Reference-free reconstruction and error correction of transcriptomes from Nanopore long-read sequencing

Ivan de la Rubia[1], Joel A. Indi, Silvia Carbonell, Julien Lagarde, M Mar Albà, Eduardo Eyras

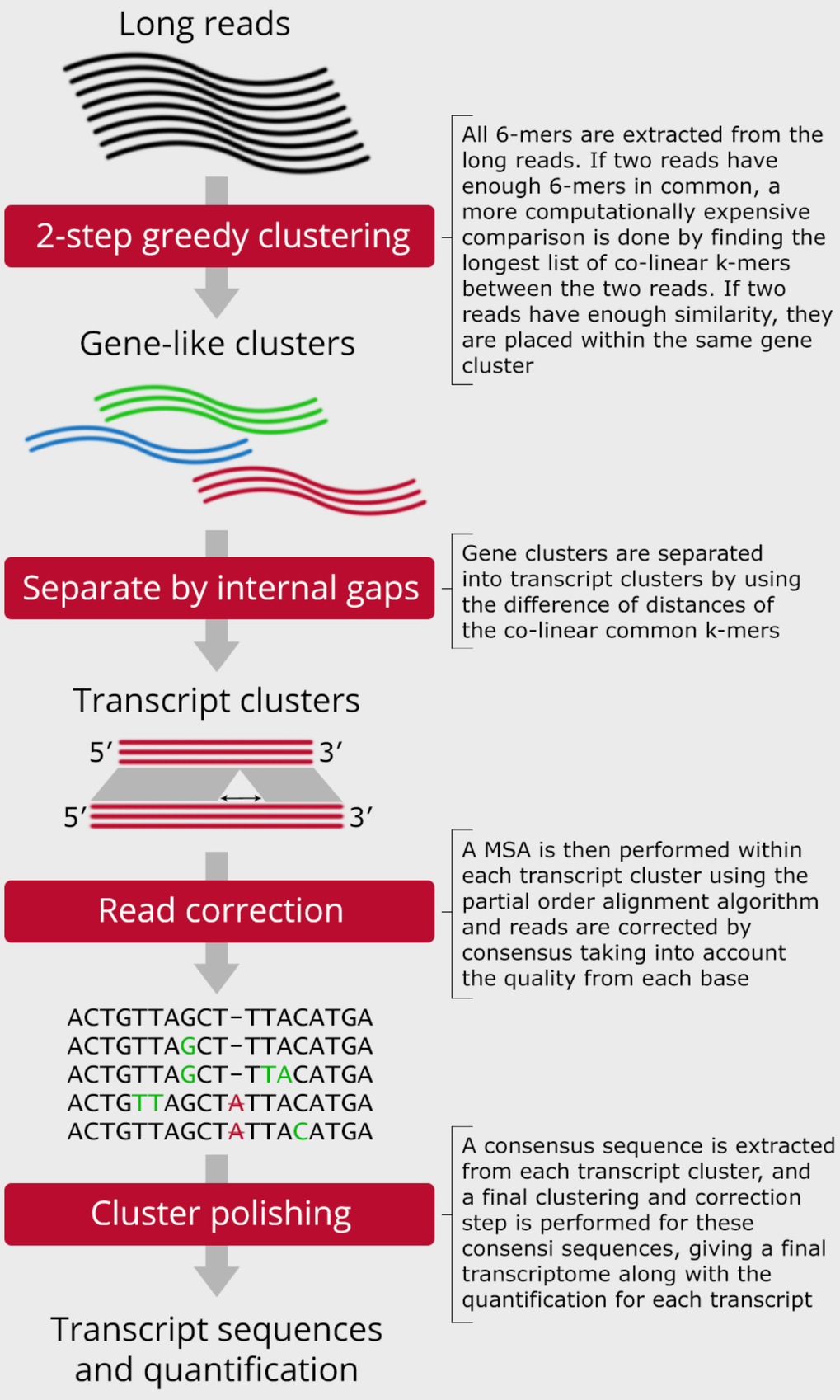
[1] Contact: ivan.delarubia@upf.edu

1 Introduction

Long read sequencing technologies still produce a considerable error rate, averaging around 10% in the case of ONT MinION. Full-length transcript sequencing in eukaryotes have the added challenge of having to deal with introns and small transcript variations (alternative splicing). This presents challenges not only for error correction but also for sequence alignment to reference genomes.

We have thus developed RATTLE, a new method for the reference-free reconstruction of transcriptomes from long-read sequencing from Nanopore. RATTLE clusters long (cDNA and RNA) reads using approximate sequence similarity, performs error correction per cluster, and outputs corrected reads and quantified transcripts.

