Piquant: a pipeline for assessing the performance of transcriptome quantification tools

Owen Dando[1,2*], Peter Kind[1,2], T. Ian Simpson[3,4]

[1]Centre for Brain Development and Repair, Institute for Stem Cell Biology and Regenerative Medicine (inStem), Bangalore
[2]Centre for Integrative Physiology, University of Edinburgh
[3]Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh
[4]Biomathematics and Statistics Scotland (BioSS)
*contact: owen.dando@ed.ac.uk @odando

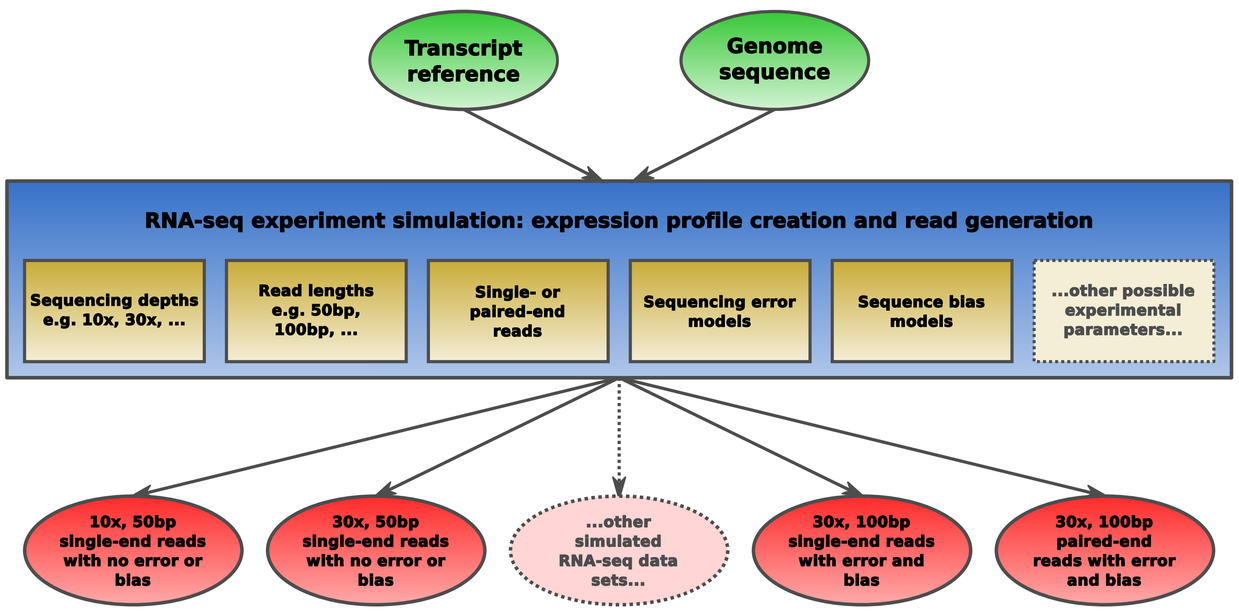
RNA-sequencing has become an important technique in cellular biology for characterising and quantifying the transcriptome, and many computational methods have been developed to reconstruct transcripts from RNA-seq data and then estimate their abundances. Gene expression estimates calculated by these methods have been shown to be relatively robust. However, for transcript quantification, complications arising from the ambiguous origin of short RNA-seq reads and from bias in their sequence composition are compounded, and thus estimates of isoform abundance may be less accurate. It is therefore important to be able to assess the conditions under which different transcriptome quantification tools perform well or more poorly, and how the many optional parameter choices available for each tool may affect their performance.

piquant is a pipeline of Python scripts to help assess the accuracy of transcriptome quantification by such tools. In its first stage, RNA-seq reads are simulated from a starting set of transcripts under specified combinations of sequencing parameters: for example, different read lengths and sequencing depths, single- and paired-end reads, reads with or without sequencing errors, and reads with or without sequence bias. In the second stage, a number of transcriptome quantification tools (or the same tool with different optional parameter choices) estimate isoform abundances for each set of simulated reads. Finally, the isoform expression estimates calculated by each tool for each RNA-seq data set are compared to the known transcript abundances used to generate the reads. The comparative accuracy of estimates calculated by each tool can then be assessed as sequencing parameters are changed, or for different groups of transcripts segregated by particular transcript classification measures, via a range of automatically generated statistics and graphs.

