants. The concentration ratio is identical for each isoform. Meanwhile, Lexogen also provides three types of annotation libraries: "corrected", with all 68 truly-expressed isoforms; "insufficient", including 43 of 68 truly-expressed isoforms; and "over-annotated", with 68 truly-expressed isoforms and an additional 32 falsely-expressed isoforms.

When illustrating the performance of Illumina short reads on isoform identification, reference-guided assembly software StringTie[41] (version 1.3.0) with default parameters was used, based on three different SIRV annotation libraries above. For all SIRV isoforms, we classified them into two groups: 1) true positive if the isoform was annotated by SIRV "correct" annotation library; and 2) false positive if not. The numbers of true positive and false positive assembled isoforms were counted when using three SIRV annotation libraries in StringTie, respectively.

When illustrating the performance of PacBio and ONT long reads on isoform identification, an isoform was considered identified when at least one long read was uniquely aligned to this isoform.

Isoform identification in hESCs by PacBio, ONT and Hybrid-Seq

Gencode (version 24) gene annotation library (https://www.gencodegenes.org/; Supplementary Table 1) was used for isoform detection. AlignQC was used to identify isoforms annotated by Gencode (version 24). Briefly, for isoforms with only one exon (singleton isoform), if 90% of the isoform length could be covered by at least one long read, it was considered identified. For isoforms with multiple exons (multi-exon isoform), we required at least one long read that covered the first and last exons and ≥ 80% mutual overlap of each internal exon.

Notably, for Hybrid-seq (PacBio+Illumina and ONT+Illumina) strategies, we combined the results mentioned above and the output of IDP[11] (version 0.1.9), which is a tool specifically for isoform detection and prediction by Hybrid-seq data. The primary parameters of IDP were "Njun_limit=10, Niso_limit=100, and FPR=0.05", using Gencode (version 24) as the primary reference, and a comprehensive transcript reference from the combination of Gencode (version 24), RefSeq (UCSC version 2015-06-03; http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/refFlat.txt.gz; Supplementary Table 1) and ESTs (downloaded from UCSC genome browser; http://hgdownload.cse.ucsc.edu/goldenPath/hg38/database/all_est.txt.gz; Supplementary Table 1).

For novel isoform identification, the output of IDP with the same parameters was used.

When investigating the accuracy of splice sites/exon boundaries within the multi-exon isoforms, we calculated the relative distance between known splice sites annotated by Gencode and detected splice sites by four strategies.

For repetitive element analysis, the lower-case sequence marked by RepeatMasker and Tandem Repeats Finder tools was used from the reference genome (UCSC hg38; http://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/; Supplementary Table 1). For each isoform, the proportion of repetitive element sequence was calculated.

Isoform abundance estimation by PacBio, ONT and Hybrid-Seq

The isoforms identified by 7 strategies (1. use Illumina data by StringTie with the "correct" SIRV annotation library; 2. use Illumina data by StringTie with the "insufficient" SIRV annotation library; use Illumina data by StringTie with the "over-annotated" SIRV annotation library; 4. use PacBio data with the "correct" SIRV annotation library; 5. use ONT data with the "correct" SIRV annotation library; 6. use PacBio+Illumina data with the "correct" SIRV annotation library; and 7. use ONT+Illumina data with the "correct" SIRV annotation library) were used to perform isoform abundance estimation.

The relative expression percentage (REP) of each isoform was calculated. Expected REP is 1/68.

For three Illumina-only strategies (Illumina data with the "correct" SIRV annotation library, Illumina data with the "insufficient" SIRV annotation library and Illumina data with the "over-annotated" SIRV annotation library), the TPM (transcripts per million) value from RSEM with default parameter was used to calculate the REP.

For two long read only strategies (PacBio data with the "correct" SIRV annotation library and ONT data with the "correct" SIRV annotation library), the read count from AlignQC was used to calculated the REP.

For two Hybrid-Seq strategies (PacBio+Illumina data with the "correct" SIRV annotation library and ONT+Illumina data with the "correct" SIRV annotation library), only the Illumina short read data was used to run RSEM with default parameters. The TPM (transcripts per million) value from RSEM was used to calculate the REP.

To compare the estimation error of 7 strategies, the euclidean distance between expected REP and estimated REP was calculated.

Complexity analysis of the hESC transcriptome

For alternative splicing analysis, LESSeq (https://github.com/gersteinlab/LESSeq) was used, following its instructions. We required a minimum frequency of each alternative event >10%.

Functional analysis of identified isoforms in hESCs

For the prediction of protein coding capability of novel isoforms, GeneMarkS-T (version 5.1; http://exon.gatech.edu/GeneMark/) with default parameters was used. For gene enrichment analysis, DAVID (version 6.8)[42] was used.

Results

Read length of PacBio and ONT data

The mappable length is a good representation of the useful length of long reads. The median mappable lengths of PacBio data are 1,299bp and 1,464bp for subreads and CCS reads, respectively. ONT data are slightly longer, with median lengths of 1,602bp and 1,754bp for 2D and 1D reads, respectively (Table 1), although size selection was performed in PacBio, but not in ONT (Methods).

The overall length distributions of the raw data and consensus data for both PacBio and ONT (subreads vs. CCS and 1D vs. 2D) are similar, while the differences between PacBio and ONT are more remarkable (Figure 1). Compared to ONT, the length distribution of PacBio data skews to the left, with many reads <1kb, which may be caused by a short size-selected fraction (<1kb) of cDNA library (see Methods, Figure 1 and Supplementary Figure S1). In addition, CCS reads have a large proportion of very long reads (>3.5kb), as the high quality of CCS reads guarantee the alignment of the full length while the other reads (e.g., subreads) are partially aligned.

ONT R9 and the previous sequencing platform R7 have similar length distributions (Figure 1, Supplementary Figure S2.1 and Supplementary Figure S2.2), while the yield of R9 is much higher (204,891±61,389 vs. 61,799±42,393 molecules were sequenced and mappable per R9 and R7 per flow cells, respectively). Thus, R9 provides a more stable and higher throughput, which will allow broader applications of ONT data (Supplementary Table 2). The length distribution of the previous PacBio C2 sequencing data skews to a shorter length, compared to P6-C4. The yield of P6-C4 increased (76,597±23,387 vs. 21,827±9,707 molecules were sequenced and mappable per P6-C4 and C2 per SMRT cells, respectively). Overall, the yield per flow cell of ONT is much higher than PacBio, because each nanopore can sequence multiple molecules, while the wells of PacBio SMRT cells are not reusable. In addition, the PacBio read lengths in each SMRT cell are consistent with the sizes selected, so the size-selection protocol works well for PacBio data (Supplementary Figure S1).

Mappability of PacBio and ONT data

Mappability of long reads is necessary to confirm repetitive elements, gene isoforms and gene fusions[11,12,21]. PacBio subreads and ONT 1D reads have similar rates of aligned reads (80.41% and 78.24%) and bases (81.80% and 81.03%) to the reference genomes (Figure 2). However, a higher proportion of PacBio CCS reads (96.15%) and bases (95.07%) can be aligned than ONT 2D reads (92.05% and 87.37%), while both are higher than their corresponding raw data (subreads and 1D). Thus, generation of consensus sequences truly improves data quality. As 2D reads only sequence target molecules twice, it is expected to have lower quality than CCS with multiple subreads.

For all types of data, we consistently observe that short read lengths (<500bp) have low alignment rates. This is likely due to a larger portion of adapter and linker sequences in this short-length data bin. In addition, although a large fraction of ONT data are defined as "fail" reads during the data pre-process and filtered out, the alignment rates are as high as 65.74% and 50.95% for 2D fail reads and 1D fail reads, respectively. These findings indicate that parts of the fail reads are informative and should be rescued (e.g., by error correction) to increase throughput.

The mappability of PacBio data is similar between the C2 and P6-C4 chemistries, while the ONT 1D reads in R9 have almost doubled the proportion of aligned bases relative to R7 (81.03% vs. 44.43%). However, the alignment rate of R9 1D reads is surprisingly slightly worse than the previous R7 data (78.24% vs. 82.19%). The improvements in total bases aligned is likely attributable to improvements in raw data quality, while relaxing criteria for calling 1D reads in R9 may explain the slight drop in the overall

Table 1. Statistics of mappable length and error rates of PacBio and ONT long reads.

Read type	Mappable length (bp)				Error rate (Proportion of overall error) (%)			
	Mean	Median	Standard deviation	Maximum	Overall	Insertion	Deletion	Mismatch
PacBio CCS	1772	1464	1132	8006	1.72	0.087 (5.06)	0.34 (19.48)	1.30 (**75.46**)
PacBio subread	1570	1299	1076	16040	14.20	5.92 (**41.71**)	3.01 (21.17)	5.27 (37.12)
ONT 2D	1861	1754	882	9126	13.40	3.12 (23.30)	4.79 (35.70)	5.50 (**40.99**)
ONT 1D	1695	1602	824	9345	20.19	2.93 (14.51)	7.52 (37.24)	9.74 (**48.25**)

The fractions of each error types are in parenthesis. The fractions of the most predominant error types in each data are in bold.

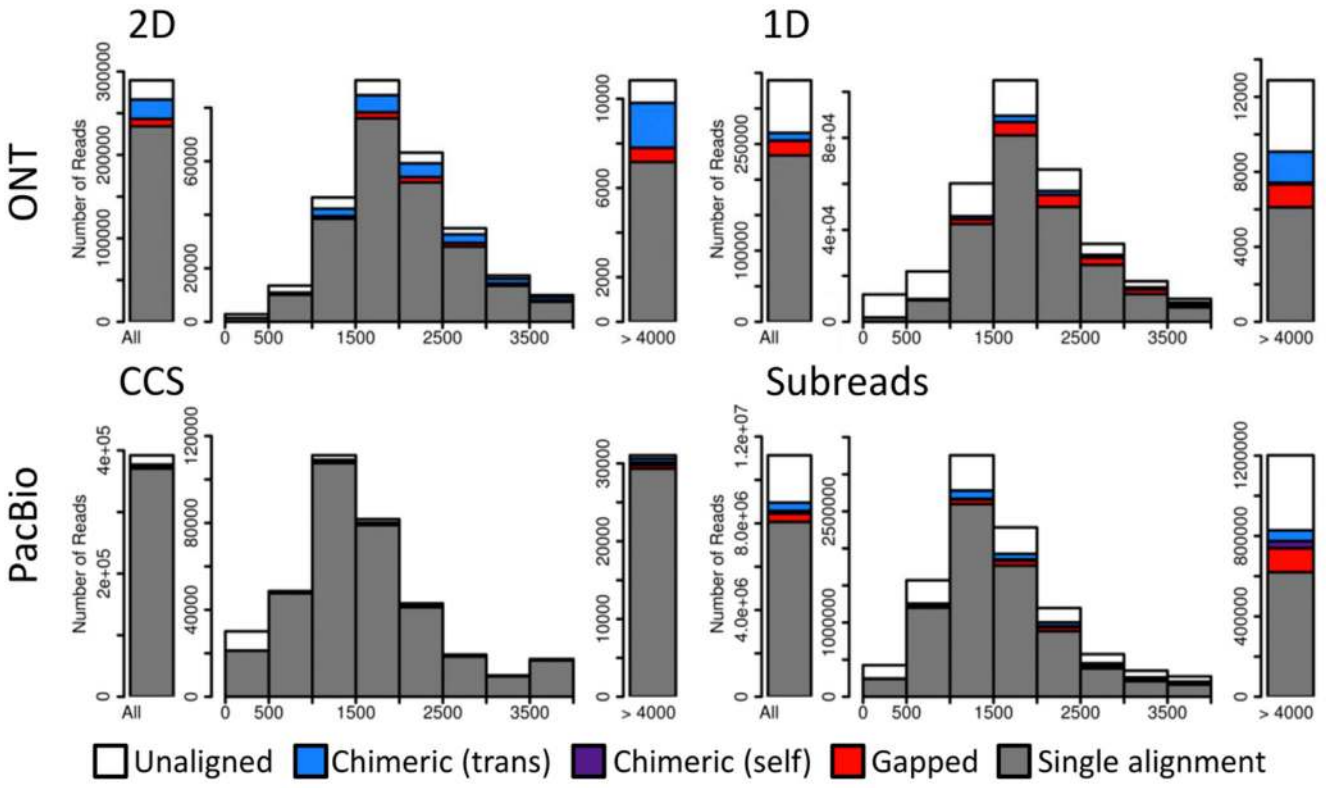

Figure 1. Length distribution of reads. The length distribution of Oxford Nanopore Technologies (ONT) 2D and 1D reads (top) and Pacific Biosciences (PacBio) CCS and subreads (bottom). Aligned reads are color-coded to indicate fraction of reads that are: single best alignments (gray), gapped alignments consisting of multiple paths (red), self-chimeric alignments (purple) where different read segments map to overlapping sequences, and trans-chimeric alignments (blue) where read segments map to different loci; white color represents unaligned reads. The leftmost bar represents all reads, the middle portion reads from 0–4kb in length, and the rightmost are reads greater than 4kb. PacBio libraries were size-selected, while ONT libraries were not; this provides PacBio with a larger proportion of longer reads. The total number of reads sequenced and the number of aligned reads from each sequencing platform are available in Supplementary Table 2.

alignment rate. The slight drop in the alignment rate accompanies a largely improved throughput of 1D reads per cell for R9 compared to R7 (181,599±54,331 vs. 55,366±26,371).

