TALC: transcription aware long read correction

Lucile Broseus[1], Aubin Thomas[1], Andrew J. Oldfield[1], Dany Severac[2], Emeric Dubois[2] and William Ritchie[1]*

[1]*Institut de Génétique Humaine, Centre National de la Recherche Scientifique (CNRS), Université de Montpellier, Montpellier, France*

[2]*MGX-Montpellier GenomiX, c/o Institut de Génomique Fonctionnelle, 141 Rue de la Cardonille, Montpellier Cedex 5 34094, France*

*To whom correspondence should be addressed. Tel:+33 4 34 35 92 40; Email: william.ritchie@igh.cnrs.fr

ABSTRACT

Long-read sequencing technologies are invaluable for determining complex RNA transcript architectures but are error-prone. Numerous "hybrid correction" algorithms have been developed for genomic data that correct long reads by exploiting the accuracy and depth of short reads sequenced from the same sample. These algorithms are not suited for correcting more complex transcriptome sequencing data. We have created a novel algorithm called TALC (Transcription Aware Long Read Correction) which models changes in RNA expression and isoform representation in a weighted De-Bruijn graph to correct long reads from transcriptome studies. We tested TALC on a dataset of short and long reads generated for this study. TALC correction results in more accurate reads with less structural errors than existing methods. TALC is implemented in C++ and available at https://gitlab.igh.cnrs.fr/lbroseus/TALC.

INTRODUCTION

Recent advances in RNA sequencing (RNA-seq) technologies have revealed that transcription is more pervasive (1), more diverse (2) and more cryptic (3) than expected. Given the major role that RNA processing plays in disease and normal biology, it is crucial to ascertain the existence of novel isoforms and to accurately quantify their abundance. Second generation RNA-seq technologies such as Illumina are well suited to the tasks of assessing gene expression levels and determining proximal exon connectivity. They produce numerous sequencing reads at a low cost ensuring sufficient representation of most transcripts. However, because the RNA or cDNA is fragmented during short RNA-seq protocols, long range connectivity can only be computationally inferred. These predictions based on short reads struggle to correctly identify transcript isoforms that contain multiple alternative exons (4) or that contain retained introns (5). In these cases, long-read (LR) sequencing technologies are invaluable because they can sequence entire molecules in one pass and thus capture long-range connectivity of complex isoforms. LR technologies however produce less reads than short read (SR) sequencing approaches (6) for similar costs and have higher error rates. In many cases these higher error rates can prevent the correct identification of isoforms.

Although several alignment software (7–9) are optimized to handle long erroneous sequences, they still display several issues which confound transcript identification and annotation. Many reads fail to be aligned and regions where the sequencing error rates are higher such as UTRs frequently produce ambiguous alignments. Secondly, they struggle to identify splice junctions, notably those flanking small structural variants such as small exons or alternative 5' and 3' splice sites. This impacts the evaluation of exon skipping events and worsens the quality of

transcript assembly (10). These drawbacks prompted the development of algorithms, referred to as *hybrid methods (11*–14), that take advantage of SR depth and accuracy to compensate LR shortcomings.

Numerous algorithms have been proposed to combine long and short reads into high-accuracy long reads. A first approach is to correct long reads by local consensus inferred from short read multi-alignments (15–18). This strategy is generally slow and computationally intensive. More importantly, it tends to show poor performances over low-expressed regions, and a risk of bias toward major isoforms. Therefore, they may not be suitable for transcriptome and single cell datasets where isoform representation may vary considerably. In the second approach (19–22), short reads are considered as the fragments of a reference transcriptome, and roughly assembled using a graph structure, onto which long reads are aligned (23). However, sequencing errors in SR datasets, complex transcript architectures and sequence duplicates often lead to extremely complex graph structures (24, 25) that elicit graph simplifications or exploration heuristics to keep computations tractable. Most hybrid correction algorithms were primarily intended to be applied to genomic data with the aim to improve the quality of genome assemblies. As such, they typically rely on assumptions that fit DNA-seq data properties and thus a linear reference genome. According to this, bifurcation nodes and tips (dead-ends in the graph) are assumed to originate mainly from sequencing errors and not transcript processing events. In addition, genome sequencing benefits from a relatively uniform read coverage and therefore the graph is simplified by discarding *a priori* all nodes whose count is below a user-defined threshold. The fact that these approaches are not adapted for RNA-seq data have been extensively discussed in previous works on transcript assembly (25) and short read correction (26, 27) and their impact on long read correction can be easily anticipated. RNA-seq data display highly uneven coverage, even across a same genomic location, thus the frequency of sequencing errors varies depending on the surrounding coverage depth. Tips may correspond to the start or end of a transcript and finally specific regions such as 3' ends of transcripts, are frequently under covered.

We have developed TALC for Transcription Aware long-read Correction which addresses RNA-seq data specificities by incorporating coverage analysis throughout the correction process. TALC takes into account transcript expression and the existence of isoforms to correct LRs. For the purpose of testing TALC, we generated Oxford Nanopore direct RNA sequencing reads and Illumina short reads on an MCF10A human cell line and downloaded LR and SR data from the GM12878 human cell line. We demonstrate that after TALC correction, long reads map with higher sequence identity and with less errors in exon assembly than currently used methods. The gains observed following TALC were tested on simulated long reads as well as on real reads where we made use of an ectopically expressed gene fusion in our cells to assess its capacity in retrieving non-reference transcripts.

MATERIAL AND METHODS

Methods overview

Figure 1 illustrates the methodology behind TALC and highlights how it takes into account transcript abundance and architecture to correct long reads. The first step of TALC graph-based procedure is to consider short reads as a raw reference transcriptome by merging them through a *weighted De-Bruijn Graph* (DBG) structure (19, 23), whose nodes represent k-mers. For each node, we record their k-mer abundance in the SR dataset. Thus, any sequence of transcripts expressed in the RNA-seq sample should appear as a unique path of the graph. The

long read is corrected by finding the right sequence of nodes to which it corresponds in the *DBG* built from short reads. Paths corresponding to true transcripts should display consistent k-mer coverage except in regions where the existence of multiple isoforms may alter the coverage such as at splice junctions (Figure 2). We thus propose a method of graph exploration that takes into account k-mer count variation consistent with transcript abundance and isoform existence (adaptive count thresholding).

