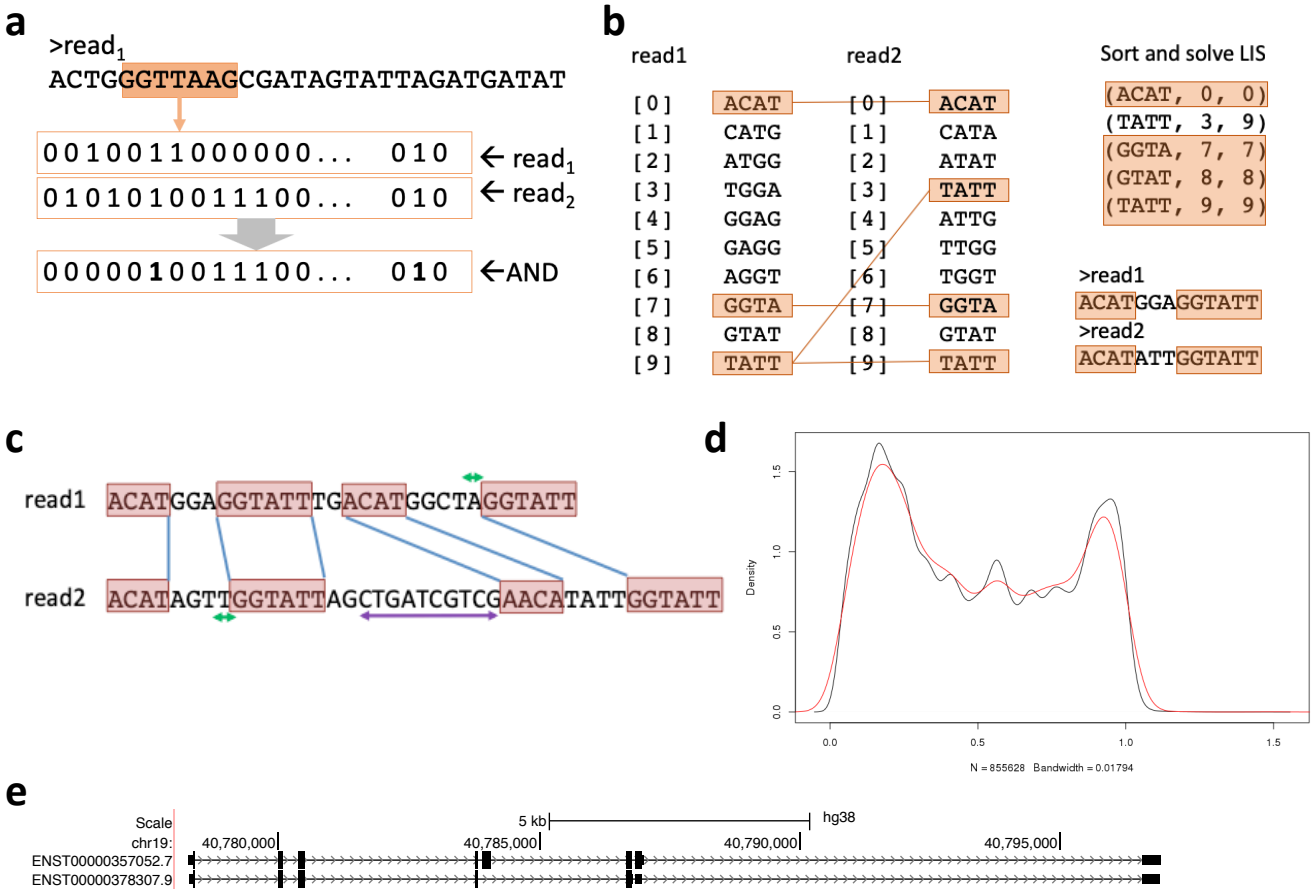
Supplementary Figures

Reference-free reconstruction and quantification of transcriptomes from long-read sequencing

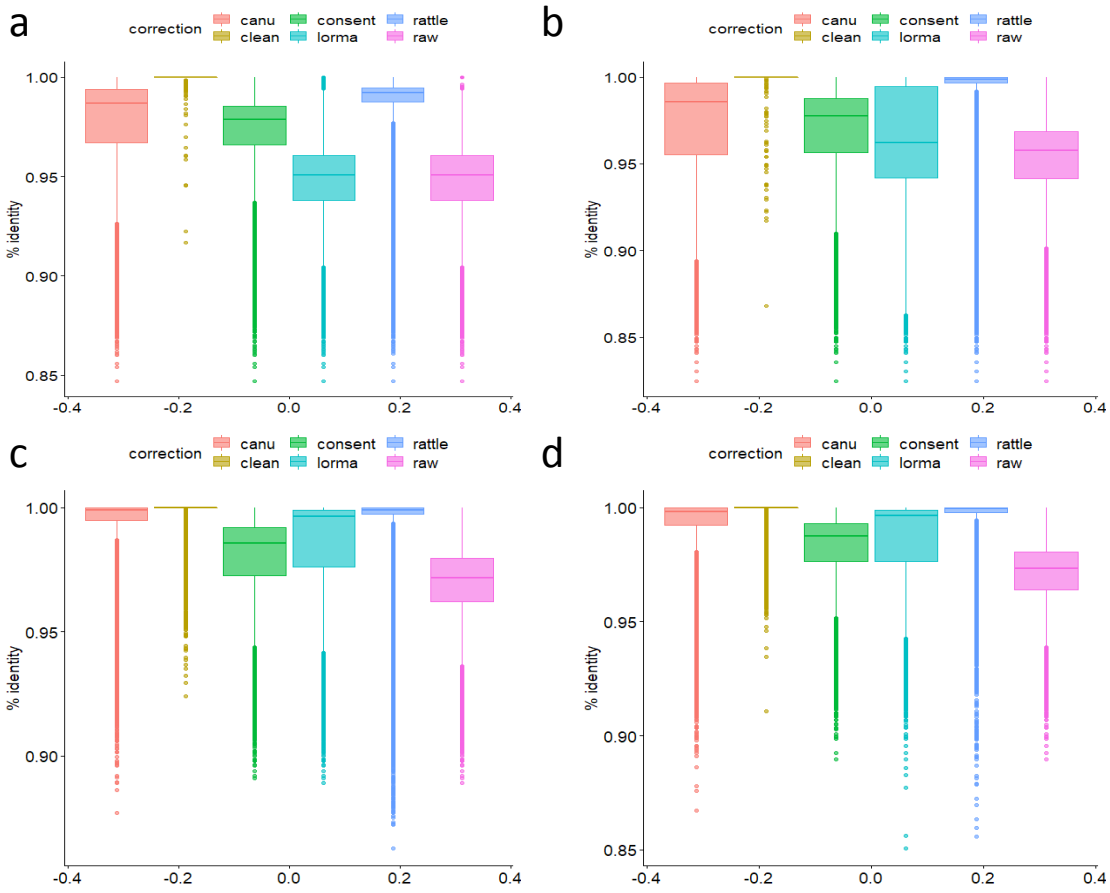
Ivan de la Rubia[1,2], Joel A. Indi[1,3], Silvia Carbonell[2,4], Julien Lagarde[2,4], M Mar Albà[2,5,6], Eduardo Eyras[1,6,7]

[1]EMBL Australia Partner Laboratory Network at the Australian National University, Acton ACT 2601, Canberra, Australia
[2]Pompeu Fabra University, E08003 Barcelona, Spain.