Chimeric and gapped alignments of PacBio and ONT data

Long reads generated from gene fusions or trans-splices can be aligned to separated genomic loci, which are denoted as "trans-chimeric". Since hESCs contain very few fusion events or trans-splices, trans-chimeric reads are likely due to library preparation artifacts. 2D data contain 8.05% trans-chimeric reads, while 1D data contain surprisingly less (3.16%). Considering they are from the same data and library preparation, the lower trans-chimeric frequency in 1D reads may be due to the very low mappability of some error-prone regions. ONT data have particularly higher trans-chimeric rates in very long reads (>4kb) (Figure 1). PacBio CCS reads have far less trans-chimeric alignments (0.93%), while 1D reads and subreads are of similar trans-chimeric fractions (3.47% vs 3.16%). Therefore, the library preparation artifact is not negligible in TGS data, and the trans-chimeric reads in non-tumor samples should be filtered before further usage. In addition, two fragments of a long read may be aligned to the same genomic locus, denoted as "self-chimeric", because of the failure of removing adaptor sequences from the raw data (e.g., PacBio CLR). Overall, self-chimeric proportion is much smaller than trans-chimeric. The chimeric reads may cause an overestimate of the lengths of DNA molecules.

Since some regions of long reads may be particularly error-prone, long reads may be aligned as separated fragments. With careful analysis, these "gapped alignments" can be used similarly with the paired-end Illumina reads. Corresponding to the high error rate, more ONT data are gapped alignments (1D: 6.10% and 2D: 2.98%) than PacBio (subreads: 3.45% and CCS: 0.48%). This rate is even more severe in the previous ONT R7 chemistry, especially for 1D reads (30.82%), while the difference between PacBio C2 and P6-C4 data is much smaller.

Error pattern of PacBio and ONT data

Whereas mappability is a metric of the fraction of useful reads, error rate and error pattern measure the quality of the data, which have a strong effect on single-nucleotide resolution analysis (e.g., SNP calling and splice detection) and design of error correction

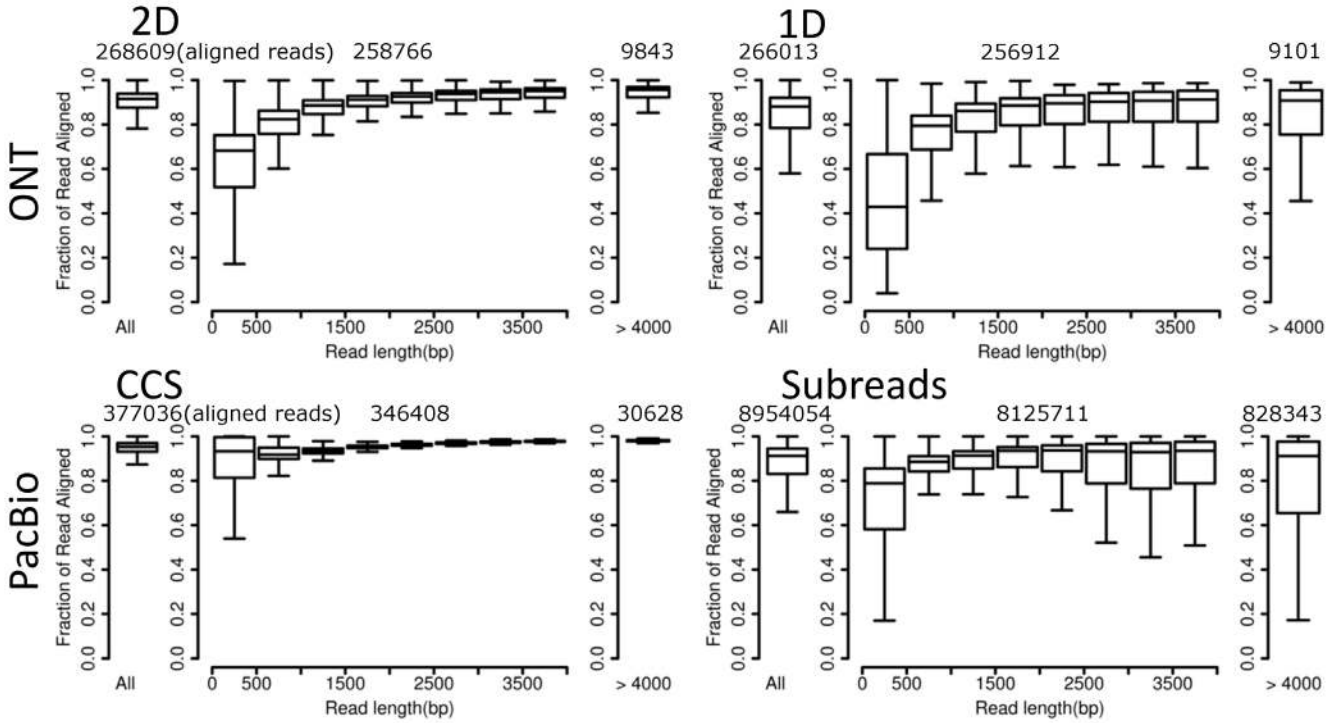

Figure 2. Mappability of different length bins. The leftmost bar represents the fraction of the mappable read length out of the total read length for all reads. The middle section shows the mappable fraction for 500bp increments ranging from 0–4kb read lengths, and the rightmost bar represents the mappable fraction of reads greater than 4kb. ONT: non-size-selected Oxford Nanopore Technologies reads; PacBio: size-selected Pacific Biosciences reads. The numbers of aligned reads contributing to the box plots in each panel are listed above each panel: total aligned reads, aligned reads <4kb and aligned reads >4kb (from left to right).

algorithms. The error rate of PacBio CCS reads is as low as 1.72%. The 14.20% error rate of subreads is consistent with previous reports (Korlach J. Understanding Accuracy in SMRT Sequencing. Pacific Biosciences; http://www.pacb.com/wp-content/uploads/2015/09/Perspective_UnderstandingAccuracySMRTSequencing1.pdf.) and is similar with ONT 2D data (13.41%). However, 1D reads have a 20.19% error rate (Table 1). Thus, the raw data and the consensus sequence of PacBio data are of higher base quality than corresponding ONT data.

Moreover, the composition of PacBio and ONT errors are different. Mismatches are the major errors in both ONT data (2D: 40.99% and 1D: 48.25%), and the proportion of deletions are also as high as >35% (Table 1). Thus, insertions are the least common errors in ONT. Insertions are also the least common in PacBio CCS reads, whereas mismatches are more predominant (75.70%), though the absolute error rate is fairly low. Conversely, the rate of insertions in subreads is the highest (41.71%), and mismatches are at a similar level (37.12%). Thus, insertions and deletions together ("indels") contribute to most errors with the exception of CCS reads.

PacBio base calling is based on distinguishing signals from the neighborhood background; ONT relies on the current signal change from the five upstream bases. Therefore, their errors may both have context-specific patterns. As the predominant error type in CCS reads, mismatches mostly arise from two context-specific events: **C**G->C**A** and **C**G->**T**G (Figure 3); however, these mismatches are likely alignment errors rather than sequencing errors as they are also observed in the alignments of high-quality Illumina data and simulation data (Supplementary Figure S3). The mismatch T**A**G->T**G**G is most striking in both ONT 2D and 1D reads, followed by T**A**C->T**G**C, while the other mismatches are far less frequent (Figure 3). In contrast, the mismatches in subreads show a clear "loose homopolymer pattern": the base is more likely mis-called as either the upstream or downstream base ("cross shape" in Figure 3). The same homopolymer pattern also exists in the indels in subreads: 46.07% indels are in a homopolymer (Figure 3). The indels prefer to occur in homopolymers in CCS and 2D reads as well, with 85.46% and 39.40% in homopolymers, respectively. In addition, both CCS and 2D reads have the same bias of homopolymer pattern to specific bases: A and T in insertions and G and C in deletions. Moreover, insertions of G and C have a "tight homopolymer pattern": both upstream and downstream bases are the same as the inserted bases ("diagonal spots" in Figure 3 and Supplementary Figure S4). Overall, the homopolymer pattern of errors is more pronounced in the raw PacBio data (subreads), but not very clear in the raw ONT data (1D reads). Regardless of the difference in sequencing platform, the overall error patterns of CCS and 2D both contain homopolymer indels, which may be due to the consensus sequence algorithm. The specific mismatches of ONT data may be caused by some difficult case contexts for the basecaller.

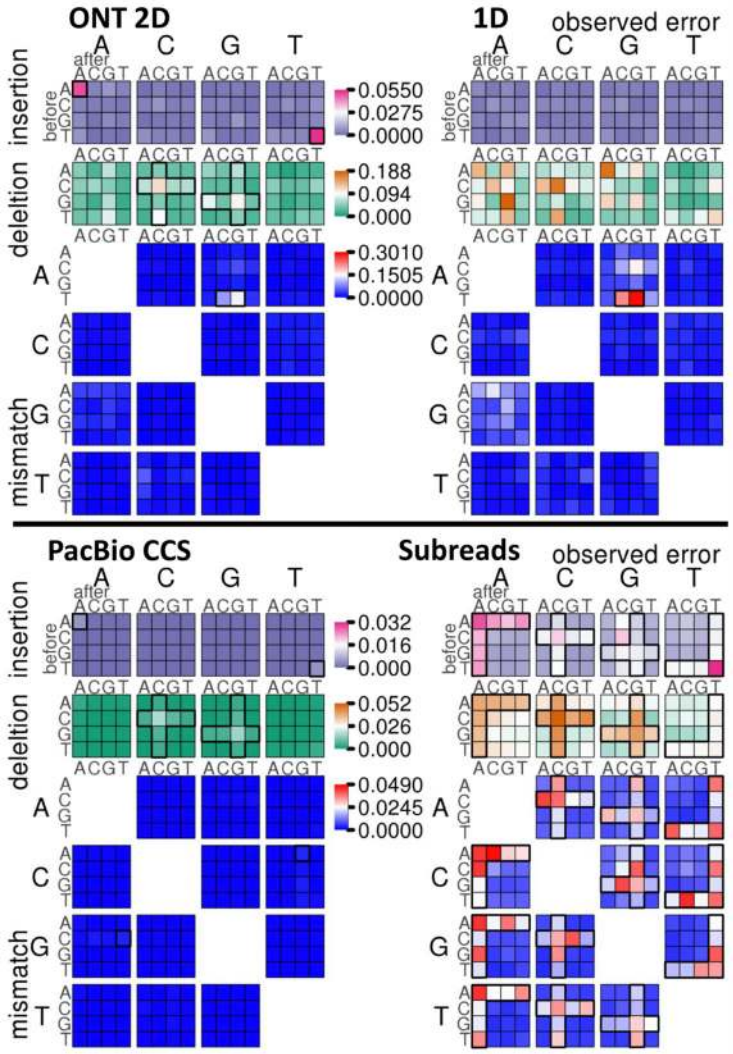

Figure 3. Context-specific errors. Context specific-errors are shown for Oxford Nanopore Technologies (ONT) 2D and 1D reads (top), and Pacific Biosciences (PacBio) CCS and subreads (bottom). The error types shown are insertions, deletions and mismatches. For insertions, the large base above the plot indicates the inserted base, and for deletions, the deleted base. For mismatch errors, the large base to the left indicates the expected reference base, and the large base above indicates the base observed in the read. A block of color tiles shows the error frequency within specific contexts for each error; the small base to the left of the tiles indicates the base preceding the error, and the small base above is the base following error. Error frequency is plotted on separate scales for insertions, deletions, and mismatches. Homopolymer error patterns are highlighted with a bold cross- or L-shaped outlines in the ONT 2D, PacBio CCS and PacBio Subreads plots. Context-specific insertions and mismatches of interest in the ONT 1D, 2D and PacBio CCS reads are highlighted by a bold outlines. For a better contrast of lower error rate in PacBio CCS reads and ONT 2D reads, Supplementary Figure S4 displays each result with its own scale.

