Theory Biosci. (2012) 131:281–285
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples

Günter P. Wagner · Koryu Kin · Vincent J. Lynch

Received: 4 June 2012 / Accepted: 24 July 2012 / Published online: 8 August 2012

Abstract Measures of RNA abundance are important for many areas of biology and often obtained from high-throughput RNA sequencing methods such as Illumina sequence data. These measures need to be normalized to remove technical biases inherent in the sequencing approach, most notably the length of the RNA species and the sequencing depth of a sample. These biases are corrected in the widely used reads per kilobase per million reads (RPKM) measure. Here, we argue that the intended meaning of RPKM is a measure of relative molar RNA concentration (rmc) and show that for each set of transcripts the average rmc is a constant, namely the inverse of the number of transcripts mapped. Further, we show that RPKM does not respect this invariance property and thus cannot be an accurate measure of rmc. We propose a slight modification of RPKM that eliminates this inconsistency and call it TPM for transcripts per million. TPM respects the average invariance and eliminates statistical biases inherent in the RPKM measure.

Keywords RNA quantification · NextGen sequencing · RPKM

Electronic supplementary material The online version of this article (doi:10.1007/s12064-012-0162-3) contains supplementary material, which is available to authorized users.

G. P. Wagner (✉) · K. Kin · V. J. Lynch
Department of Ecology and Evolutionary Biology,
Yale Systems Biology Institute, Yale University,
300 Heffernan Drive, West Haven, CT 06516, USA
e-mail: gunter.wagner@yale.edu

Introduction

Measuring and comparing transcript abundance are critical for the study of gene regulation, assessing the effect of experimental treatments on gene expression, and the evolution of gene regulation. Ideally, measurements of gene expression would directly estimate the concentration of different RNA species at the site of their biological function, i.e., their availability at ribosomes or other locations in the cell. Measuring mRNA abundance at their site of biological function would require not only measurement of mRNA amounts but also cell number, cell volume and sub-cellular localization. In most cases, we have access to various measures of RNA abundance but little or no information about the other variables. In the absence of information on cell number, volume and sub-cellular localization, the most we can hope for is a consistent measurement of the relative molar concentration (rmc) of each mRNA species. The rmc of a gene g is

$$\mathrm{rmc}_g = \frac{[\mathrm{mRNA}_g]}{[\mathrm{mRNA}_{\mathrm{total}}]}$$

$$[\mathrm{mRNA}_{\mathrm{total}}] = \sum_{g \in G} [\mathrm{mRNA}_g]$$

where G stands for the set of all genes determined in that experiment and g is an index for a gene. Since this measure is a ratio of two concentrations the denominator of the concentration measure cancels, and information about cell number or cell volume becomes irrelevant.

All commonly used techniques to measure mRNA abundance, including qPCR, microarray signals, as well as reads per kilobase per million reads (RPKM) for RNAseq data (Mortazavi et al. 2008), aim at estimating a statistic that is as closely proportional to the relative molar concentration as possible. Here, we discuss estimating rmc from mRNA-seq data.

An invariance property of rmc measures

A useful property of rmc is that, within each sample, the average of rmc across genes, $\langle \mathrm{rmc} \rangle_G$, is a constant that only depends on the number transcript types determined, i.e., the number of genes mapped in a mRNA-seq data set. This can easily be seen from the definition of average rmc

$$\langle \mathrm{rmc} \rangle_G = \frac{1}{\| G \|} \sum_{g \in G} \mathrm{rmc}_g = \frac{1}{\| G \|} \sum_{g \in G} \frac{[\mathrm{mRNA}_g]}{[\mathrm{mRNA}_{\mathrm{total}}]} = \frac{1}{\| G \|} \frac{\sum_{g \in G} [\mathrm{mRNA}_g]}{[\mathrm{mRNA}_{\mathrm{total}}]} = \frac{1}{\| G \|}$$

or if we set $\| G \| = N$, we obtain

$$\langle \mathrm{rmc} \rangle_G = \frac{1}{N}$$

This means that the average rmc for each and every sample of RNA-seq data mapped to the same genome is the same constant value.

