mmquant and mmannot: How to handle multiple-mapping reads in (s)RNA-Seq

Matthias Zytnicki, Christine Gaspin

mmquant and mmannot: How to handle multiple-mapping reads in (s)RNA-Seq

Matthias Zytnicki, Christine Gaspin
INRA MIAT, Toulouse

September 7, 2023

1 Introduction

RNA-Seq and small RNA-Seq (sRNA-Seq) are currently used routinely, and they provide accurate information on gene and small-RNA transcription. However, the methods cannot accurately estimate duplicated transcript expression. Several strategies have been previously used (drop duplicated transcripts, distribute uniformly the reads, or estimate expression), but all of them provide biased results. We provide here two tools, called mmquant[1] and mmannot[2], for computing expression of genes and small RNAs respectively, including duplicated elements. mmquant is available at `https://bitbucket.org/mzytnicki/multi-mapping-counter`, and mmannot is available at `https://sourcesup.renater.fr/wiki/mmannot`.

2 mmquant: a strategy for the gene quantification including multi-mapping reads

In general, RNA-Seq quantification reads a set of reads files (one file per sample) and a annotation file that lists the set of known genes. It produces a count table, where a row is a gene, a column is a sample, and each cell provides the number of reads matching a gene in a sample. The aim is usually to find differentially expressed genes, i.e. the set of genes that are more, or less, expressed in a subset of the samples when compared to the other samples.

So far, three strategies are used when a read may map at several positions:

- a "unique" method: discard multi-mapping reads,
- a "random" method: use a random hit,

[1] mmquant has been published [Zyt17].
[2] mmannot has been submitted for publication.

- a "ratio" method: weight each hit (if a read maps n times, each hit counts for $1/n$).

mmquant implements an other strategy, firstly presented in [RW15]. If a read maps at different positions, mmquant detects that the corresponding genes are duplicated; it merges the genes and creates a "merged gene" feature, which appears as a new line in the count table. As a result, the differentially expression test can be performed similarly on the regular genes and on the merged genes. mmquant is a drop-in replacement of the widely used tools htseq-count [APH15] and featureCounts [LSS14] that handle multi-mapping reads in an unbiased way. The tool supports paired-end reads, and checks that both ends may match the same transcript, in a way that is consistent with the sequencing strategy (forward–reverse, reverse–forward, etc.). The fragments (i.e. the pairs of reads) are then counted for quantification.

We tested our method on several data sets on different species, but for space reasons we focus on the human data set, taken from [ABJ$^+$14]. Briefly, this study uses RNA-Seq of human brain to find genes that are differentially expressed in individuals diagnosed with bipolar disorder. Admittedly, this dataset is challenging because duplicated genes are known to play a major role in human brain.

Strikingly, the p-values obtained with the three different quantification strategies show a great variability. htseq-count, featureCounts and mmquant (excluding merged genes) gave 734, 835 and 763 differentially expressed genes respectively. Most of the differences comes from the way reads are assigned to the genes.

mmquant found that 5–6% of the reads where multi-mapped and could be attributed to several genes. As a consequence, it found 254 additional differentially expressed merged genes, involving 516 new genes. Note that one fourth of the differentially expressed genes is merged.

We then considered the 33 merged genes with adjusted p-value $< 1\%$, which represented very good candidates. These merged genes included 75 genes that were not detected otherwise (neither by htseq-count nor featureCounts, nor in the non-merged genes found by mmquant). This gene list includes new excellent candidates with putative links to bipolar disorder, including ADK, GTF2I, hnRNP-A1, HTRA2, PKD1 and RERE, which have been linked to various brain-related diseases. Some of these genes have complex regulation systems in cis: ADK and HTRA2 contain overlapping processed pseudogenes and antisense transcripts or genes, and mmquant merges these annotations on the fly. Other genes, like GTF2I, hnRNP-A1, PKD1, and RERE, are duplicated genes, or have produced a pseudo-gene in another locus. It is out of the scope of this study to validate these genes, but we would like to emphasize that, because these genes are duplicated, or overlap with other genes, they have been removed from the standard analysis.

Concerning time, featureCounts is the fastest tool, taking 8–11min per sample; mmquant is second with 21–29 min (+1–3 min if the reads are not sorted); htseq-count, written in Python, takes 4h15min–5h29min. mmquant is slower

than featureCounts because it has to store (and look up) all the reads that have been mapped several times. We obtained this results allocating one thread per BAM file, but featureCounts can be further accelerated by allocating more than one thread per input file, whereas mmquant and htseq-count cannot.

3 mmannot: a strategy to quantify repetitive small non coding RNAs

Small non coding RNAs gather a very wide collection of classes, such as microRNAs, tRNA-derived fragments, small nucleolar RNAs and small nuclear RNAs, to name a few. As usual in RNA-seq studies, the sequencing step is followed by a feature quantification step: when a genome is available, the reads are aligned to the genome, and the corresponding features are quantified.

The sRNA classes are then quantified by counting the number of reads co-localizing with the members of each class. Although simple and widely used, this strategy does not work in several ambiguous cases.

1. A read maps at several loci: if two different regions of the genome are identical (usually after a genome duplication), a read may map equally well at different locations.

2. Frequently in sRNA-Seq, two different annotations overlap in the genome and a hit (i.e. a read mapping) overlaps both: In this case the hit may be attributed to either annotation

3. A hit co-localize two different annotations, even though the annotations do not overlap themselves: The hit is usually at the frontier of the annotations.

