A strategy to reduce technical variability and bias in RNA sequencing data

Francesca Finotello[1]✉, Enrico Lavezzo[2], Luisa Barzon[2], Paolo Mazzon[1], Paolo Fontana[3], Stefano Toppo[2], Barbara Di Camillo[1]

[1]Department of Information Engineering, University of Padova, Padova, Italy
[2]Department of Molecular Medicine, University of Padova, Padova, Italy
[3]Edmund Mach Foundation, San Michele all'Adige, Trento, Italy

Motivation and Objectives

In the last decade, Next-Generation Sequencing (NGS) technologies have been extensively applied to quantitative transcriptomics, making RNA sequencing (RNA-seq) a valuable alternative to microarrays for measuring and comparing gene transcription levels (Wang et al., 2009). In this framework, the millions of sequences obtained through NGS are aligned to a reference genome or transcriptome, and *counts*, i.e. the number of reads aligned to each gene, give a digital measure of gene expression. Given that longer genes are more likely to be sequenced than shorter ones, gene counts depend not only on the true gene expression, but also on its sequence length. Several approaches have been explored to reduce length bias a posteriori, namely after that read counts have been computed (Mortazavi et al., 2008; Bullard et al., 2010; Hansen et al. 2012; Risso et al., 2011), or to provide a direct and unbiased estimate of transcript abundances (Trapnell, 2010). In addition, counts are biased toward highly transcribed genes, so most of the reads sequenced in a sample arise from a restricted subset of highly expressed genes (Robinson and Oshlack, 2009).

The present work is aimed at assessing technical variability and biases of RNA-seq counts, and exploring an alternative measure of exon expression, which is less biased toward long or highly expressed genes, thus requiring no length normalization, and characterized by a lower technical variability.

Methods

We consider two different experiments (Bullard et al., 2010; Griffith et al., 2010) with multiple technical replicates. Raw reads were aligned to the reference genomes using TopHat v1.2.0 (Langmead et al., 2009) and summarized on Ensembl exons using bedtools 2.15.0 (Quinlan and Hall, 2010) to compute read counts. We consider exon counts rather than transcript counts to avoid introducing biases when dealing with alternatively spliced exons. We computed counts as the total number of reads that align to an exon (referred as *totcounts* in the following). As an alternative approach, we exploited the per-base read coverage to obtain counts for every position along each exon sequence. The measure of gene expression assigned to an exon, called *maxcounts* from here on, was then calculated as the maximum of its per-base counts. Both *totcounts* and *maxcounts* were normalized with the Trimmed Mean of M-values approach (TMM, Robinson and Oshlack, 2009) to correct differences in sequencing depth across libraries. In addition, we computed *Reads Per Kilobase of exon model per Million mapped reads* (RPKM, Mortazavi et al, 2008), calculated by dividing *totcounts*, not normalized via TMM, by the total number of reads mapped in each library, in millions, and by exon length, in kilobases.

Results and Discussion

To investigate the bias due to highly expressed exons, we computed cumulative counts for all replicates in MAQC-2 and Griffith's data sets. In MAQC-2 data (results not shown), when considering *totcounts*, about 3-5% of exons account for 50% of total exon counts and 27-32% of exons account for 90% of total exon counts, showing that a great fraction of counts belong to a restricted subset of exons. Differently, *maxcounts* are more evenly distributed across exons: 7-8% of exons account for 50% of total counts and 44-45% of exons account for 90% of total counts. RPKM distribution lies in between that of *maxcounts* and *totcounts*, with 5-7% of exons accounting for 50% of total RPKMs and 36-38% of exons accounting for 90% of total RPKMS. Also with Griffith's data (Figure 1A), *maxcounts* have the less steep cumulative distribution curves.

We also investigated length bias at single-exon level using smoothed scatter plots of counts/RPKMs versus exon-length, in log-log scale (see