Determining anchor points

As in (19), the LR sequence is first divided into *solid regions* (stretches of k-mers shared with the SRs) interspersed with w*eak regions* (stretches of likely erroneous k-mers).

To crop background noise that frequently surround solid regions, we first estimate the frequency of k-mers resulting from sequencing errors in the SR. The count of k-mers containing a sequencing error in SR is assumed to follow a Poisson distribution whose mean should be less than the average noise level λ. An estimate of λ is taken as the average coverage depth across the LR multiplied by the per base error rate in SR, whose estimates usually range from 1% to 3% in RNA-seq datasets (26).

A robust estimate of the average depth is derived *a priori* from the shared k-mer counts: the 15% smallest and 10% highest counts are removed and the mean is calculated from all the remaining k-mers. K-mers whose count falls under the upper bound of a confidence interval from the Poisson distribution Poisson(λ) (cf: Supplementary Materials) are not considered solid and will not be used as anchor points.

We then split solid regions into anchor points with contrasting k-mer coverage. To this end, we inspect the regions coverage and each k-mer at which an abrupt count variation is detected is selected (cf Competing paths section below). K-mers at the tips of a region are always selected, and additional k-mers may be picked so as to have at least three distinct anchor points per solid region.

Once the long read is anchored, the DBG is explored in-between any couple of consecutive anchors, in search for the error-free path matching the inner weak region.

Selective Exploration of the DBG

In TALC, a path in the DBG is defined as an ordered list of connected k-mers (nodes) which are weighted by their number of occurrence in the SR dataset. When a k-mer is mapped to a LR, we attribute its weight to the mapped region of the LR and use the term coverage. Thus, when we speak of coverage of a LR, it is in effect the weight of the k-mer that mapped to that region.

Abrupt changes in the coverage depth can be caused by alternative splicing events, duplicate k-mers, or sequencing bias in SR. These abrupt changes are termed split-coverage.

Nodes in a *DBG* are either *simple*, which means there exists only one successor k-mer in the data, or admit at least two successors, and thus can be extended by several distinct paths. We will refer to them as *bifurcation nodes*. Exploration is implemented as a breadth-first approach where at each bifurcation, only the next nodes whose count is considered consistent with the coverage of the current path will be followed. More precisely, at a given step n0, the k-mers database is queried for the counts of its four possible successors: (nA, nC, nG, nT). And we want to assess which of these make up valid extensions of the current path.

We need to infer two things: whether n0 is close to a split-coverage event, in which case the exploration might be split into several competing paths; and which nodes are likely erroneous and should be filtered-out. To this end, we use the following decision rules (cf: Supplementary Materials)**:**

1) If there is only one non-zero k-mer, there is no decision to make.

2) If there are at least two non-zero k-mers, we postulate that most of the time, there is no real split coverage event at n0, so that its count provides a good estimate of the downstream coverage depth, from which both the correct successor (say nL) and a local noise threshold can be inferred.

Accordingly, allowing to the consistence of the coverage, we pose as a null model:

$$nL \mid (n0 \ \& \text{ no split coverage event}) \sim \text{Poisson}(n0)$$

and

$$\text{sum}(nJ, J!=L) \sim \text{Poisson}(n0*\varepsilon)$$

When the counts (nA, nC, nG, nT) do not fit this null model, we infer the current bifurcation is due to a change-point and the exploration is split into all downstream nodes. At each new node, the expected count and the noise level are thus re-estimated, allowing an adaptive filtering-out of erroneous k-mers.

Therefore, a true path is expected to be mostly made of count-consistent nodes possibly with a few change-point nodes. And the more a path admits change-points the less likely it is to represent the sequence of a true transcript.

Competing paths

TALC favours coverage-consistent exploration of the DBG. All paths passing the test described above are explored in parallel, according to a breadth-first approach.

When the number of parallel paths exceeds a specified threshold, an evaluation is performed: all on-going paths are compared to the LR sequence and the least similar ones are stopped.

All paths which successfully bridge both Solid Regions are scored by their edit distance with the LR. The most similar one is considered suitable for further validation. Its sequence is aligned against the LR's; if the percent identity score with the LR is higher than a user-specified threshold (by default 70%), we assume we have likely found the best candidate.

Border exploration

Compared to inner weak regions (flanked on both sides by a solid region), the correction of weak border regions (flanked only on one side by a solid region) raises additional problems. Firstly, there is no clear targeted anchor point at which to stop exploration. Secondly, UTR sequences often contain low complexity regions and duplicated k-mers, which leads to increased complexity in the DBG. Finally, borders of transcripts suffer from higher sequencing bias in short reads (notably in 3' ends of transcripts). The two last points make k-mer coverage more erratic and over dispersed. In certain cases, incomplete coverage UTR extremities (eg: rare longer UTR forms) can sometime prevent a full-length correction entirely. For these reasons, we rely more heavily on sequence length and similarity between the visited paths and the LR's to direct graph exploration across those regions. More exactly, graph exploration is monitored in a same manner as described above (cf: **Selective Exploration of the DBG)** until the paths's length matches the border's length. When the number of consecutive errors exceeds a given threshold, the corresponding paths are stopped.

When there are no more branches to investigate (that is all possibilities have ended in dead-ends of the graph or have been stopped), the very last error-prone bases of all interrupted paths are trimmed (cf: Supplementary Materials). To elect the best border path, we first search for it among the paths that were at least as long as the portion of the read we are attempting to correct, despite the constraints on sequence similarity. We choose the one having the smallest edit distance with the LR, and the most consistent coverage if there are *ex-aequo*. If no long border path could be found, shorter paths are compared and the one with highest similarity to the LR is selected.

Choosing between multiple paths whose scores are tied

Paths in the *DBG* that successfully bridge two Solid Regions are first ranked according to their sequence similarity with the Weak Region's. The similarity is computed as the edit distance between sequences.

We notice that the exploration stage often provides several solution paths having the exact same alignment score (this occurs for example when there are variants in sub-sequences that have been inserted or deleted from the LR), so that we cannot decide between them based on sequence information only. In order to choose between these equally-scoring paths, we perform a second ranking using count variation: the path with the least change-points wins.

RESULTS

Evaluation datasets

We compared the corrections made by TALC with six other state of the art hybrid correctors: FMLRC (22), CoLoRMap (17), Jabba (20), Hercules (18), LoRDEC (19) and LSC (15). We chose FMLRC, Jabba, LoRDEC and ColoRMap as they were found to be the top performers according to a recent benchmark on genomic data (28). We added Hercules because it was a very recent method not yet benchmarked. We also included LSC for it is (with LoRDEC) currently applied in numerous transcriptomic studies combining long and short reads (29–32).