The first source of ambiguity arise often, because some sRNAs are known to co-localize with duplicated regions (such as piRNAs or siRNAs), or to be included into duplicated genes (miRNAs and tRFs).

mmannot implements a strategy similar to mmquant, that compares all the reads that map at several positions, and their annotations when available. In many cases, all the hits co-localize with the same feature annotation (a duplicated miRNA or a duplicated gene, for instance). When different annotations exist for a given read, we propose to merge existing features and provide the counts for the merged features.

A configuration file is required to select the annotations that should be quantified. The configuration file ranks the annotation by order of priority, but ties are accepted. Using the exon annotation usually provided in the annotation file, mmannot automatically extracts introns, coding sequences, 5' and 3' untranslated regions (UTRs), down- and upstream regions of the features selected by the users (e.g. coding genes, non-coding genes), and adds them in the in-memory annotation dataset. The user can also specify a strand orientation of the read with respect to the annotation (collinear or antisense).

The process of read annotation proceeds in two steps. The first step aims at finding the matching annotations of a given hit. If a hit matches several different annotations (e.g. miRNA and intron), the annotation with highest priority is kept (here, the miRNA). If several annotations have the same priority, then all of them are kept and the hit is already ambiguous.

The last step resolves the ambiguities. If a read maps uniquely, with no ambiguity, the count of the corresponding annotation is incremented. If a read maps at different locations, but all the hits match the same annotation, we declare that the read is rescued and the corresponding annotation count is also incremented. Likewise, if a read maps only one annotation and intergenic regions, we consider that the read belongs to this annotation, and the read is rescued. If the read or a hit overlaps several annotations, the annotations with highest priority are kept. If there is only one annotation with highest priority, the read or the hit is not ambiguous. Otherwise, there is an ambiguity: we create a new annotation type, called a merged annotation, which is the concatenation of the matching annotations, and its count is incremented. For instance, if a read maps to a 3'UTR and a miRNA, the count of the 3'UTR–miRNA will be incremented. After having scanned all the reads, the quantification table is produced.

We compared the results of the three strategies mentioned previously with the strategy implemented in mmannot. We used datasets of experiments already published, covering several eukaryotic organisms, but we will focus on an *Arabidopsis thaliana* data set [VMK+13]. Our aim was to show how each strategy impacts the results of quantification in each class. We found that, the "unique" method, arguably the most used one, provides very biased results in terms of representative percentage of the class due the strategy used. This strategy annotates around 40% of reads as miRNAs, whereas the other strategies, when considering multi-mapping reads, annotate only around 20% of the reads in the miRNA class. For all datasets, using multi-mapping strategies increases considerably the percentage of annotated reads, showing that the repertoire of expressed regions is largely associated to repeated regions in all genomes. Some multi-mapping reads may have hits that do not co-localize with any annotation. These reads may be unannotated by the "random" strategy, and the associated weight in the "ratio" strategy is lost. As a consequence, some of these reads are not quantified, resulting in less annotated reads.

We focused on ambiguous reads to analyze the origin of reads annotated as ambiguous in that organism. Most of them involve downstream regions, upstream regions, or introns, and they are probably intergenic duplicated regions. Then, the most frequent ambiguous annotations involve a transcribed region within the downstream, upstream, or intron regions. These elements might be produced by some unannotated regions that target genes, e.g. siRNAs, or belong to a non-functional genomic duplication. Interestingly, the most frequent ambiguous annotation involving two transcribed regions is miRNA–gene (−). In plants, a miRNA and its target may be 100% identical, and thus pose a real problem to the annotation. We studied the pairs of miRNAs–genes that were involved in this class, with at least 100 reads supporting this association. We

found well-known miRNAs and their targets at mapped loci: miR156/miR157 with SPL, miR163 with PXMT1, miR171 with ATHAM, miR400 with PPR1, miR403 with Ago2 and miR824 with AGL. Note that these reads cannot be correctly annotated by any other method, and the expression of the miRNAs are thus under-estimated.

4 Conclusion

Transcript quantification is an essential step of many RNA-Seq analyses. Yet, the assumption used by the quantification tools is not always fully understood, especially concerning multi-mapping reads. With mmquant and mmannot, we provide simple tools, that include these reads in the quantification step, with no assumption on the read distribution. We hope that that these tools could be used as a drop-in replacement of previous tools, and that part of the genomic "dark matter" will be at last explored.

References

[ABJ+14] N Akula, J Barb, X Jiang, J Wendland, KH Choi, SK Sen, L Hou, DTW Chen, G Laje, K Johnson, BK Lipska, JE Kleinman, H Corrada-Bravo, S Detera-Wadleigh, PJ Munson, and FJ McMahon. Rna-sequencing of the brain transcriptome implicates dysregulation of neuroplasticity, circadian rhythms and gtpase binding in bipolar disorder. *Molecular psychiatry*, 19:1179–1185, 2014.

[APH15] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. Htseq–a python framework to work with high-throughput sequencing data. *Bioinformatics*, 15:166–169, 2015.

[LSS14] Yang Liao, Gordon Keith Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30:923–930, 2014.

[RW15] Christelle Robert and Mick Watson. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biology*, 16:177, 2015.

[VMK+13] Elena A Vidal, Tomás C Moyano, Gabriel Krouk, Manpreet S Katari, Milos Tanurdzic, W Richard McCombie, Gloria M Coruzzi, and Rodrigo A Gutiérrez. Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in *Arabidopsis thaliana* roots. *BMC Genomics*, 14:701, 2013.

[Zyt17] Matthias Zytnicki. mmquant: how to count multi-mapping reads? *BMC Bioinformatics*, 18:411, 2017.