ReorientExpress: reference-free orientation of nanopore cDNA reads with deep learning

Angel Ruiz-Reche[1], Joel A. Indi[1,2], Ivan de la Rubia[1], Eduardo Eyras[3,4,5,*]

[1]Pompeu Fabra University, E08003, Barcelona, Spain

[2]Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

[3]Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain

[4]IMIM, E08003, Barcelona, Spain

[5]John Curtin School of Medical Research, The Australian National University, Canberra, Australia

Long-read sequencing technologies allow the systematic interrogation of transcriptomes from any species. However, functional characterization requires the determination of the correct 5'-to-3' orientation of reads. Oxford Nanopore Technologies (ONT) allows the direct measurement of RNA molecules in the native orientation (Garalde et al. 2018), but sequencing of complementary-DNA (cDNA) libraries yields generally a larger number of reads (Workman et al. 2018). Although strand-specific adapters can be used, error rates hinder their detection. Current methods rely on the comparison to a genome or transcriptome reference (Wyman and Mortazavi 2018; Workman et al. 2018) or on the use of additional technologies (Fu et al. 2018), which limits the applicability of rapid and cost-effective long-read sequencing for transcriptomics beyond model species. To facilitate the interrogation of transcriptomes de-novo in species or samples for which a genome or transcriptome reference is not available, we have developed ReorientExpress (https://github.com/comprna/reorientexpress), a new tool to perform reference-free orientation of ONT reads from a cDNA library, with our without stranded adapters. ReorientExpress uses a deep neural network (DNN) to predict the orientation of cDNA long-reads independently of adapters and without using a reference.

ReorientExpress approach builds on the hypothesis that, despite the sequencing errors, reads obtained from transcripts should still retain multiple sequence motifs corresponding to a strand-specific RNP code that is responsible for RNA metabolism (Rissland 2017; Hentze et al. 2018; Gerstberger et al. 2014). The model can be trained from any set of RNA sequences with known orientation, like a transcriptome annotations or ONT direct RNA sequencing (DRS) reads. ReorientExpress calculates a matrix of normalized k-mer counts from the input sequences, with k from 1 to any specified length (Supp. Methods). A DNN model is then trained to classify long reads into being in the correct 5'-to-3' orientation or in the reverse-complemented orientation. We used a DNN with 5 hidden layers, 500 nodes in the first one and half the amount at each subsequent layer (i.e. 500, 250, 125, 62, 1). The last layer has only one node and applies a sigmoid function to approximate a probability from its input score. This probability represents the certainty that the read

is not in the right orientation. Additionally, the DNN uses dropout layers to reduce overfitting. ReorientExpress implementation allows other DNN architectures.

To test ReorientExpress we trained a model on the human transcriptome using k-mers from k=1 to k=5 (Supp. Methods). Testing this model on DRS reads from the Nanopore Sequencing Consortium (Workman et al. 2018) showed 0.86 precision and recall (Supp. Table 1). We trained a similar model using the *Saccharomyces cerevisiae* transcriptome, which yielded 0.88 precision and recall values on DRS reads from *S. cerevisiae* (Garalde et al. 2018) (Supp. Table 2). To evaluate the accuracy of ReorientExpress on ONT cDNA-seq data with no known orientation, we first mapped human (Workman et al. 2018) and *S. cerevisae* (Garalde et al. 2018) ONT cDNA reads to their respective annotated transcriptomes using minimap2 (Li 2018), selecting only primary mappings. These mappings showed consistency for the orientation of the read with all possible matching transcripts, which in human mostly corresponded to multiple isoforms from the same gene. We assigned to each read the strand according to the best matching transcript annotation according to a score calculated as the fraction of the read length corresponding to matching bases. Using these orientations, we tested ReorientExpress models trained on the transcriptome annotations on the cDNA datasets. This yielded a precision of 0.84 and a recall of 0.83 for human (Figs. 1a and 1b) (Supp. Table 3), and a precision and recall of 0.93 for *S. Cerevisiae* (Figs. 1c and 1d) (Supp. Table 4).

Remarkably, training with ONT DRS reads and testing with the ONT cDNA reads yielded a lower accuracy (Supp. Table 5). This may suggest an effect due to the presence of adapters or differences in poly-A tail lengths. To evaluate this, we trimmed a number of nucleotides from both ends of the test cDNA reads. The same human model showed a similar accuracy values at different trimmings (10nt - precision = 0.84, 20nt - precision = 0.84, 40nt - precision = 0.83; 80nt – precision = 0.83; 100nt – precision = 0.81, 200nt – precision = 0.79) (Supp. Table 6). This indicates that long reads contain signals beyond the poly-A tail that are specific of the 5’-to-3’ orientation of RNA molecules.

