Comparative assessment of long-read error-correction software applied to RNA-sequencing data

Leandro Lima [1,2,3,*], Camille Marchet [4], Ségolène Caboche [5], Corinne Da Silva [6], Benjamin Istace [6], Jean-Marc Aury [6], Hélène Touzet [4] and Rayan Chikhi [4]

[1]Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558 F-69622 Villeurbanne, France
[2]EPI ERABLE - Inria Grenoble, Rhône-Alpes, France
[3]Università di Roma ”Tor Vergata”, Roma, Italy
[4]CNRS, Université de Lille, CRIStAL UMR 9189, Lille, France
[5]Université de Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019, UMR8204, Center for Infection and Immunity of Lille, Lille, France
[6]Genoscope, Institut de biologie Francois-Jacob, Commissariat à l’Energie Atomique (CEA), Université Paris-Saclay, Evry, France

Abstract

Motivation: Long-read sequencing technologies offer promising alternatives to high-throughput short read sequencing, especially in the context of RNA-sequencing. However these technologies are currently hindered by high error rates that affect analyses such as the identification of isoforms, exon boundaries, open reading frames, and the creation of gene catalogues. Due to the novelty of such data, computational methods are still actively being developed and options for the error-correction of RNA-sequencing long reads remain limited.

Results: In this article, we evaluate the extent to which existing long-read DNA error correction methods are capable of correcting cDNA Nanopore reads. We provide an automatic and extensive benchmark tool that not only reports classical error-correction metrics but also the effect of correction on gene families, isoform diversity, bias toward the major isoform, and splice site detection. We find that long read error-correction tools that were originally developed for DNA are also suitable for the correction of RNA-sequencing data, especially in terms of increasing base-pair accuracy. Yet investigators should be warned that the correction process perturbs gene family sizes and isoform diversity. This work provides guidelines on which (or whether) error-correction tools should be used, depending on the application type.

Benchmarking software: `https://gitlab.com/leoisl/LR_EC_analyser`

Key words: Long reads, RNA-sequencing, Nanopore, Error correction, Benchmark

1 INTRODUCTION

Recent advances in long-read sequencing technology have enabled the sequencing of RNA molecules, using either cDNA-based or direct RNA protocols from Oxford Nanopore (referred to as *ONT* or *Nanopore*) and Pacific Biosciences (*PacBio*). The Iso-Seq protocol from PacBio consists in a size selection step, sequencing of cDNAs, and finally a set of computational steps that produce sequences of full-length transcripts. ONT has three different experimental protocols for sequencing RNA molecules: cDNA transformation with amplification, direct cDNA (with or without amplification), and direct RNA.

Long-read sequencing is increasingly used in transcriptome studies (Sedlazeck *et al.*, 2018; Wang *et al.*, 2016; Byrne *et al.*, 2017; Oikonomopoulos *et al.*, 2016) as they better describe exon/intron combinations (Sedlazeck *et al.*, 2018). For instance the Iso-seq protocol has been used for isoform identification, including transcripts identification (Wang *et al.*, 2016), de novo isoform discovery (Li *et al.*, 2017) and fusion transcript detection (Weirather *et al.*, 2015). Nanopore has recently been used for isoform identification (Byrne *et al.*, 2017) and quantification (Oikonomopoulos *et al.*, 2016).

The sequencing throughput of long-read technologies is significantly increasing over the years. It is now conceivable to sequence a full eukaryote transcriptome using either only long reads, or a combination of high-coverage long and short (Illumina) reads. Unlike the Iso-Seq protocol that requires extensive *in silico* processing prior to primary analysis (Sahlin *et al.*, 2018), raw Nanopore reads can in principle be readily analyzed. Direct RNA reads also permit the analysis of base modifications (Workman *et al.*, 2018), unlike all other cDNA-based sequencing technologies. There also exist circular sequencing techniques for Nanopore such as INC-Seq (Li *et al.*, 2016) which aim at reducing error rates, at the expense of a special library preparation. With raw long reads, it is up to the primary analysis software (typically a mapping algorithm) to deal with sequences that have significant per-base error rate, currently around 13% (Weirather *et al.*, 2017).

In principle, a high error rate complicates the analysis of transcriptomes especially for the accurate detection of exon boundaries, or the quantification of similar isoforms and paralogous genes. Reads need to be aligned unambiguously and with high base-pair accuracy to either a reference genome or transcriptome. Indels (i.e. insertions/deletions) are the main type of errors produced by long-read technologies, and they confuse aligners more than substitution errors (Sović *et al.*, 2016). Many methods have been developed to correct errors in RNA-seq reads, mainly in the short-read era (Tong *et al.*, 2016; Song and Florea, 2015). They no longer apply to long reads because they were developed to deal with low error rates, and principally substitutions. However, a new set of methods have been proposed to correct genomic long reads. There exist two types of long-read error-correction algorithms, those using information from long reads only (*self* or *non-hybrid* correction), and those using short reads to correct long reads (*hybrid* correction). In this article, we will report on the extent to which state-of-the-art tools enable to correct long noisy RNA-seq reads produced by Nanopore sequencers.

Several tools exist for error-correcting long reads, including ONT reads. Even if the error profiles of Nanopore and PacBio reads are different, the error rate is quite similar and it is reasonable to expect that tools originally designed for PacBio data to also perform well on recent Nanopore data. There is, to the best of our knowledge, very little prior work that specifically addresses error-correction of RNA-seq long reads. A notable exception is the PBcR tool, which is mainly designed for genomes but is also evaluated on a Iso-Seq transcriptome (Koren *et al.*, 2012). Here we will take the standpoint of evaluating DNA long-read error-correction tools on RNA-seq data, an application that was likely not considered by the respective tools authors.

We evaluate the following DNA hybrid correction tools: LoRDEC (Salmela and Rivals, 2014), NaS (Madoui *et al.*, 2015), PBcR (Koren *et al.*, 2012), proovread (Hackl *et al.*, 2014); and the following DNA self-correction tools: Canu (Koren *et al.*, 2017), daccord (Tischler and Myers, 2017), LoRMA (Salmela *et al.*, 2016), MECAT (Xiao *et al.*, 2017), pbdagcon (Chin *et al.*, 2013). A majority of hybrid correction methods employ mapping strategies to place short fragments on long reads and correct long read regions using the related short read sequences. But some of them rely on graphs to create a consensus that is used for correction. These graphs are either k-mer graphs (de Bruijn graphs), or nucleotide graphs resulting from multiple alignments of sequences (partial order alignment). For self-correction methods, strategies using the aforementioned graphs are the most common. LSCPlus, a RNA-seq correction tool designed for PacBio reads, was not evaluated as the software webpage was unreachable (Hu *et al.*, 2016). We have selected what we believe is a representative set of tools but there also exist other tools that were not evaluated in this study, e.g. HALC (Bao and Lan, 2017), Falcon_sense (Chin *et al.*, 2016), HG-Color (Morisse *et al.*, 2018), HECIL (Choudhury *et al.*, 2018), MIRCA (Kchouk and Elloumi, 2016), Jabba (Miclotte *et al.*, 2016), nanocorr (Goodwin *et al.*, 2015), nanopolish (Loman *et al.*, 2015), and Racon (Vaser *et al.*, 2017).

Other works have evaluated error correction tools in the context of DNA sequencing. LRCStats evaluates error-correctors in a simulated framework, without the need to align corrected reads (La *et al.*, 2017). A technical report from Bouri and Lavenier (2017) provides an extensive evaluation of PacBio/Nanopore error-correction tools, in the context of de novo assembly. Perhaps the closest work to ours is the AlignQC software (Weirather *et al.*, 2017), which provides a set of metrics for the evaluation of RNA-sequencing long-read dataset quality. In Weirather *et al.* (2017) a comparison is provided between Nanopore and PacBio RNA-sequencing datasets in terms of error patterns, isoform identification and quantification. While Weirather *et al.* (2017) did not compare error-correction tools, we will use and extend AlignQC metrics for that purpose.

In this article, we will focus on the qualitative and quantitative measurements of error-corrected long reads, with transcriptomic features in mind. First we examine basic metrics of error-correction, e.g. mean length, base accuracy, homopolymers errors, and performance (running time, memory) of the tools. Then we ask several questions that are specific to transcriptome applications: (i) how is the number of detected genes, and more precisely the number of genes within a gene family, impacted by read error correction? (ii) Can error correction significantly change the number of reads mapping to genes or transcripts, possibly affecting downstream analysis based on these metrics? (iii) Do error-correction tools perturb isoform diversity, e.g. by having a correction bias towards the major isoform? (iv) What is the impact of error correction on identifying splice sites? To answer these questions, we provide an automatic framework (LC_EC_analyser, see Methods) for the evaluation of transcriptomic error-correction, that we apply to nine different error-correction tools.

