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

Leandro Lima\textsuperscript{1,2,3,*}, Camille Marchet\textsuperscript{4}, Ségolène Caboche\textsuperscript{5}, Corinne Da Silva\textsuperscript{6}, Benjamin Istace\textsuperscript{6}, Jean-Marc Aury\textsuperscript{6}, Hélène Touzet\textsuperscript{4} and Rayan Chikhi\textsuperscript{4,7}

\textsuperscript{1}Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558 F-69622 Villeurbanne, France
\textsuperscript{2}EPI ERABLE - Inria Grenoble, Rhône-Alpes, France
\textsuperscript{3}Università di Roma "Tor Vergata", Rome, Italy
\textsuperscript{4}CNRS, Université de Lille, CRISAL UMR 9189, Lille, France
\textsuperscript{5}Université de Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, U1019, UMR8204, Center for Infection and Immunity of Lille, Lille, France
\textsuperscript{6}Genoscope, Institut de biologie Francois-Jacob, Commissariat à l’Energie Atomique (CEA), Université Paris-Saclay, Evry, France
\textsuperscript{7}Institut Pasteur, C3BI - USR 3756, 25-28 rue du Docteur Roux, 75015 Paris, France

\*Corresponding author. Tel.: +33 (0) 4 26 23 44 60; E-mail: leandro.ishi.lima@gmail.com

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Nanopore long-read sequencing technology offers promising alternatives to high-throughput short read sequencing, especially in the context of RNA-sequencing. However this technology is currently hindered by high error rates in the output data 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 Nanopore 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 towards 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 Nanopore 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

Supplementary information: Supplementary data are available at Briefings in Bioinformatics online.

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

The most commonly used technique to study transcriptomes is through RNA sequencing. As such, many tools were developed to process Illumina or short RNA-seq reads. Assembling a transcriptome from short reads is a central task for which many methods are available. When a reference genome or reference transcriptome is available, reference-based assemblers can be used (such as Cufflinks [1], Scallop [2], Scripture [3], and StringTie [4]). When no references are available, de novo transcriptome assembly can be performed (using tools such as Oases [5], SOAPdenovo-Trans [6], Trans-ABySS [7] and Trinity [8]). Potential disadvantages of reference-based strategies include: i) the resulting assemblies might be biased towards the used reference, and true variations might be discarded in favour of known isoforms; ii) they are unsuitable for samples with a partial or missing reference genome [9]; iii) such methods depend on correct read-to-reference alignment, a task that is complicated by splicing, sequencing errors, polyploidism, multiple read mapping, mismatches caused by genome variation, and the lack or incompleteness of many reference genomes [10]; iv) sometimes, the model being studied is sufficiently different from the reference because it comes from a different strain or line such that the mappings are not altogether reliable [11].

On the other hand, some of the shortcomings of de novo transcriptome assemblers are: i) low-abundance transcripts are likely to not be fully assembled [12]; ii) reconstruction heuristics are usually employed, which may lead to missing alternative transcripts, and highly similar transcripts are likely to be assembled into a single transcript [13]; iii) homologous or repetitive regions may result in incomplete assemblies [14]; (iv) accuracy of transcript assembly is called into question when a gene exhibits complex repetitive regions may result in incomplete assemblies [14]; (v) isoform expression [15]; (vi) reconstruction heuristics are usually employed, which may lead to missing alternative transcripts, and highly similar transcripts are likely to be assembled into a single transcript [13]; (vii) homologous or repetitive regions may result in incomplete assemblies [14]; (viii) isoform expression [15].

Recent advances in long-read sequencing technology have enabled longer, up to full-length sequencing of RNA molecules. This new approach has the potential to eliminate the need for transcriptome assembly, and thus also eliminate from transcriptome analysis pipelines all the biases caused by the assembly step. Long read sequencing can be done 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, not just to prevent problems caused by short-read transcriptome assembly, but also for several of the following reasons. Mainly, long reads can better describe exon/intron combinations [16]. The Iso-seq protocol has been used for isoform identification, including transcripts identification [17], de novo isoform discovery [18] and fusion transcript detection [19]. Nanopore has recently been used for isoform identification [20] and quantification [21].

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 [22], raw Nanopore reads can in principle be readily analyzed. Direct RNA reads also permit the analysis of base modifications [23], unlike all other cDNA-based sequencing technologies. There also exist circular sequencing techniques for Nanopore such as INC-Seq [24] 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% [25].

In principle, a high error rate in the data 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 [26]. Many methods have been developed to correct errors in RNA-seq reads, mainly in the short-read era [27-29]. 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. Notable exceptions include: a) LSC [30], which is designed to correct PacBio RNA-seq long reads using Illumina RNA-seq short reads; b) PBcR [31] and c) HALC [32], which are mainly designed for genomes but are also evaluated on transcriptomic data. Here we will take the standpoint of evaluating long-read error-correction tools on RNA-seq data, most of which were designed to process DNA sequencing data only.

We evaluate the following DNA hybrid correction tools: HALC [32], LoRDEC [33], NaS [34], PBcR [35], proovread [36], and the following DNA self-correction tools: Canu [37], daccord [38], LoRMA [39], MECAT [40], pbChinese [41]. We also evaluate an additional hybrid tool, LSC [30], the only one specifically designed to error correct (PacBio) RNA-seq long reads. 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 Kruijn 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. We have also considered evaluating nanocorr [42], Nanopolish [43], Falcon_sense [44], and LSCPlus [45], but some tools were deprecated, not suitable for read correction, or unavailable. Our detailed justifications can be found in Section S1.12 of the Supplementary Material. We have selected what we believe is a representative set of tools but there also exist other tools that were not considered in this study, e.g. HG-Color [46], HECIL [47], MIRCA [48], Jabba [49], nanocorr [50], and Ronac [51].

Other works have evaluated error correction tools in the context of DNA sequencing. LRCStats [52], and more recently ELECTOR [53] provide automated evaluations of genomic long read correction using a simulated framework. A technical report from Boui and Lavener [54] provides an extensive evaluation of PacBio/Nanopore error-correction tools, in the context of de novo assembly. This analysis is completed with more recent results in Fu et al. [55] on hybrid correction methods. Perhaps the closest work to ours is the AlignQC software [56], which provides a set of metrics for the evaluation of RNA-sequencing long-read dataset quality. In Weirather et al. [57] a comparison is provided between Nanopore and PacBio RNA-sequencing datasets in terms of error patterns, isoform identification and quantification. While Weirather et al. [57] did not compare error-correction tools, we will use and extend AlignQC metrics for that purpose.
Comparative assessment of long-read error-correction software applied to Nanopore RNA-sequencing data

In this article, we will focus on the qualitative and quantitative measurements of Nanopore 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 methods, that we apply to eleven different error-correction tools.