For comparison, we trained a Markov model based on k-mer frequencies (k=1,…,5) to predict cDNA read orientation (Supp. Methods). This yielded worse accuracy compared to ReorientExpress (Precision and recall = 0.72) (Supp. Tables 7 and 8). This indicates that although k-mers are informative for long-read orientation and localization, more complex models, like the one implemented in ReorientExpress, are needed to capture the complex associations of RNA motifs that are characteristic of 5’-to-3’ orientation.

To demonstrate the suitability of ReorientExpress for samples without a genome reference available, we mimicked this situation by building a DNN model in one species and testing on a different one. We trained a model (k=1,…,5) with the mouse transcriptome after removing pseudogenes (Supp. Material), and tested it on human

ONT cDNA reads. This showed a comparable accuracy to the model trained on human data (precision and recall = 0.83) (Fig. 1) (Supp. Table 9), and higher accuracy values when tested on ONT DRS reads (precision and recall = 0.87) (Supp. Table 10). We also trained a model (k=1,…,5) with the transcriptome annotation for the fungus *Candida glabrata*. When tested on *S. cerevisae* ONT cDNA reads, this model yielded accuracy values as high as for the previous *S. cerevisae* model (precision and recall = 0.94) (Fig. 1) (Supp. Table 11). A similar situation was found when the model was tested on DRS reads (precision and recall = 0.87) (Supp. Table 12). Testing more distant cross-species models yielded slightly lower accuracies (Supp. Table 13).

To further demonstrate the potential impact of ReorientExpress for the reference-free interrogation of transcriptomes with long-reads, we performed clustering of the cDNA reads using IsONclust (Sahlin and Medvedev 2018). For the majority of clusters with >2 reads (89.9% for human and 96.2% for yeast) ReorientExpress predicted correctly more than 50% of the reads in a cluster (Figs. 1e and 1f) (Supp. Fig 1). We thus applied a majority vote per cluster to set the orientation of all reads in each cluster and set the orientation of all reads to be the majority label predicted by ReorientExpress in the cluster (Supp. Methods). With this, ReorientExpress established the right orientation for the majority of cDNA reads tested (92% for human and 97% for yeast) (Fig. 1g).

On the basis of the above results, we conclude that ReorientExpress facilitates the interpretation of transcriptomes from long-read cDNA sequencing data without the need of a genome reference or an additional technology. ReorientExpress provides a crucial aid in the interpretation of transcripts using cDNA long-reads in samples for which the genome reference is not known, as it is the case for many non-model organisms, but also for long-reads from unstranded libraries in general in human and model organisms beyond the available references. This is particularly relevant considering the differences observed between individuals at large and small genomic scales (Sherman et al. 2019; Dashnow et al. 2018). Our analyses show that ReorientExpress can be very valuable in combination with long-read clustering methods (Marchet et al. 2018; Sahlin and Medvedev 2018) to facilitate more accurate downstream analyses of transcriptomes, like the prediction of open reading frames (Watson and Warr 2019). The ability to predict the 5'-to-3' orientation of cDNA nanopore reads using models trained on related species, makes ReorientExpress a key processing tool for the study of transcriptomes from non-model organisms with long-reads.