2 RESULTS

2.1 Error-correction tools

Tables 1 and 2 present the main characteristics of respectively the hybrid and non-hybrid error-correction tools that were considered in this study. For the sake of reproducibility, in the Supplementary Material Section S1 are described all the versions, dependencies, and parameters. Note that these error-correction tools were all tailored for DNA-seq data except for PBcR. PBcR was ran only in hybrid mode, as the authors suggest using Canu over the non-hybrid mode.

2.2 Evaluation datasets

Our evaluation dataset consists of a single 1D Nanopore run using the cDNA preparation kit of RNA material taken from a mouse brain. We obtained 1,256,967 Nanopore 1D reads representing around 2 Gbp of data with an average size of 1650 bp and a N50 of 1885 bp. An additional Illumina dataset containing 58 million paired-end 151 bp reads was generated using a different cDNA protocol. The Nanopore and Illumina reads from the mouse RNA sample are available in the ENA repository under the following study: PRJEB25574.

2.3 Error-correction improves base accuracy and affects the number of detected genes

Tables 3 and 4 show an evaluation of error-correction based on AlignQC results, for the hybrid and non-hybrid tools, respectively. The per-base error rate is 13.7% in raw reads, 0.3-4.5% for reads corrected using hybrid methods and 2.9-6.4% with self-correctors. As expected the correction rate is better for hybrid correctors leading to a per-base error rate lower than 1% (except for LoRDEC and Proovread/untrimmed, which was equal to 4.5% and 2.6% resp.) because they use additional information from short Illumina reads

Table 1. Main characteristics of the hybrid error correction tools considered in this study

	LoRDEC	NaS	PBcR	Proovread
Reference	Salmela and Rivals (2014)	Madoui *et al.* (2015)	Koren *et al.* (2012)	Hackl *et al.* (2014)
Context	DNA	DNA	mRNA or DNA	DNA
Technology	PacBio or ONT	ONT	PacBio or ONT	PacBio
Main algorithmic idea	Construction of short read dBG, path search between k-mers in long reads	Recruitment of short reads by alignment to long reads, assembly of short reads to correct the long reads	Alignment of short reads to long reads and consensus.	Alignment of short reads to long reads and consensus.

Table 2. Main characteristics of the non-hybrid (self) error correction tools considered in this study

	Canu	daccord	LoRMA	MECAT	pbdagcon
Reference	Koren *et al.* (2017)	Tischler and Myers (2017)	Salmela *et al.* (2016)	Xiao *et al.* (2017)	Chin *et al.* (2013)
Context	DNA	DNA	DNA	DNA	DNA
Technology	PacBio or ONT	PacBio	PacBio or ONT	PacBio or ONT	PacBio
Main algorithmic idea	All-versus-all read overlap, filtering, alignment, DAG from the alignments, highest weight path search.	Multiple dBGs built from overlapping window of long reads alignments, consensus per window	Path search in dBG and multi-iterations.	k-mer based read matching, pairwise alignment between matched reads, alignment-based consensus calling on trivial regions, local POG-based consensus calling on complicated regions.	Align long reads to ”backbone” sequences, correction by iterative directed acyclic graph consensus calling from the multiple sequence alignments.

to correct the long reads. The error rate is around 4-6% for self-correction algorithms, except for LoRMA that reached 2.91%. A detailed error-rate analysis will be carried in Section 2.4.

In terms of number of reads after the correction step, LoRDEC, Proovread/untrimmed, daccord/untrimmed, and pbdagcon returned a number of reads similar to that of the uncorrected (raw) reads. All other softwares split and/or discard reads, likely because full-length error-correction was deemed impossible in some reads. PBcR and LoRMA tend to split reads into two or more shorter reads during the correction step, as they return ∼2x more reads after correction that are also shorter (mean length of respectively 776bp and 497bp, versus 2011bp in raw reads) and overall have significantly less bases in total (loss of respectively 298Mbp and 553Mbp). Canu and MECAT mostly discarded reads (30-33%) resulting in 14-25% less bases in total, with comparable mean length to other tools. Apart from LoRDEC, Proovread/untrimmed, and daccord (trimmed and untrimmed) for which only 85-94% of reads were mapped, corrected reads from all the other tools were mapped at a rate of 98.2-99.4%, showing a significant improvement over raw reads (mapping rate of 83.5%).

Apart from Canu, tools with high mean read length (*i.e.* LoRDEC, Proovread/untrimmed, daccord/untrimmed) showed the lowest percentages of mapped reads, indicating that trimming, splitting or discarding reads seems necessary in order to obtain shorter but overall less error-prone reads. A similar conclusion can be reached by comparing the results of trimmed and untrimmed versions of the same tool: reads corrected with Proovread and daccord in trimmed versions showed higher numbers of mapped reads and bases, and lower per-base error rates. However trimmed reads become 300-600 bases shorter on average, and around 2,000 genes are no longer detected. Therefore it is unclear whether trimming should always be performed by error-correctors in a transcriptomic context.

An important observation is that almost all tools, except for LoRDEC and Proovread/untrimmed, lost at least 1,000 genes after correction. Moreover, three of the tools that have the

Table 3. Statistics of hybrid error correction tools on the 1D run RNA-seq dataset. To facilitate the readability of this table and the next ones, we highlighted values that we deemed satisfactory in green colour, borderline in brown, and unsatisfactory in red, noting that such a classification is somewhat arbitrary.

	Raw	LoRDEC	NaS	PBcR	Proovread	Proovread trimmed
nb of reads	741k	741k	619k	1321k	738k	626k
mapped reads	83.5%	85.5%	98.7%	99.2%	85.5%	98.9%
mean length	2011	2097	1931	776	2117	1796
nb of bases	1313M	1394M	1179M	1015M	1400M	1112M
mapped bases^a	89.0%	90.6%	97.5%	99.2%	92.4%	99.5%
per-base error rate^b	13.72%	4.50%	0.38%	0.67%	2.65%	0.33%
nb of detected genes	16.8k (33.9%)	16.8k (33.9%)	15.0k (30.2%)	15.6k (31.4%)	16.6k (33.4%)	14.6k (29.5%)

^{a}As reported by AlignQC. Percentage of bases aligned among mapped reads, taken by counting the M parts of CIGAR strings in the BAM file. Bases in unmapped reads are not counted.
^{b}As reported by AlignQC, using a sample of 1 million bases from aligned reads segments.

Table 4. Statistics of non-hybrid error correction tools on the 1D run RNA-seq dataset.

	Raw	Canu	daccord	daccord trimmed	LoRMA	MECAT	pbdagcon
nb of reads	741k	519k	675k	840k	1540k	495k	778k
mapped reads	83.5%	99.1%	92.5%	94.0%	99.4%	99.4%	98.2%
mean length	2011	2193	2102	1476	497	1995	1473
nb of bases	1313M	1126M	1350M	1212M	760M	980M	1137M
mapped bases^a	89.0%	92.0%	92.5%	94.7%	99.2%	96.9%	97.0%
per-base error rate^b	13.72%	6.43%	5.20%	4.12%	2.91%	4.49%	5.65%
nb of detected genes	16.8k (33.9%)	12.4k (24.9%)	15.5k (31.3%)	13.9k (28.1%)	6.8k (13.7%)	10.4k (20.9%)	13.2k (26.5%)

^{a}As reported by AlignQC. Percentage of bases aligned among mapped reads, taken by counting the M parts of CIGAR strings in the BAM file. Bases in unmapped reads are not counted.
^{b}As reported by AlignQC, using a sample of 1 million bases from aligned reads segments.

highest number of detected genes (LoRDEC, Proovread/untrimmed, daccord/untrimmed) also have the lowest percentage of mapped reads, hinting that error correction might reduce gene diversity in favor of lower error-rate. It is noteworthy that for some tools (e.g. Canu, MECAT, LoRMA), the number of detected genes can drop by 26%-59% compared to the number of genes reported in raw reads.

Overall, no correction tool outperforms the others across all metrics. We note that a reasonable balance appears to be achieved by NaS and Proovread/trimmed, and that overall, hybrid correctors tend to outperform self-correctors.

2.4 Detailed error-rate analysis

The high error-rate of transcriptome long reads significantly complicates their primary analysis (Križanović *et al.*, 2018). While Section 2.3 presented a general per-base error rate, this section breaks down sequencing errors into several types and examines how each error-correction tool deals with them. The data presented here is a compilation of AlignQC results. Note that AlignQC computed the following metrics only on reads that could be aligned, thus unaligned reads are not counted, yet they may possibly be the most erroneous ones. AlignQC also subsampled aligned reads to around 1 million number of bases to calculate the presented values.