In spite of the higher overall error rate, the error pattern of the PacBio C2 data is almost the same as P6-C4 data, while the C2 CCS reads have a "loose" rather than the "tight" homopolymer pattern of P6-C4 data for indels (Supplementary Figure S4). Compared to ONT R9 data, the error patterns of R7 data (both 2D and 1D reads) are mosaic, with a few predominant errors (Supplementary Figure S4). Only the "tight homopolymer pattern" of indels is observed in R7 2D reads. Therefore, PacBio and ONT data have been improved dramatically, except for some systematic errors at homopolymers and specific contexts.

Isoform identification in SIRVs by Illumina, PacBio and ONT

Our next goal was to investigate the advantages of PacBio and ONT long reads for transcriptome analysis over Illumina short reads. We first compared the performance of gene isoform identification using the gold standard Spike-In RNA Variant Control mixes (SIRVs), which contain 68 isoforms of 7 genes with various splicing complexity and known abundance. This allows the evaluation of isoform recall by PacBio, ONT and Illumina data. We reconstructed isoforms from Illumina short reads by the reference-guided mode of StringTie[41] with three types of SIRV annotation libraries: the

"correct" library containing all 68 truly-expressed isoforms, the "insufficient" library containing 43 of 68 truly-expressed isoforms, and the "over-annotated" library containing 68 truly-expressed isoforms and 32 additional unexpressed isoforms (see Methods). None were able to report all 68 truly expressed isoforms (44, 63 and 62, respectively; Table 2). When the reconstruction was guided by the "insufficient" SIRV annotation library, only 20.00% (5 of 25) of missing isoforms were rescued, along with 33 false positive predictions. When guided by the "over-annotated" SIRV annotation library, 46.87% (15 of 32) of unexpressed, but annotated, isoforms were incorrectly reported, with an additional 24 false positive predictions. Even if the assembly was guided by the "correct" SIRV annotation library, which is rarely available in practical transcriptome analysis, short reads identified 92.65% (63 of 68) annotated isoforms, but with 27 false positive predictions. These results demonstrated the incompleteness or high false positive rate of isoform reconstruction by short reads. In contrast, ONT directly detected all 68 expressed isoforms, and PacBio missed only one, isoform SIRV618, which is 219 bp and may be filtered out by size selection in PacBio library preparation. Thus, PacBio and ONT long reads show a far superior performance in isoform identification over short reads.

Isoform identification in hESCs by PacBio, ONT and Hybrid-Seq

We further evaluated the performance of PacBio and ONT in identifying isoforms from hESCs (H1 cell line, see Methods). In total, 919,158 mappable PacBio reads and 923,671 mappable ONT reads were used. A total of 57,868 and 59,098 Gencode-annotated isoforms were detected by PacBio and ONT reads, including 23,067 and 21,196 full-length isoform detection, respectively (Figure 4A and Supplementary Figure S5). The full-length isoform identification rates were 47.14% and 44.79%, respectively. For the >1kb isoforms that are difficult to detect by short reads, PacBio and ONT directly detected 15,764 and 14,669 full length transcripts (Figure 4A). Thus, ONT shows comparable sensitivity with PacBio for full-length isoform detection.

Table 2. Performance of Illumina, PacBio and ONT on isoform identification in the gold standard SIRVs.

Strategy (SIRV annotation library)	True positive	False positive
Illumina (with "insufficient" SIRV annotation library)	39 + 5*	33
Illumina (with "correct" SIRV annotation library)	63	27
Illumina (with "over-annotated" SIRV annotation library)	62	24 + 15**
PacBio (with "correct" SIRV annotation library)	67	-
ONT (with "correct" SIRV annotation library)	68	-

*In the "insufficient" SIRV annotation library, 25 isoforms are not included but are truly-expressed. Of these 25 isoforms, 5 isoforms were rescued when using Illumina short reads data.

**In the "over-annotated" SIRV annotation library, 32 isoforms are included but are not truly-expressed. Of these 32 isoforms, 15 isoforms were assembled.

Next, we identified isoforms from two Hybrid-Seq datasets: PacBio+Illumina and ONT+Illumina. Firstly, the long reads were corrected by LSC (version 1 beta)[18] and Illumina reads, and the number of mappable reads increased to 951,258 and 933,762 for PacBio and ONT, respectively (see Methods). Furthermore, error correction greatly improved overall error rates and context-specific errors patterns (Supplementary Figure S4). By inputting the corrected long reads and Illumina reads to IDP, 26,325 and 23,340 Gencode-annotated isoforms were identified by full length by PacBio+Illumina and ONT+Illumina, respectively (Figure 4A), demonstrating the superior sensitivity of Hybrid-Seq over long reads only to identify isoforms. For multi-exon isoforms that are difficult to be constructed by short reads alone, the full-length isoform identification ratios were as high as 92.82% and 91.48% for PacBio+Illumina and ONT+Illumina, respectively (Figure 4B). Whereas 16,711 isoforms were identified by both Hybrid-Seq datasets, the overlap ratios of identified isoforms were not very high (PacBio+Illumina: 63.48% and ONT+Illumina: 71.60%; Figure 4C). That is, the two Hybrid-Seq datasets rescued significant numbers of isoforms that were missed by the other (9,614 and 6,629 for PacBio+Illumina and ONT+Illumina, respectively). These discordant isoforms were mostly multi-exon isoforms (Supplementary Figure S6).

Imperfect alignments of error-prone long reads subsequently result in ambiguous determination of splice sites/exon boundaries within the multi-exon isoforms. Using splice sites annotated by the reference library and or detected by short reads as the gold standard, 14.72% and 30.82% splice sites were incorrectly identified by PacBio and ONT, respectively (Figure 4D and Supplementary Figure S7). By contrast, by correcting long reads with short reads and integrating short reads in isoform identification (i.e., by the tool IDP), the incorrectly identified rates were decreased to 7.05% and 19.94% for PacBio+Illumina and ONT+Illumina, respectively. Thus, Hybrid-Seq provides a higher resolution of the exon-intron structures within each identified isoform. In addition, PacBio showed a better performance of splice site determination for both long read only and Hybrid-Seq strategies, which is consistent with the lower error rates than ONT.

With the determination of high-resolution exon-intron structure and the consistent evidence from both TGS and SGS data, we can discover and annotate significant amounts of novel multi-exon isoforms accurately: 2,712 and 2,095 by PacBio+Illumina and ONT+Illumina, respectively (Figure 4E). Compared with the overlap of annotated isoform detection (Figure 4C), only a minority of novel isoforms (467) were identified by both Hybrid-Seq strategies (Figure 4E). Besides the possible technological difference, the distinct coverage of novel isoforms by our PacBio and ONT data may be attributable to sampling differences.

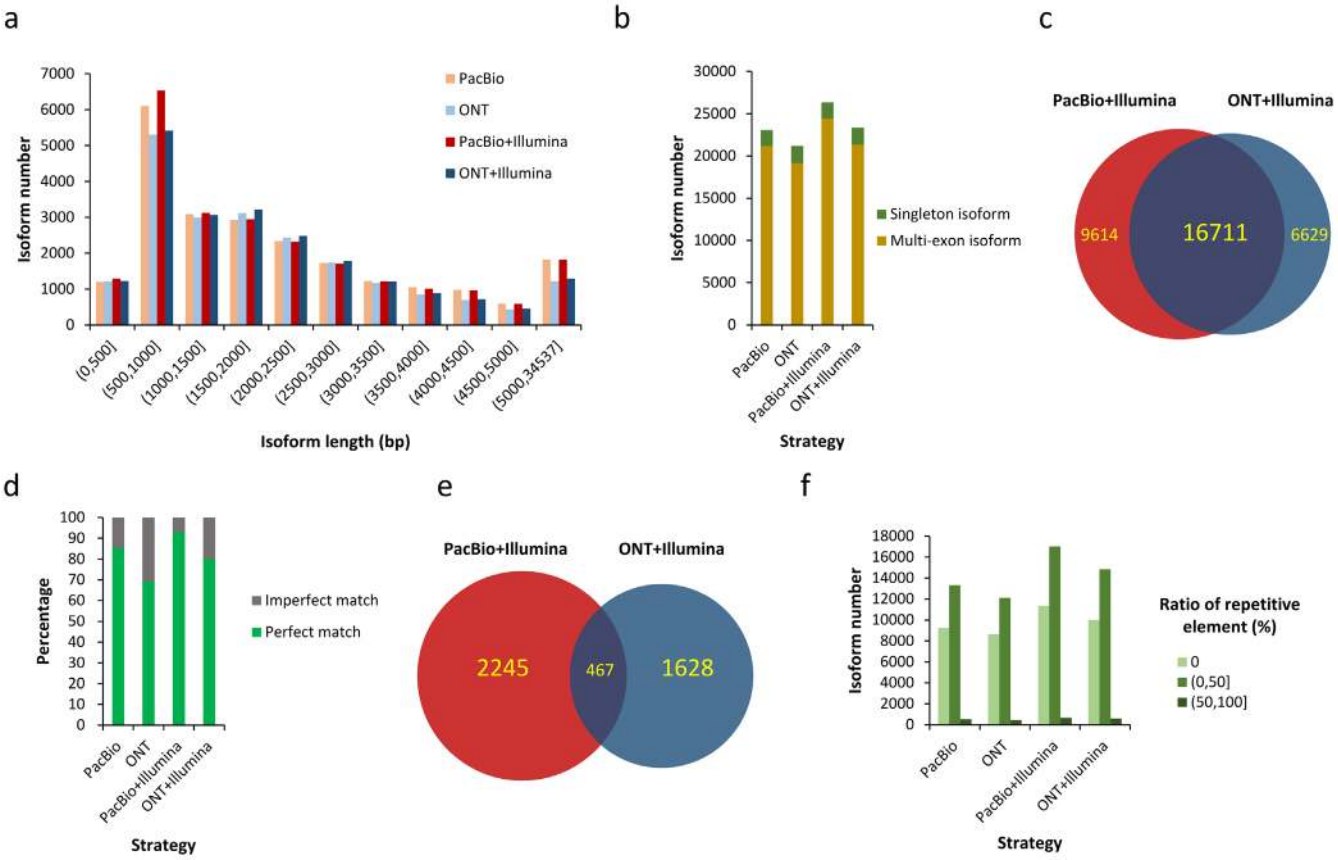

Figure 4. Isoform identification in human embryonic stem cells. (**a**) Length distribution of isoforms identified by full-length by long read only and Hybrid-Seq strategies. (**b**) Numbers of identified isoforms with single exon (singleton isoform) and multiple exons (multi-exon isoform). (**c**) Overlap between isoforms identified by two Hybrid-Seq strategies. (**d**) Accuracy of splice sites detected by four strategies. Perfect means the detected splice sites exactly match known splice sites annotated by Gencode (version 24). Imperfect means the detected splice sites are shorter or longer than known splice sites annotated by Gencode (version 24). (**e**) Overlap between novel isoforms identified by two Hybrid-Seq strategies. (**f**) Numbers of identified isoforms with different ratios of repetitive elements. ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

We also illustrated the utility of long reads to identify isoforms with repetitive elements (see Methods). Approximately 60% of isoforms identified by PacBio (13,830; 59.96%), ONT (12,559; 59.25%), PacBio+Illumina (17,672; 60.86%) and ONT+Illumina (15,426; 60.65%) contained repetitive elements, and in particular, a significant amount of isoforms identified contained > 50% repetitive elements (516, 451, 665 and 593, respectively; Figure 4F). Reconstruction of isoforms with repetitive elements is difficult for short reads[43], while it is relatively easily and accurately accomplished using long reads.