All jobs were run using 15 CPUs with 512 gigabytes of available RAM. If a job was not finished after 1 week, it was terminated (Jabba failed to finish within a week and was eventually excluded from the evaluation and, as already observed in (18), LSC could not scale to the big size real LR dataset). LoRDEC and TALC were run with the recommended size of k-mers set to 21 (19, 33).

To evaluate the efficiency of these algorithms, we performed ONT direct RNA sequencing and Illumina mRNA sequencing in parallel on the MCF10A cell line (cf: Supplementary Material). We chose this model because numerous transcriptional isoforms with key biological functions have been described in this cell line (34, 35). We also made use of publicly available SR and LR data generated on the GM12878 B-Lymphocyte cell line. By using real datasets to compare these methods instead of a computationally generated one we are testing them in authentic conditions. However, the drawback of using real data is that it is impossible to measure precisely the number of good corrections performed because the entire repertoire of transcripts and their true abundance is unknown; there is no reference transcriptome that perfectly matches our MCF10A and GM12878 cells. To work around this problem, we used the LR data to measure the abundance of transcripts in MCF10A and GM12878 cells (Supplementary Materials). From this LR-derived transcriptome, we simulated long ONT reads that respect the error rates that we observed in the sequencing data. Thus, we have a dataset of reads that are derived from

the MCF10A and GM12878 transcriptomes and for which we know the exact abundance of each isoform. In addition to this simulated ONT dataset, we made use of a unique feature of our cells: the ER-Snail gene fusion. The MCF10A cells sequenced in this experiment have been transfected with a gene fusion construct between the estrogen receptor (ER) and the Snail-1 gene (36). This construct allows an inducible expression of the Snail-1 gene which alters MCF10A cell morphology from an epithelial to a mesenchymal one; this is known as the epithelial to mesenchymal transition (EMT). We can make use of the ectopic expression of the non-reference ER-Snail construct in our cells to evaluate the efficiency of the different correction methods. Thus, we were able to first evaluate performances on a simulated LR dataset derived from our MCF10A and GM12878 transcriptomes and then on real ONT data by assessing the capacity of each correction method for retrieving the ER-Snail fusion.

Measures of evaluation

To assess the behaviour and performance of the 7 correction methods, we computed multiple alignment-free and alignment-dependent indicators that are summarized in Tables 1-3 and Supplementary tables 1-3. These indicators can be broken down into two categories: sequence quality indicators and structure accuracy indicators.

Sequence quality indicators include the identity score between the LR and the molecule from which the LR was generated (this is independent of alignment) and the number and percentage of LRs with at least one alignment to the reference genome.

Structure accuracy requires LRs to be aligned to a reference and measures the proportion of long reads whose transcript structure could be correctly elucidated by aligners and their ability to preserve true exon connectivity and to improve false ones.

Sequence quality of simulated long reads: TALC correction results in more mapped reads and higher identity scores

Our first measure of performance was the number of mapped reads following correction (Table 1). This was evaluated for all 7 methods using three different aligners: Minimap2 (9), GMAP (8) and GraphMap (7). For all three aligners, we kept default parameters suggested for ONT data by their respective authors. We chose Minimap2 and GraphMap as they were specifically designed for long mRNA reads and GMAP because it was considered the best splice-aware aligner for long RNA data according to a recent benchmark (37). TALC shows consistently high mapping numbers regardless of the aligner used. It’s closest competitor in this aspect, LoRDEC, has a high number of mapped reads with Minimap2 and GMAP but this drops to the lowest number of all methods with GraphMap.

We then measured the percentage of identity between the corrected reads and the reference sequence to which they match. This percentage of identity can be calculated in two manners: the number of matched base-pairs between corrected reads and their reference can be divided by the length of the reference sequence or by the length of the read. Thus, if one of the reference or the read is longer than the other, the two methods to calculate identity percentages will differ. The results (Table 1 and Supplementary Table 1A and 1B) of this analysis show that TALC globally performs best on this measure. It’s closest competitor for identity percentage is CoLoRMap

which displays excellent results when the identity is based on the length of the reference transcript but drops to the worst method when based on the length of reads. This is because CoLoRMap tends to extend reads during correction. We noticed a similar trend at the gene-level assessment where LoRDEC correctly associated LRs with their cognate gene when Minimap2 or GMAP were used but was one of the worst performers with GraphMap (Supplementary Table 2).

To assess whether methods performances depended on the coverage in SRs or diverged in presence of many isoforms, we also computed the summary statistics of identity scores by coverage depth in SRs and by splicing complexity (Supplementary tables 3 and 4). TALC globally performed best across the entire range of SR coverage and splicing complexities tested.

Structure accuracy of simulated long reads: TALC correction results in fewer structural errors

LR technologies are able to capture the full connectivity between exons of transcripts. Although this does not necessarily require nucleotide resolution of the transcript, the error rate of current LR technologies is sufficiently high to confound exon connectivity analysis (Figure 3A). Here we define "structural errors" as the incorrect inclusion or deletion of an exon or the incorrect identification of exonic boundaries. We developed a method to systematically identify these errors (Figure 3B and Supplementary material) and evaluated the impact of LR correction methods on them. Because the mapping approach used to assign LRs to a given transcript may impact on structural errors we again tested three mapping approaches: mapping of LRs to a reference genome (hg38) using GMAP (8) and Minimap2 (9), mapping of LRs to a reference transcriptome (hg38) using GraphMap (7). In our evaluation, we measured for each exon, how many were properly identified after correction (Tables 2A and 2B). If these exons were already properly identified before correction, we say that they were preserved by the correction algorithm; if they were not, we say that they were clarified by the correction algorithm (Table 2B).

Regardless of the mapping strategy, we found that TALC is systematically in the top two best algorithms. Again, we notice that other algorithms may compete with TALC on specific criteria given specific aligners but their performance drops drastically in other conditions. For example, when using the Minimap2 aligner, LSC slightly outperforms TALC in preserving the number of properly assigned exons (99.3% versus 98.5% preserved exons) but is very poor at correcting exons that were initially incorrect (14.8% versus 56.6% clarified exons).