2 Results

2.1 Error-correction tools

Table 1 presents the main characteristics of 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 LSC.

The output of each error correction method can be classified into one of the four following types: full-length, trimmed, split, and micro-assembly. Usually, due to methodological reasons, extremities of long reads are harder to correct. As an example, hybrid correctors based on mapping short to long reads, and calling a consensus from the mapping, have difficulties aligning short reads to the extremities of long reads. As such, some methods output trimmed error-corrected reads, i.e. error-correction tools such that their uncorrected ends are removed. Examples of methods producing this type of output considered in this study are HALC, LoRDEC, LSC, proovread, daccord, and pb-dagcon. Sometimes, internal parts of long reads can also be hard to correct, due to a lack of coverage of short reads, or a drop of sequencing quality, or due to mapping issues. Some algorithms thus output split error-correction reads, splitting one long read into several well-corrected fragments, such as HALC, LoRDEC, PBKR, and LoRMA.

Finally, some tools decide to not trim nor split the original reads, outputting full-length error-corrected reads. Examples include HALC, LoRDEC, LSC, proovread, canu, daccord, MECAT, and pb-dagcon. NaS does not fit the previous three categories, as it uses a micro-assembly strategy. Outputs that have no suffixes are considered full-length corrections. For example, HALC denotes the HALC full-length error-corrected reads, HALC(t), the HALC trimmed output, and HALC(s), the HALC split output. As we will see, there is no type of output that outperforms all the others in all metrics. Choosing the appropriate type of output is heavily dependent on the application.

2.2 Evaluation datasets

Our main evaluation dataset consists of a single 1D Nanopore run using the cDNA preparation kit of RNA material taken from a mouse brain, containing 740,776 long reads. An additional Illumina dataset containing 58 million paired-end 151 bp reads was generated on the same sample but using a different cDNA protocol. For more details on the sequencing protocol, see Section 3. The Nanopore and Illumina reads from the mouse RNA sample are available in the ENA repository under the following study: PRJEB25574. In this paper, we provide a detailed analysis of this dataset, from Section 2.3 to Section 2.11.

In order to obtain a more comprehensive understanding of the evaluated tools, we further analysed the correction of the methods on one human Nanopore direct RNA sequencing data from the Nanopore-WGS-Consortium (dataset fromcentre Bham, run#, sample type RNA, kit SQK-RNA001, pore R9.4, available at https://github.com/nanopore-wgs-consortium/NA12878/raw/1dbb/master/nanopore-human-transcriptome/fastq/NA12878.fastq_bulk.md). We concatenated the fail and pass RNA-direct reads from the aforementioned dataset, obtaining 894,289 reads. Further, to correctly run all tools, we transformed bases U into T.

2.3 Error-correction improves base accuracy and splits, trims, or entirely removes reads

Table 2 shows an evaluation of error-correction based on AlignQC results, for the hybrid and non-hybrid tools. The error rate is 13.72% in raw reads, 0.33-5.45% for reads corrected using hybrid methods and 2.91-6.43% with self-correctors. Notably, the hybrid tools NaS(t), Proovread(t), and HALC(s) output micro-assembled, trimmed and split error-corrected reads, respectively, with the lowest error rates (0±5%). We observe that HALC produced the full-length error-corrected reads with the lowest error-rate (1.85%), but that is still significantly higher than the error-rate of the three aforementioned methods. This is expected, as micro-assembling, trimming or splitting reads usually do not retain badly corrected regions of the reads, lowering the error rate. LoRMA(s), which is the only split self-correction tool, was the one that decreased the error-rate the most among non-hybrid tools, but still just managed to reach 2.91%, one order of magnitude higher than the best hybrid correctors. If we look at non-split outputs among the self-correctors, MECAT and daccord(t) obtained the lowest error rates for full-length and trimmed error-corrected reads, respectively, but still presenting an error-rate higher that 4%. It is not a surprise that the best error correctors are hybrid when looking at the error rates, given their usage of additional high-quality Illumina reads. As expected, trimming and splitting error-corrected reads reduces a lot the error-rate, e.g. LoRDEC reduced the error rate from 4.5% to 3.73% by trimming, and to 1.59% by splitting. As such, the split output consistently outperformed trimmed and full-length outputs, regarding the error-rate.

A detailed error-rate analysis will be carried in Section 4.2.3.

In terms of throughput after the correction step, tools that do not trim nor split reads tend to return a number of reads similar to that of the uncorrected (raw) reads. Notably, HALC and LoRDEC returned exactly the same number of reads, and Proovread returned just 3k less reads. On the other hand, Canu and MECAT decreased a lot the number of output reads, probably due to post-filtering procedures. Moreover, many of full-length outputs (HALC, LoRDEC, LSC, proovread, and daccord) increased the mean length of the raw reads while also increasing the number of output bases, showing that they tend to further extend the information contained in the long reads.

Trimming almost always decreased the number of output reads, like in HALC(t), LoRDpec(t), Proovread(t), and pb-dagcon(t), probably due to post-filtering procedures. However, in LSC(t), trimming has no effect on the number of reads, and in daccord(t), trimming actually increased the number of reads. In half of the trimmed outputs (HALC(t), LoRDEC(t), and LSC(t)), the mean length of the reads was usually preserved, decreasing only by around 100bps. However, in the other half
Table 1. Main characteristics of the error correction tools considered in this study
(A) Hybrid tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Context</th>
<th>Technology</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>HALC</td>
<td>DNA</td>
<td>PacBio</td>
<td>[27]</td>
</tr>
<tr>
<td>LoRDEC</td>
<td>DNA</td>
<td>PacBio</td>
<td>[28]</td>
</tr>
<tr>
<td>LSC</td>
<td>DNA</td>
<td>PacBio or ONT</td>
<td>[25]</td>
</tr>
<tr>
<td>NaS</td>
<td>DNA</td>
<td>ONT</td>
<td>[29]</td>
</tr>
<tr>
<td>PlcR</td>
<td>DNA</td>
<td>PacBio or ONT</td>
<td>[30]</td>
</tr>
<tr>
<td>Proovread</td>
<td>DNA</td>
<td>PacBio</td>
<td></td>
</tr>
</tbody>
</table>

(B) Non-hybrid tools

<table>
<thead>
<tr>
<th>Tool</th>
<th>Context</th>
<th>Technology</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canu</td>
<td>DNA</td>
<td>PacBio or ONT</td>
<td>[31]</td>
</tr>
<tr>
<td>daccord</td>
<td>DNA</td>
<td>PacBio</td>
<td>[32]</td>
</tr>
<tr>
<td>LoRMA</td>
<td>DNA</td>
<td>PacBio or ONT</td>
<td>[33]</td>
</tr>
<tr>
<td>MECAT</td>
<td>DNA</td>
<td>PacBio or ONT</td>
<td>[34]</td>
</tr>
<tr>
<td>pbdagcon</td>
<td>DNA</td>
<td>PacBio</td>
<td>[35]</td>
</tr>
</tbody>
</table>

These observations indicate that care should be taken when considering which type of output should be used. For example, all split and half of the trimmed outputs should not be used in applications trying to describe the full transcript structure, or distant exons coupling, as the long read connectivity is lost in many cases in these types of outputs.