2.4.1 Deletions are the most problematic sequencing errors

Table 5 shows the error rate in the raw reads and in the corrected reads for each tool. In raw reads, deletions are the most prevalent type of errors (7.4% of bases), closely followed by subsitutions (5.1%), then insertions (1.2%). LoRDEC is the least capable of correcting mismatches (2% of them remaining), even though it is a hybrid tool. This is possibly related to the large amount of uncorrected reads in its output, 90k reads out of 741k (12%, as computed by exactly matching raw reads to corrected reads). The other hybrid tools result in less than 1% of substitution errors. Surprisingly, the non-hybrid tools also presented very low mismatches rates: all of them showed rates lower than 1%, except for Canu (1.33%) and daccord/untrimmed (1.1%). This suggests that the rate of systematic substitution errors in ONT data is low, as self-correctors were able to achieve results comparable to the

Table 5. Error rate in the raw reads and in the corrected reads for each tool, on the 1D run RNA-seq dataset, computed from 1M random aligned bases.

	Raw	LoRDEC	NaS	PBcR	Proovread	Proovread trimmed	Canu	daccord	daccord trimmed	LoRMA	MECAT	pbdagcon	pbdagcon trimmed
Error rate	13.72%	4.50%	0.38%	0.67%	2.65%	0.33%	6.43%	5.20%	4.12%	2.91%	4.49%	5.65%	5.71%
Mismatch	5.11%	2.04%	0.20%	0.18%	0.93%	0.13%	1.33%	1.10%	0.67%	0.37%	0.35%	0.50%	0.49%
Deletion	7.40%	2.15%	0.09%	0.30%	1.51%	0.18%	4.82%	3.82%	3.27%	2.51%	4.08%	5.06%	5.17%
Insertion	1.20%	0.32%	0.08%	0.19%	0.22%	0.03%	0.28%	0.28%	0.19%	0.03%	0.06%	0.09%	0.05%

Table 6. Homopolymer error rate in the raw reads and in the corrected reads for each tool, on the 1D run RNA-seq dataset, computed from 1M random aligned bases.

	Raw	LoRDEC	NaS	PBcR	Proovread	Proovread trimmed	Canu	daccord	daccord trimmed	LoRMA	MECAT	pbdagcon	pbdagcon trimmed
Homop. deletion	2.96%	0.77%	0.02%	0.10%	0.46%	0.04%	2.46%	2.14%	2.05%	1.82%	2.05%	2.26%	2.26%
Homop. insertion	0.38%	0.08%	0.01%	0.02%	0.06%	0.01%	0.08%	0.06%	0.03%	0.01%	0.01%	0.02%	0.01%

hybrid ones, even without access to Illumina reads. Still, the three best performing tools were all hybrid (Proovread/trimmed, PBcR and NaS), which should therefore be preferred for applications that require very low mismatch rates.

The contrast between self and hybrid tools is more visible on deletion errors. All hybrid tools outperformed the non-hybrid ones. Although in the hybrid ones, LoRDEC (2.15%) and Proovread/untrimmed (1.51%) still showed moderate rates of deletions, NaS, Proovread/trimmed and PBcR were able to lower the deletion error rate from 7.4% to less than 0.3%. All non-hybrid tools presented a high rate (3% or more) of deletion errors, except LoRMA (2.51%). This comparison suggests that ONT reads exhibit systematic deletions, that cannot be corrected without the help of Illumina data. The contribution of homopolymer errors will be specifically analyzed in Section 2.4.2. Considering insertion errors, all tools performed equally well. It is worth noting that more non-hybrid tools (LoRMA, pbdagcon/untrimmed, pbdagcon/trimmed and MECAT) achieved sub-0.1% insertions than hybrid tools (NaS and Proovread/trimmed).

Overall, hybrid tools outperformed non-hybrid ones in terms of error-rate reduction. However, the similar results obtained by both types of tools when correcting mismatches and insertions, and the contrast in correcting deletions, seem to indicate that the main advantage of hybrid correctors over self-correctors is the removal of systematic errors using Illumina data.

2.4.2 Homopolymer insertions are overall better corrected than deletions In this section we further analyze homopolymers indels, *i.e.* insertion or deletion errors consisting of a stretch of the same nucleotide. Table 6 shows that homopolymer deletions are an order of magnitude more abundant in raw reads than homopolymer insertions. It is worth noting that, by comparing the values for the raw reads in Tables 5 and 6, homopolymers are involved in 40% of all deletions and 31% of all insertions.

A closer look at Table 6 reveals that hybrid error correctors outperform non-hybrid ones, as expected, mainly as homopolymer indels are likely systematic errors in ONT sequencing. Hybrid correctors correct them using Illumina reads that do not contain such biases. Moreover, all tools performed well on correcting homopolymer insertions, reducing the rate from 0.38% to less than 0.1%. In particular, the hybrid tools NaS and Prooovread/trimmed, as well as the non-hybrid ones LoRMA, MECAT and pbdagcon/trimmed reached 0.01% homopolymer insertion error rate. Regarding homopolymer deletions, hybrid tools return less than 0.5% of them, except LoRDEC (0.77%). Non-hybrid tools performed more pooly, returning 1.8-2.4% of homopolymers deletion errors – a small improvement over the raw reads.

NaS and Proovread/trimmed showed the best reduction of homopolymers indels. It is also worth noting that hybrid correctors are able to correct homopolymer deletions even better than non-homopolymer deletions. For instance the ratio of homopolymer deletions over all deletions is 40% in raw reads, and decreases for all hybrid correctors, dropping to 20.2% for NaS and 25.6% for Proovread/trimmed, but increases to at least 43.8% (pbdagcon/trimmed) up to 72.6% (LoRMA) in non-hybrid tools (see Supplementary Material Section S3).

2.5 Error-correction perturbs the number of reads mapping to the genes and transcripts

Downstream RNA-sequencing analyses typically rely on the number of reads mapping to each gene and transcript for quantification, differential expression analysis, etc. In the rest of the paper, we define the **coverage** of a gene or a transcript as the