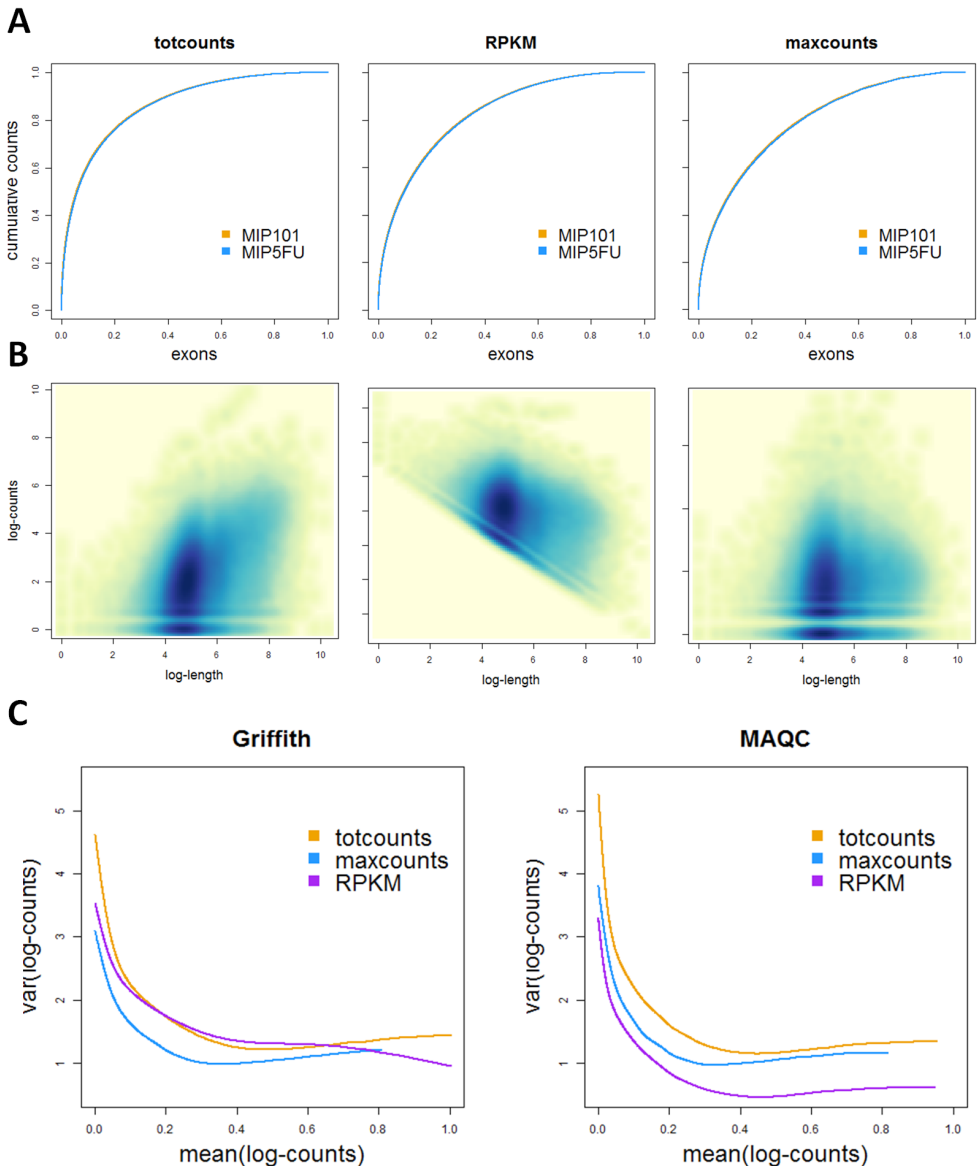

Figure 1: Diagnostic plots of *totcounts*, RPKMs and *maxcounts*: (A) distribution of exon counts/RPKMs in Griffith's data; (B) smoothed scatter plots showing dependence of counts/RPKMs over exon length for one Griffith's library; (C) variance of counts/RPKMs across technical replicates.

Figure 1B for results on Griffith's data). These plots show an increasing pattern of *totcounts* in dependence of exon-length, meaning that longer exons tend to have higher counts than shorter ones (Pearson's correlation r=0.38 for MAQC-2 and r=0.43 for Griffith). On the contrary, *maxcounts* are not correlated with exon-length (Pearson's correlation r=0.10 for MAQC-2 and r=0.01 for Griffith). RPKMs do not show the increasing pattern of *totcounts*, and are in fact characterized by negative correlation with exon length (Pearson's correlation r=-0.28 for MAQC-2 and r=-0.29 for Griffith), meaning that dividing by exon length over-corrects length bias in shorter exons. Plots are reported for one library of Griffith's data set, but the same patterns are confirmed across all libraries of the two data sets (results not shown).

Finally, we assessed variance of *totcounts*, RPKMs and *maxcounts* across technical replicates, using a cubic-spline fit of the variance versus the mean of log-counts/log-RPKMs (Figure 1C): in both data sets maxcounts have a lower variance with respect to *totcounts*. Anyway, on

MAQC-2, RPKMs provide the lowest technical variance.

In summary, we confirm that *totcounts* strongly depends on the length of the feature they are summarized on, even when considering exons in place of genes. Using RPKMs, that normalize *totcounts* by exon length and sequencing depth, reduces technical variability but does not completely remove exon length bias. We propose an alternative measure of exon expression, *maxcounts*, which is less biased toward long or highly expressed genes than *totcounts* and RPKMs, and whose technical variability is lower than or comparable to that of *totcounts* and RPKMs, respectively.

We are now working on a refinement of this measure, to make it more robust to sequencing and mapping biases. In addition, we are assessing the accuracy and precision of totcounts and maxcounts in assessing the real RNA abundances using publicly available data sets for which spike-in RNAs measures are available. Future studies will focus on the definition of transcriptional models that could be used to aggregate maxcounts at gene or transcript level.

Acknowledgements

This research is supported by PRAT 2010 CPDA101217, "Models of RNA sequencing data variability for quantitative transcriptomics", and AACSE Project, "Algorithms and Architectures for Computational Science and Engineering".

References

1. Bullard JH, Purdom E, Hansen KD, et al. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. BMC Bioinformatics 11, 94. doi: 10.1186/1471-2105-11-94
2. Griffith M, Griffith OL, Mwenifumbo J, et al. Alternative expression analysis by RNA sequencing. Nat Methods 7, 843. doi: 10.1038/nmeth.1503
3. Hansen KD, Irizarry RA, Wu Z (2012) Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics 13, 204. doi: 10.1093/biostatistics/kxr054
4. Langmead B, Trapnell C, Pop M, et al. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25. doi: 10.1186/gb-2009-10-3-r25. doi: 10.1186/gb-2009-10-3-r25
5. Mortazavi A, Williams BA, McCue K, et al. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat methods 5, 621. doi: 10.1038/nmeth.1226
6. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841. doi: 10.1093/bioinformatics/btq033
7. Risso D, Schwartz K, Sherlock G, et al. (2011) GC-Content Normalization for RNA-Seq Data. BMC Bioinformatics 12, 480. doi: 10.1186/1471-2105-12-480
8. Robinson MD and Oshlack A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol, 11, R25.
9. Trapnell C, Williams BA, Pertea G, et al. (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511. doi: 10.1038/nbt.1621
10. Wang Z, Gerstein M, Snyder M. (2009) RNA-seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 10, 57. doi:10.1038/nrg2484