Stages of the *piquant* pipeline

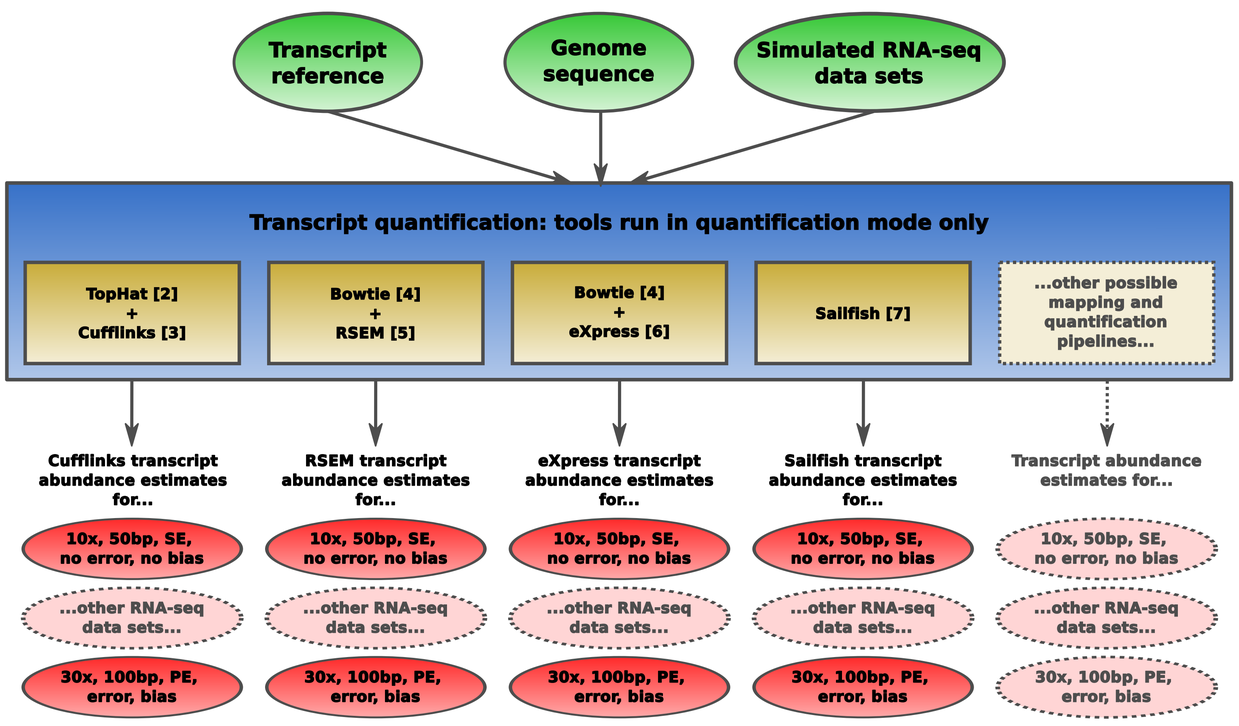
1. Simulation

piquant uses the Flux Simulator [1] RNA-seq experiment simulator to generate sets of RNA-Seq data.