Isoform abundance estimation by PacBio, ONT and Hybrid-Seq

We evaluated the performance of PacBio, ONT, Hybrid-Seq and Illumina data on isoform quantification, using the gold standard SIRVs (see Methods). The abundance of all 68 SIRVs are the same and here we evaluated the estimation of their uniform relative abundance (1/68≈0.15). We first tested Illumina data, with the isoform library reconstructions guided by the three aforementioned SIRV annotation libraries. The median estimation errors were 0.12, 0.18 and 0.12 for "correct", "insufficient" and "over-annotated" annotation libraries, respectively (Figure 5). It suggests isoform abundance estimation is less accurate when expressed, but unannotated isoforms are missed in isoform identification (e.g. the "insufficient" library). In contrast, when isoforms were identified and quantified by Hybrid-Seq, the median estimation errors were as low as 0.06 for PacBio+Illumina and 0.05 for ONT+Illumina. Additionally, we also observed high median estimation errors when using long reads only (0.15 for PacBio and 0.13 for ONT). This reflects the drawbacks of TGS long reads in quantitative analysis, such as low throughput and bias, yet a better isoform library can be obtained than with short reads only. Overall, although the errors from all estimation methods are of the same order of magnitude of the relative abundance (0.15), Hybrid-Seq provides a better strategy to fully utilize PacBio and ONT long reads in transcriptome analysis.

Complexity of the hESC transcriptome

Alternative splicing and alternative polyadenylation, produce a substantial number of isoforms with different lengths, exon usage and polyadenylation sites, which greatly enriches the complexity of the human transcriptome[44–46]. The average lengths of identified isoforms were 1,759bp, 1,670bp, 1,848bp and 1,747bp for PacBio, ONT, PacBio+Illumina and ONT+Illumina, respectively (Supplementary Figure S8). The longest isoform (Gencode ID:

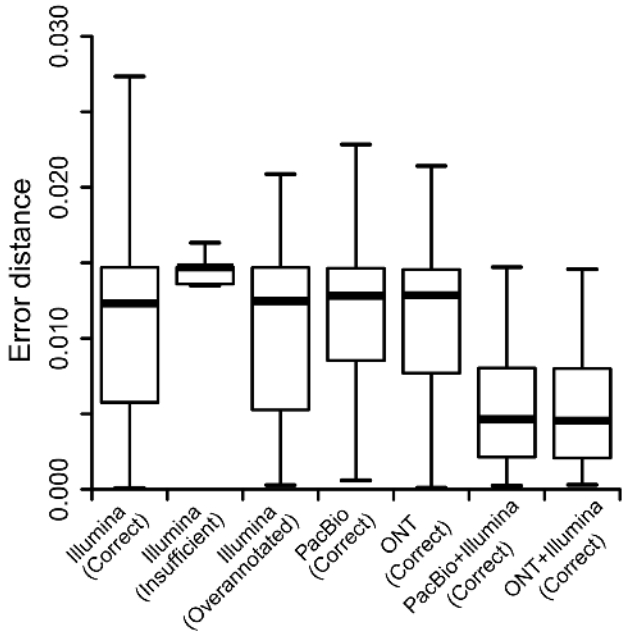

Figure 5. Estimation errors of isoform abundance estimation in Spike-in RNA Variant data. The X axis shows 7 strategies. The label "correct", "insufficient" and "over-annotated" in parentheses represent three different SIRV annotation libraries, respectively. The Y axis shows the euclidean distance between real relative expression percentage (1/68≈0.15) and estimated relative expression percentage (for more details see Methods). ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

ENST00000262160.10), which was simultaneously identified by all four strategies, was 34,537bp.

For multi-exon isoforms, an average of ~8 exons in each isoform was identified by each of the four strategies (Supplementary Figure S9). However, the largest numbers of exons contained within single isoforms differed among PacBio, ONT, PacBio+Illumina and ONT+Illumina datasets: 64, 49, 67 and 52, respectively. When considering the isoforms with ≥30 exons, both PacBio (243) and PacBio+Illumina (367) were capable of identifying more isoforms than ONT (84) and ONT+Illumina (169). These results indicate both technologies can identify isoforms with many exons, but there is not sufficient evidence to reveal conclusive difference between the sequencing platforms. PacBio being size-selected and ONT's lack of size-selection may also have contributed to the observed differences.

Alternative splicing events lead to the diversity of isoform expression. PacBio, ONT, PacBio+Illumina and ONT+Illumina identified 1,076, 1,003, 1,476 and 1,370 alternative splicing events, respectively (Figure 6). On average, the most frequent alternative splicing events identified were exon skipping (37.96%), followed by intron retentions (25.77%), alternative 3' splicing sites (18.62%) and alternative 5' splicing sites (17.07%). A few mutually exclusive exons events (0.57%) were also discovered.

As reported recently, PacBio data can identify alternative polyadenylation sites[23]. In our data, poly(A/T) tails were detected in

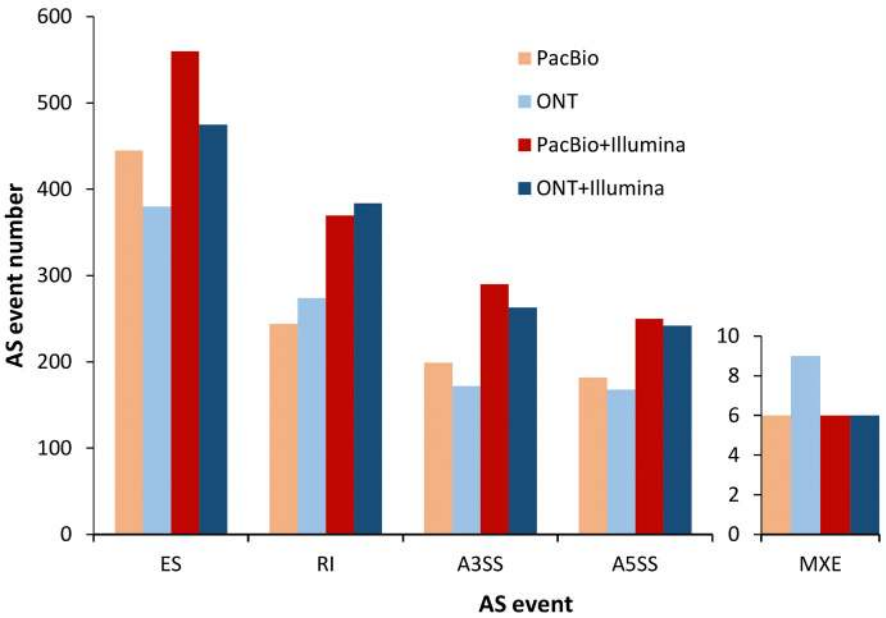

Figure 6. Numbers of different alternative splicing (AS) events in human embryonic stem cells transcriptome. A5SS: alternative 5' splicing site; A3SS: alternative 3' splicing site; ES: exon skipping; RI: retained intron; MXE: mutually exclusive exons; ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

76.71% PacBio CCS reads and 59.75% ONT 2D reads. It shows the comparable potential of ONT to identify alternative polyadenylation sites as PacBio.

Functional analysis of identified isoforms in hESCs

For the Gencode-annotated isoforms identified by PacBio, ONT, PacBio+Illumina and ONT+Illumina, 42.51%, 41.87%, 44.06% and 43.78% were protein-coding, respectively (Figure 7A) and the ratios of pseudogenes were 28.38%, 29.99%, 26.38% and 28.46%. Some isoforms were annotated as retained introns (9.48%, average), lincRNA (4.47%, average) and antisense transcripts (3.02%, average).

For novel isoforms identified by Hybrid-Seq, we evaluated the protein coding potential by GeneMarkS-T (see Methods). Open reading frames (ORFs) with >97 amino acids were found in 92.59% (2,511/2,712) and 89.40% (1,873/2,095) novel isoforms identified by PacBio+Illumina and ONT+Illumina, respectively, with average lengths of 516 and 427 amino acids (Figure 7B). The longest ORFs were 2,302 and 1,980 amino acids, respectively.

We performed gene enrichment analysis for those genes with ≥1 novel isoform. Most genes were enriched in transcription regulation, DNA binding and metal ion binding processes (Figure 7C), which are likely important for human embryonic development. Some other enriched genes have protein kinase activity and are associated with DNA damage response, cell division, cell cycle and RNA processing processes.

Furthermore, 26 hESC-relevant genes expressed ≥1 novel isoform, which was supported by PacBio or ONT full length data (Supplementary Table 3). For example, five novel isoforms (red track in Figure 7D) were full-length identified by both PacBio and ONT long reads in *ESRG* (Embryonic Stem Cell Related Gene), which is required for maintenance of human embryonic stem cell pluripotency[47]. These isoforms were not annotated by the existing annotation libraries (Gencode, Ensembl

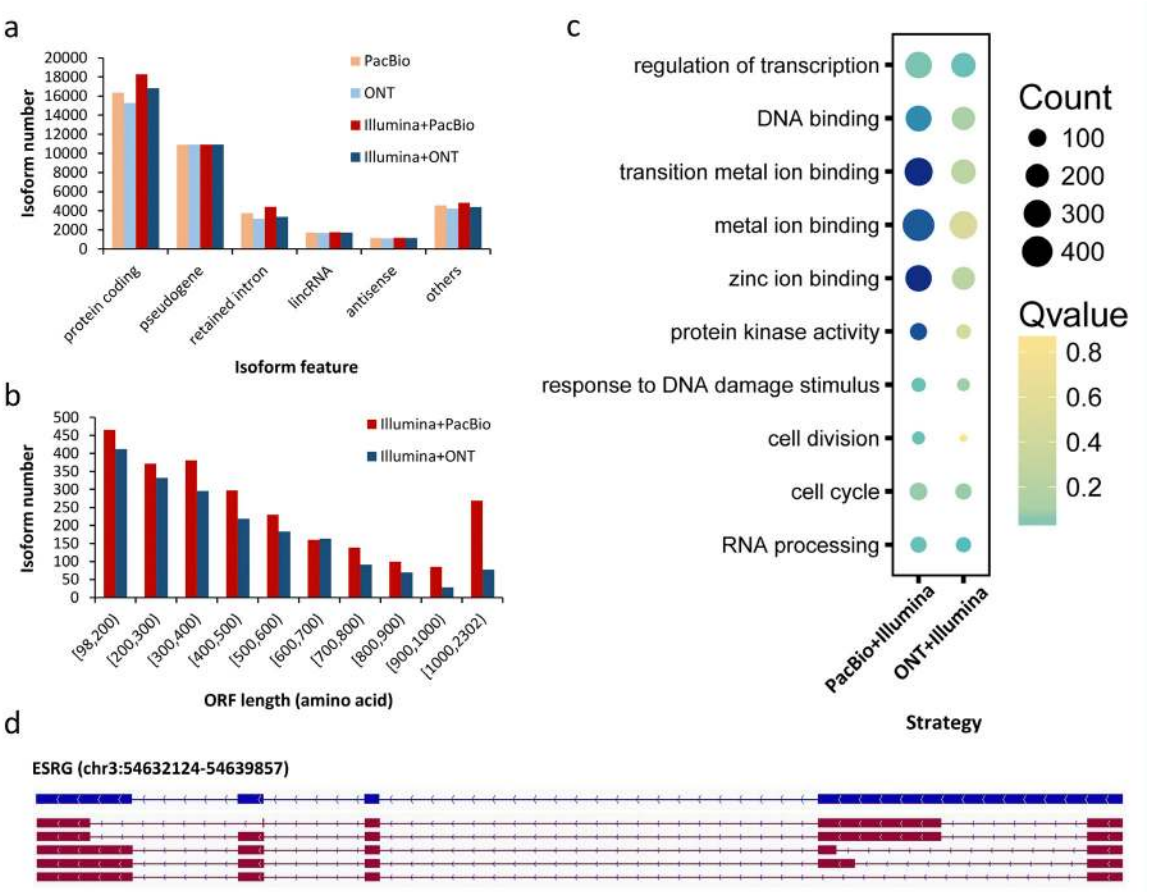

Figure 7. Functional analysis of identified isoforms. (**a**) Feature statistics of isoforms annotated by Gencode (version 24). (**b**) Length distribution of open reading frames (ORFs) of novel isoforms identified by two Hybrid-Seq strategies. (**c**) Gene enrichment analysis of genes with at least one novel isoform identified by two Hybrid-Seq strategies. (**d**) Five novel isoforms (red tracks) of the human embryonic stem cell-relevant gene *ESRG* were identified by two Hybrid-Seq strategies. The topmost isoform (blue track) is annotated by Gencode (version 24). ESRG: Embryonic Stem Cell Related Gene; ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

or RefSeq) and contained alternative 5' splicing sites and alternative 3' splicing sites.