Overall, no correction tool outperforms all the others across the metrics analysed in this section. However, hybrid correctors are systematically better than self-correctors at decreasing the error-rate (and preserving the transcriptome diversity, as we will discuss in the next Section). Trimming and splitting usually increase the read accuracy (and also mapping rate, as we see next), but decrease the total amount of bases in the read set and the mean read length, which can lead to loss of long-range information that was present in the raw reads.

2.4 Error-correction facilitates mapping yet generally lowers the number of detected genes

Apart from HALC, LoRDEC, Proovread, and daccord, for which only 85-92% of reads were mapped, corrected reads from all the other tools were mapped at a rate of 94-99%, showing a significant improvement over raw reads (mapping rate of 83.5%). We observe that these four tools...
Table 2. Statistics of 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 classification is somewhat arbitrary.

### (A) Hybrid tools

<table>
<thead>
<tr>
<th>Raw</th>
<th>HALC</th>
<th>HALC(t)</th>
<th>HALC(s)</th>
<th>LoRDEC</th>
<th>LoRDEC(t)</th>
<th>LoRDEC(s)</th>
<th>LSC</th>
<th>LSC(t)</th>
<th>NaS(µ)</th>
<th>PBcR(s)</th>
<th>Proovread</th>
<th>Proovread(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nb reads</td>
<td>741k</td>
<td>741k</td>
<td>709k</td>
<td>914k</td>
<td>741k</td>
<td>677k</td>
<td>1388k</td>
<td>619k</td>
<td>619k</td>
<td>619k</td>
<td>1321k</td>
<td>738k</td>
</tr>
<tr>
<td>mapped reads</td>
<td>83.5%</td>
<td>88.1%</td>
<td>95.6%</td>
<td>98.8%</td>
<td>85.5%</td>
<td>95.5%</td>
<td>97.5%</td>
<td>97.1%</td>
<td>96.7%</td>
<td>98.7%</td>
<td>99.2%</td>
<td>85.5%</td>
</tr>
<tr>
<td>mean length</td>
<td>2011</td>
<td>2174</td>
<td>1926</td>
<td>1378</td>
<td>2097</td>
<td>1953</td>
<td>816</td>
<td>2212</td>
<td>1901</td>
<td>1931</td>
<td>776</td>
<td>2117</td>
</tr>
<tr>
<td>nb bases</td>
<td>1313M</td>
<td>1469M</td>
<td>1334M</td>
<td>1245M</td>
<td>1394M</td>
<td>1289M</td>
<td>1106M</td>
<td>1332M</td>
<td>1151M</td>
<td>1179M</td>
<td>1015M</td>
<td>1400M</td>
</tr>
<tr>
<td>mapped bases</td>
<td>89.0%</td>
<td>90.3%</td>
<td>96.6%</td>
<td>99.2%</td>
<td>90.6%</td>
<td>95.9%</td>
<td>99.1%</td>
<td>90.9%</td>
<td>97.5%</td>
<td>97.5%</td>
<td>99.2%</td>
<td>92.4%</td>
</tr>
<tr>
<td>error rate</td>
<td>13.72%</td>
<td>1.85%</td>
<td>1.32%</td>
<td>0.44%</td>
<td>4.5%</td>
<td>3.73%</td>
<td>1.59%</td>
<td>4.54%</td>
<td>4.36%</td>
<td>0.38%</td>
<td>0.68%</td>
<td>2.65%</td>
</tr>
<tr>
<td>nb detected genes</td>
<td>16.8k</td>
<td>17.6k</td>
<td>16.3k</td>
<td>16.4k</td>
<td>16.3k</td>
<td>16.4k</td>
<td>16.3k</td>
<td>16.3k</td>
<td>16.2k</td>
<td>15.0k</td>
<td>15.6k</td>
<td>16.6k</td>
</tr>
</tbody>
</table>

Note: "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) Non-hybrid tools

<table>
<thead>
<tr>
<th>Raw</th>
<th>Canu</th>
<th>daccord</th>
<th>daccord(t)</th>
<th>LoRMA(s)</th>
<th>MECAT</th>
<th>pbdagcon</th>
<th>pbdagcon(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>nb reads</td>
<td>741k</td>
<td>519k</td>
<td>675k</td>
<td>840k</td>
<td>1540k</td>
<td>495k</td>
<td>778k</td>
</tr>
<tr>
<td>mapped reads</td>
<td>83.5%</td>
<td>99.1%</td>
<td>92.5%</td>
<td>94.0%</td>
<td>99.4%</td>
<td>99.4%</td>
<td>98.2%</td>
</tr>
<tr>
<td>mean length</td>
<td>2011</td>
<td>2192</td>
<td>2102</td>
<td>1476</td>
<td>497</td>
<td>1992</td>
<td>1473</td>
</tr>
<tr>
<td>nb of bases</td>
<td>1313M</td>
<td>1125M</td>
<td>1350M</td>
<td>1212M</td>
<td>760M</td>
<td>980M</td>
<td>1136M</td>
</tr>
<tr>
<td>mapped bases</td>
<td>89.0%</td>
<td>92.6%</td>
<td>92.5%</td>
<td>94.7%</td>
<td>99.2%</td>
<td>96.9%</td>
<td>97.0%</td>
</tr>
<tr>
<td>error rate</td>
<td>13.72%</td>
<td>6.43%</td>
<td>5.2%</td>
<td>4.12%</td>
<td>2.91%</td>
<td>4.57%</td>
<td>5.65%</td>
</tr>
<tr>
<td>nb detected genes</td>
<td>16.8k</td>
<td>12.4k</td>
<td>15.5k</td>
<td>13.9k</td>
<td>6.8k</td>
<td>10.4k</td>
<td>13.2k</td>
</tr>
</tbody>
</table>

Note: "As reported by AlignQC, using a sample of 1 million bases from aligned reads segments.

with the lowest percentages of mapped reads had high mean read length, indicating that trimming, splitting or discarding reads seems necessary in order to obtain shorter but overall less error-prone reads. In general in all tools (except pbdagcon), trimming and splitting increased the proportion of mapped reads and bases, sometimes significantly (e.g. Proovread). However some tools which do not offer the option to trim or split reads, such as Canu and MECAT, showed very high mapping rate with high mean read length and error-rate. This is related to their aggressive post-filtering measure, which removed a significant portion of the reads (29-33%).