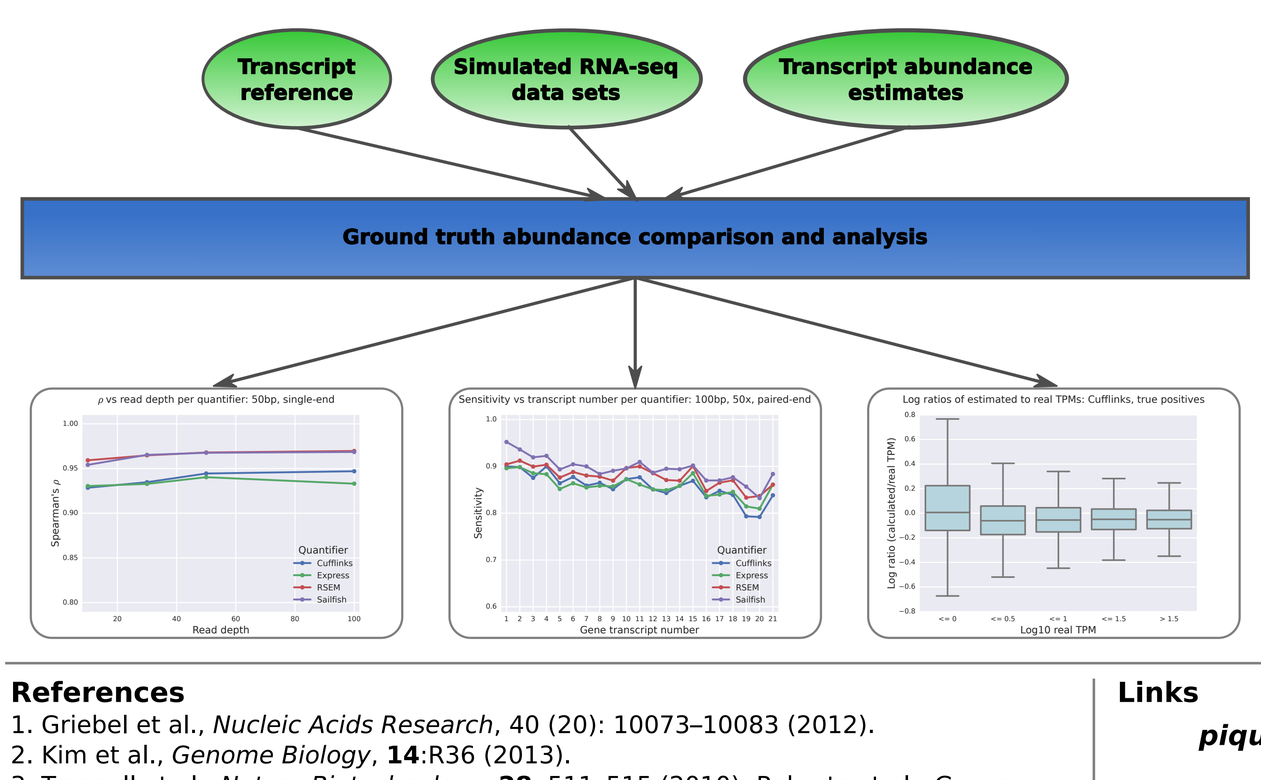
2. Quantification

For each set of RNA-seq data, simulated reads are mapped to the genome or transcriptome, and quantification tools estimate transcript abundances.



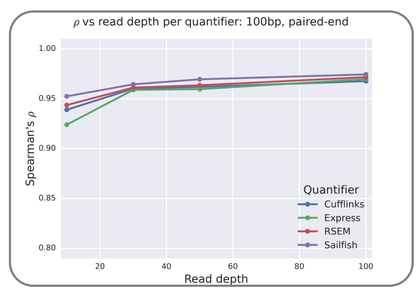
3. Analysis

For each set of RNA-seq data and each quantification tool, estimated transcript abundances are compared with the ground truth values for accuracy, and a rich collection of statistics and graphs are automatically generated.

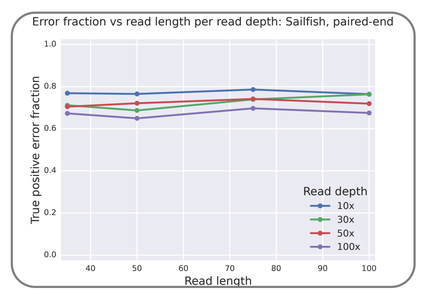


Statistics

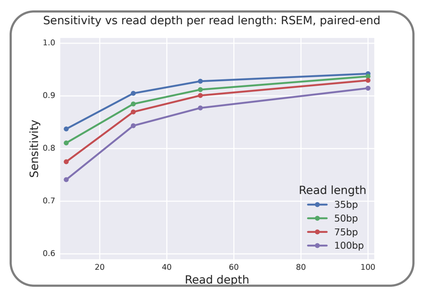
piquant calculates a range of statistics and draws graphs to aid the assessment of transcript quantification performance. Additional statistics, and their resultant graphs, are easily added with just a few lines of code.



Spearman's ρ: the rank correlation coefficient between estimated and ground truth abundances



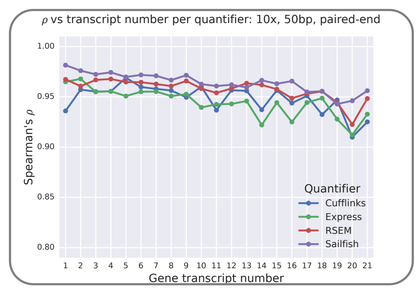
Error fraction: the fraction of transcripts whose estimated abundance is >10% higher or lower than the ground truth value



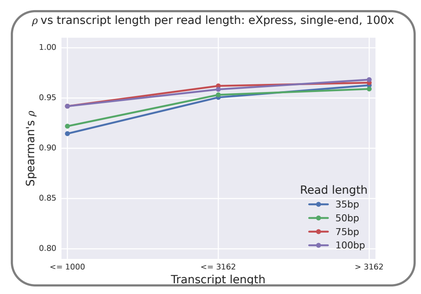
Sensitivity and specificity: the proportion of transcripts correctly identified as being present or not

Classification of transcripts

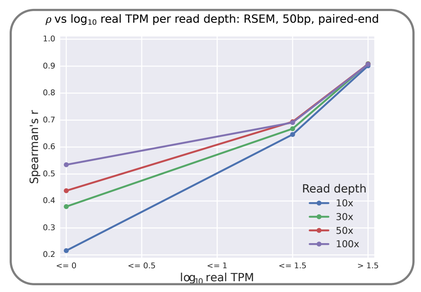
Classifiers split the whole set of input transcripts into discrete groups sharing similar properties. Such divisions allow quantification performance to be assessed across different types of transcript. Again, additional classifiers are added with a few lines of code.