Discussion

Overall, PacBio and ONT are similar: long read length, high error rate and relatively low throughput. However, they have distinct aspects, such as homopolymer error in PacBio and context-specific mismatches in ONT. PacBio sequences a molecule multiple times to generate high-quality consensus data, while ONT can only sequence a molecule twice. Together with the higher quality of the raw data, PacBio can generate extremely-low-error-rate data for high-resolution studies, which is not feasible for ONT. PacBio has better data quality for most aspects, such as error rate and mappability, especially for the consensus data (CCS vs. 2D). However, ONT has a few advantages: in addition to slightly longer mappable length, ONT MinION provides very high throughput as the nanopores can sequence multiple molecules. The cost for our ONT data generation was 1,000–2,000USD. Since sequencing cost is a significant obstacle of TGS application, the relatively high throughput and affordability makes ONT promising for many applications, especially for genome-wide and transcriptome-wide studies, requiring large amounts of data.

With a comprehensive understanding of the data features of PacBio and ONT, we can perform better data analysis and bioinformatics method development. We found a significant number of chimeric reads, which may be generated by either library preparation artifacts or failure of removing adaptors. Thus, it is important to filter these problematic long reads before further analyzing TGS data. However, we cannot filter the data using a simple cutoff: though the subreads and 1D reads are not as accurate as CCS and 2D reads, they are useful because of their reasonable mappability. In particular, error correction by short reads can improve the error rates and increase the mappability. The subreads and 1D reads consist of ~50–60% of the total data provided from the machines, and moreover, many ONT "fail" reads are also mappable, though they are often discarded. Therefore, sophisticated data analysis and bioinformatics methods, such as error correction, are required to rescue or to make better use of these data. The specific error pattern lays the groundwork for better method development. Similarly, the studies of error pattern can also benefit the development and applications of both long-reads only and Hybrid-Seq approaches for nucleotide analysis, such as SNP calling. We notice that our results are subjected to a compound workflow, including library preparation, sequencing, base calling, and analysis software. However, as we used standard protocols/analyses, these results can still serve as an informative reference.

In fact, studies concerning ONT have recently validated its utility in genome assembly[48]. For transcriptome analysis, we demonstrated the capability of both ONT and PacBio to provide precise and complete isoform identification of a small gold standard library SIRVs. For complicated transcriptomes (e.g., hESCs), ONT also provided comparable results to PacBio. However, with the higher data quality, PacBio has a slightly better overall performance, such as discovery of transcriptome complexity and sensitive identification of isoforms. Furthermore, we successfully improved the overall transcriptome analysis by ONT+Illumina, which is the first study to use ONT data in the Hybrid-Seq strategy. This similar improvement is also observed in PacBio Hybrid-Seq over PacBio alone, as reported previously[11], because short reads not only correct the errors of long reads, but also improve abundance estimation and splice site determination. Abundance estimation could be also benefit from a more precise isoform library by Hybrid-Seq. In addition, the requirement of consistency between TGS and SGS data could also filter out many false positives, such as false gene fusion detection from library preparation artifacts. It is notable that PacBio and ONT have their unique discoveries missed by each other, such as novel isoforms.

Additionally, we established that the technology improvements from the previous to the latest sequencing models of both PacBio and ONT are significant, including error rates and yields (especially for ONT). Therefore, the applications of both PacBio and ONT are expected to increase dramatically in the near future, and the results and the comparisons above provide a reference for analyzing PacBio and ONT data. This study also provides an informative paradigm for the application of PacBio and ONT to analyze transcriptomes by long reads only and their corresponding Hybrid-Seq strategies.

Software and data availability

The AlignQC software described herein is freely available for use and can be downloaded from: http://www.healthcare.uiowa.edu/labs/au/AlignQC/.

Source code available from: https://github.com/jason-weirather/AlignQC

Archived source code as at time of publication: doi, 10.5281/zenodo.224125[49] (https://zenodo.org/record/224125#.WHUFN-1WLTcs)

License: Apache 2.0

Reference sequence and annotation versions are described in Supplementary Table 1.

Author contributions

KFA and DB designed the experiments. JLW wrote the AlignQC software. JLW and YW analyzed the data. MC and PP prepared samples for sequencing and performed all ONT sequencing. VS cultured the H1 cell line. XW contributed critical intellectual content. KFA, MC, YW, and JLW wrote the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the National Human Genome Research Institute [R01HG008759 to KFA, YW and JLW]; the institutional fund of Department of Internal Medicine, University of Iowa [to KFA YW and JLW]. The Multidisciplinary Lung Research Career Development Program [T32HL007638 to JLW]; and the National Natural Science Foundation of China [91540204 to XW].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

Thank you to Kristina Thiel, Ph.D., Research Scientist at the University of Iowa Carver College of Medicine, for critical reading of the manuscript.

Supplementary material

Supplementary Figures 1–9 (in zipped file; Click here to access the data.**):**

Figure S1. Read length distribution per-cell. The read length distribution per Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) cells and Oxford Nanopore Technologies (ONT) flow cell are shown. Read counts are plotted from bins of 200bp. Panels represent each type of read generated by the platforms: circular consensus sequence (CCS) and subreads for PacBio, and for ONT there are 2D, 1D template and 1D complement. ONT reads classified as pass or fail are plotted as different colors. Note* that PacBio C2 subreads are plotted on a different y-axis scale than PacBio P6-C4 for visibility of the lower per-cell throughput from the older C2 chemistry. The number of cells (n) and number of reads (**r**) for each technology is listed at the top of each panel.

Figure S2.1. Length distribution of reads and mappability. The length distribution of alignments and their mapped portions shown for Oxford Nanopore Technologies (ONT) 2D, and 1D (template strand) reads and Pacific Biosciences (PacBio) circular consensus sequence (CCS) and subreads in the main text are supplemented here with additional read sets of interest. High quality reads represent ONT 2D reads and PacBio CCS reads. The Raw reads represent ONT template strand reads and PacBio subreads. The 1D complement strand of ONT is now included. Columns have also been added for side-by-side comparison with LSC-corrected reads. Rows contain results for both current PacBio P6-C4 and ONT R9 along with older PacBio C2 and ONT R7 chemistries.

Figure S2.2. Length distribution of reads and mappability for ONT 'fail'. These results show the mappability of 'fail' classified ONT reads.

Figure S3. Mismatch error pattern regardless of platform. A 'C' followed by a 'G' followed by any base or the reverse complement that sequence is a pattern observed cross the low error rate Illumina and is also observed even when reads perfectly match the reference sequence. PacBio: Pacific Biosciences.

Figure S4. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) context-specific errors. Along with the context-specific errors plotted in the main text, this plot adds side-by-side comparisons of LSC-corrected data reads. Each error type ('insertion', 'deletion', and 'mismatch') for each result is individually scaled for better resolution of errors present in each result.

Figure S5. Length statistics of 'partial' isoforms detected by four strategies. ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

Figure S6. Overlap between isoforms identified by two Hybrid-Seq strategies. (**a**) Overlap between multi-exon isoforms. (**b**) Overlap between singleton isoforms. ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

Figure S7. Statistics of splice site accuracy. This figure does not include the perfectly matched splice sites (relative distance is equal to 0). The negative values and positive values represent the truncated (shorter than known splice sites) and elongated (longer than known splice sites) nucleotide distance from the reference splice site, respectively. ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

Supplementary Figure S8. Length statistics of identified isoforms. ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

Supplementary Figure S9. Exon number statistics of identified isoforms. ONT: Oxford Nanopore Technologies; PacBio: Pacific Biosciences.

Supplementary Table 1. Reference sequences and annotations. The source, address, version, and accession numbers are provided, when available, for reference sequences and annotations.

Click here to access the data.

Supplementary Table 2. Summary statistics comparing technologies. The statistics of reads outputted by each technology are organized by row. The colored columns, A–F, represent the subset of the technology being shown. These variables include the Platform (A), Chsemistry (B), and Correction status (C). GeneralType (D) describes whether reads are high quality (HQ) consensus sequences, the raw nucleotides (possibly multiple per molecule), or the single-best aligned sequence representing a single molecule. The ReadType (E) more specifically defines the GeneralType based on the platform-specific outputs. Finally column (F) specifies whether reads were called as 'pass' or 'fail' for the Oxford Nanopore Technologies (ONT) platform. The remaining columns provide yield, length, mappability, and error rate information.

Click here to access the data.

Supplementary Table 3. Novel human embryonic stem cell (hESC)-relevant isoforms. The novel isoforms of 22 hESC-relevant genes are shown along with a functional description, the number of supporting full-length Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) long reads. Mapping information to the hg38 genome is also provided.

Click here to access the data.

References

1. McCarthy A: **Third generation DNA sequencing: pacific biosciences' single molecule real time technology.** *Chem Biol.* 2010; **17**(7): 675–6.
PubMed Abstract | Publisher Full Text

2. Laver T, Harrison J, O'Neill PA, *et al.*: **Assessing the performance of the Oxford Nanopore Technologies MinION.** *Biomol Detect Quantif.* 2015; **3**: 1–8.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Rhoads A, Au KF: **PacBio Sequencing and Its Applications.** *Genomics Proteomics Bioinformatics.* 2015; **13**(5): 278–89.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Lu H, Giordano F, Ning Z: **Oxford Nanopore MinION Sequencing and Genome Assembly.** *Genomics Proteomics Bioinformatics.* 2016; **14**(5): 265–79.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Reuter JA, Spacek DV, Snyder MP: **High-throughput sequencing technologies.** *Mol Cell.* 2015; **58**(4): 586–97.
PubMed Abstract | Publisher Full Text | Free Full Text

6. van Dijk EL, Auger H, Jaszczyszyn Y, *et al.*: **Ten years of next-generation sequencing technology.** *Trends Genet.* 2014; **30**(9): 418–26.
PubMed Abstract | Publisher Full Text

7. Liu L, Li Y, Li S, *et al.*: **Comparison of next-generation sequencing systems.** *J Biomed Biotechnol.* 2012; **2012**: 251364.
PubMed Abstract | Publisher Full Text | Free Full Text

8. McCoy RC, Taylor RW, Blauwkamp TA, *et al.*: **Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements.** *PLoS One.* 2014; **9**(9): e106689.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Zheng GX, Lau BT, Schnall-Levin M, *et al.*: **Haplotyping germline and cancer genomes with high-throughput linked-read sequencing.** *Nat Biotechnol.* 2016; **34**(3): 303–11.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Pendleton M, Sebra R, Pang AW, *et al.*: **Assembly and diploid architecture of an individual human genome via single-molecule technologies.** *Nat Methods.* 2015; **12**(8): 780–6.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Au KF, Sebastiano V, Afshar PT, *et al.*: **Characterization of the human ESC transcriptome by hybrid sequencing.** *Proc Natl Acad Sci U S A.* 2013; **110**(50): E4821–30.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Weirather JL, Afshar PT, Clark TA, *et al.*: **Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing.** *Nucleic Acids Res.* 2015; **43**(18): e116.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Deonovic B, Wang Y, Weirather JL, *et al.*: **IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing.** *Nucleic Acids Res.* 2016; pii: gkw1076.
PubMed Abstract | Publisher Full Text

14. Ip CL, Loose M, Tyson JR, *et al.*: **MinION Analysis and Reference Consortium: Phase 1 data release and analysis [version 1; referees: 2 approved].** *F1000Res.* 2015; **4**: 1075.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Quick J, Quinlan AR, Loman NJ: **A reference bacterial genome dataset generated on the MinION™ portable single-molecule nanopore sequencer.** *Gigascience.* 2014; **3**: 22.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Feng Z, Fang G, Korlach J, *et al.*: **Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic.** *PLoS Comput Biol.* 2013; **9**(3): e1002935.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Koren S, Schatz MC, Walenz BP, *et al.*: **Hybrid error correction and *de novo* assembly of single-molecule sequencing reads.** *Nat Biotechnol.* 2012; **30**(7): 693–700.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Au KF, Underwood JG, Lee L, *et al.*: **Improving PacBio long read accuracy by short read alignment.** *PLoS One.* 2012; **7**(10): e46679.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Salmela L, Rivals E: **LoRDEC: accurate and efficient long read error correction.** *Bioinformatics.* 2014; **30**(24): 3506–14.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Tevz G, McGrath S, Demeter R, *et al.*: **Identification of a novel fusion transcript between human relaxin-1 (*RLN1*) and human relaxin-2 (*RLN2*) in prostate cancer.** *Mol Cell Endocrinol.* 2016; **420**: 159–68.
PubMed Abstract | Publisher Full Text