[3]Universidade de Lisboa, Lisboa, Portugal
[4]CRG, E08001 Barcelona, Spain
[5]ICREA, E08010 Barcelona, Spain
[6]IMIM, E08001 Barcelona, Spain
[7]Australian National University, Acton ACT 2601, Canberra, Australia

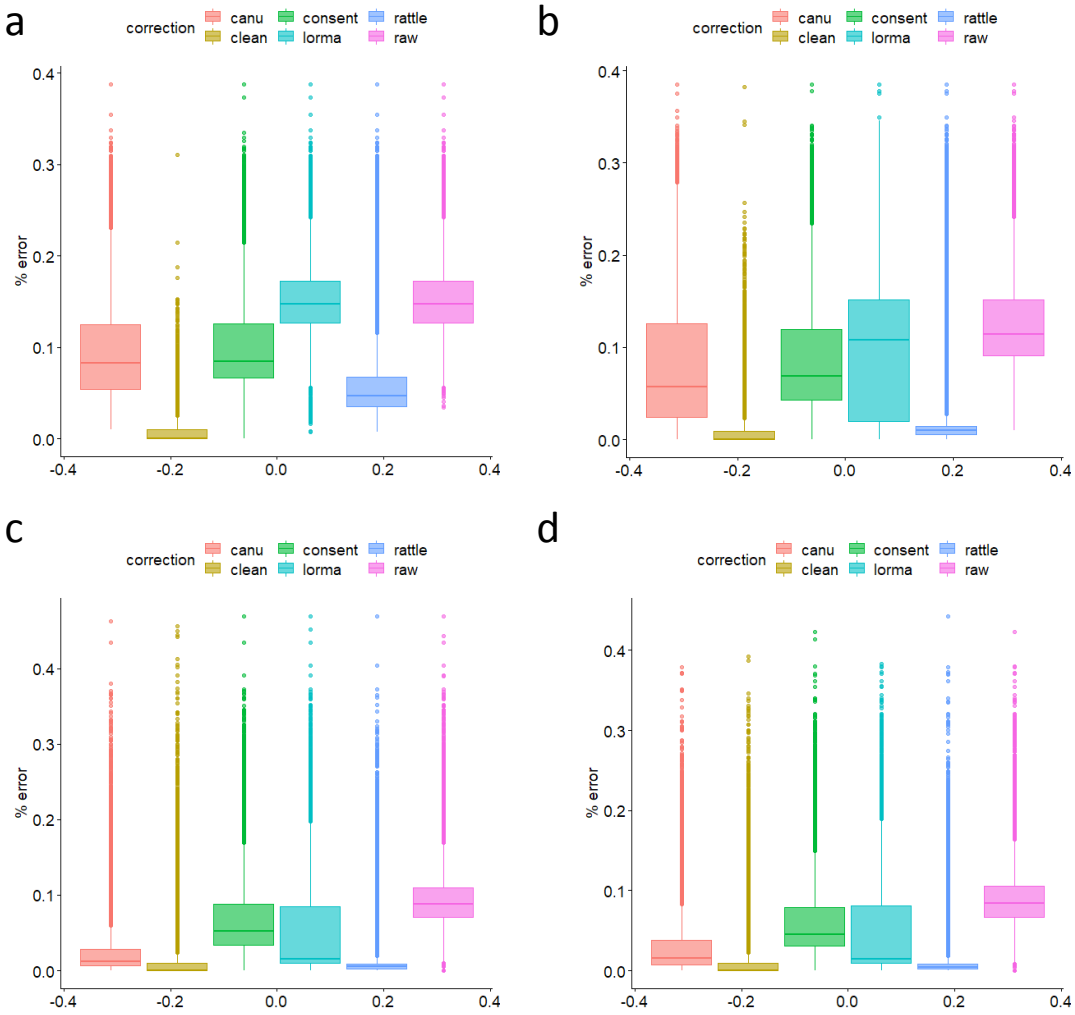
Correspondence to: eduardo.eyras@anu.edu.au



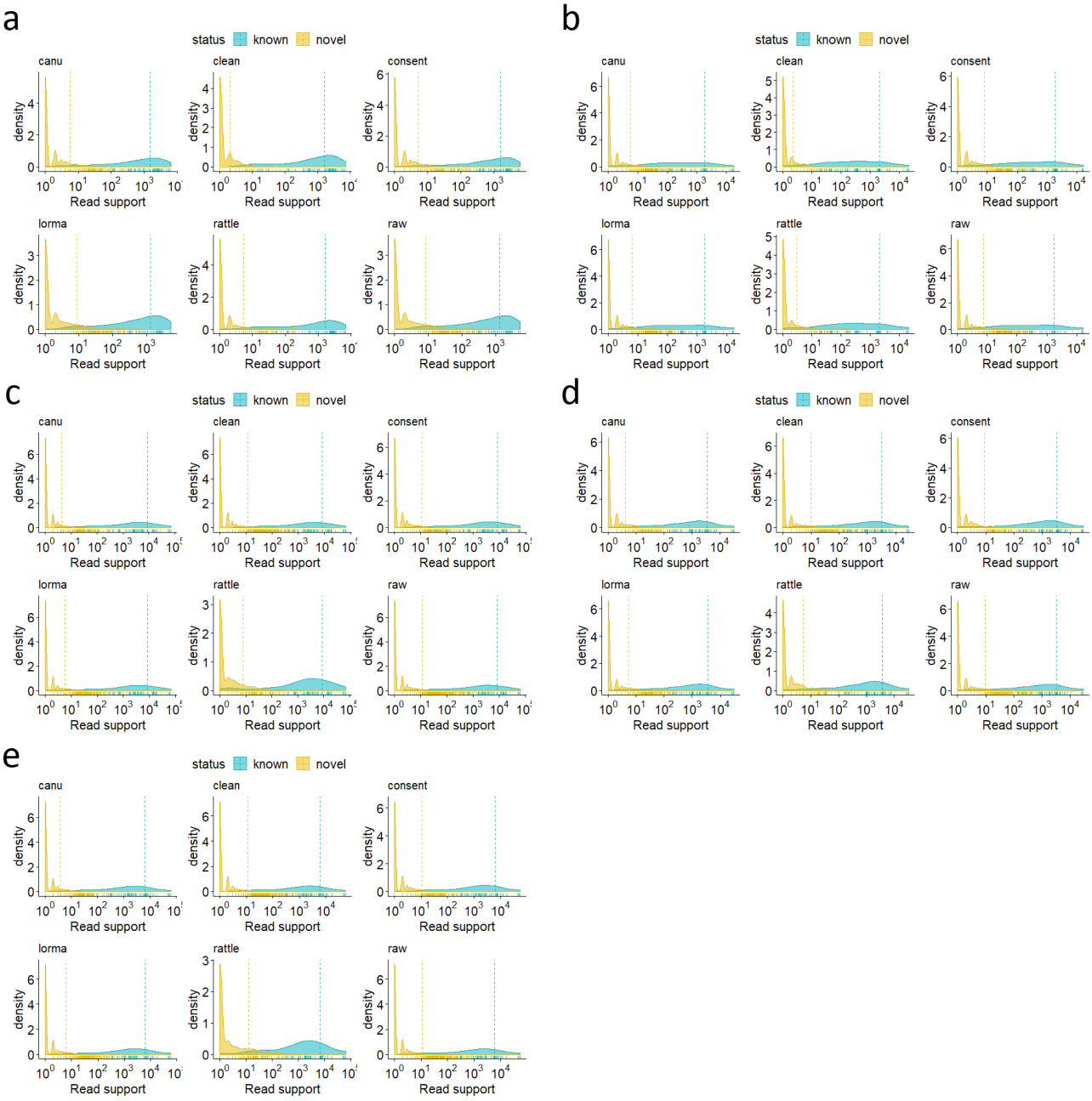
Supplementary Figure 1. RATTLE calculates the similarity between two reads in two steps. The first measure **(a)** computes the number of common k-mers using k-mer hashing, and a logical AND between binary representations of the k-mer content. The first similarity score is calculated as the fraction of unique k-mers in common over the maximum of unique k-mers in the two reads. If this similarity exceeds a pre-defined threshold, a second comparison is performed between both reads. **(b)** In the second step all k-mers occurring in both reads and their positions are extracted. This list is sorted by the positions in the first read, and the Longest Increasing Subsequence (LIS) problem is solved with a dynamic programming algorithm for the position