Sequence quality and structure accuracy on real data: TALC correction results in higher numbers of alignments, higher average sequence identity and recovers true ER-Snail fusion transcripts

In our final series of tests, we wished to compare TALC with other correction methods using real LR and SR data. However, when evaluating correction methods on real data, performance metrics can be unintentionally skewed because the molecule that was sequenced is unknown. For example, a correction method that would simply replace a LR by a known transcript sequence to which it was most similar could obtain a high percentage of mapped reads and low error rates but would not correctly represent the sequenced transcriptome. Thus, to evaluate the 7 correction methods in this paper, we calculated statistics based on the genome alignment of our

MCF10A cells and the GM12878 cells (Table 3A) but we also made use of an exogenous transcript present at high levels in our cells: the ER-Snail gene fusion. The ER-Snail construct serves as a good control because it is a non-reference transcript that is expressed in the sequenced cells. Using the Minimap2 aligner, which is capable of reliably mapping ONT data to fusion events (38), we were thus able to measure mismatches, deletions, insertions and the number of identified transcripts from the ER-Snail fusion (Table 3B). The results show that TALC again is amongst the best algorithms. FMLRC which was one of the worst performing algorithms on transcript structure does very well on the genome and ER-Snail fusion alignment statistics.

Table 1: General alignment statistics for the 7 correction methods tested. (real SR data, simulated LR data). Only the best reported alignment is counted. The two best performance scores are colored in green and the worst in red. The median identity percentages are calculated by using the number of matches between the corrected read and the true reference sequence and dividing this either by the length of the alignment, the length of the reference or by the length of the read.

MCF10A

Method	Number of unique alignments			Median identity % between reads and true reference sequences		
	Minimap2	GMAP	GraphMap	Overall sequence accuracy (#matches/aligned length)	#matches/length of reference transcript	#matches/length of read
Raw	102087	93015	102276	91.6	84.8	91.7
FMLRC	**101511**	**98302**	**102464**	96.4	94.1	**97.6**
CoLoRMap	**103379**	100622	**84315**	**98.8**	**98.6**	94.7
Hercules	Job killed (>1 week)					
LSC	102367	98589	102101	**93.5**	**88**	**93.5**
LoRDEC	**103237**	**101922**	95171	96.9	94.9	96.1
TALC	102729	**100967**	**102534**	**98.2**	**96.8**	**97.7**

GM12878

Raw	161710	159072	160807	93.6	89.3	93.6
FMLRC	**160669**	158276	160894	**99**	**97.7**	**98.5**
CoLoRMap	161824	**148162**	**161600**	97.8	97.2	**96.2**
Hercules	**161959**	**161164**	160263	**97.6**	**94**	97.6
LSC	Job killed (>1week)					
LoRDEC	**161912**	**161270**	**159596**	98	97.3	97.3
TALC	161284	159141	**161317**	**98.7**	**98.1**	**98.3**

Table 2A: Overall impact of hybrid correction methods on transcript structure (real SR data, simulated LR data). For each aligner, we consider all exons from the transcripts that generated the LR if the LR mapped to the correct gene (#Exons evaluated). We computed the proportion of exons that were correctly mapped (%Identified), and the fraction of full-length LRs for which all exons could be accurately identified (*error-free* structure). The set of exons that could be evaluated varies between tools as it depends on the reads that could be mapped to the correct gene of origin. An exon is considered as identified if its genomic position is fully overlapped and the exon boundaries recognized by the aligner differ by no more than 10bp. False exons are estimated by measuring sequences longer than 25bp introduced in the intronic region of the transcript of origin.

MCF10A

Method	GraphMap			GMAP			Minimap2		
	#Exons evaluated %Identified	Average # of false exons per 100 LRs	%evaluated LR with error-free structure	#Exons evaluated %Identified	Average # of false exons per 100 LRs	%evaluated LR with error-free structure	#Exons evaluated %Identified	Average # of false exons per 100 LRs	%evaluated LR with error-free structure
Raw	784 270 84.4	2.82	90.4	687 948 45.5	10	7300.00 %	768 693 82.7	8.1	72.3
FMLRC	784 201 86.5	**9.24**	**82.1**	733 559 76.6	**14.26**	**70.3**	761 309 83.7	**13.8**	**69.6**
CoLoRMap	615 660 **89.8**	7.7	85.4	760 483 **81.4**	8.69	78.6	775784 **88.7**	10.67	**78.4**
Hercules	Job killed (>1 week)								
LSC	782 747 86.2	**3.17**	**90.7**	730 408 **52.5**	8.54	74.9	775 160 83.1	**7.3**	74.8
LoRDEC	754 764 **84.8**	**2.67**	**90.3**	763 388 77.9	**3.47**	**82.2**	777 363 **87.4**	**4.36**	78.3
TALC	784 471 **89.7**	5.23	90.2	762 436 **80.8**	**6.68**	**82.9**	774 031 **87.4**	7.5	**81.4**

GM12878

Raw	1 146 151 88.9	1.47	94.6	1 091 800 63.4	3.99	86.9	1 124 919 89	3.49	84.1
FMLRC	1 146 613 91	**3.83**	**91.5**	1 108 002 85.5	5.27	**85.8**	1 112 787 88.9	**5.54**	**82.3**
CoLoRMap	1 144 432 **92.9**	2.90	93.6	1 054 935 84.1	**5.39**	88	1 119 129 **90.3**	5.28	85.9
Hercules	1142374 87.8	2.72	92.6	1128603 **73.6**	**1.93**	90.8	1128585 85.9	**2.65**	87.7
LSC	Job killed (>1 week)								
LoRDEC	1 140 409 89.2	**1.24**	**95.8**	1 138 913 90.2	**0.91**	**93.9**	1 124 919 89	**1.54**	**90.1**
TALC	1 147 448 **92.6**	**2.40**	**95.5**	1 123 019 **87.8**	2.46	**93.3**	1 122 112 **91.1**	2.78	**91.6**

Table 2B: Improvement of hybrid correction methods on transcript structure compared to raw data (real SR data, simulated LR data). To evaluate the improvement on exon identification, we computed the number of exons properly mapped in raw data that are still properly mapped after correction (number of exons preserved) and the number of exons missed by the aligner in the raw data that are properly assigned after correction (number of exons clarified). For each correction method, we only consider LRs that aligned to the correct genomic locus both in the raw and in the corrected data set.