On verifying if error-correctors are able to preserve transcriptome diversity, we can see a striking difference between hybrid and self-correctors: in general, hybrid correctors present a far higher number of detected genes than the self ones. Interestingly, HALC was able to even increase the number of detected genes by 221 with regard to the raw reads, indicating that some genes were maybe not detected before due to imperfect mapping caused by the high error rate. We also found that 72 genes were detected in the raw reads but not in any of the error-corrected outputs. Furthermore, 354 genes are absent from the results of nearly all correction methods (≥ 16 out of 19).

Overall, all hybrid tools presented a satisfactory amount of detected genes, except for NaS(µ), PBcR(s) and Proovread (t), while self-correctors did not present any satisfactory results, with Canu, LoRMA(s) and MECAT reducing by 35%-59% the number of detected genes reported in raw reads. We can also note that trimming and splitting systematically resulted in a loss of the sensitivity to detect new genes. Moreover, except for HALC(s), tools with very high percentage of mapped reads (NaS(µ), PBcR(s), Proovread(t), Canu, LoRMA(s), MECAT, pbdagcon, pbdagcon(t)) had the largest losses in number of detected genes, hinting that error correction can reduce gene diversity in favor of lower error-rate, and/or that clusters of similar genes (e.g. paralogous) are corrected towards a single gene. Therefore, if preserving the transcriptome diversity is required for the downstream application, self-correctors should be avoided altogether, along with some hybrid correctors (NaS(µ), PBcR(s), and Proovread(t)).
2.5 Detailed error-rate analysis

The high error-rate of transcriptomic long reads significantly complicates their primary analysis [23]. 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. A general, and expected, trend that we find in all tools and in all types of errors is that trimming and splitting the reads result in less substitutions, deletions and insertion errors. We will therefore focus in other aspects in this analysis. 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 bases to calculate the presented values.

2.5.1 Deletions are the most problematic sequencing errors

Table 3 shows the 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.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Raw</th>
<th>HALC</th>
<th>HALC(s)</th>
<th>HALC(t)</th>
<th>LoRDEC</th>
<th>LoRDEC(s)</th>
<th>LoRDEC(t)</th>
<th>LSC</th>
<th>LSC(t)</th>
<th>NaS(µ)</th>
<th>PbR(µ)</th>
<th>Proovread</th>
<th>Proovread(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate</td>
<td>13.72%</td>
<td>1.85%</td>
<td>1.32%</td>
<td>0.44%</td>
<td>4.5%</td>
<td>3.73%</td>
<td>1.59%</td>
<td>5.45%</td>
<td>4.16%</td>
<td>0.38%</td>
<td>0.68%</td>
<td>2.65%</td>
<td>0.33%</td>
</tr>
<tr>
<td>Mismatch</td>
<td>5.11%</td>
<td>0.79%</td>
<td>0.54%</td>
<td>0.22%</td>
<td>2.04%</td>
<td>1.36%</td>
<td>1.13%</td>
<td>2.35%</td>
<td>2.01%</td>
<td>0.2%</td>
<td>0.18%</td>
<td>0.93%</td>
<td>0.13%</td>
</tr>
<tr>
<td>Deletion</td>
<td>7.41%</td>
<td>0.85%</td>
<td>0.64%</td>
<td>0.17%</td>
<td>2.15%</td>
<td>1.73%</td>
<td>0.39%</td>
<td>2.64%</td>
<td>1.94%</td>
<td>0.09%</td>
<td>0.3%</td>
<td>1.51%</td>
<td>0.18%</td>
</tr>
<tr>
<td>Insertion</td>
<td>1.2%</td>
<td>0.21%</td>
<td>0.14%</td>
<td>0.05%</td>
<td>0.32%</td>
<td>0.24%</td>
<td>0.07%</td>
<td>0.47%</td>
<td>0.4%</td>
<td>0.08%</td>
<td>0.19%</td>
<td>0.22%</td>
<td>0.03%</td>
</tr>
</tbody>
</table>

(A) Hybrid tools

(B) Non-hybrid tools

The contrast between self and hybrid tools is more visible on deletion errors. In general, all hybrid tools outperformed the non-hybrid ones (the only exception is LSC (2.64%), with higher deletion error rate than LoRMA(s) (2.51%)). Although in the hybrid ones, LoRDEC (2.15%), LSC (2.64%), LSC(t) (1.94) and Proovread (1.51%) still showed moderate rates of deletions, all the other seven corrected outputs were able to lower the deletion error rate from 7.4% to less than 1%. Notably, HALC(s) and Proovread(t) to less than 0.2%, and NaS(µ) to less than 0.1%. All non-hybrid tools presented a high rate (3% or more) of deletion errors, except LoRMA(s) (2.51%). This comparison suggests that ONT reads exhibit systematic deletions, that cannot be well corrected without the help of Illumina data. The contribution of homopolymer errors will be specifically analyzed in Section 2.5.2. Considering insertion errors, all tools performed equally well. It is worth noting that several hybrid (HALC(s), LoRDEC(s), NaS(µ), and Proovread(t)) and non-hybrid tools (LoRMA(s), MECAT, pbdagcon, and pbdagcon(t)) achieved sub-0.1% insertion rate errors. 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.

2.5.2 Homopolymer insertions are overall better corrected than deletions

In this section we further analyze homopolymer indels, i.e. insertion or deletion errors consisting of a stretch of the same nucleotide. Table 4 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 3 and 4, homopolymer deletions are an order of magnitude more abundant than homopolymer insertions. Still, the three best performing tools were all hybrid (NaS(µ), PbR(µ), and Proovread(t)), which should therefore be preferred for applications that require very low mismatch rates.