of the k-mers in the second read. This yields a common set of co-linear k-mers between two reads and the number of bases covered by this set is used as similarity value. The similarity value is defined as the number of bases covered by these k-mers over the length of the shortest read in the pair. **(c)** Figure illustrating how reads in a gene-cluster are separated into transcript-clusters. Blocks of adjacent matching k-mers (highlighted in orange) are separated by gaps. There are length differences in gaps due to base-calling errors (indicated in green) and due to a different exonic content (highlighted in purple). The variance of these gap differences between two reads from two different transcripts is expected to be larger than for reads from the same transcript. **(d)** Length distribution of real cDNA reads (black line) and of simulated reads (red line). **(e)** Exonic structure for the two transcripts ENST00000378307.9 and ENST00000357052.7 used to compare the variance in the distribution of gap-length differences in adjacent matching k-mers (Fig. 1c).



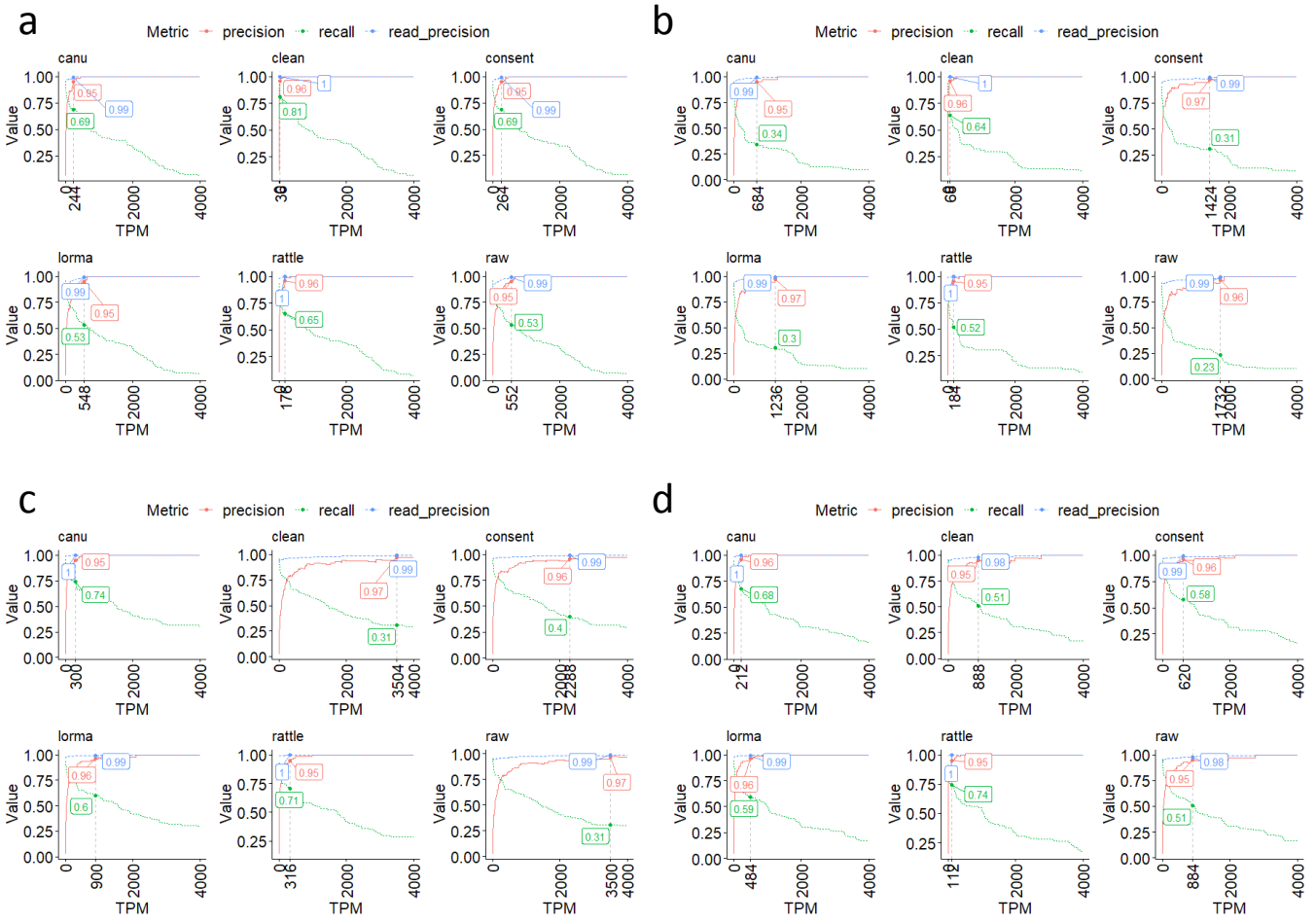
Supplementary Figure 2. Distribution of percentage identity for SIRV reads before (raw) and after correcting with RATTLE, CONSENT, Lorma, Canu and TranscriptClean (clean) for the samples of (a) Mouse brain Nanopore RNA-seq (ERR2680375), (b) Mouse brain Nanopore cDNA-seq (ERR2680377), (c) Human brain Nanopore cDNA-seq (Brain 2), and (d) Human heart Nanopore cDNA-seq (Heart). Percentage identity was calculated as the number of correct matches divided by the total length of the aligned region.



Supplementary Figure 3. Distributions of error rates for SIRV reads before (raw) and after correcting with RATTLE, CONSENT, Lorma, Canu and TranscriptClean (clean) for the same samples: (a) Mouse brain Nanopore RNA-seq (ERR2680375), (b) Mouse brain Nanopore cDNA-seq (ERR2680377), (c) Human brain Nanopore cDNA-seq (Brain 2), and (d) Human heart Nanopore cDNA-seq (Heart). Error rate was calculated as the sum of insertions, deletions and substitutions divided by the length of the read.



Supplementary Figure 4. Density profiles (y axis) for the read support (x axis) for true positive introns (known) and false positive introns (novel) using the SIRV annotations as reference using RATTLE, CONSENT, Canu, LORMA and TranscriptClean (clean) for all samples tested: **(a)** Mouse brain Nanopore RNA-seq (ERR2680375), **(b)** Mouse brain Nanopore cDNA-seq (ERR2680377), **(c)** Human brain Nanopore cDNA-seq (Brain 1), **(d)** Human heart Nanopore cDNA-seq (Heart), and **(e)** Human brain Nanopore cDNA-seq (Brain 2).



Supplementary Figure 5. We plot the recall (green), precision (red) and read-precision (blue) of the SIRV introns, as a function of an expression cut-off (x axis) in terms of the number of reads supporting the introns. We indicate for each case the threshold at which a precision of approximately 0.95 is achieved. For that threshold we indicate the corresponding recall, precision, and read-precision values. The plot corresponds to the samples **(a)** Mouse brain Nanopore RNA-seq (ERR2680375), **(b)** Mouse brain Nanopore cDNA-seq (ERR2680377), **(c)** Human brain Nanopore cDNA-seq (Brain 2), **(d)** and Human brain Nanopore cDNA-seq (Brain 2).