21. Sharon D, Tilgner H, Grubert F, *et al.*: **A single-molecule long-read survey of the human transcriptome.** *Nat Biotechnol.* 2013; **31**(11): 1009–14.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Tilgner H, Grubert F, Sharon D, *et al.*: **Defining a personal, allele-specific, and single-molecule long-read transcriptome.** *Proc Natl Acad Sci U S A.* 2014; **111**(27): 9869–74.
PubMed Abstract | Publisher Full Text | Free Full Text

23. Abdel-Ghany SE, Hamilton M, Jacobi JL, *et al.*: **A survey of the sorghum transcriptome using single-molecule long reads.** *Nat Commun.* 2016; **7**: 11706.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Minoche AE, Dohm JC, Schneider J, *et al.*: **Exploiting single-molecule transcript sequencing for eukaryotic gene prediction.** *Genome Biol.* 2015; **16**: 184.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Thomas S, Underwood JG, Tseng E, *et al.*: **Long-read sequencing of chicken transcripts and identification of new transcript isoforms.** *PLoS One.* 2014; **9**(4): e94650.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Xu Z, Peters RJ, Weirather J, *et al.*: **Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis.** *Plant J.* 2015; **82**(6): 951–61.
PubMed Abstract | Publisher Full Text

27. Shi L, Guo Y, Dong C, *et al.*: **Long-read sequencing and *de novo* assembly of a Chinese genome.** *Nat Commun.* 2016; **7**: 12065.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Gordon SP, Tseng E, Salamov A, *et al.*: **Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing.** *PLoS One.* 2015; **10**(7): e0132628.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Treutlein B, Gokce O, Quake SR, *et al.*: **Cartography of neurexin alternative splicing mapped by single-molecule long-read mRNA sequencing.** *Proc Natl Acad Sci U S A.* 2014; **111**(13): E1291–9.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Larsen PA, Heilman AM, Yoder AD: **The utility of PacBio circular consensus sequencing for characterizing complex gene families in non-model organisms.** *BMC Genomics.* 2014; **15**(1): 720.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Zhang W, Ciclitira P, Messing J: **PacBio sequencing of gene families - a case study with wheat gluten genes.** *Gene.* 2014; **533**(2): 541–6.
PubMed Abstract | Publisher Full Text

32. Bolisetty MT, Rajadinakaran G, Graveley BR: **Determining exon connectivity in complex mRNAs by nanopore sequencing.** *Genome Biol.* 2015; **16**: 204.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Oikonomopoulos S, Wang YC, Djambazian H, *et al.*: **Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations.** *Sci Rep.* 2016; **6**: 31602.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Sebastiano V, Zhen HH, Haddad B, *et al.*: **Human *COL7A1*-corrected induced pluripotent stem cells for the treatment of recessive dystrophic epidermolysis bullosa.** *Sci Transl Med.* 2014; **6**(264): 264ra163.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Sebastiano V, Maeder ML, Angstman JF, *et al.*: ***In situ* genetic correction of the sickle cell anemia mutation in human induced pluripotent stem cells using engineered zinc finger nucleases.** *Stem Cells.* 2011; **29**(11): 1717–26.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Lamble S, Batty E, Attar M, *et al.*: **Improved workflows for high throughput library preparation using the transposome-based Nextera system.** *BMC Biotechnol.* 2013; **13**: 104.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Picelli S, Faridani OR, Björklund AK, *et al.*: **Full-length RNA-seq from single cells using Smart-seq2.** *Nat Protoc.* 2014; **9**(1): 171–81.
PubMed Abstract | Publisher Full Text

38. Kent WJ, Sugnet CW, Furey TS, *et al.*: **The human genome browser at UCSC.** *Genome Res.* 2002; **12**(6): 996–1006.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** 2011; **17**(1).
Publisher Full Text

40. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics.* 2005; **21**(9): 1859–75.
PubMed Abstract | Publisher Full Text

41. Pertea M, Pertea GM, Antonescu CM, *et al.*: **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol.* 2015; **33**(3): 290–5.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protoc.* 2009; **4**(1): 44–57.
PubMed Abstract | Publisher Full Text

43. Loomis EW, Eid JS, Peluso P, *et al.*: **Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene.** *Genome Res.* 2013; **23**(1):

121–8.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Bentley DL: **Coupling mRNA processing with transcription in time and space.** *Nat Rev Genet.* 2014; **15**(3): 163–75.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Keren H, Lev-Maor G, Ast G: **Alternative splicing and evolution: diversification, exon definition and function.** *Nat Rev Genet.* 2010; **11**(5): 345–55.
PubMed Abstract | Publisher Full Text

46. Elkon R, Ugalde AP, Agami R: **Alternative cleavage and polyadenylation: extent, regulation and function.** *Nat Rev Genet.* 2013; **14**(7): 496–506.
PubMed Abstract | Publisher Full Text

47. Wang J, Xie G, Singh M, *et al.*: **Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells.** *Nature.* 2014; **516**(7531): 405–9.
PubMed Abstract | Publisher Full Text

48. Goodwin S, Gurtowski J, Ethe-Sayers S, *et al.*: **Oxford Nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome.** *Genome Res.* 2015; **25**(11): 1750–6.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Weirather J: **jason-weirather/AlignQC: Current version code accompanying publication [Data set].** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 04 August 2017

https://doi.org/10.5256/f1000research.12836.r23594

✔ **Jingyi Jessica Li**
Department of Statistics, University of California, Los Angeles, Los Angeles, CA, USA

The authors have done a careful revision to successfully address all my questions about the previous version.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 31 July 2017

https://doi.org/10.5256/f1000research.12836.r23595

✔ **Hagen Tilgner**
Weill Cornell Medical College , New York City, NY, USA

Weirather and coworkers have much improved their manuscript. Specifically the rewritten version of the abstract has lowered the probability of misleading readers drastically. There are still three small issues that I'd ask the authors to take care of. Since they are straight-forward, I think the authors can take care of this without further review. So, I'll say congratulations (from my side).

Point 1 (I touched on that in my first review): Regarding Figure 5: Figure 5's legend says "the Y axis shows the euclidean distance between real relative expression percentage and estimated relative expression percentage". Therefore I expect one single value for each dataset (the euclidean

distance taking into account multiple isoforms). But I see a boxplot (representing of course multiple values).
Please clarify what the different data points in each boxplot are. The seven genes? Different subsamples of reads? Something else? I bet that adding one or two phrases to the legend of figure 5 will make it obvious, for readers who were not involved in the research.

Point 2: The text and the legend of Figure 5 says "(1/68≈0.15)", which is off by an order of magnitude.

Point 3 (I touched on that in my first review): In the section about alternative splicing quantification (Figure 6), the 10% cutoff, that I asked the authors to use appears to have changed the results, so that they are consistent with my expectations. Thus, this cutoff should be mentioned for reproducibility (apologies if I overlooked it) and the two publications I mentioned in my previous review cited.

Again, I do not think this requires re-review.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 01 March 2017

https://doi.org/10.5256/f1000research.11392.r20639

?

Hagen Tilgner
Weill Cornell Medical College , New York City, NY, USA

The manuscript by Weirather and coworkers compares two third generation sequencing protocols (Pacific Biosciences – PacBio as well as Oxford Nanopore technologies – ONT) in terms of their performance for RNA sequencing. It concludes that both technologies can be used for transcriptome analysis with PacBio having advantages in terms of sequencing errors and consequently alignability while ONT gives higher sequencing throughput.

Generally speaking, this is an important topic, which many readers will find of interest. The manuscript has a lot of very informative information that can guide researchers in their experiments.
On the flipside there are also a couple of instances where readers might be misled if they are not specialists in the field. I will detail these points of the manuscript below and what I think should be

done in order to address them. The authors should be able to address these issues without many difficulties. This will then be an important contribution to the field.

Strengths:

1) The demonstration of the dependence of sequencing quality (or the Fraction of read aligned) on read length (figure 2) both for single pass reads (subreads for PacBio and 1D for ONT) and for multi-pass consensus reads (CCS for PacBio and 2D reads for ONT) is very useful. Future readers will be able to have a good estimate of what they might expect for their genes of interest.

2) The comparison of the type of error (figure 3) is very useful.

3) Likewise the chimera analysis is useful to understand the limitations one must be aware of when planning experiments.

Weaknesses and solutions and other questions:

1a) The first drawback is that the experiment for PacBio and ONT is not exactly identical. PacBio libraries underwent size selection, whereas ONT libraries did not (as the authors indicate in an upfront way), although in theory, I do not see why this could not have been done for ONT. The reason, I would guess, is that for ONT size fractions are not required (just as they were not in our 2015 synthetic long read isoform paper[1]). Nevertheless, this leaves us with the problem that we cannot exactly understand what are characteristic differences between ONT and PacBio and what may be linked to size selection. I think the authors should indicate in a prominent place (e.g. the abstract) that this is a comparison of a "PacBio experiment using size fractions" and a non-size-selected Oxford Nanopore experiment. This is of importance because many readers may only read the abstract and look at the figures – and the current version could cause them to miss this point.
1b) From the above drawback, it follows that for PacBio the authors need to choose how much sequencing is devoted to the four size bins (1,4,4 and 3 SMRT cells, I believe are chosen) but for ONT this is not done. Therefore the length profile in figure 1 (top) is a function of the Oxford Nanopore system and the cDNA sample only, but the distributions (bottom) for PacBio also depend on the employed size selections and SMRT cell numbers. In principle one could (if one wanted to) make the 500bp-1kb bin the most prominent bin in the PacBio length distribution, by also using 4 SMRT cells for this bin. Conversely one could give more weight to other bins. On the upside, this means one can zoom in on sizes of interest. On the downside, one must carefully consider the implications for the transcriptome of interest. The important point here is, again, that all of this could also have been done for ONT. I suggest to make readers aware of this in an obvious way in the legend of figure 1.
1c) Additionally, in figure 2, I would remove the leftmost boxplot for each panel (the overall Fraction of Read aligned), because in the case of PacBio this would change, if one were to use different amounts of sequencing for different bins (because these bins differ, as the authors show, in terms of alignability). The "Fraction of Read Aligned" broken up by length bins however is highly informative. Please do keep this by all means!

2a) Regarding isoform abundance estimation from SIRVs (figure5): The authors employed the E0 mix of the SIRVs, in which all different isoforms are of equal abundance. This is very different from real-world situations, in which different genes but also different isoforms from the same gene can be of very different expression level. The authors note earlier that ONT has advantages in

sequencing depth (at the cost of quality), which (we would hope) would lead to better isoform quantification for lowly expressed genes and minor isoforms– but using the E0 mix we cannot tell (while we could have, I think with the E1 mix). Reading the paper, I was searching for the use of the E1 and E2 mixes which could have answered these questions. It would be good to point out that lowly expressed gene and minor isoform quantification were not addressed here.
2b) Also, regarding the isoform abundance estimation, my first impression was "these are actually very small errors" when looking at the y-axis of figure 5. My current understanding of the situation is however different: As the authors point out, the actual expression of each isoform is 1/68~=0.015, meaning that the errors are of the same order of magnitude as the (uniformly) expressed transcripts – and a bit less for error corrected reads. If my reading of the situation is accurate, then this should be noted somewhere.

3) Other points:

- Page 8 left column: fig 2 is referenced for "ONT data have particularly higher trans-chimeric rates in very long reads (>4kb)". Shouldn't this be fig. 1 ?

- Page 10, right column, end of first paragraph: when referring to table 2, it is not obvious (apologies, if I missed it) what kind of long-reads (ONT-1D vs. ONT-2D vs. ONT-errorCorrection and PacBio-CCS vs. PacBio-subread vs. PacBio-errorCorrected) are used. Earlier parts of the paper use abbreviations like ONT-1D or PacBio-subreads, but not here.

- A similar statement is true for figure 5 and the corresponding text ("when using long reads only"): it is not clear if PacBio-CCS or PacBio-subreads are used (and the same for ONT) when comparing to the error-corrected subreads.