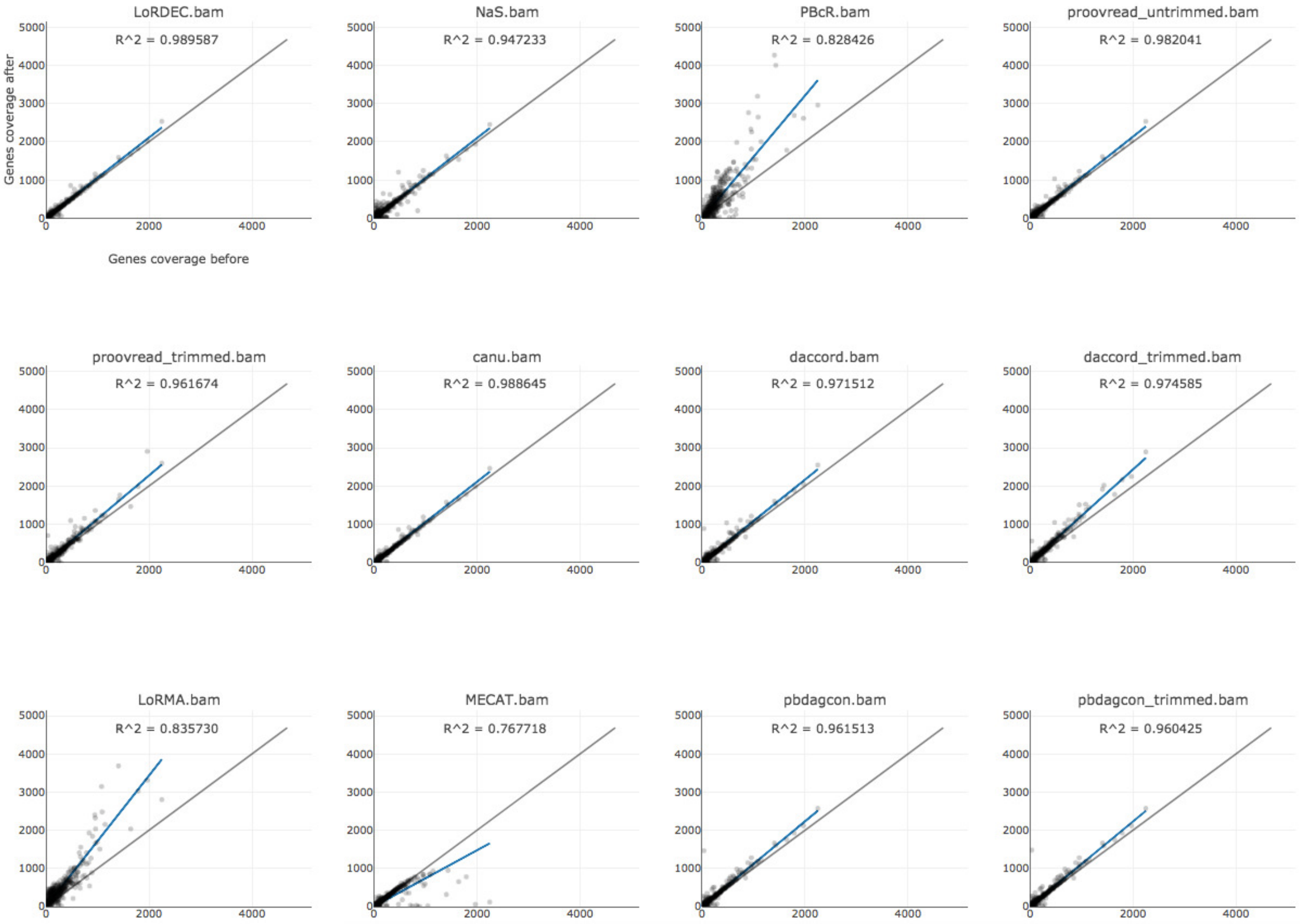

Fig. 1. Number of reads mapping to genes (C_G) before and after correction for each tool. The genes taken into account here were expressed in either the raw dataset or after the correction by the given tool.

number of reads mapping to it. For short we will refer to those coverages as C_G and C_T, respectively. In this section we investigate if the process of error correction can perturb C_G and C_T, which in turn would affect downstream analysis. Note that error correction could potentially slightly increase coverage, as uncorrected reads that were unmapped can become mappable after correction. Figure 1 shows the C_G before and after correction for each tool. PBcR is the only hybrid corrector that significantly inflates C_G, probably due to read splitting (see Section 2.3). Among self-correctors, LoRMA also inflates this value (also due to read splitting), while MECAT presents the lowest correlation and a significant drop in C_G. Besides these three tools, all the others present good correlation and the expected slight increase in C_G due to better mapping. All tools systematically presented a similar trend and lower correlation values on C_T (see Supplementary Material Figure S1), in comparison to C_G. This is expected, as it is harder for a tool to correct a read into its true isoform than into its true gene. The behaviour of the tools in the isoform level are in coherence with their behaviour in the gene level (C_G): PBcR and LoRMA inflates C_T; MECAT deflates; and all the others present a slight increase.

2.6 Error-correction perturbs gene family sizes

Tables 3 and 4 indicate that error correction results in a lower number of detected genes. In this section we explore the impact of error-correction on paralogous genes. By paralogous **gene family**, we denote a set of paralogs computed from Ensembl (see Section 4.3). Figure 2 represents the changes in sizes of paralogous gene families before and after correction for each tool, in terms of number of genes expressed within a given family. Overall, error-correctors do not strictly preserve the sizes of gene families. Correction more often shrinks families of paralogous genes than it expands them, likely due to erroneous correction in locations that are different between paralogs. In summary, 36-86% of gene families are kept of the same size by correctors, 1-12% are expanded and 6-61% are shrunk. Supplementary Material Figure S2 shows the magnitude of expansion/shrinkage for each gene family.

2.7 Error-correction perturbs isoform diversity

We further investigated whether error-correction introduces a bias towards the major isoform of each gene. Note that AlignQC does not directly address this question. To answer it, we computed the

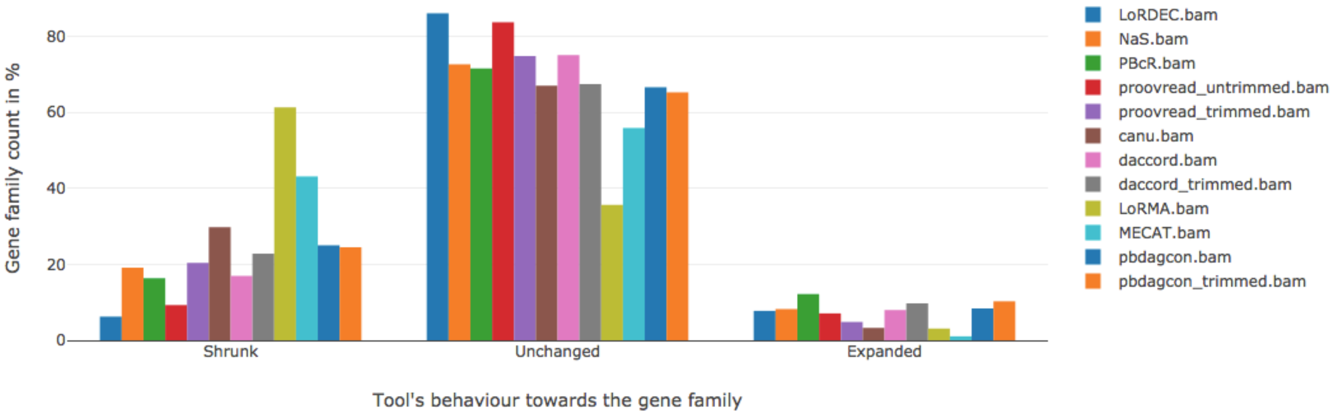

Fig. 2. Summary of gene family size changes across error-correction tools.

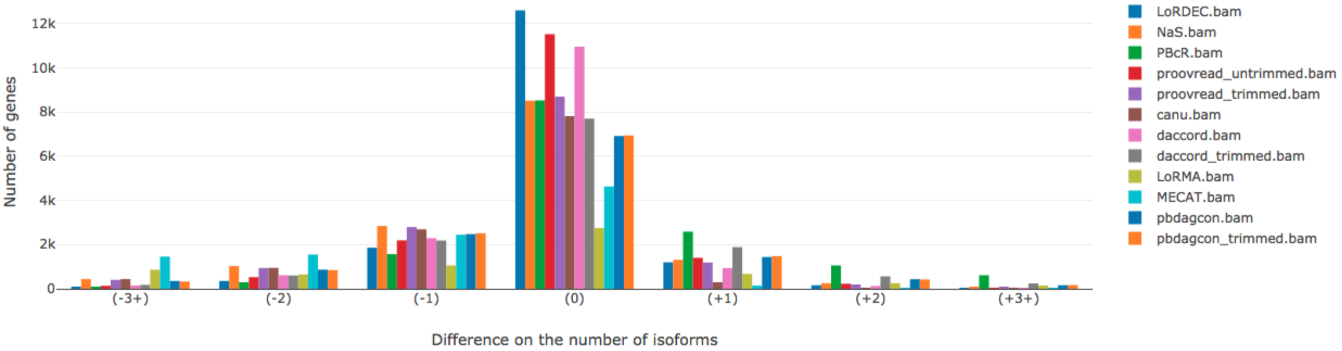

Fig. 3. Histogram of genes having more or less isoforms after error-correction.

following metrics: number of isoforms detected in each gene before and after correction by alignment of reads to genes, coverage of lost isoforms in genes having at least 2 expressed isoforms, and coverage of the major isoform before and after correction.

2.7.1 The number of isoforms varies before and after correction

Figure 3 shows the number of genes that have the same number of isoforms after correction, or a different number of isoforms (-3, -2, -1, +1, +2, +3). In this Figure, only the genes that are expressed in both the raw and the corrected reads (for each tool) are taken into consideration. The negative (resp. positive) values indicate that isoforms were lost (resp. gained). We observe that a considerable number of genes (1k-3k) lose at least one isoform in all tools, which suggests that current methods reduce isoform diversity during correction. NaS and MECAT tend to lose isoforms the most, and PBcR identifies the highest number of new isoforms after correction. It is however unclear whether these lost and new isoforms are real (present in the sample) or due to mapping ambiguity. For instance, PBcR splits corrected reads into shorter sequences that may map better to other isoforms.

Overall, the number of isoforms is mostly unchanged in daccord/untrimmed, LoRDEC and Proovread/untrimmed. We observe that, counter-intuitively, trimming has a slight effect on the number of detected isoforms for Proovread and daccord but not for pbdagcon.

2.7.2 Multi-isoform genes tend to lose lowly-expressed isoforms after correction

Figure 4 explores the relative coverage of isoforms that were possibly lost after correction, in genes having two or more expressed isoforms. The relative coverage of a transcript is the number of raw reads mapping to it over the number of raw reads mapping to its gene in total. Only the genes that are expressed in both the raw and the error-corrected reads (for each tool) are taken into consideration here. We anticipated that raw reads that map to a minor isoform are typically either discarded by the corrector, or modified in such a way that they now map to a different isoform, possibly the major one. The effect is indeed relatively similar across all correctors, except for MECAT that tends to remove a higher fraction of minor isoforms, and LoRDEC that tends to be the most conservative. This result suggests that current error-correction tool overall do not conservatively handle reads that belong to low-expression isoforms.