<table>
<thead>
<tr>
<th>Tool</th>
<th>Raw</th>
<th>Canu</th>
<th>daccord</th>
<th>daccord(t)</th>
<th>LoRMA(s)</th>
<th>MECAT</th>
<th>pbdagcon</th>
<th>pbdagcon(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate</td>
<td>13.72%</td>
<td>6.43%</td>
<td>5.2%</td>
<td>4.12%</td>
<td>2.91%</td>
<td>4.57%</td>
<td>5.65%</td>
<td>5.71%</td>
</tr>
<tr>
<td>Mismatch</td>
<td>5.11%</td>
<td>3.33%</td>
<td>1.1%</td>
<td>0.67%</td>
<td>0.37%</td>
<td>0.33%</td>
<td>0.49%</td>
<td>0.49%</td>
</tr>
<tr>
<td>Deletion</td>
<td>7.40%</td>
<td>4.82%</td>
<td>3.82%</td>
<td>3.27%</td>
<td>2.51%</td>
<td>4.18%</td>
<td>5.06%</td>
<td>5.17%</td>
</tr>
<tr>
<td>Insertion</td>
<td>2.00%</td>
<td>0.28%</td>
<td>0.28%</td>
<td>0.19%</td>
<td>0.04%</td>
<td>0.06%</td>
<td>0.09%</td>
<td>0.05%</td>
</tr>
</tbody>
</table>
from 0.38% to less than 0.11%. In particular, the hybrid tools HALC(s), NaS(µ), and Proovread(t), as well as the non-hybrid ones LoRMA(s), MECAT and pbdagcon(t) reached 0.01% homopolymer insertion error rate. Regarding homopolymer deletions, the majority of hybrid tools returned less than 0.5% of them, except LoRDEC (0.77%), LoRDEC(t) (0.63%), and LSC (0.62%). Notably, HALC(s), NaS(µ), and Proovread(t) presented less than 0.05% of homopolymer deletion error rate. Non-hybrid tools performed more poorly, returning 1.8-2.4% of homopolymers deletion errors—a small improvement over the raw reads. HALC(s), NaS(µ) and Proovread(t) 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 39.9% in raw reads, and decreases for all hybrid correctors, as 17.6% for HALC(s), and 22.2% for the non-hybrid ones LoRMA(s), MECAT and pbdagcon(t), but increases to at least 43.9% (pbdagcon(t)) up to 72.5% (LoRMA(s)) in non-hybrid tools (see Supplementary Material Section S2).

### 2.7 Error-correction perturbs gene family sizes

Table 4 indicates that error correction generally 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 3.5). Figure 3 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, 36-87% of gene families are kept of the same size by correctors, 1-17% are expanded and 6-61% are shrunk. Supplementary Material Figure S2 shows the magnitude of expansion/shrinkage for each gene family.

### 2.8 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 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.

Table 2 indicates that error correction generally results in a lower number of isoforms. To answer this question, we computed the 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.

#### 2.8.1 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). The behaviour of the tools in the isoform level are in coherence with their behaviour in the gene level (C_T): split outputs inflate C_T; MECAT deflates it; and all the others present a slight increase.

| Raw HALC HALC(t) HALC(s) LoRDEC LoRDEC(t) LoRDEC(s) LSC LSC(t) NaS(µ) pbdagcon Proovread Proovread(t) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Homop. deletion | 0.26% | 0.19% | 0.03% | 0.77% | 0.63% | 0.19% | 0.62% | 0.42% | 0.02% | 0.1% | 0.46% | 0.04% |
| Homop. insertion | 0.38% | 0.06% | 0.03% | 0.09% | 0.07% | 0.02% | 0.11% | 0.09% | 0.01% | 0.02% | 0.06% | 0.01% |

#### (B) Non-hybrid tools

| Raw Canu daccord daccord(t) LoRMA(s) MECAT pbdagcon pbdagcon(t) |
|---|---|---|---|---|---|---|---|---|
| Homop. deletion | 2.46% | 2.14% | 2.05% | 1.82% | 2.08% | 2.26% | 2.27% |
| Homop. insertion | 0.38% | 0.08% | 0.06% | 0.03% | 0.01% | 0.01% | 0.02% | 0.01% |

### Table 4: 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.

### Section S2

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 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.

#### 2.8.1 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 (~1.9% for LoRDEC(t), LSC and pbdagcon(t), and ~5.4k for MECAT) lose at least one isoform in all tools, which suggests that current methods reduce isoform diversity during correction. NaS(µ), Proovread(t), Canu, and MECAT tend to lose isoforms the most, and HALC(s), LoRDEC(s), and pbdagcon(t) identify the highest
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.
Comparative assessment of long-read error-correction software applied to Nanopore RNA-sequencing data

2.8.2 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 and LSC, that tend to be the most conservatives. We can also note that trimming and splitting reads increase even further isoform losses in all tools, except for pbdagcon. This can be explained by the fact that lowly-expressed isoforms possibly share regions (e.g. common exons) with highly-expressed isoforms, and these shared regions are usually better corrected than regions that are unique to the lowly-expressed isoforms. If read splitting then takes places, such unique regions will then be removed from the output. Even if there are variations between a highly-expressed isoform I and a lowly-expressed isoform i, if these variations are relatively small (e.g. a small exon skipping) and are flanked by long shared regions, it is probable that the methods will truncate the variation, correcting the unique fragment of i into I, and potentially losing the signal that i is expressed (this is explored in details in Section 2.8.4). This result suggests that current error-correction tools overall do not conservatively handle reads that belong to low-expression isoforms.
2.8.3 Minor isoforms are corrected towards major isoforms

We define a major isoform of a gene as the isoform with the highest coverage of that gene in the raw dataset, all other isoforms are considered to be minor. To follow-up on the previous subsection, we investigate whether correctors tend to correct minor isoforms towards major isoforms. We do so by comparing the difference of coverage of the major and the minor isoforms before and after correction. In Figure 5, we observe that the coverage of the major isoform generally slightly increases after correction. The exceptions are tools that split reads (HALC(s), LoRDEC(s), PBcR(s), and LoRMA(s)), where the coverage is increased significantly, and MECAT, where the coverage decreases significantly, likely due to a feature of MECAT’s own correction algorithm. Since these 5 tools seem to heavily distort the coverage of isoforms due to aggressive splitting or filtering steps, we will focus now on the 14 other results. The slight increase of a transcript coverage after correction is expected, as already discussed in Section 2.6: uncorrected reads that were unmapped can become mappable after correction. Therefore, the effect presented in Figure 5 could be simply due to reads being corrected to their original respective isoforms, instead of correction inducing a switch from a minor isoform to the major isoform. To verify this hypothesis, Supplementary Material Figure S4 shows that the coverage of the minor isoforms usually decreases after correction ($R^2 \in [0.5,0.8]$), except for: i) tools that split reads (HALC(s), LoRDEC(s), PBcR(s), LoRMA(s)), which keeps even more the coverage of minor isoforms, and ii) both HALC and HALC(t). This indicates that error-correction tools tend to correct reads towards the major isoforms. It is worth noting that the increase of the coverage of the major isoform is not pronounced. This is expected, as the sum of the expression of the minor isoforms is, by nature, a small fraction of the total gene expression. On the other hand, the correlation of the coverage of the minor isoforms before and after correction are far more spurious, suggesting a stronger effect. 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.