MCF10A

Method	GraphMap		GMAP		Minimap2	
	#Raw True #Preserved %preserved	#Raw False #Clarified %Clarified	#Raw True #Preserved %preserved	#Raw False #Clarified %Clarified	#Raw True #Preserved %preserved	#Raw False #Clarified %Clarified
FMLRC	655610 633408 96.6	10070 4854 48.2	297882 279929 **94**	26839 13434 50	617044 591339 **95.8**	34182 13871 40.6
CoLoRMap	511900 496570 97	7838 5027 **64**	302854 287853 95	27471 15789 **57.5**	26080 606432 96.9	34597 20573 **59.5**
Hercules	Job killed					
LSC	657401 654678 **99.6**	10030 2107 **21**	305775 290263 94.9	27226 6248 **22.9**	630779 626339 **99.3**	34638 5111 **14.8**
LoRDEC	632786 608172 **96.1**	9769 5695 58.3	306702 301385 **98.3**	27320 13652 50	628474 624944 **99.4**	34550 12556 36
TALC	655488 647047 **98.7**	10051 6588 **65.5**	303886 296934 **97.7**	27367 16731 **61.1**	624881 615650 98.5	34452 19512 **56.6**

GM12878

FMLRC	1012185 991313 97.9	8519 4840 56.8	669328 646408 96.6	18031 8565 47.5	977535 953327 97.5	27356 9159 33.5
CoLoRMap	1011390 998538 **98.7**	8549 4914 57.5	639460 61722 96.5	17146 10018 58.4	983379 959073 97.5	27093 13154 **48.6**
Hercules	101033 992817 98.2	8493 2772 **32.6**	681744 640724 **94**	18107 8123 **44.9**	989334 965503 97.6	27592 8937 **32.4**
LSC	Job killed					
LoRDEC	1006423 984650 **97.8**	8499 5603 **65.9**	684944 678472 **99.1**	18416 11591 **62.9**	989733 985834 **99.6**	27657 13055 47.2
TALC	1012506 1003879 **99.1**	8600 6118 **71.1**	674515 666202 **98.8**	18273 11949 **65.4**	983063 974064 **99.1**	27494 16366 **59.5**

Table 3A: Number of alignments and average identity with a reference genome (real SR data, real LR data).

Method	MCF10A						GM12878					
	GraphMap		GMAP		Minimap2		GraphMap		GMAP		Minimap2	
	#unique best alignments	Average sequence identity	#unique best alignments	Average sequence identity	#unique best alignments	Average sequence identity	#unique best alignments	Average sequence identity	#unique best alignments	Average sequence identity	#unique best alignments	Average sequence identity
Raw	296941	89.4	231 161	82.5	284 374	87	471171	87.5	437109	84.1	473231	86
FMLRC	**297 927**	**93.1**	272 691	**90.9**	**287 398**	**92.5**	**473576**	**90.9**	**467511**	**89.8**	**473550**	**90.5**
CoLoRMap	270705	91.6	**294 198**	90.4	**298865**	91.5	**404291**	88.7	**453310**	**89.8**	**478657**	90.2
Hercules	296941	90	**271331**	**83.7**	290 827	**87.7**	>1 week					
LoRDEC	**262712**	**89.6**	**286 105**	87.8	**295969**	89.7	465020	**87.7**	**475502**	**88.4**	**478206**	**89.2**
TALC	**299 500**	**93.1**	285 645	**90.6**	294 491	**92.6**	**474672**	**90.8**	462231	**90**	474113	**90.9**

Table 3B: Error rates and number of known ER-Snail sequences retrieved (Minimap2)

Method	Mismatch rate	Deletion rate	Insertion rate	# ER-Snail Fusions detected
Raw	0.139	0.103	0.033	57
FMLRC	0.091	0.026	0.009	262
CoLoRMap	0.113	0.034	0.011	**235**
Hercules	**0.137**	**0.069**	**0.029**	238
LoRDEC	0.122	0.049	0.015	250
TALC	**0.097**	**0.021**	**0.007**	**269**

DISCUSSION

High error rates of long read sequencing technologies can substantially bias the assignment of reads to unique transcripts and can also introduce major structural errors in *de novo* transcript prediction. According to a recent study (10), the use of short read sequencing to correct these errors is the most straightforward and efficient approach to retrieve the correct chain of exons in a transcript. Although numerous algorithms have been developed to implement this hybrid correction, mostly for genome assembly, none of them make use of structural and quantitative information on transcripts to guide their correction.

We devised a hybrid correction method tailored to RNA sequencing that takes into account transcript abundance and the possible existence of splicing junctions to correct long reads. This information is used to guide the exploration of a graph structure by eliminating edges that are not consistent in terms of transcript abundance while simultaneously using an adaptive threshold to account for the existence of multiple transcript isoforms. By eliminating inconsistent nodes, TALC reduces the exploration space in the graph and thus reduces the probability of exploring false paths. By integrating coverage information, TALC can efficiently account for the existence of multiple transcript isoforms.

We tested the efficiency of our approach on data we generated for the purpose of this paper. Globally, TALC shows better correction than currently used methods both in total sequence identity with a reference transcript

but more importantly it conserves the exon connectivity in transcripts. TALC is notably better than other algorithms at correcting whole exons and non-reference transcripts. Although other state of the art correction algorithms compete with and even outcompete TALC on specific criteria, they have clear deficiencies on other important criteria. FMLRC for example is one of the best algorithms in terms of sequence quality indicators but has one of the poorest structural accuracies. CoLoRMap performs well on real data with very few sequence and structural errors but has the lowest number of uniquely mapped reads and is one of the slowest methods. The only drawback to TALC is its high memory requirement (up to 34Gb, Supplementary Table 5) which is increased due to the more complex data structure required to take into account transcript changes.

Given the recent discovery of numerous novel functional transcripts in multiple organs (3, 39). TALC's capacity to correct the entire gamut of transcripts, and preserve the correct transcript structure may prove increasingly valuable.

DATA AVAILABILITY AND IMPLEMENTATION

TALC is written in C++ and uses the SeqAn library (40). The program is freely available under the CECILL licence in the GitLab repository (https://gitlab.igh.cnrs.fr/lbroseus/TALC). Currently, the De Bruijn graph is built from k-mer counts files as output by Jellyfish2 (41).

Direct RNA Nanopore and Illumina RNA-seq MCF10A samples have been deposited on GEO under accession number GSE126638.