The mathematical constraint on rmc shown above implies a similar constraint for any statistic S that is meant to estimate a value proportional to rmc. Specifically, if we assume that S is proportional to rmc, i.e., there exists a positive non-zero number k such that $S = k$*rmc, then it is easy to see that the average statistic has to be

$$\langle S \rangle_G = \frac{k}{N}$$

for each and every sample mapped to the same genome. We can use the equation above to test candidate rmc measures for their consistency across samples. Only a statistic that meets that criterion can be a legitimate estimate of rmc.

RPKM as a measure of rmc

The most frequently used measure of mRNA abundance based on RNA-seq data is RPKM. It is calculated from the number of reads mapped to a particular gene region g, r_g, and the feature length, fl_g, which is the number of nucleotide in a mapable region of a gene (Mortazavi et al. 2008). It is calculated as

$$\mathrm{RPKM}_g = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

where R is the total number of reads from the sequencing run of that sample, $R = \sum_{g \in G} r_g$. RPKM accommodates two biases that the number of reads mapped for each gene r_g introduces compared to the actual transcript abundance. At the one hand, differences in the feature length lead to different expected read counts from Illumina sequencing runs, even for genes with the same transcript abundance. One expects more reads to be produced from longer transcripts because the transcripts are transcribed to cDNA and then broken into smaller pieces accessible to sequencing. This normalization is achieved by dividing the number of reads by the feature length and multiplying them with 1,000:

$$\frac{r_g}{\mathrm{fl}_g} 10^3$$

Another factor that influences the number of reads obtained for each gene is the sequencing depth, i.e., the total number of reads obtained in one sequencing run. Even when the transcript abundance is the same in two samples for the same gene one expects more reads from the sample that has been sequenced to greater depth. To accommodate this bias, the RPKM measure normalizes by the total number of reads R, divided by 10^6, to obtain reads for each gene per million reads,

$$\frac{R}{10^6}$$

leading to the canonical formula for RPKM

$$\mathrm{RPKM}_g = \frac{\frac{r_g 10^3}{\mathrm{fl}_g}}{\frac{R}{10^6}} = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

Table 1 summarizes the average RPKM values from a number of RNA-seq data for cultured human cell lines from our lab. As can be seen from this data, the average RPKM is similar among technical replicates from the same sample but substantially different among samples (F ratio = 119.61; $df = 4$ and 5, $p = 3.77 \times 10^{-5}$). Note that these samples have all been mapped to the same version of the human genome (hg18, canonical) and thus the differences cannot be explained by different gene sets used for the mapping of the sequence reads. The variation of average RPKM among samples raises doubt about the appropriateness of RPKM as a measure of relative molar RNA concentration.

The reason for the inconsistency of RPKM across samples arises from the normalization by the total number of reads. While rmc as well as qPCR results are ratios of transcript concentrations, the RPKM normalizes a proxy for transcript number by $r_g \times 10^3/\mathrm{fl}_g$ the number of sequencing reads in millions, $R/10^6$. The latter, however, is not a measure of total transcript number. The relationship between R and the total number of transcripts sampled depends on the size distribution of RNA transcripts, which can differ between samples. In a sample with, on average, longer transcripts the same number of reads represents fewer transcripts.

Transcripts per million (TPM): an alternative to RPKM

Here a slightly modified measure of transcript abundance is introduced, the TPM. TPM is calculated as

Table 1 Comparison of average TPM and RPKM among different cells types and samples (see supplementary material and Wang et al. 2011)

Species	Tissue/cell type	Replicate	AvTPM	AvRPKM	Scaling f
Human	Differentiated decidual cells	1	46.518	15.94	2.92
		2	46.518	16.13	2.83
Human	Un-differentiated dec. cells	1	46.518	15.27	3.05
		2	46.518	15.22	3.06
Human	Myofibroblast cells	1	46.518	17.66	2.62
		2	46.518	17.65	2.62
Human	Chondrocyte cells	1	46.518	16.57	2.81
		2	46.518	16.57	2.81
Human	Myometrial cells	1	46.518	17.77	2.62
		2	46.518	17.79	2.61
Chicken	Forelimb digit 1 stage 28–29	–	65.527	28.35	2.31
Chicken	Forelimb digit 1 stage 31	–	65.527	28.56	2.29