2.8.4 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 scenarios 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. Thus, regardless of the abundance of inclusion reads in the dataset to be corrected, 99% of reads from inclusion are corrected to inclusion form for an exon size of 10, and 100% of reads from inclusion are corrected to inclusion form for exon sizes of 50 and 100. 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 Sections S7 and S8). Self-correctors generally have a minimum coverage threshold (only daccord could be ran on the 10x coverage dataset of long reads, with rather erratic results, see Supplementary Material Section S8). 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.

2.9 Error-correction affects splice site detection

The identification of splice sites from RNA-seq data is an important but challenging task [50]. 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.
Comparative assessment of long-read error-correction software applied to Nanopore RNA-sequencing data

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. Blue line: regression, black line: x=y.
With long reads, the high error rate makes precise splice site detection even more challenging, as indels (see Section 2.3) confuse aligners, shifting predicted spliced alignments away from the true splice sites.

In this section, we evaluate how well splice sites are detected before and after error-correction. Figure 6 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 less than 1.2% in 8 hybrid corrected outputs: HALC, HALC(t), HALC(s), LoRDEC(s), PBcR(s), Proovread and Proovread(t). Notably, Proovread(t) presented only 0.28% incorrectly mapped splice sites. LoRDEC (2.43%) and LoRDEC(t) (2.12%) presented higher rates, but still manageable, but LSC (7.27%) and LSC(t) (5.68%) underperformed among the hybrid correctors. Among self-correcting tools, LoRMA presented the lowest proportion of incorrectly detected splice sites (3.04%), however it detects −6.7 times less splice sites (~200k) than the raw reads (~1.9M), due to read splitting. The other non-hybrid tools incorrectly detected splice sites at a rate between 5.61% (diaccord(t)) and 11.95% (Cam). It is worth noting that trimming usually decreased the proportion of incorrectly mapped splice sites, with a very slight impact on the total amount of identified splice sites. On the other hand, the three tools with lowest number of identified splice sites output split reads (LoRDEC(s), PBcR(s), and LoRMA(s)), identifying less than ~1.1M splicing sites, compared to the ~1.9M in the raw reads, and thus not being adequate for splice sites analyses. Additional detailed plots on incorrectly mapped splice sites can be found in the Supplementary Material Section S9.

2.10 Running time and memory usage of error-correction tools

Table 5 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 7 hours, except for LSC, NaS, PBcR, and Proovread, which took 63-116 hours, but also achieved some of the lowest post-correction error rates in Table 2 (except for LSC). In terms of memory usage, all tools required less than 10 GB of memory except for HALC, 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 memory usage and correction for.

2.11 Using a different read aligner mildly but not significantly affects the evaluation

We chose GMAP (version 2017-05-08 with parameters −n 30) to perform long reads mapping to the Ensembl r87 Mus Musculus unmasked reference genome in our analysis, since Kržanović et al. show it produces the best alignment results between five alignment tools. The GMAP parameters are those from the original AlignQC publication. However, Minimap2 is not evaluated in that study, and it is also widely used, being the default long-read mapper in several studies. In this subsection, we verify to which extent the differences between GMAP and Minimap2 can influence our evaluation. To try to highlight such differences, we chose some correctors with the worst and best performances in some measures presented in the previous analysis (made with GMAP). We thus further mapped with Minimap2 (version 2.14-r883 with parameters −ax splice) the following read datasets: i) raw reads; ii) LoRDEC, the
correttore con il minore numero di mappature, ma che conserva
bene la diversità transcriptionale, ii) LoRMA, il tool con la più
alta ratio di mappature e numero di mappature, ma con il più
basso numero di sequenze mappate, lunghezza, numero di
detected genes, e numero di bases; iii) Proovread(t),
la metà dei più bassi error rate, e più alto della più
correttamente mappato splice site; iv) Daccord, il
corrector con il più basso error rate e il più
basso della più correttamente mappato splice site.

Gli esperti presenti in Tab. 5 il principale
differenza tra Minimap2 e
GMAP sulla richiesta di correzioni. Posso notare che Minimap2
è in grado di mappare più sequenze di GMAP attraverso tutti i tool, ma
Proovread(t). Tuttavia, quando entrambi i mappatori sono in grado di mappare quasi
95% (≥98.9%),
la differenza mappatura tra di loro è inferiore (≤1%). Le più
grande differenza è quando entrambi i mappatori non eseguono bene, mappatura meno
90% delle sequenze. In questo caso, Minimap2 dà una miglior mappatura, mappatura
3.3% e 1.9% più mappatura di GMAP nella sequenza della
LoRDEC, rispettivamente, ricevendo notevole, ma non alto. Una simile conclusione può
essere ottenuta guardando il numero di mappature. Il rapporto errore mappatura è
è molto simile tra entrambi i mappatori, con una maggiore differenza (0.32%) nelle sequenze
mappate. Vi è una notevole differenza, tuttavia, quando
sebbene minimap2 report successivo più insertioni e inserzioni. Anche se possiamo notare
le differenze sono notevoli nelle sequenze mappate. Il numero in cui
GMAP evidentemente superava Minimap2 è il numero di detected genes,
identificando 453 (Proovread(t)) e 723 (LoRDEC) più genes,
che possono essere considerati significativi. Diversamente, Minimap2
è in modo considerabilmente meglio che GMAP sulla mappatura splice site correttamente.
La differenza tra entrambi i mappatori quando hanno eseguito bene, mappatura
più 96.9% (≤1.15%). Le più grandi discrepanze possono essere visti sulla mappatura splice site delle
raw e Canu reads, con Minimap2 correttamente aligning 18.49% e 6.64% più
splice site di GMAP.

Possiamo concludere che Minimap2 era lievemente migliore sulla mappatura
zygotes and base, e significativamente meglio alla causa di mappatura splice site.
Sul lato GMAP, minima considerazione alla detezione genes.
La differenza mappatura delle sequenze mappate da entrambi i mappatori è notevole, con
il rapporto errore mappatura GMAP più del rapporto Minimap2, e Minimap2 report successivo più
insertioni. Tuttavia, queste differenze sono principalmente concentrati su mappatura sequenza, la quale è il
meno accurato.