2.7.3 Coverage of the major isoform before and after correction

To follow-up on the previous subsection, we investigate whether the coverage of the **major isoform**, *i.e.* the isoform with the highest expression in the raw dataset, increased after correction. In Figure 5, We observe that the coverage of the major isoform generally slightly increases after correction, except for MECAT, where its coverage decreases, likely due to a feature of MECAT's own correction algorithm. This indicates that error-correction tools

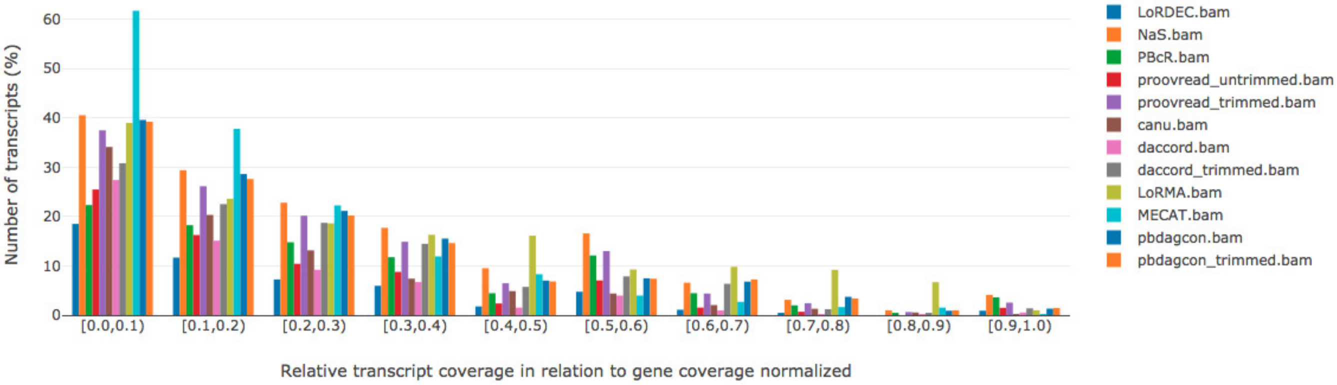

Fig. 4. Histogram of isoforms that are lost after correction, in relation to their relative transcript coverage, in genes that have 2 or more isoforms. The y axis reflects the percentage of isoforms lost in each bin. Absolute values can be found in the Supplementary Material Figure S3.

tend to correct reads towards the major isoform, but the effect is not pronounced. This is expected as the sum of expression of minor isoforms is, by nature, a small fraction of the total gene expression. Apart from LoRMA, MECAT and PBcR, where the correlations of the major isoform coverages are spurious (r^2 ¡ 0.77), other correctors tend to preserve this coverage after correction (r^2=0.90-0.96), with LoRDEC and Canu showing the highest correlations (96%). It is noteworthy that correction biases with respect to the major isoform do not appear to be specific to self correctors nor to hybrid correctors, but an effect that happens in both types of correctors.

2.7.4 Correction towards the major isoform is more prevalent when the alternative exon is small In order to observe if particular features of alternative splicing have an impact on error-correction methods, we designed a simulation over two controlled parameters: skipped exon length and isoform relative expression ratio. Using a single gene, we created a mixture of two simulated alternative transcripts: one constitutive, one exon-skipping. Several simulated read datasets were created with various relative abundances between major and minor isoform (in order to model a local differential in splicing isoform expression), and sizes of the skipped exon. Due to the artificial nature and small size of the datasets, many of the error-correction methods could not be run. We thus tested these scenarii on a subset of the correction methods.

In Figure 6, we distinguish results from hybrid and self-correctors, presented with respectively 100x coverage of short reads and 100x coverage of long reads, and only 100x coverage of long reads. Results on more shallow coverage (10x) and impact of simulation parameters on corrected reads sizes are presented in Supplementary Material Sections S7 and S8. Overall, hybrid correctors are less impacted by isoform collapsing than self-correctors. LoRDEC shows the best capacity to preserve isoforms in presence of alternatively skipped exons. However with less coverage, *e.g.* due to low-expressed genes and rare transcripts, all tools tend to mis-estimate the expression of isoforms (see Supplementary Material). Self-correctors generally have a minimum coverage threshold (only daccord could be run on the 10x coverage dataset of long reads, with rather erratic results, see Supplementary Material). Even with higher coverage, not all correctors achieve to correct this simple instance. Among all correctors, only LoRDEC seems to report the expected number of each isoforms consistently in all scenarios. We could not derive any clear trend concerning the relative isoform ratios, even if the 90% ratio seems to be in favor of overcorrection towards the major isoform. Skipped exon length seems to impact both hybrid and self correctors, small exons being a harder challenge for correctors.

2.8 Error-correction affects splice site detection

The identification of splice sites from RNA-seq data is an important but challenging task (Kaisers *et al.*, 2017). When mapping reads to a (possibly annotated) reference genome, mapping algorithms typically guide spliced alignments using either a custom scoring function that takes into account common splices sites patterns (e.g. GT-AG), and/or a database of known junctions. With long reads, the high error rate make precise splice site detection even more challenging, as indels (see Section 2.4) confuse aligners, shifting predicted spliced alignments away from true splice sites.

In this section, we evaluate how well splice sites are detected before and after error-correction. Figure 7 shows the number of correctly and incorrectly mapped splice sites for the raw and corrected reads, as computed by AlignQC. One would expect that a splice site is correctly detected when little to no errors are present in reads mapping around it. Thus, as expected, the hybrid error correction tools present a clear advantage over the non-hybrid ones, as they better decrease the per-base error rate. In the uncorrected reads, 27% of the splice sites were incorrectly mapped, which is brought down to between 0.28% (Proovread/trimmed) and 2.43% (LoRDEC) with hybrid correction tools. Among non-hybrid tools, LoRMA presented the lowest proportion of incorrectly detected splice sites (3.04%), however it detects 3.5-7x less splice sites (280k) than the other tools (which detect around 1-2 million splice sites). The other non-hybrid tools incorrectly detected splice sites at a rate between 5.61% (daccord/trimmed) and 11.95% (Canu). A detailed analysis of the incorrectly mapped splice sites can be found in the Supplementary Material Section S9.