If a measure of RNA abundance is proportional to rmc, then their average should be the same for all samples since all these samples are from the same species, human, and the RNA sequence reads were mapped to the same genome. Note that the average TPM is in fact identical among the human samples but different from the chicken sample as expected. In contrast, the average RPKM varies among samples even for the same genome, i.e., the different human cell types. The ANOVA for AvRPKM among human samples is highly significant with an $F(4,5) = 119.6$ and $p = 3.78 \times 10^{-5}$. The degree of difference can be seen in the scaling factors f, which is the factor that converts corresponding RPKM into TPM values. Note that within a sample the RPKM and the TPM are proportional, and the scaling factor f is the coefficient of proportionality. The higher average TPM and RPKM values from the chicken samples are due to the differences in genome annotation in the chicken versus the human genome. The lower number of annotated genes in the chicken genome leads to higher average TPM/RPKM

$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T}$$

$$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

where rl is the read length, i.e., the average number of nucleotides mapped per read. The rationale for this calculations is the following. The value

$$\frac{r_g \times \mathrm{rl}}{\mathrm{f}l_g}$$

is a proxy for the number of transcript samples by r_g sequencing reads. This is a simplification, which will be discussed separately below. This value is the number of mapped read divided by the length of the transcript. T is the total number of transcripts sampled in a sequencing run. It is easy to see that TPM fulfills the invariant average criterion:

$$\langle \mathrm{TPM} \rangle_g = \frac{10^6}{N}$$

and thus can be proportional to the average rmc. Table 1 shows that the average TPM is in fact invariant among samples, as is necessary for mathematical reasons.

Within one sample TPM and RPKM are proportional. This is seen in Fig. 1 for RNA from human chondrocyte. The proportionality between TPM and RPKM for a given sample can also be deduced from the equations defining RPKM and TPM:

$$\mathrm{RPKM}_g = \frac{T \times 10^3}{R \times \mathrm{rl}} \times \mathrm{TPM}_g$$

Note that the proportionality coefficient only contains values constant among genes for the same sample, T, R, rl. However, the scaling factor between RPKM and TPM differs between samples (Table 1) (see supplemental material for details about the data acquisition and for the chicken data see (Wang et al. 2011)). Among our collection of five different sample types (cell types or different stages of differentiation) the scaling factor varies between 2.6 and 3, and thus the difference of between samples can be as

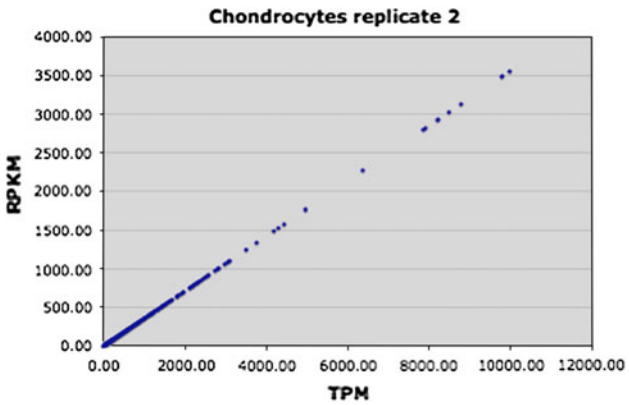

Fig. 1 Relationship between RPKM and TPM in data from RNA abundance in cultured human chondrocytes (ATCC, Cat. No. CRL-2847). RPKM and TPM are proportional to each other within a given sample, but see Table 1 for variation between samples

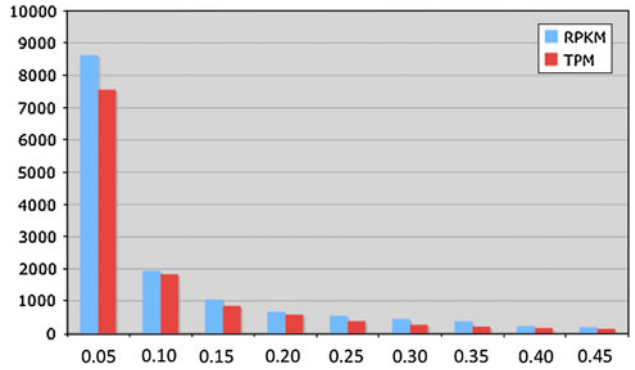

Fig. 2 p value distribution of RNA abundance data from human chondrocytes and myometrial cells for data expressed in RPKM and TPM. The p values were calculating from two-tailed t test assuming different variances. The p values are binned in 0.05 bins. Note that t tests using RPKM lead to higher number of low p values as expected if RPKM introduces artifactual differences in RNA abundance measures between samples

much as much as 14 %. This means that a gene which may have the same gene expression level in terms of rmc could differ by as much as 14 % in terms of RPKM and thus could suggest differences where there are none. Or the gene expression level in terms of RPKM could suggest no difference, even though there are differences in transcript abundance.