I dati mostrati e discussi in questa sezione rivelano che la
principale conclusione che entrambi i mappatori, conservando ulteriori dettagli, possono essere trovati nel
support page of our method:
https://leoisl.gitlab.io/LR_analysis/EC_analyser_support/
2.12 Analysing human Nanopore direct RNA-sequencing data

We further analysed a human Nanopore direct RNA-sequencing dataset from the Nanopore-WGS-Consortium (see Section 2.2 for details). Since there was no corresponding Illumina sequencing for this dataset, we were able to evaluate only the non-hybrid error correction tools. Moreover, although LoRMA could be successfully executed, AlignQC could not process its output so we removed LoRMA from the evaluation. Table 7 presents some main statistics of non-hybrid error correction tools on the aforementioned dataset. In the rest of this section, we highlight the major differences and similarities between cDNA and direct RNA datasets, keeping in mind that they were performed on two different species (human and mouse, respectively).

In general, self-correctors discarded more reads on the direct RNA dataset than on the 1D cDNA dataset. Daccord(t) discarded the least number of reads (102k), while Canu and MECAT discarded a considerable amount of reads (361k and 670k, respectively), due to post-correction filtering. Due to our choice of parameters, the shortest reads in Canu and MECAT outputs were of lengths 101 and 100 bases, respectively. However, the removal of shorter reads explains only a fraction of the difference in the mapping rate of error-corrected reads. The mapping rate of error-corrected reads was generally higher. Notably, 97.5% of the daccord reads (resp. 98.2% for daccord(t)) were mapped, as opposed to 92.5% (resp. 94%) in the 1D cDNA dataset. The mean length of the corrected reads when compared to the mean length of the raw reads was also in general higher in the direct RNA dataset, which translated into tools having a number of output bases more similar to the number of bases in the raw reads in this dataset, except for MECAT.

The error rate of the raw direct RNA reads was 14.61%, 0.89% higher than in the raw 1D cDNA reads. As expected, the error rates in all tools were also higher in the direct RNA dataset, leading to worse results. The largest difference is with pbdagcon(t), where the error rate after correction of direct RNA reads is 2.55% higher than on the 1D cDNA dataset. The distribution of errors in the direct RNA dataset was more balanced, with mismatches and deletions having almost the same representation (around 5.85%), but insertions still being less represented (2.87%). The correction behaviour of the tools is similar across both datasets: insertion is the best corrected type of error, followed by mismatches, with satisfactory results, and deletions, in which the methods overall did a poor job. In particular, pbdagcon and pbdagcon(t) even increased the deletion error rate by 0.84% and 0.99%, respectively. The behaviour was similar on correcting homopolymer errors: homopolymer deletions were poorly corrected, with MECAT, pbdagcon and pbdagcon(t) not reducing the homopolymer deletion error rate at all, while homopolymer insertions were well corrected. The number of detected genes in the raw direct RNA dataset (14.1k) is less than in the raw 1D cDNA dataset (16.8k), although this is consistent with the difference in human/mouse genes count. Moreover, the tools also lose more genes in the direct RNA dataset. In particular, MECAT loses 6.4k genes in the 1D cDNA dataset, and 7.8k in the direct RNA dataset. The rate of correctly mapped splice sites was slightly higher in the raw direct RNA reads (76.95% vs 72.94%), but in the error corrected reads, this rate was highly similar (the largest difference was 0.44% in the daccord(t) correction).

As a result of our evaluation, and in accordance with the cDNA analysis, care should be taken when applying self-correctors to remove errors from Nanopore direct RNA-seq data. For example, Canu and MECAT present the undesirable side effect of discarding a lot of input reads, thus reducing the amount of information in the long reads, and decreasing considerably the number of detected genes. Although the tools perform well at correcting mismatches and insertions, they have trouble correcting deletions. In particular, MECAT, pbdagcon, and pbdagcon(t) perform rather poorly, with the last two even increasing the deletion error rate.

The full report of the analysis output by our method on this dataset, containing further details, can be found at https://leoisl.gitlab.io/LR_EC_analyser_support/.
Comparative assessment of long-read error-correction software applied to Nanopore RNA-sequencing data

Table 7. Statistics of non-hybrid error correction tools on one human Nanopore direct RNA-sequencing data from the Nanopore-WGS Consortium.

<table>
<thead>
<tr>
<th>Tool</th>
<th>nb reads</th>
<th>mapped reads</th>
<th>mean length</th>
<th>nb of bases</th>
<th>error rate</th>
<th>mismatches</th>
<th>deletions</th>
<th>insertions</th>
<th>homop. deletions</th>
<th>homop. insertions</th>
<th>nb detected genes</th>
<th>correctly mapped ss(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>894k</td>
<td>76.0%</td>
<td>739</td>
<td>507M</td>
<td>14.61%</td>
<td>5.79%</td>
<td>5.95%</td>
<td>2.87%</td>
<td>2.18%</td>
<td>1.13%</td>
<td>14.1k</td>
<td>76.95%</td>
</tr>
<tr>
<td>Camu</td>
<td>533k</td>
<td>99.2%</td>
<td>917</td>
<td>484M</td>
<td>7.61%</td>
<td>2.12%</td>
<td>4.77%</td>
<td>4.13%</td>
<td>2.01%</td>
<td>0.3%</td>
<td>8.8k</td>
<td>87.97%</td>
</tr>
<tr>
<td>daccord</td>
<td>744k</td>
<td>97.5%</td>
<td>849</td>
<td>602M</td>
<td>6.03%</td>
<td>1.56%</td>
<td>4.13%</td>
<td>3.9%</td>
<td>1.84%</td>
<td>0.14%</td>
<td>12.3k</td>
<td>92.52%</td>
</tr>
<tr>
<td>daccord(t)</td>
<td>792k</td>
<td>98.2%</td>
<td>755</td>
<td>591M</td>
<td>5.47%</td>
<td>1.3%</td>
<td>5.9%</td>
<td>3.9%</td>
<td>1.81%</td>
<td>0.11%</td>
<td>10.8k</td>
<td>93.94%</td>
</tr>
<tr>
<td>MECAT</td>
<td>224k</td>
<td>99.2%</td>
<td>1095</td>
<td>243M</td>
<td>6.36%</td>
<td>0.7%</td>
<td>5.5%</td>
<td>5.5%</td>
<td>2.24%</td>
<td>0.09%</td>
<td>6.3k</td>
<td>92.86%</td>
</tr>
<tr>
<td>pbdagcon</td>
<td>772k</td>
<td>98.6%</td>
<td>666</td>
<td>510M</td>
<td>8.1%</td>
<td>0.7%</td>
<td>6.79%</td>
<td>6.79%</td>
<td>2.18%</td>
<td>0.06%</td>
<td>10.0k</td>
<td>91.16%</td>
</tr>
<tr>
<td>pbdagcon(t)</td>
<td>771k</td>
<td>98.5%</td>
<td>675</td>
<td>516M</td>
<td>8.26%</td>
<td>1.19%</td>
<td>6.94%</td>
<td>6.94%</td>
<td>2.22%</td>
<td>0.07%</td>
<td>10.1k</td>
<td>91.14%</td>
</tr>
</tbody>
</table>

\(^a\) As reported by AlignQC, using a sample of 1 million bases from aligned reads segments.