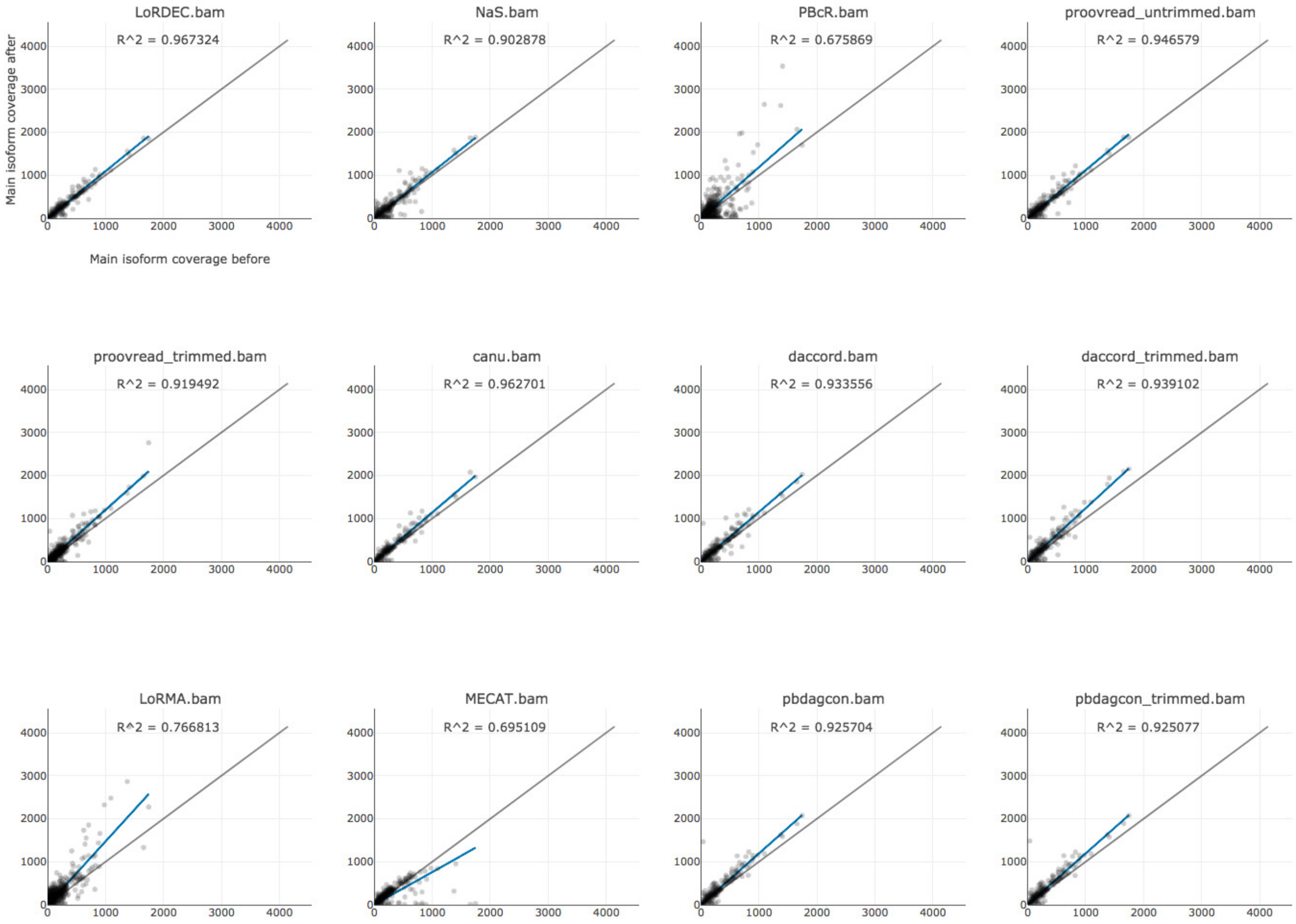

Fig. 5. Coverage of the major isoform of each gene before and after error-correction. The x-axis reflects the number of reads mapping to the major isoform of a gene before correction, and the y-axis is after correction.

2.9 Running time and memory usage of error-correction tools

Table 7 shows the running time and memory usage of all evaluated tools, measured using `GNU time`. The running time shown is the elapsed wall clock time (in hours) and the memory usage is the maximum resident set size (in gigabytes). All tools were ran with 32 threads. Overall, all tools were able to correct the dataset within 0.3-7 hours except for PBcR, NaS and Proovread, which took 63-116 hours, but also achieved the three lowest post-correction error rates in Table 3. In terms of memory usage, all tools required less than 10 GB of memory except PBcR, proovread and LoRMA, which required 53-166 GB. It is worth noting, however, that hybrid error correctors have to process massive Illumina datasets, which contributes to them taking higher CPU and memory usage for correction.

3 DISCUSSION

This work shed light on the versatility of long-read DNA error-correction methods, which can be successfully applied to error-correction of RNA-sequencing data as well. In our tests, error rates can be reduced from 13.7% in the original reads down to as low as 0.3% in the corrected reads. This is perhaps an unsurprising realization as the error-correction of RNA-sequencing data presents similarities with DNA-sequencing data, however this comes with a collection of caveats that we described in the Results section. Most importantly, the number of genes detected by alignment of corrected reads to the genome was reduced significantly by most error-correction methods. Furthermore, depending on the method, error-correction results have a more or less pronounced bias towards correction to the major isoform for each gene, jointly with a loss of the most lowly-expressed isoforms. We provided a software that enables automatic benchmarking of long-read RNA-sequencing error-correction software, in the hope that future error-correction methods will take advantage of it to avoid biases.

The summary statistics of error-corrected data (number of corrected reads, mean length, percentage of mapped reads, per-base error rate, number of detected genes) reveal that no tool outperforms

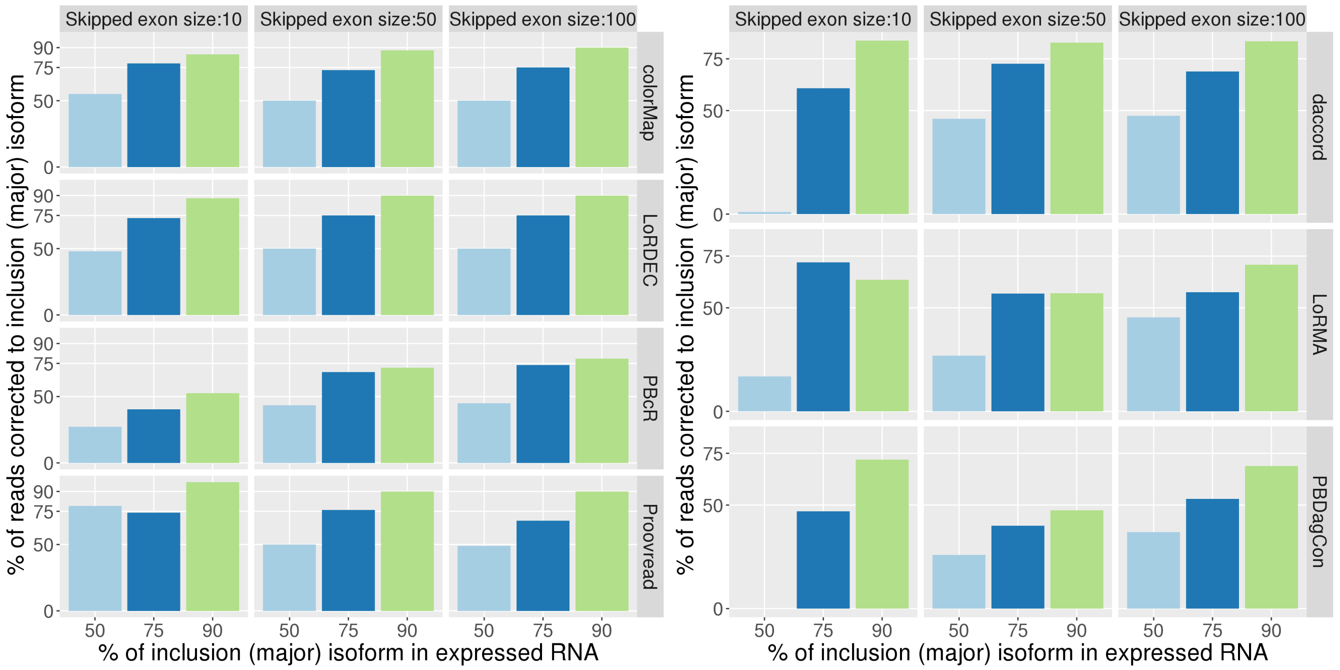

Fig. 6. Mapping of simulated raw and error-corrected reads to two simulated isoforms, and measurements of the percentage of reads mapping to the major isoform. The two isoforms represent an alternatively skipped exon of variable size: 10 bp, 50 bp, 100bp. Left: isoform structure conservation using 100X short reads coverage and 10X long reads, using three error-correction programs, one per row: LoRDEC, PBcR, proovread. Right: same with three self-correctors and 100X long reads: daccord, LoRMA and pbdagcon. Columns are alternative exon sizes. Bars are plots for each isoform ratio (50%; 75% and 90%) on the x-axis. On the y-axis, the closer a bar is to its corresponding ratio value on the x, the better. For instance, the bottom left light blue bar corresponds to a 50% isoform ratio with an exon of size 10, and we do not retrieve a 50% ratio after correction with Proovread (the bar does not go up to 50% on the vertical axis, but around 75% instead). The same layout applies to the right plot, where self-correctors are presented.

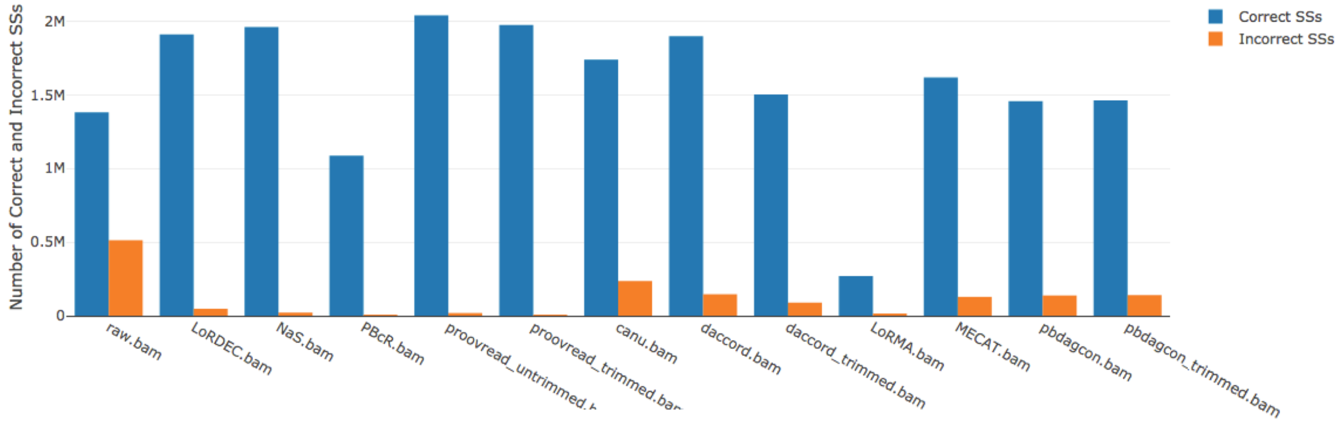

Fig. 7. Statistics on the correctly and incorrectly mapped splice sites (abbreviated SSs) for the uncorrected (raw) and corrected reads.

the others across all metrics, yet a reasonable balance is achieved by NaS and Proovread/trimmed, and that hybrid correction tools generally outperformed the self-correctors.