The above described differences among samples in the proportionality of TPM and RPKM suggest that between sample comparisons would lead, on average, to inflated statistical significance or lower p values than is justified by differences of transcript abundance. In Fig. 2 the p values for a comparison of RNA abundance in two cell types are plotted. For each gene, the p value was estimated from a t test, using the two replicates for estimating variance. As can be seen, the number of low p values is higher if the comparison is done in terms of RPKM compared to when TPM values are used. This is expected if RPKM causes artifactual differences in the mean as suggested by the differences in the proportionality between TPM and RPKM between different samples.

Remark about estimating alternative transcript abundance

Both, TPM as well as RPKM, rely on feature length to correct read numbers for differences in transcript size. Often, feature length is estimated as the total length of the exonic region. However, there is a well-recognized problem with this approach, because cell types can differ in the splicing variant of the transcript they express (Stamm et al. 2005). There are various approaches to solve that problem (Jiang and Wong 2009; Wang et al. 2009; Ozsolak and Milos 2011). These either require pre-processing of the RNA sample to focus only on sequences from the transcriptional start site, or prior knowledge of all possible splice variants. At this time, knowledge of splice variants can only be approximate, since there is no guarantee that all possible splice variants have been described, in particular for a specific cell or tissue type. Here, we suggest an alternative approach that relies on post hoc validation rather than a priori transcript modeling.

One could proceed by mapping genes and quantifying RNA transcript abundance in one of the acceptable ways, say with TPM and test for differences between samples or treatments. If a significant difference is found, the difference can have two reasons. At the one hand, the actual transcript abundance can be different between the samples, or, on the other hand, the same level of transcript abundance for the gene exists, but the two samples contain different splice variants of substantially different length. The sample with the smaller transcript is expected to lead to lower estimates of the transcript abundance.

To validate a suspected difference in RNA abundance based on TPM one can test the inference by comparing the read coverage of the gene in the two samples. If the TPM difference between samples is, at least partially, due to the expression of different splice variants, one expects that the sequencing coverage of the coding region differs between the samples. If there is no significant difference in sequence coverage it is likely that the detected TPM difference in fact reflects a difference in RNA transcript expression. If there are differences in read overage, special procedures need to be developed to assess how much of the difference is due to differential splicing. Exactly how read distribution across a transcript is to be compared between samples needs to be worked out.

Conclusions

Here, we argue that the intended meaning of many RNA abundance measures, in particular RPKM, is to measure the relative molar concentration of a RNA species. We presented evidence that RPKM is an inconsistent measure of relative molar concentration and suggests a closely related alternative, TPM, which is not biased in the way RPKM is. We show that the RPKM measures can differ substantially between samples and thus has the potential to cause inflated statistical significance values. At a conceptual level the problem with RPKM can be traced back to an issue of meaningfulness (Narens 2002; Houle et al. 2011). In measurement theory, the notion of meaningfulness is the question what the physical or biological interpretation of the numerical measure is. In the case of RPKM, the problem originates from the fact that there is no biological

interpretation of the denominator R, the total number of reads. It is a variable that characterizes a particular sequencing run, but does not correspond in any direct way to a biological variable, like the total RNA abundance. In other words, the units of RNA abundance in terms of RPKM differ between samples, i.e. it behaves like a "rubber measuring tape." Hence, the inconsistencies highlighted here can be traced back to insufficient attention to issues of meaningfulness of quantitative measures frequently found in biology (Houle et al. 2011).

References

Houle D, Pelabon C, Wagner GP, Hansen TF (2011) Measurement and meaning in biology. Q Rev Biol 86:3–34

Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-seq. Bioinformatics 25:1026–1032

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5:621–628

Narens L (2002) Theories of meaningfulness. Lawrence Erlbaum Associates, Mahwah

Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12:87–98

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H (2005) Function of alternative splicing. Gene 344:1–20

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wang Z, Young RL, Xue H, Wagner GP (2011) Transcriptomic analysis of avian digits reveals conserved and derived digit identities in birds. Nature 477:583–586