\(^b\) ss stands for splice sites.

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 Nanopore 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, however a collection of caveats are described in the Results section. Most importantly, the number of genes detected by alignment of corrected reads to the genome was reduced significantly by several error-correction methods, mainly the non-hybrid ones. 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.

Detailed error-rate analysis showed that while hybrid correctors have lower error rates than self-correctors, the latter achieved comparable performance to the former in correcting substitutions and insertions. Deletions appear to be caused by systematic sequencing errors of the Nanopore technology, making them fundamentally hard (or even impossible) to address in a self-correction setting. Moreover PBrR, NaS, and Proovread are one of the most resource-intensive error-correction tools, but also are some of the few correctors able to reduce the base error rate below 0.7%. The only exception to this is HALC, which presents a low running time, and <0.5% error rate in its split output.

We observe that hybrid correctors were able to better preserve the number of detected genes than self-correctors. The large majority of the hybrid corrections (9/12) were able to identify an amount of genes similar to the raw reads, with only NaS(µ), PBcR(s), and Proovread(t) being less sensitive, but still obtaining reasonable results. On the other hand, daccord was the only self-correction tool that reached the same gene identification level of the three aforementioned hybrid tools, while the others heavily truncated the transcriptome diversity. HALC, LoRDEC, LSC, Proovread (only in full-length mode) and daccord (only in full-length mode) appear to also better preserve the number of detected isoforms better than other correctors (Section 2.6). All tools tend to lose lowly-expressed isoforms after correction (Section 2.8.4). Several tools also tend to correct minor isoforms towards major isoforms (Section 2.9.3) mainly when the variation between them is small (Section 2.9.4). These points are expected, as most 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 (Section 2.9).

As a result, we conclude that no evaluated corrector outperforms all the others across all metrics and is the most suited in all situations, and the choice should be guided by the downstream analysis, yet hybrid correction tools generally outperformed the self-correctors. For quantification, we have shown that error-correction introduces undesirable coverage biases, as per Section 2.8 it is then recommended avoiding this step altogether. For isoform detection, HALC, LoRDEC, LSC and Proovread (only in full-length mode) appear to be the methods of choice as they result in the highest number of detected genes in Table 2 and also preserve the number of detected isoforms as per Section 2.8.2. These points are expected, as most 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 (Section 2.9).
low general error rate. Finally for all other applications, we make some general recommendations. A reasonable balance appears to be achieved by tools with no unsatisfactory values in Table SI, HALC(t), NaS(t), and Proove(t). If Illumina reads are unavailable, then the best overall self correctors seem to be daccord(t), pbdragon and pbdragon(t). Moreover, trimming and splitting usually increase the mapping rate and the read accuracy, but can lead to loss of information that was present in the raw reads, complicating the correct identification of genes.

Our analyses relied on a single mapping software (GMAP [51]) to align raw and error-corrected reads, as in previous benchmarks [23, 59]. However, we were also able to verify that Minimap2 [52], another widely used mapper, produces similar results than GMAP (see Section 2.11) and thus the main messages of the analyses presented in this paper should not change by replacing GMAP by Minimap2.

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 LoRDEC(t) (0.2%), PBRC(t) (0.1%), Proove(t) (0.1%), LoRMA(s) (0.04%), and MECAT (0.2%), which either correctly split reads or discarded chimeric ones. We observe that HALC (4.2%), HALC(t) (3.9%), and daccord (1.7%) increased considerably the proportion of chimeric reads in the output.

Furthermore, we focused our evaluation on a single technology: Nanopore. We did an extensive analysis of 1D cDNA Nanopore data, using Illumina data for hybrid correction. We also performed a brief analysis of Nanopore direct RNA-seq data. 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. 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 also note 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 coverage. But we note that in current RNA-seq analysis protocols, transcript/gene expression is related the results of error-correction to transcript/gene counts, but we note that it may be a critical factor for some tools (e.g. Canu) on larger datasets. We also note 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.

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, Cintoch, Cat# 636601), according to the Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) protocol "1D cDNA by ligation (SQD-LSK108)". The data generated by MinION software (MinKNOW 1.1.21, MinKNOW 1.1.21) were stored and organized using a Hierarchical LSK108). The data generated by MinION software (MinKNOW 1.1.21, Technologies Ltd, Oxford, UK) protocol "1D cDNA by ligation (SQK-

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.

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 paralogue gene families can also be provided if plots on gene families should be created. In our main analysis, gene families were computed by selecting all paralogs from Ensemble 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. The main processing of our method involves running the AlignQC software [21] (https://github.com/jason-weiher/AlignQC) on the input BAMs and parsing its output to create custom plots. It then aggregates information into a HTML report. For example, Tables 5-6 are compilations from AlignQC results, as well as Tables 2 and 4. Figures 1-6 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://t/leoisl.gitlab.io/LR_EC_analyser_support/.

4.4 Simulation framework for biases evaluation

In the simulation framework of Section 2.7 exons length and number were chosen according to what is reported in eukaryotes (9 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 [59], 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
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

The authors thank the ASTER consortium, in particular Camille Sesegolo and Vincent Lacroix, for fruitful discussions. LL acknowledges CNPq/Brazil for the support. This work was performed using the computing facilities of the CC LIBE/PRABI and the France Gréontomique e-infrastructure (ANR-10-INBS-09-08).

Funding

This work was supported by the French National Research Agency [ANR ASTER, ANR-16-CE23-0001]; the INCEPTION project [PIA/ANR-16-CONV-0005]; 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, Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq) scholarship grant [process number 2033622014-4 to LL].

Competing interests

JMA is one of the authors of the NaSe error-correction tool [29]. 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


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

Figure legends

Fig. 1. Number of reads mapping to genes (CG) 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.

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.

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.

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. Blue line: regression, black line: x=y.

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, and Proovread. Right: same with three self-correctors and 100X long reads: dacord, 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.