- For figure 5, it is somewhat difficult to understand, what was exactly done. The authors say that, they used the "Euclidean distance" between REP and estimated REP. The way I understand it, is that the authors calculated REP and estimated REP for each transcript, and then calculated the Euclidean distance for each isoform. In this case (one dimension only) the Euclidean distance reduces to the absolute value of REP minus estimated REP. If this was done, this simpler way of saying it, is advantageous, I believe. Using the word "Euclidean distance" makes me expect multidimensionality. This would suggest that the authors have a vector of isoform expression values for each gene (or maybe multiple samples)? That would imply that the boxplots only represent 7 dots for the 7 SIRV genes...please clarify so that there is no doubt.

- The section "Isoform Identification in hESCs by PacBio, ONT and Hybrid-Seq" is difficult to read. This may stem from the terms "full length rates" and "full-length isoform identification rates". It is not fully clear, if they mean the same or different things; What is exactly meant? Is it "fraction of discovered annotated isoforms that are seen at least once in a full length read" or "fraction of reads that are judged as full-length" or something else? Please clarify.

- Page 12, the third paragraph, regarding the discovery of isoforms with >=30 exons. The correct finding of isoforms with lots of exons of course depends on error-rate (which is linked to getting all splice sites correctly) and having long enough reads. In the absence of a size selection experiment for the Minion, one cannot prove that the observed difference between PacBio and Minion would not be rendered smaller (probably not totally removed though – because of the higher Minion error rate), with a size selection experiment for the

Minion. I would mention that.

- Regarding the quantification of alternative splicing events ... there are many publications that suggested exon skipping is the most frequent type of alternative splicing in humans. There are reports that have reported high occurrence of intron retention – but to my knowledge not that intron retention is more frequent than exon skipping. For example Braunschweig et al, Genome Research 2014[2] using short reads and our own paper Tilgner et al, Nature Biotech, 2015[1] using long reads. The authors could also use a minimum frequency of each single alternative event (e.g. 10% as in the papers referenced above) to distinguish splicing errors and few intermediate RNA molecules from "real" isoforms. This may change the relative abundance of each type of splicing event.

- In the last paragraph on page 12, the word "alterative" is used. I assume this should be "alternative". If this is a spelling mistake, there may be more.

References

1. Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, et al.: Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat Biotechnol*. 2015; **33** (7): 736-42 PubMed Abstract | Publisher Full Text
2. Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, et al.: Widespread intron retention in mammals functionally tunes transcriptomes.*Genome Res*. 2014; **24** (11): 1774-86 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 Jun 2017

Kin Fai Au, Ohio State University, USA

We greatly appreciate your time and thoughtful questions and critiques of our manuscript "Comprehensive comparison of PacBio and Oxford Nanopore Technologies and their applications to transcriptome analysis." These are addressed in this point by point response and in the corresponding manuscript revisions.

HT: The manuscript by Weirather and coworkers compares two third generation sequencing protocols (Pacific Biosciences – PacBio as well as Oxford Nanopore technologies – ONT) in terms of their performance for RNA sequencing. It concludes that both technologies can be used for transcriptome analysis with PacBio having advantages in terms of sequencing errors and consequently alignability while ONT gives higher sequencing throughput.

Generally speaking, this is an important topic, which many readers will find of interest. The manuscript has a lot of very informative information that can guide researchers in their

experiments.
On the flipside there are also a couple of instances where readers might be misled if they are not specialists in the field. I will detail these points of the manuscript below and what I think should be done in order to address them. The authors should be able to address these issues without many difficulties. This will then be an important contribution to the field.

Strengths:

1) The demonstration of the dependence of sequencing quality (or the Fraction of read aligned) on read length (figure 2) both for single pass reads (subreads for PacBio and 1D for ONT) and for multi-pass consensus reads (CCS for PacBio and 2D reads for ONT) is very useful. Future readers will be able to have a good estimate of what they might expect for their genes of interest.

2) The comparison of the type of error (figure 3) is very useful.

3) Likewise the chimera analysis is useful to understand the limitations one must be aware of when planning experiments.

Weaknesses and solutions and other questions:

HT (1a): The first drawback is that the experiment for PacBio and ONT is not exactly identical. PacBio libraries underwent size selection, whereas ONT libraries did not (as the authors indicate in an upfront way), although in theory, I do not see why this could not have been done for ONT. The reason, I would guess, is that for ONT size fractions are not required (just as they were not in our 2015 synthetic long read isoform paper[1]). Nevertheless, this leaves us with the problem that we cannot exactly understand what are characteristic differences between ONT and PacBio and what may be linked to size selection. I think the authors should indicate in a prominent place (e.g. the abstract) that this is a comparison of a "PacBio experiment using size fractions" and a non-size-selected Oxford Nanopore experiment. This is of importance because many readers may only read the abstract and look at the figures – and the current version could cause them to miss this point.

Thank you for strongly making this point. The fact that PacBio was size selected and ONT was not deserves discussion and consideration. In fact, we did try size selection with ONT, but unfortunately it did not work in our hands and we haven't figured out the reason. Size selection is officially recommended for PacBio Iso-Seq protocol and has been validated by many published works, while there is so far no "official" protocol released by ONT. Therefore, transcriptome data collection without size selection was the only successful way that we could perform with ONT platform. We strongly encourage more follow-up studies to figure out an optimal protocol to generate transcriptome data with ONT platform.

Nevertheless, we agree size selection is a critical difference between the two sequencing data collections in our work and needs prominent mention in the manuscript. To this end, we have modified the Abstract, Introduction, and first two figures to make specific mention of this difference. We hope this change will make readers more clearly aware of this difference.

HT (1b): From the above drawback, it follows that for PacBio the authors need to choose how much sequencing is devoted to the four size bins (1,4,4 and 3 SMRT cells, I believe are chosen) but for ONT this is not done. Therefore the length profile in figure 1 (top) is a function of the Oxford Nanopore system and the cDNA sample only, but the distributions (bottom) for PacBio also depend on the employed size selections and SMRT cell numbers. In principle one could (if one wanted to) make the 500bp-1kb bin the most prominent bin in the PacBio length distribution, by also using 4 SMRT cells for this bin. Conversely one could give more weight to other bins. On the upside, this means one can zoom in on sizes of interest. On the downside, one must carefully consider the implications for the transcriptome of interest. The important point here is, again, that all of this could also have been done for ONT. I suggest to make readers aware of this in an obvious way in the legend of figure 1.

We modified the legend of Figure 1 to point out the size selection step in PacBio data. As mentioned above, we did not have a successful experiment doing size-selection of ONT or have an official protocol recommendation. To be clear, we do not want our lack of success in working size-selection into the ONT protocol to be misinterpreted as deficiency in the ONT platform. Rather, we would prefer to defer topic of size selection in ONT until it has been better explored by ourselves or others in the community.

HT (1c): Additionally, in figure 2, I would remove the leftmost boxplot for each panel (the overall Fraction of Read aligned), because in the case of PacBio this would change, if one were to use different amounts of sequencing for different bins (because these bins differ, as the authors show, in terms of alignability). The "Fraction of Read Aligned" broken up by length bins however is highly informative. Please do keep this by all means!

Thank you for this suggestion. While we agree that the most informative parts of the plot are the center and left panels, we feel the leftmost (all aligned reads) plot is somewhat useful for providing an overall view of alignability and would prefer to keep it. In response to the other reviewer's comment, this plot was supplemented with the aligned read counts, which should improve the overall readability.

HT (2a): Regarding isoform abundance estimation from SIRVs (figure5): The authors employed the E0 mix of the SIRVs, in which all different isoforms are of equal abundance. This is very different from real-world situations, in which different genes but also different isoforms from the same gene can be of very different expression level. The authors note earlier that ONT has advantages in sequencing depth (at the cost of quality), which (we would hope) would lead to better isoform quantification for lowly expressed genes and minor isoforms– but using the E0 mix we cannot tell (while we could have, I think with the E1 mix). Reading the paper, I was searching for the use of the E1 and E2 mixes which could have answered these questions. It would be good to point out that lowly expressed gene and minor isoform quantification were not addressed here.

Thank you for the suggestions. We elected to use the E0 mix to have as many fixed variables as we possibly could to get a simple and clear readout on performance. We aimed to evaluate how isoform identification and different types of sequencing

coverage (by long reads or short reads) affect the isoform quantification. For example, hybrid sequencing strategies had better isoform identification by long reads (PacBio or ONT) and better quantitative information from short-read coverage (Illumina) in the statistical model, so they had better accuracy. We agree that including E1 and E2 is good to explore more issues in isoform quantification, such as the lowly-expressed ones. For example, lower sequencing coverage of lowly-expressed transcripts could contribute to the variance of abundance estimation. We could consider a separate manuscript to study all problems of isoform abundance estimation thoroughly.

HT (2b): Also, regarding the isoform abundance estimation, my first impression was "these are actually very small errors" when looking at the y-axis of figure 5. My current understanding of the situation is however different: As the authors point out, the actual expression of each isoform is 1/68~=0.015, meaning that the errors are of the same order of magnitude as the (uniformly) expressed transcripts – and a bit less for error corrected reads. If my reading of the situation is accurate, then this should be noted somewhere.

We are sorry for the unclear description of Figure 5. We revised the section "Isoform abundance estimation by PacBio, ONT and Hybrid-Seq" and the legend of Figure 5 to clarify this issue.

HT (3): Other points:
HT (3a): Page 8 left column: fig 2 is referenced for "ONT data have particularly higher trans-chimeric rates in very long reads (>4kb)". Shouldn't this be fig. 1 ?

Yes, thank you so much for pointing this out. We have made this correction in the manuscript.

HT (3b): Page 10, right column, end of first paragraph: when referring to table 2, it is not obvious (apologies, if I missed it) what kind of long-reads (ONT-1D vs. ONT-2D vs. ONT-errorCorrection and PacBio-CCS vs. PacBio-subread vs. PacBio-errorCorrected) are used. Earlier parts of the paper use abbreviations like ONT-1D or PacBio-subreads, but not here.

We are sorry for confusing labels in Table 2 and main text. In Table 2, "correct" means one of three SIRV annotation libraries ("correct", "insufficient" and "over-annotated"). However, in the end of first paragraph, right column, Page 10, the word "corrected"/"correction" means the sequencing long reads that are corrected by short reads using error-correction software (e.g., LSC). We have added some annotation for Table 2 for better understanding.

HT (3c): A similar statement is true for figure 5 and the corresponding text ("when using long reads only"): it is not clear if PacBio-CCS or PacBio-subreads are used (and the same for ONT) when comparing to the error-corrected subreads.

We are sorry for the unclear figure legend of Figure 5. The x-axis shows the strategy of isoform identification and quantification. Here, the words "correct", "insufficient" and "over-annotated" inside the parentheses represents three different SIRV annotation libraries that were used in the "reference-annotation-guided" mode of StringTie. They

do not represent the types of sequencing reads. We have modified the figure legends to clarify this issue.

In addition, we have updated the section "Short read and long read data processing and alignment" to describe more details about which long reads were used in the analyses. Reads used in the technical comparisons are defined specifically throughout as being either consensus or raw reads (e.g. CCS or subreads). For the transcriptome analyses, both PacBio and ONT reads were comprised of "best reads". These were constructed with the goal of 1) having each molecule represented in the dataset once and only once and 2) choosing the best read of each molecule for transcriptome analysis. Below is the priority order of reads to be selected as the "best read" for each molecule in different analysis strategies:

PacBio (long reads only)

1. **The best aligned CCS read (defined by the number of bases in the read mapped to the reference genome)**
2. **Otherwise, the best aligned subread**

PacBio (long and short reads combined, Hybrid-Seq)

1. **The best aligned CCS read with >2 passes and accuracy greater than 95 (estimated by SMRT Analysis software). Corrected reads were not used here because the consensus is already exceeding typical short read correction.**
2. **Otherwise, the best aligned CCS read corrected by short reads.**
3. **Otherwise, the best aligned subread corrected by short reads.**

ONT (long reads only)

1. **The best aligned 2D read**
2. **Otherwise, the best aligned 1D template read**
3. **Otherwise, the best aligned 1D complement read**

ONT (long and short reads combined, Hybrid-Seq)

1. **The best aligned 2D read corrected by short reads**
2. **Otherwise, the best aligned 1D template read corrected by short reads**
3. **Otherwise, the best aligned 1D complement read corrected by short reads**

So, for example of the "long read only" analysis of ONT, a 2D read was aligned, its best alignment would be used, and 1D reads would not be used.