AUTHORS' CONTRIBUTIONS

LB developed and implemented the algorithm. LB and WR devised the analyses and planned the experiments. AT provided valuable suggestions for the implementation. AJO cultured MCF10A cells and extracted their RNA. DS carried out the sequencing, and ED performed the base calling of the MiNION sample. LB and WR wrote the article. All authors read and approved the final manuscript.

ACKNOWLEDGEMENT

We wish to acknowledge the Genotoul platform (genotoul.fr) for providing us with calculation time on their servers. We thank Alban Mancheron and Jean-Philippe Villemin for their advice.

FUNDING

This work was supported by the Agence Nationale de la Recherche [ANRJCJC - WIRED], the Labex EpiGenMed [ANR-10-LABX-12-01] and the MUSE initiative [GECKO].

DS acknowledges financial support from France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche [ANR-10-INBS-09].

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

1. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C., *et al.* (2005) The Transcriptional Landscape of the Mammalian Genome. *Science*, **309**, 1559–1563.

2. The FANTOM Consortium and the RIKEN PMI and Clst (dgt), Forrest,A.R.R., Kawaji,H., Rehli,M., Kenneth Baillie,J., de Hoon,M.J.L., Haberle,V., Lassmann,T., Kulakovskiy,I.V., Lizio,M., *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.

3. 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.

4. Bolisetty,M.T., Rajadinakaran,G. and Graveley,B.R. (2015) Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology*, **16**, 204.

5. Middleton,R., Gao,D., Thomas,A., Singh,B., Au,A., Wong,J.J.-L., Bomane,A., Cosson,B., Eyras,E., Rasko,J.E.J., *et al.* (2017) IRFinder: assessing the impact of intron retention on mammalian gene expression. *Genome Biology*, **18**, 51.

6. Shendure,J., Balasubramanian,S., Church,G.M., Gilbert,W., Rogers,J., Schloss,J.A. and Waterston,R.H. (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345–353.

7. 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.

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

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

10. Kuosmanen,A., Norri,T. and Mäkinen,V. (2018) Evaluating approaches to find exon chains based on long reads. *Brief Bioinform*, **19**, 404–414.

11. Fertin,G., Jean,G., Radulescu,A. and Rusu,I. (2015) Hybrid de novo tandem repeat detection using short and long reads. *BMC Medical Genomics*, **8**, S5.

12. 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 Res*, **43**, e116–e116.

13. Deonovic,B., Wang,Y., Weirather,J., Wang,X.-J. and Au,K.F. (2017) IDP-ASE: haplotyping and quantifying allele-specific expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Res*, **45**, e32–e32.

14. Fu,S., Ma,Y., Yao,H., Xu,Z., Chen,S., Song,J. and Au,K.F. (2018) IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics*, **34**, 2168–2176.

15. Au,K.F., Underwood,J.G., Lee,L. and Wong,W.H. (2012) Improving PacBio Long Read Accuracy by Short Read Alignment. *PLOS ONE*, **7**, e46679.

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

17. Haghshenas,E., Hach,F., Sahinalp,S.C. and Chauve,C. (2016) CoLoRMap: Correcting Long Reads by Mapping short reads. *Bioinformatics*, **32**, i545–i551.

18. Firtina,C., Bar-Joseph,Z., Alkan,C. and Cicek,A.E. (2018) Hercules: a profile HMM-based hybrid error correction algorithm for long reads. *Nucleic Acids Res*, **46**, e125–e125.

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

20. 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**, 10.

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

22. Wang,J.R., Holt,J., McMillan,L. and Jones,C.D. (2018) FMLRC: Hybrid long read error correction using an FM-index. *BMC Bioinformatics*, **19**, 50.

23. Limasset,A., Cazaux,B., Rivals,E. and Peterlongo,P. (2016) Read mapping on de Bruijn graphs. *BMC Bioinformatics*, **17**, 237.

24. Lima,L., Sinaimeri,B., Sacomoto,G., Lopez-Maestre,H., Marchet,C., Miele,V., Sagot,M.-F. and Lacroix,V. (2017) Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. *Algorithms for Molecular Biology*, **12**, 2.

25. Peng,Y., Leung,H.C.M., Yiu,S.-M., Lv,M.-J., Zhu,X.-G. and Chin,F.Y.L. (2013) IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, **29**, i326–i334.

26. Le,H.-S., Schulz,M.H., McCauley,B.M., Hinman,V.F. and Bar-Joseph,Z. (2013) Probabilistic error correction for RNA sequencing. *Nucleic Acids Res.*, **41**, e109.

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

28. Fu,S., Wang,A. and Au,K.F. (2019) A comparative evaluation of hybrid error correction methods for error-prone long reads. *Genome Biology*, **20**, 26.

29. Sahraeian,S.M.E., Mohiyuddin,M., Sebra,R., Tilgner,H., Afshar,P.T., Au,K.F., Asadi,N.B., Gerstein,M.B., Wong,W.H., Snyder,M.P., *et al.* (2017) Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications*, **8**, 59.

30. Filichkin,S.A., Hamilton,M., Dharmawardhana,P.D., Singh,S.K., Sullivan,C., Ben-Hur,A., Reddy,A.S.N. and Jaiswal,P. (2018) Abiotic Stresses Modulate Landscape of Poplar Transcriptome via Alternative Splicing, Differential Intron Retention, and Isoform Ratio Switching. *Front. Plant Sci.*, **9**.

31. Zhao,Y., Zhang,Z., Li,M., Luo,J., Chen,F., Gong,Y., Li,Y., Wei,Y., Su,Y. and Kong,L. (2019) Transcriptomic profiles of 33 opium poppy samples in different tissues, growth phases, and cultivars. *Scientific Data*, **6**, 66.

32. Lian,B., Hu,X. and Shao,Z. (2019) Unveiling novel targets of paclitaxel resistance by single molecule long-read RNA sequencing in breast cancer. *Scientific Reports*, **9**, 6032.

33. Philippe,N., Boureux,A., Bréhélin,L., Tarhio,J., Commes,T. and Rivals,É. (2009) Using reads to annotate the genome: influence of length, background distribution, and sequence errors on prediction capacity. *Nucleic Acids Res*, **37**, e104–e104.