Detailed error-rate analysis showed that while hybrid correctors have lower error rates than self-correcters, the latter achieved comparable performance to the former in correcting substitutions and insertions. Deletions appear to be caused by systematic sequencing errors, making them fundamentally hard (or even impossible) to address in a self-correction setting. Moreover PBcR, NaS, and Proovread are the most resource-intensive error-correction tools, but also are the only correctors able to reduce base error rate below 0.7%.

We note that LoRDEC, PBcR, Proovread/untrimmed, daccord/untrimmed, and to a lower extent NaS, were able to preserve the number of detected genes better than other correctors. Among those, LoRDEC, Proovread/untrimmed and daccord/untrimmed appear to

Table 7. Running time and memory usage of error-correction tools on the 1D run RNA-seq dataset

	LoRDEC	NaS[1]	PBcR[2]	Proovread	Canu	daccord[3]	daccord trimmed[3]	LoRMA[4]	MECAT	pbdagcon[3]	pbdagcon[3] trimmed
Running time	2.4h	63.2h	116h	107.1h	0.7h	6.9h (7.4h)	6.6h (7.1h)	3.4h	0.3h	5.7h (6.2h)	5.6h (6.1h)
Memory usage	5.6GB	3GB	166.5GB	53.6GB	2.2GB	6.9GB (27.2GB)	6.8GB (27.2GB)	79GB	9.9GB	6.4GB (27.2GB)	6.4GB (27.2GB)

[1] NaS was ran in batches on a different system (TGCC cluster) than other tools; total running time was estimated based on subset of batches.

[2] PBcR was ran on a machine different from the others.

[3] daccord and pbdagcon need DAZZ_DB and DALIGNER to be ran before performing their correction. DAZZ_DB execution time and memory usage was disregarded due to being negligible. DALIGNER, however, took 0.5h and 27.2Gb of RAM. The runtime in parenthesis denotes the runtime of the tool + DALIGNER. The memory usage in parenthesis denotes the maximum memory usage between the tool and DALIGNER.

[4] LoRMA was using more than its allocated 32 cores in some (short) periods of time during the run.

also better preserve the number of detected isoforms better than other correctors. All tools tend to lose lowly-expressed isoforms after correction. This is expected, as these tools were mainly tailored to process DNA data where heterogeneous coverage is not expected. Furthermore, hybrid correctors outperformed self-correctors in the correction of errors near splice site junctions.

As a result, we conclude that no evaluated corrector is the most suited in all situations, and the choice should be guided by the downstream analysis. For quantification, we have shown that error-correction introduces undesirable coverage biases, as per Section 2.5, therefore we would recommend avoiding this step altogether. For isoform detection, LoRDEC, Proovread/untrimmed (hybrid) and daccord/untrimmed (non-hybrid) appear to be the methods of choice as they result in the the highest number of detected genes in Tables 3 and 4 and also preserve the number of detected isoforms as per Section 2.7. For splice site detection, we recommend using hybrid correctors, preferably NaS, PBcR or Proovread, as per Section 2.8. The same three tools (however, Proovread should be in trimmed mode) are also recommended if downstream analyses require very low general error rate. Finally for all other applications, NaS and Proovread/trimmed achieve a reasonable balance across all metrics.

In our analysis, we used a single mapping software (GMAP) to align raw and error-corrected reads, as in previous benchmarks (Weirather *et al.*, 2017; Križanović *et al.*, 2018). We note that other long-read mapping software have since been published, e.g. minimap2 (Li, 2018), which may increase the percentage of mapped read across all methods.

Furthermore, we only focused our evaluation on a single data type: 1D cDNA Nanopore data, using Illumina data for hybrid correction. While it would be natural to also evaluate PacBio data, we note that data from the PacBio Iso-Seq protocol is of different nature as the reads are pre-corrected by circular consensus.

As a side note, AlignQC reports that raw reads contained 1% of chimeric reads, i.e. either portions of reads that align to different loci, or to overlapping loci. The number of chimeric reads after error-correction remains in the 0.7%-1.3% range except for PBcR (0.1%), Proovread/trimmed (0.1%), MECAT (0.1%) and LoRMA (0.04%), which either correctly split reads or discarded chimeric ones.

In the evaluation of tools, we did not record the disk space used by each method, yet we note that it may be a critical factor for some tools (e.g. Canu) on larger datasets. We note also that genes that have low Illumina coverage are unlikely to be well corrected by hybrid correctors. Therefore our comparison does not take into account differences in coverage biases between Illumina and Nanopore data, which may benefit self-correctors. Finally, transcript and gene coverages are derived from the number of long reads aligning to a certain gene/transcript. This method enables to directly relate the results of error-correction to transcript/gene counts, but we note that in current RNA-seq analysis protocols, transcript/gene expression is still generally evaluated using short reads.

4 METHODS

4.1 Nanopore library preparation and sequencing

RNA MinION sequencing cDNA were prepared from 4 aliquots (250ng each) of mouse commercial total RNA (brain, Clontech, Cat# 636601), according to the Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) protocol ”1D cDNA by ligation (SQK-LSK108)”. The data generated by MinION software (MinKNOW 1.1.21, Metrichor 2.43.1) were stored and organized using a Hierarchical Data Format. FASTA reads were extracted from MinION HDF files using poretools (Loman and Quinlan, 2014).

4.2 Illumina library preparation and sequencing

RNA-Seq library preparations were carried out from 500 ng total RNA using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA, USA), which allows mRNA strand orientation (sequence reads occur in the same orientation as anti-sense RNA). After quantification by qPCR, each library was sequenced using 151 bp paired end reads chemistry on a HiSeq4000 Illumina sequencer. Reads were filtered *in silico* to remove mtRNA and rRNA using `BLAT` and `est2genome`.

4.3 Reference-based evaluation of long read error correction

A tool coined LR_EC_analyser, available at `https://gitlab.com/leoisl/LR_EC_analyser`, was developed using the Python language to analyze the output of long reads error correctors. The required arguments are the BAM files of the raw and corrected reads aligned to a reference annotated genome, as well as the reference genome in

Fasta file format and the reference annotation in GTF file format. A file specifying the paralogous gene families can also be provided if plots on gene families should be created. The main processing involves running the AlignQC software (Weirather *et al.*, 2017) (`https://github.com/jason-weirather/AlignQC`) on the input BAMs and parsing its output to create custom plots. It then aggregates information into a HTML report. For example, Tables 3 − 6 are compilations from AlignQC results, as well as Figure 7. Figures 1 − 5 were created processing text files built by AlignQC called ”Raw data” in their output. In addition, an in-depth gene and transcript analysis can be performed using the IGV.js library (`https://github.com/igvteam/igv.js`). In this paper, we did not include all plots and tables created by the tool. To visualise the full latest reports, visit `https://leoisl.gitlab.io/LR_EC_analyser_support/`.

More specifically, in this work we aligned the raw and corrected reads to the Ensembl r87 Mus Musculus unmasked reference genome using the GMAP software (version 2017-05-08 with parameters `-n 10`) (Wu and Watanabe, 2005). The GMAP parameters map those from the original AlignQC publication (Weirather *et al.*, 2015). Gene families were computed by selecting all paralogs from Ensembl r87 mouse genes with 80%+ identity. Note that paralogs from the same family may have significantly different lengths, and no threshold was applied with respect to coverage. The complete selection procedure is reported here: `https://gitlab.com/leoisl/LR_EC_analyser/blob/master/GettingParalogs.txt`.

4.4 Simulation framework for biases evaluation

In the simulation framework of Section 2.7.4, exons length and number were chosen according to resemble what is reported in eukaryotes (Sakharkar *et al.*, 2004) (8 exons, 200 nucleotides). A skipped exon, whose size can vary, was introduced in the middle of the inclusion isoform. Skipped exon can have a size of 10, 50 or 100 nt. We also allowed the ratio of minor/major isoforms (M/m) to vary. For a coverage of C and a ratio M/m, the number of reads coming from the major isoform is MC and the number of minor isoform reads is mC. We chose relative abundances ratios for the inclusion isoform as such: $90/10$, $75/25$ and $50/50$. All reads are supposed to represent the full-length isoform. Finally for hybrid correction input, short reads of length 150 were simulated along each isoform, with 10X and 100X coverage.