HT (3d): For figure 5, it is somewhat difficult to understand, what was exactly done. The authors say that, they used the "Euclidean distance" between REP and estimated REP. The way I understand it, is that the authors calculated REP and estimated REP for each transcript, and then calculated the Euclidean distance for each isoform. In this case (one dimension only) the Euclidean distance reduces to the absolute value of REP minus estimated REP. If this was done, this simpler way of saying it, is advantageous, I believe. Using the word "Euclidean distance" makes me expect multidimensionality. This would suggest that the authors have a vector of isoform expression values for each gene (or maybe multiple samples)? That would imply that the boxplots only represent 7 dots for the 7 SIRV genes...please clarify so that there is no doubt.

Thank you for the question. The "Euclidean distance" is the aggregated measure of errors that are the differences between the expected relative expression percentage (REP) and observed REP.

We calculated "Euclidean distance" with multiple dimensions, where each transcript represents one dimension. The expected REP of each transcript is 1/68. The observed REP was calculated by dividing a transcript TPM (or read counts) by the sum of all observed TPMs (or read counts) of 68 SRIV transcripts. Below is the formula:

Total_expression = Isoform1_TPM + Isoform2_TPM + … + Isoform68_TPM

Euclidean_distance = sqrt((Isoform1_TPM/Total_expression-1/68)^2+(Isoform2_TPM/Total_expression-1/68)^2+…+(Isoform68_expression/Total_expression-1/68)^2)

HT (3e): The section "Isoform Identification in hESCs by PacBio, ONT and Hybrid-Seq" is difficult to read. This may stem from the terms "full length rates" and "full-length isoform identification rates". It is not fully clear, if they mean the same or different things; What is exactly meant? Is it "fraction of discovered annotated isoforms that are seen at least once in a full length read" or "fraction of reads that are judged as full-length" or something else? Please clarify.

We are sorry for the unclear description. The terms "full length rates" and "full-length isoform identification rates" mean the same things: *"fraction of discovered annotated isoforms that are seen at least once in a full length read"*. We changed "full length rates" to "full-length isoform identification rates" to for consistency. Please find the detailed definition of "full-length isoform identification rates" in Methods section ("Isoform identification in hESCs by PacBio, ONT and Hybrid-Seq").

HT (3f): Page 12, the third paragraph, regarding the discovery of isoforms with >=30 exons. The correct finding of isoforms with lots of exons of course depends on error-rate (which is linked to getting all splice sites correctly) and having long enough reads. In the absence of a size selection experiment for the Minion, one cannot prove that the observed difference between PacBio and Minion would not be rendered smaller (probably not totally removed though – because of the higher Minion error rate), with a size selection experiment for the Minion. I would mention that.

We agree that it is important to mention the size-selection difference in two sequencing experiments, since it could affect these numbers. We have adjusted the manuscript text accordingly to report the observations, and not to draw conclusions about the technologies relative capabilities.

HT (3g): Regarding the quantification of alternative splicing events … there are many publications that suggested exon skipping is the most frequent type of alternative splicing in humans. There are reports that have reported high occurrence of intron retention – but to my knowledge not that intron retention is more frequent than exon skipping. For example Braunschweig et al, Genome Research 2014[2] using short reads and our own paper Tilgner et al, Nature Biotech, 2015[1] using long reads. The authors could also use a minimum frequency of each single alternative event (e.g. 10% as in the papers referenced above) to distinguish splicing

errors and few intermediate RNA molecules from "real" isoforms. This may change the relative abundance of each type of splicing event.

Thanks for your suggestions. Based on this suggestion, we calculated the minimum frequency of each single alternative splicing event and took 10% as the cut-off. The results showed that exon skipping is the most frequent AS event as the reviewer expected (see the updated Figure 6). We have also updated our analyses in Results section "Complexity of the hESC transcriptome".

HT (3h): In the last paragraph on page 12, the word "alterative" is used. I assume this should be "alternative". If this is a spelling mistake, there may be more.

Thank you for pointing out this typo. We have fixed this in the manuscript.

Competing Interests: No competing interests were disclosed.

Reviewer Report 27 February 2017

https://doi.org/10.5256/f1000research.11392.r19894

?

Jingyi Jessica Li
Department of Statistics, University of California, Los Angeles, Los Angeles, CA, USA

In this paper, the authors provides comprehensive analyses to compare two third-generation sequencing technologies (PacBio and ONT) for RNA sequencing. The comparison was conducted in many aspects, including read lengths, mappability, chimeric and gapped alignments, error patterns, isoform identification, and isoform abundance estimation. To my knowledge, this paper is the first comparison of PacBio and ONT and using each of them in hybrid with Illumina, and its comparison results will provide valuable information about these two third-generation technologies to the transcriptomics field. My comments/questions about some contents in this paper are summarized below.

1. In the isoform identification task, it is unclear how the authors defined "true positive and false positive isoforms" assembled by StringTie from Illumina reads?

2. In Figure 1, why does ONT 2D have more reads than ONT 1D?

3. In the comparison of error patterns, the definition of "homopolymer pattern" is unclear.

4. In Figure 2, only the percentages of mapped reads of each read category are shown. While this is important information, it would be also important to know the absolute numbers of mapped reads in each category.

5. In Figure 3, the row containing labels “A C G T ” above the insertion row should be better placed above the mismatch row.

6. In Table 2, the top row labeling is confusing. It would be clearer to remove "Over-annotated library (100)", "Correct library (68)", and "Insuf cient library (43)" from the top row. Also why does the "Illumina+Insufficient" row have one additional cell?

7. In Figure 4, it would be better to make the circles in Venn diagrams proportional to the numbers?

8. It is unclear why the authors included the insufficient annotation and the overannotated cases in the study of isoform identification and isoform abundance estimation. Since they are only applicable to the Illumina data using StringTie but not relevant to the PacBio or the ONT data, including them seems deviation from the theme of the paper.

9. In Figure 7d, are the seven novel isoforms verified?

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 07 Jun 2017

Kin Fai Au, Ohio State University, USA

We greatly appreciate your time and thoughtful questions and critiques of our manuscript “Comprehensive comparison of PacBio and Oxford Nanopore Technologies and their applications to transcriptome analysis.” These are addressed in this point by point response and in the corresponding manuscript revisions.

JJL: In this paper, the authors provides comprehensive analyses to compare two third-generation sequencing technologies (PacBio and ONT) for RNA sequencing. The comparison was conducted in many aspects, including dread lengths, mappability, chimeric and gapped alignments, error patterns, isoform identification, and isoform abundance estimation. To my knowledge, this paper is the first comparison of PacBio and ONT and using each of them in hybrid with Illumina, and its comparison results will provide valuable information about these two third-generation technologies to the transcriptomics field. My comments/questions about some contents in this paper are summarized below.

JJL (1): In the isoform identification task, it is unclear how the authors defined "true positive and false positive isoforms" assembled by StringTie from Illumina reads?

Thank you for your question. Our original manuscript did not adequately describe the criteria used in this analysis and has been modified accordingly in the revision (see the

second paragraph in Methods "Isoform identification in SIRVs by Illumina, PacBio, ONT"): "For all SIRV isoforms, we classified them into two groups: 1) true positive if the isoform was annotated by SIRV "correct" annotation library; and 2) false positive if not.

JJL (2): In Figure 1, why does ONT 2D have more reads than ONT 1D?

We apologize if the axis labeling and scaling of Figure 1 made this point unclear; raw counts associated with this figure are available in the Supplemental Table 2. As expected, ONT 2D have less reads (289430) than the ONT 1D reads (339651). The Figure 1 legend has been modified to appropriately refer readers to Supplementary Table 2.

JJL (3): In the comparison of error patterns, the definition of "homopolymer pattern" is unclear

Thank you for give us a chance to clarify definitions of "loose" and "tight" homopolymer patterns. If an error rate is much higher when both surrounding bases are the same as the mismatched, inserted or deleted bases, then it indicates that these errors are mostly occurring in a homopolymer runs. In the "loose" homopolymer error pattern, the error rate is high if either surrounding bases are the same as the mismatched, inserted or deleted bases. The context requirement is "looser" than "tight" homopolymer patterns. This is observed as the cross-shaped higher error rates (in the context of Figure 3).

JJL (4): In Figure 2, only the percentages of mapped reads of each read category are shown. While this is important information, it would be also important to know the absolute numbers of mapped reads in each category.

We agree that total number of aligned reads represented in each category would be a very useful addition, and have updated Figure 2 accordingly and added to the figure legend.

JJL (5): In Figure 3, the row containing labels "A C G T " above the insertion row should be better placed above the mismatch row.

.

Thank you for the suggestion. To improve the visual cues in the figure, we have filled out the labeling in Figure 3 around the mismatch patterns.

JJL (6): In Table 2, the top row labeling is confusing. It would be clearer to remove "Over-annotated library (100)", "Correct library (68)", and "Insuf cient library (43)" from the top row. Also why does the "Illumina+Insufficient" row have one additional cell?

Thanks for your suggestion. We revised Table 2 to clearly show our results on isoform identification.

JJL (7): In Figure 4, it would be better to make the circles in Venn diagrams proportional to the numbers?

Yes, we agree with you. For Figure 4e and 4f, the circles in Venn diagrams were made to be proportional to the numbers.

JJL (8): It is unclear why the authors included the insufficient annotation and the overannotated cases in the study of isoform identification and isoform abundance estimation. Since they are only applicable to the Illumina data using StringTie but not relevant to the PacBio or the ONT data, including them seems deviation from the theme of the paper.

For isoform identification, a "reference-annotation-guided" mode is recommended by most short read-based method (e.g. Cufflinks and StringTie). The performance could strongly rely on the reliability and completeness of the reference annotation library. To consider different scenarios, we included three types of reference annotation libraries in the comparison. In detail, we want to prove two points:

1) First, for most non-model organisms, isoform annotation libraries are incompletely annotated and thus insufficient for transcriptome analysis. Recovering un-annotated isoforms that are expressed in the given sample is therefore challenging. As shown in Table 2, StringTie (by Illumina data and "insufficient library") only rescued 5 of 25 un-annotated but truly-expressed isoforms. Second, for well-studied species like human, not all annotated isoforms are expressed in a given sample. Thus, the isoform annotation libraries are usually "over-annotated" for a given sample (e.g., a specific cell line or tissue). Using the "over-annotated library" with 32 unexpressed isoforms, StringTie incorrectly assembled 15 unexpressed isoforms, which highly increased the false positive rate. Therefore, short read-based strategies have an inherent disadvantage to long read-based strategies, and prediction alone is insufficient to overcome this.

2) Long read sequencing technologies (PacBio or ONT) can directly detect expressed isoforms. As shown in Table 2, PacBio and ONT detected 67 and 68 expressed isoforms, respectively. Therefore, both long read-based strategies overcome shortcomings of prediction through direct detection.

JJL (9): In Figure 7d, are the seven novel isoforms verified?

In this study, we did not verify identified novel isoforms. In our future work, we will verify these novel isoforms, especially for those isoforms that are associated with embryonic stem cell identity. To increase the reliability of novel isoforms, we have updated Figure 7d and only retained those novel isoforms that are supported by both PacBio and ONT full-length long reads. The number of supporting long reads can be found in updated Supplementary Table 3. The manuscript has also been updated to reflect this change (see Results "Functional analysis of identified isoforms in hESCs").

Competing Interests: No competing interests were disclosed.

Comments on this article

Version 2

Reader Comment 07 Dec 2018

Julien Lagarde, CRG

Dear authors,
Thank you for this interesting study.
One issue I have is that I could not find a clear mention of the number of false positives for PacBio and ONT in **Table 2** ("**Performance of Illumina, PacBio and ONT on isoform identification in the gold standard SIRVs**"). Or am I to understand that you found none? I would be very interested to see overall sensitivity/precision values for each of the platforms, where any transcript model falling outside the SIRV "correct" set would be considered a false positive (as you did with StringTie models).
Best regards,
Julien Lagarde

Competing Interests: No competing interests were disclosed.

Reader Comment 07 Jul 2017

Anthony Bayega, McGill University, Canada

How does PacBio and Nanopore compare regarding resolution of long homopolymers? I found it hard to determine in many cases if Nanopore data presented was in reference to R7 or R9 Nanopore chemistry, this could be made very clear.

Competing Interests: No competing interests were disclosed.