34. Shapiro,I.M., Cheng,A.W., Flytzanis,N.C., Balsamo,M., Condeelis,J.S., Oktay,M.H., Burge,C.B. and Gertler,F.B. (2011) An EMT–Driven Alternative Splicing Program Occurs in Human Breast Cancer and Modulates Cellular Phenotype. *PLOS Genetics*, **7**, e1002218.

35. Yang,Y., Park,J.W., Bebee,T.W., Warzecha,C.C., Guo,Y., Shang,X., Xing,Y. and Carstens,R.P. (2016) Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition. *Molecular and Cellular Biology*, **36**, 1704–1719.

36. Javaid,S., Zhang,J., Anderssen,E., Black,J.C., Wittner,B.S., Tajima,K., Ting,D.T., Smolen,G.A., Zubrowski,M., Desai,R., *et al*. (2013) Dynamic Chromatin Modification Sustains Epithelial-Mesenchymal Transition following Inducible Expression of Snail-1. *Cell Reports*, **5**, 1679–1689.

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

38. Gamaarachchi,H., Parameswaran,S. and Smith,M.A. (2019) Featherweight long read alignment using partitioned reference indexes. Sci Rep, 9, 1–12.

39. Clark,M., Wrzesinski,T., Garcia-Bea,A., Kleinman,J., Hyde,T., Weinberger,D., Haerty,W. and Tunbridge,E. (2018) Long-read sequencing reveals the splicing profile of the calcium channel gene CACNA1C in human brain. *bioRxiv*, 10.1101/260562.

40. Reinert,K., Dadi,T.H., Ehrhardt,M., Hauswedell,H., Mehringer,S., Rahn,R., Kim,J., Pockrandt,C., Winkler,J., Siragusa,E., *et al.* (2017) The SeqAn C++ template library for efficient sequence analysis: A resource for programmers. *Journal of Biotechnology*, **261**, 157–168.

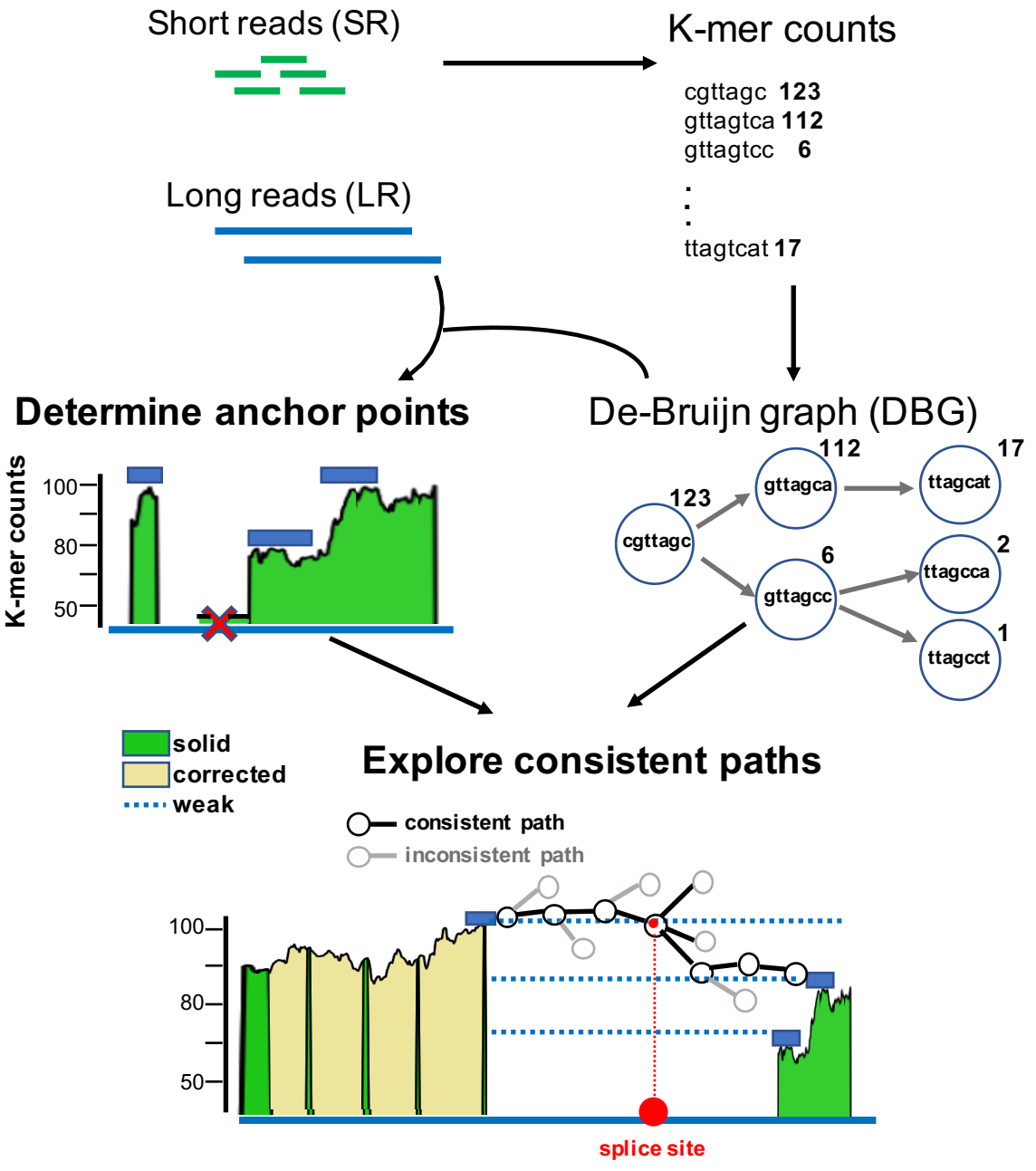
41. Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