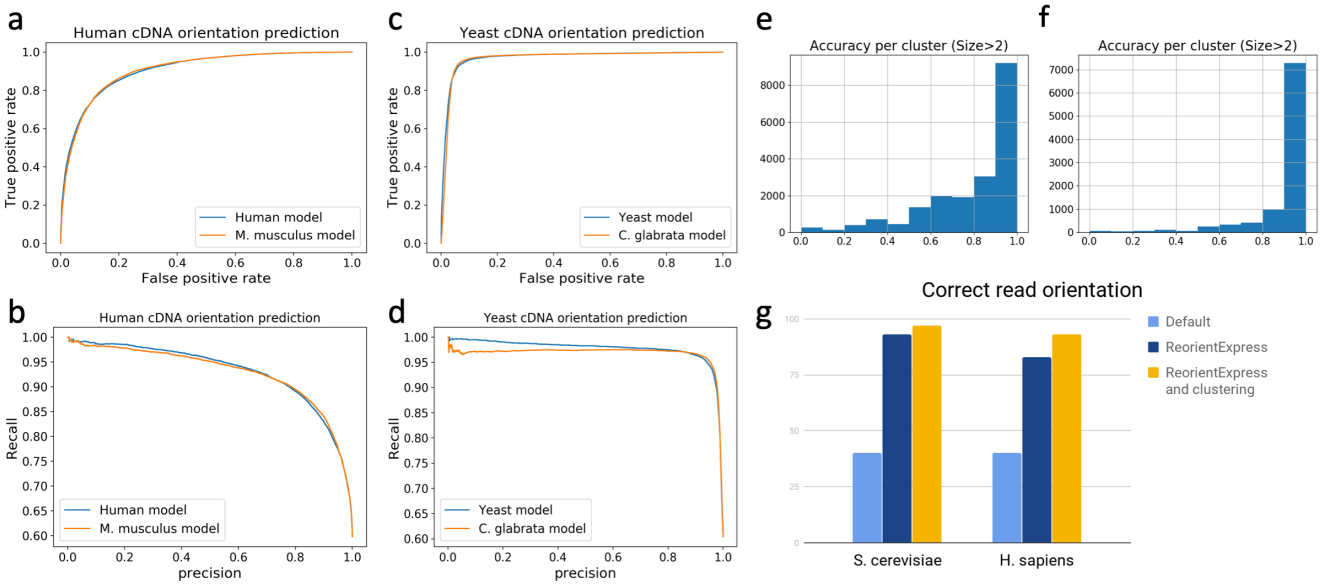

Figure 1. Accuracy analysis of ReorientExpress. (a) Receiving Operating Characteristic (ROC) curves, representing the false positive rate (x axis) versus the true positive rate (y axis) for the orientation of human cDNAs**.** We show the curves for the models trained on the human (blue) and mouse (orange) transcriptomes and tested on human ONT cDNA reads. The curves are built by varying the minimum score threshold required for the orientation prediction as produced by ReorientExpress. **(b)** Precision (x axis) and recall (y axis) (same as true positive rate) curve for the same models in (a). **(c)** As in (a) but tested on *S. cerevisiae* cDNAs. The curves show the accuracy for the models trained on the *S. cerevisiae* (blue) and the *C. glabrata* (orange) transcriptomes and tested on *S. cerevisiae* ONT cDNA reads. **(d)** Precision (x axis) and recall (y axis) curve for the same models in (b). **(e)** Distribution of clusters according to the proportion of human ONT cDNA reads in each cluster with orientation correctly predicted by ReorientExpress (x axis) trained on the human transcriptome. Only clusters with more than 2 reads are shown. Similar plots with all clusters and clusters with more than 1 read are shown in Supp. Fig. 1. **(f)** Same as in (e) but for *S. cerevisiae* ONT cDNA reads (Garalde et al. 2018) and for the model trained on the *S. cerevisiae* annotated transcriptome. **(g)** Comparison of the proportion of human or yeast ONT cDNA reads correctly oriented in three cases: taking the default orientation from the FASTQ file (Default), using ReorientExpress (ReorientExpress), and using a majority vote in clusters to predict the orientation of all reads in each cluster (ReorientExpress and clustering).

References

Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, Davis M, Lamont P, Clayton JS, Laing NG, et al. 2018. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol* **19**: 121. http://www.ncbi.nlm.nih.gov/pubmed/30129428.

Fu S, Ma Y, Yao H, Xu Z, Chen S, Song J, Au KF. 2018. IDP-denovo: de novo

transcriptome assembly and isoform annotation by hybrid sequencing. *Bioinformatics* **34**: 2168–2176. http://www.ncbi.nlm.nih.gov/pubmed/29905763.

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* **15**: 201–206.

Gerstberger S, Hafner M, Tuschl T. 2014. A census of human RNA-binding proteins. *Nat Rev Genet* **15**: 829–45. http://www.ncbi.nlm.nih.gov/pubmed/25365966.

Hentze MW, Castello A, Schwarzl T, Preiss T. 2018. A brave new world of RNA-binding proteins. *Nat Rev Mol Cell Biol* **19**: 327–341. http://www.ncbi.nlm.nih.gov/pubmed/29339797.

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

Marchet C, Lecompte L, Silva C Da, Cruaud C, Aury J-M, Nicolas J, Peterlongo P. 2018. De novo clustering of long reads by gene from transcriptomics data. *Nucleic Acids Res*. http://www.ncbi.nlm.nih.gov/pubmed/30260405.

Rissland OS. 2017. The organization and regulation of mRNA-protein complexes. *Wiley Interdiscip Rev RNA* **8**. http://www.ncbi.nlm.nih.gov/pubmed/27324829.

Sahlin K, Medvedev P. 2018. De novo clustering of long-read transcriptome data using a greedy, quality-value based algorithm. *bioRxiv* 463463. https://www.biorxiv.org/content/early/2018/11/06/463463.

Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, et al. 2019. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet* **51**: 30–35. http://www.ncbi.nlm.nih.gov/pubmed/30455414.

Watson M, Warr A. 2019. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* **37**: 124–126. http://www.ncbi.nlm.nih.gov/pubmed/30670796.

Workman RE, Tang A, Tang PS, Jain M, Tyson JR, Zuzarte PC, Gilpatrick T, Razaghi R, Quick J, Sadowski N, et al. 2018. Nanopore native RNA sequencing of a human poly(A) transcriptome. *bioRxiv*.

Wyman D, Mortazavi A. 2018. TranscriptClean: Variant-aware correction of indels, mismatches, and splice junctions in long-read transcripts. *Bioinformatics*. http://www.ncbi.nlm.nih.gov/pubmed/29912287.