During the simulation, we produced two versions of each read. The reference read is the read that represents exactly its isoform, without errors. The uncorrected read is the one in which we introduced errors. We used an error rate and profile that mimics observed R9.4 errors in ONT reads (total error rate of ∼13%, broken down as ∼5% of substitutions, ∼1% of insertions and ∼7% of deletions). After each corrector was applied to the read set, we obtained a triplet (reference, uncorrected, corrected) read that we used to assess the quality of the correction under several criteria.

We mapped the corrected reads on both exclusion and inclusion reference sequences using a fast Smith-Waterman implementation (Zhao *et al.*, 2013), from which we obtained a SAM file. It is expected that exclusion corrected reads will map on exclusion reference with no gaps, and that a deletion of the size of the skipped exon will be reported when mapping them to the inclusion. For each read, if it could be aligned to one of the two reference sequences in one block (according to the CIGAR), then we assigned it to to this reference. If more blocks were needed, we assigned the read to the reference sequence with which the cumulative length of gaps is the lowest. We also reported the ratio between corrected reads size of each isoform kind and the real expected size of each reference isoform.

KEY POINTS

- Long-read transcriptome sequencing is hindered by high error rates that affect analyses such as the identification of isoforms, exon boundaries, open reading frames, and the creation of gene catalogues.
- This review evaluates the extent to which existing long-read DNA error correction methods are capable of correcting cDNA Nanopore reads.
- Existing tools significantly lower the error rate, but they also significantly perturb gene family sizes and isoform diversity.

ACKNOWLEDGEMENTS

LL acknowledges CNPq/Brazil for the support. This work was performed using the computing facilities of the CC LBBE/PRABI and the France Génomique e-infrastructure (ANR-10-INBS-09-08).

FUNDING

This work was supported by the French National Research Agency [ANR ASTER, ANR-16-CE23-0001]; and the Brazilian Ministry of Science, Technology and Innovation (in portuguese, Ministério da Ciência, Tecnologia e Inovação - MCTI) through the National Counsel of Technological and Scientific Development (in portuguese, Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq), under the Science Without Borders (in portuguese, Ciências Sem Fronteiras) scholarship grant [process number 203362/2014-4 to LL].

COMPETING INTERESTS

JMA is one of the authors of the NaS error-correction tool (Madoui *et al.*, 2015). However, this study was designed and performed with no bias towards this particular tool. JMA is part of the MinION Access Programme (MAP) and received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences.

BIOGRAPHICAL NOTE

All authors are part of the ASTER project (ANR ASTER) with the purpose of developing algorithms and software for analyzing third-generation sequencing data.

REFERENCES

Bao, E. and Lan, L. (2017). HALC: High throughput algorithm for long read error correction. *BMC Bioinformatics*, **18**(1), 204.

Bouri, L. and Lavenier, D. (2017). Evaluation of long read error correction software. Technical report, INRIA Rennes - Bretagne Atlantique ; GenScale.

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., DuBois, R. M., Forsberg, E. C., Akeson, M., and Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, **8**, 16027.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., and Korlach, J. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, **10**(6), 563–569.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O’Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., and Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, **13**(12), 1050–1054.

Choudhury, O., Chakrabarty, A., and Emrich, S. J. (2018). HECIL: A Hybrid Error Correction Algorithm for Long Reads with Iterative Learning. *Scientific Reports*, **8**(1), 9936.

Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., and McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, **25**(11), 1750–6.

Hackl, T., Hedrich, R., Schultz, J., and Förster, F. (2014). proovread : large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, **30**(21), 3004–3011.

Hu, R., Sun, G., and Sun, X. (2016). LSCplus: a fast solution for improving long read accuracy by short read alignment. *BMC bioinformatics*, **17**(1), 451.

Kaisers, W., Ptok, J., Schwender, H., and Schaal, H. (2017). Validation of Splicing Events in Transcriptome Sequencing Data. *International journal of molecular sciences*, **18**(6).

Kchouk, M. and Elloumi, M. (2016). Efficient Hybrid De Novo Error Correction and Assembly for Long Reads. In *2016 27th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 88–92. IEEE.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., and Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, **30**(7), 693–700.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive ¡i¿k¡/i¿ -mer weighting and repeat separation. *Genome Research*, **27**(5), 722–736.

Križanović, K., Echchiki, A., Roux, J., and Šikić, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. *Bioinformatics*, **34**(5), 748–754.

La, S., Haghshenas, E., and Chauve, C. (2017). LRCstats, a tool for evaluating long reads correction methods. *Bioinformatics*, **33**(22), 3652–3654.

Li, C., Chng, K. R., Boey, E. J. H., Ng, A. H. Q., Wilm, A., and Nagarajan, N. (2016). INC-Seq: accurate single molecule reads using nanopore sequencing. *GigaScience*, **5**(1), 34.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100.

Li, J., Harata-Lee, Y., Denton, M. D., Feng, Q., Rathjen, J. R., Qu, Z., and Adelson, D. L. (2017). Long read reference genome-free reconstruction of a full-length transcriptome from Astragalus membranaceus reveals transcript variants involved in bioactive compound biosynthesis. *Cell Discovery*, **3**, 17031.

Loman, N. J. and Quinlan, A. R. (2014). Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*, **30**(23), 3399–3401.

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods*, **12**(8), 733–735.

Madoui, M.-A., Engelen, S., Cruaud, C., Belser, C., Bertrand, L., Alberti, A., Lemainque, A., Wincker, P., and Aury, J.-M. (2015). Genome assembly using Nanopore-guided long and error-free DNA reads. *BMC genomics*, **16**(1), 327.

Miclotte, G., Heydari, M., Demeester, P., Rombauts, S., Van de Peer, Y., Audenaert, P., and Fostier, J. (2016). Jabba: hybrid error correction for long sequencing reads. *Algorithms for Molecular Biology*, **11**(1), 10.

Morisse, P., Lecroq, T., and Lefebvre, A. (2018). Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph. *Bioinformatics*.

Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., and Ragoussis, J. (2016). Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. *Scientific Reports*, **6**(1), 31602.

Sahlin, K., Tomaszkiewicz, M., Makova, K. D., and Medvedev, P. (2018). Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. *Nature Communications*, **9**(1), 4601.

Sakharkar, M. K., Chow, V. T. K., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. *In silico biology*, **4**(4), 387–93.

Salmela, L. and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, **30**(24), 3506–3514.

Salmela, L., Walve, R., Rivals, E., and Ukkonen, E. (2016). Accurate self-correction of errors in long reads using de Bruijn graphs. *Bioinformatics*, **33**(6), btw321.

Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics*, **19**(6), 329–346.

Song, L. and Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, **4**(1), 48.

Sović, I., Šikić, M., Wilm, A., Fenlon, S. N., Chen, S., and Nagarajan, N. (2016). Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nature communications*, **7**, 11307.

Tischler, G. and Myers, E. W. (2017). Non Hybrid Long Read Consensus Using Local De Bruijn Graph Assembly. *bioRxiv*, page 106252.

Tong, L., Yang, C., Wu, P.-Y., and Wang, M. D. (2016). Evaluating the impact of sequencing error correction for RNA-seq data with ERCC RNA spike-in controls. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, volume 2016, pages 74–77. IEEE.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research*, **27**(5), 737–746.

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., Lu, Z., Olson, A., Stein, J. C., and Ware, D. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nature Communications*, **7**, 11708.

Weirather, J. L., Afshar, P. T., Clark, T. A., Tseng, E., Powers, L. S., Underwood, J. G., Zabner, J., Korlach, J., Wong, W. H., and Au, K. F. (2015). Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Research*, **43**(18), e116–e116.

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, **6**, 100.

Workman, R. E., Tang, A., Tang, P. S., Jain, M., Tyson, J. R., Zuzarte, P. C., Gilpatrick, T., Razaghi, R., Quick, J., Sadowski, N., Holmes, N., Jesus, J. G. d., Jones, K., Snutch, T. P., Loman, N. J., Paten, B., Loose, M. W., Simpson, J. T., Olsen, H. E., Brooks, A. N., Akeson, M., and Timp, W. (2018). Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*, page 459529.

Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**(9), 1859–1875.

Xiao, C.-L., Chen, Y., Xie, S.-Q., Chen, K.-N., Wang, Y., Han, Y., Luo, F., and Xie, Z. (2017). MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nature Methods*, **14**(11), 1072–1074.

Zhao, M., Lee, W.-P., Garrison, E. P., and Marth, G. T. (2013). SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. *PLoS ONE*, **8**(12), e82138.