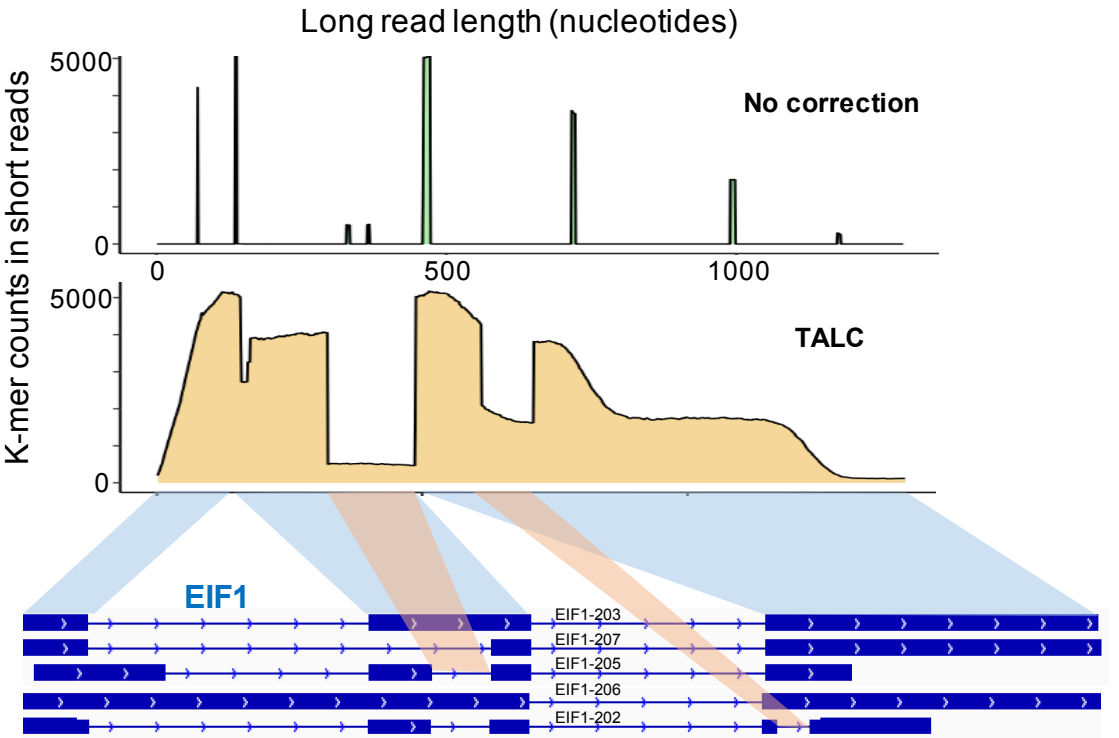
FIGURE LEGENDS

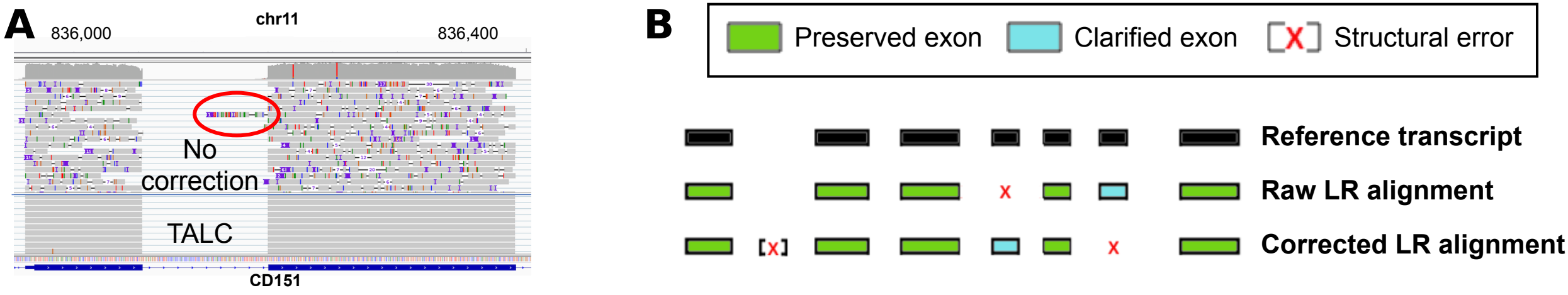
Figure 1: Overview of the TALC algorithm. TALC's correction procedure begins by creating a weighted DBG (DBG) from the k-mer counts of the SRs. It then searches for common sequences between the LRs and the k-mers of the DBG. Stretches of common k-mers are called *Solid Regions* and are assumed to be error-free parts of the long read (green peaks). From the solid regions, TALC will then **determine anchor points**. These will be used as entry points into the DBG; regions between these anchor points are called *weak regions* and will be corrected. To extract anchor points from the solid regions TALC will first crop background noise that frequently surround solid regions (red cross). The second step in determining anchors is to search for sudden changes in coverage within a solid region. These changes may correspond to divergent transcript architectures and should be explored separately. Thus, a solid region will be split into as many anchors as there are changes in coverage. Finally, to correct weak paths between anchors, TALC will **explore consistent paths.** TALC explores the DBG, following paths of similar k-mer coverage. When the exploration reaches a potential transcriptional event such as a splice site (red dot), the exploration will branch out to account for the existence of multiple isoforms with varying coverage.

Figure 2: Example of K-mer coverage across a long read before and after TALC correction. We mapped the k-mers discovered in the short-read (SR) sequencing of MCF10A cells onto one of the long reads (LR) sequenced from the same sample. The x-axis represents each nucleotide of the LR, the y-axis shows the abundance of the mapped k-mers in the SR sample. Blue shading at the bottom of the figure highlights the correspondence between the transcript to which the long read was mapped and the read itself; red shading highlights alternative transcript architectures that could explain the sudden changes in the k-mer coverage of the graph. Alternative isoforms are annotated using ENSEMBL transcript annotation.

Figure 3: Structural errors in long reads after different correction methods. A) Screenshot of the IGV viewer showing long reads aligned to the genome with GMAP. A structural error (insertion) has occurred in the non-corrected long reads (red circle). **B)** Overview of our approach to finding structural errors which are either deletions or insertions in LRs computationally generated from a given transcript.


Short reads (SR)
K-mer counts
cgttagc 123
gttagtca 112
gttagtcc 6
ttagtcat 17
Long reads (LR)
Determine anchor points
De-Bruijn graph (DBG)
K-mer counts
100
80
50
123
cgttagc
112
gttagca
17
ttagcat
6
gttagcc
2
ttagcca
1
ttagcct
solid
corrected
weak
Explore consistent paths
consistent path
inconsistent path
100
80
50
splice site


Long read length (nucleotides)
K-mer counts in short reads
5000
0
No correction
0
500
1000
5000
0
TALC
EIF1
EIF1-203
EIF1-207
EIF1-205
EIF1-206
EIF1-202


A
836,000
chr11
836,400
No correction
TALC
CD151
B
Preserved exon
Clarified exon
Structural error
Reference transcript
Raw LR alignment
Corrected LR alignment