Isoforms per gene: Quantification accuracy decreases as the number of transcripts per gene grows



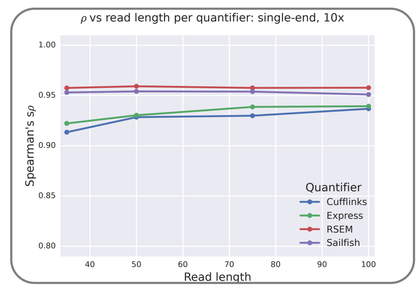
Transcript length: Quantification accuracy is better for long transcripts than for shorter isoforms



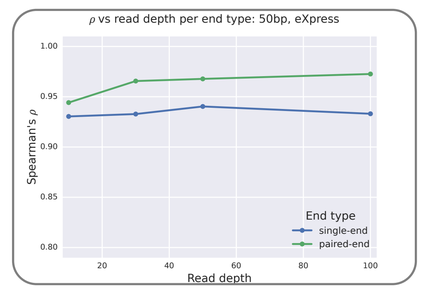
Transcript prevalence: Quantification accuracy is better for isoforms with high abundance than rare, lowly-expressed transcripts

Quantification tools

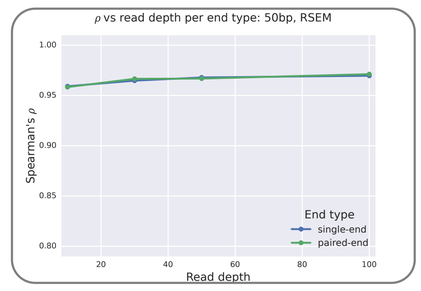
Out of the box, *piquant* runs four mapping and quantification pipelines: TopHat+Cufflinks, Bowtie+RSEM, Bowtie+eXpress, and the alignment-free Sailfish algorithm. Additional pipelines are easy to add, or the current configurations edited to test alternate optional parameter choices.



***piquant* allows the relative strengths of quantification pipelines to be assessed as sequencing parameters change**



For Cufflinks and eXpress, paired-end reads improve quantification accuracy over single-end reads



However, for RSEM and Sailfish this benefit is less consistent or not present

Challenges and next steps

An abiding risk when using simulated data is its potential failure to truly capture the characteristics of real experiments: here, the analyses produced by *piquant* will be of limited value if their verdicts about quantification tool performance fail to translate to the real world. An additional problem in the assessment of transcript quantification from RNA-seq is the relative paucity of gold-standard experimental data by which to tests our conclusions. One such data set, produced as part of the MAQC project [8], consists of ~1000 gene abundances measured by TaqMan PCR for two reference RNA samples.

While noting that the relation between quantification accuracy at the level of genes and transcripts is itself not necessarily straightforward, we compared the performance of quantification tools on both simulated and TaqMan-validated RNA-seq data. In consequence, we will next investigate the effect on quantification of the following improvements to the RNA-seq simulation model:

- a biologically-realistic distribution of gene abundances
- reads arising from transcripts not present in the reference
- reads arising from chimeric transcripts

References

1. Griebel et al., *Nucleic Acids Research*, 40 (20): 10073–10083 (2012).
2. Kim et al., *Genome Biology*, **14**:R36 (2013).
3. Trapnell et al., *Nature Biotechnology*, **28**, 511–515 (2010); Roberts et al., *Genome Biology*, **12**:R22 (2011).
4. Langmead et al., *Genome Biology*, **10**:R25 (2009).
5. Li and Dewey, *BMC Bioinformatics*, **12**:323 (2011).
6. Roberts and Pachter, *Nature Methods*, **10**, 71–73 (2013).
7. Patro et al., *Nature Biotechnology*, **32**, 462–464 (2014).
8. Shi et al., *Nature Biotechnology*, **24**, 1151–1161 (2006).

Links

***piquant* code repository**

https://github.com/lweasel/piquant

***piquant* documentation**

http://piquant.readthedocs.org/en/latest/index.html

Neuroregulatory Genomics

http://neuroregulatorygenomics.org