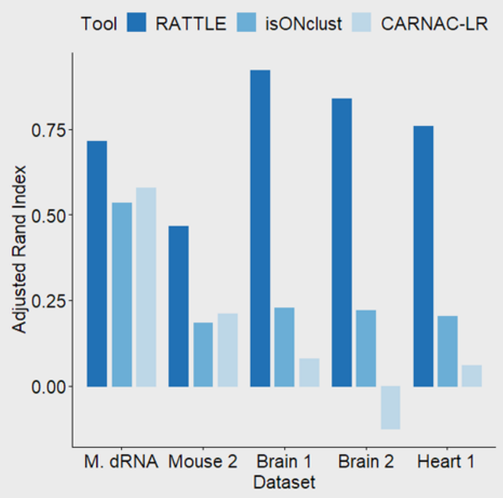
2 Methods



3 Results

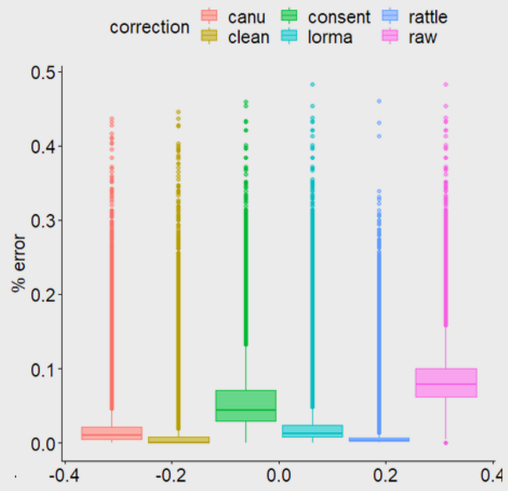
Clustering accuracy

Clustering accuracy using Spike-in RNA Variant (SIRV) genes as reference. The plot shows the adjusted rand index (y axis) for RATTLE, CARNAC and isONclust for each sample (x axis): Nanopore direct RNA sequencing from mouse brain tissue (M. dRNA) and cDNA sequencing from human (Brain1, Brain2, Heart) and mouse (Mouse 2) tissues



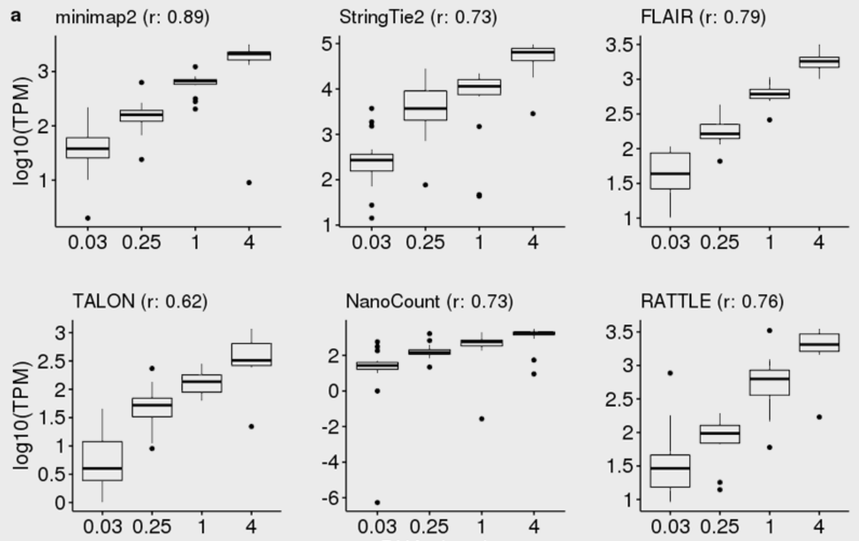
Error rate

Error rate distribution of SIRV reads before and after correction with RATTLE, CONSENT, Lorma, Canu and TranscriptClean (clean). Error rate was calculated as the sum of insertions, deletions and substitutions divided by the length of the read



Transcriptome reconstruction & quantification

Comparison of the predicted transcript abundances (y axis) by RATTLE, FLAIR, StringTie2, TALON, NanoCount, and selecting the best match from minimap2, with abundances of the SIRV transcript isoforms (x axis). In each panel we show the Pearson correlation R from the comparison of the abundance values for that method.



4 Conclusion

We have presented a new method for clustering, error correction and transcriptome reconstruction and quantification of cDNA and RNA Nanopore long reads without using a reference transcriptome or genome, which is specially useful in species where no such reference is available.

Our method outperforms other clustering and error correction methods (even those reference-based) both in simulated and experimental MinION data, and improves mapping quality to a reference genome or transcriptome after error correction has been performed.

Source code available at GitHub:
https://github.com/comprna/RATTLE

DOI:
10.1101/2020.02.08.939942