OXFORD UNIVERSITY PRESS | Bioinformatics

Accurate Modeling and Correction of GC Content and Gene Length Bias from RNA-seq Data

Journal:	*Bioinformatics*
Manuscript ID	BIOINF-2016-0539
Category:	Original Paper
Date Submitted by the Author:	06-Apr-2016
Complete List of Authors:	Rehrauer, Hubert; University of Zurich, Functional Genomics Center Zurich Dimitrieva, Slavica; University of Zurich, Functional Genomics Center Zurich Schlapbach, Ralph; Functional Genomics Center Zurich
Keywords:	HiTSeq, Data analysis, Gene expression, Next-generation sequencing

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

Accurate Modeling and Correction of GC Content and Gene Length Bias from RNA-seq Data

Hubert Rehrauer[1,*], Slavica Dimitrieva[1] and Ralph Schlapbach[1]

[1]Functional Genomics Center Zurich, ETH Zurich and University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Several approaches exist that deal with GC content and gene length dependent biases in RNA-seq measurements of transcript abundances. However, when computing correction factors, existing approaches do not consider the potential interdependence and interaction of both effects. Additionally, they do not deal satisfactorily with genes that have partially or entirely zero counts. Thirdly, the biases may affect a large fraction of genes such that the assumed global normalization schemes are invalid.
Results: We present a novel method for library bias correction (LBC) with a 2D function that simultaneously depends on GC content and gene length. Our approach is unique by modeling not the absolute biases but the sample-specific deviations from the data-set wide bias. The computed correction factors are precise even in the case of partial zero counts. The presented method is useful for correcting data sets with subsets of deviating samples as well as for the joined analysis of different data sets generated with different biases.
Availability: The algorithm is implemented as an R script and is available in the supplementary material.
Contact: hubert.rehrauer@fgcz.uzh.ch
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA sequencing (RNA-seq) is a widely used high-throughput technique for investigation of the transcriptional activity of cells by quantifying the abundance levels of transcripts/genes. In the last few years a number of different sequencing platforms and library preparation protocols have been introduced (Li, et al., 2014; Nagalakshmi, et al., 2010). Generally, RNA-seq protocols include a series of technical steps ranging from RNA extraction, fragmentation, adaptor ligation, reverse transcription, PCR amplification, cluster amplification to sequencing of amplified fragments. Each of these technical procedures is susceptible to experimental conditions and can introduce sample specific systematic biases that can affect the representation of the transcripts in the sequencing output. Among them, library amplification by PCR plays a major role in generating biases, as it may lead to an under-representation of GC-poor and GC-rich regions (Aird, et al., 2011). Such technology related artifacts and biases are especially evident in single-cell RNA sequencing experiments due to the limited amounts of starting RNA material (Stegle, et al., 2015). Accurate modeling and correction of such biases is of vital importance for ensuring accurate inference of expression levels and downstream analysis.

The most well-documented biases in RNA-seq experiments are due to the variable GC-content, i.e. the proportion of G and C nucleotides in a region, and gene length. As previous studies have shown (Hansen, et al., 2012; Pickrell, et al., 2010), these two variables can have a strong impact on the quantification of the expression fold change and failure to adjust for these biases can mislead the differential expression analyses. Thus, bias modeling and correction is crucially important in standard RNA-seq experiments, and mandatory in RNA-seq projects where different RNA-seq protocols have been used for RNA library preparation.

In the past few years a number of normalization approaches have been developed to address the major causes of systematic biases in RNA-seq experiments (Filloux, et al., 2014; Finotello, et al., 2014; Hansen, et al., 2012; Leek, 2014; Li, et al., 2014; Love, et al., 2015; Risso, et al., 2014; Risso, et al., 2011). Hansen et al., (2012) have developed a procedure known as *conditional quantile normalization* (*CQN*) that models GC-content and gene length effects as smooth functions using natural cubic splines and estimates the biases using robust quantile regression. This approach models the read counts dependence on the two factors: GC-content and length (and in principle other sources of bias), but does not consider their interaction. Furthermore, the bias estimates are computed from the median of the expression values. However, as the expression levels have a wide range of values, such estimates are not very precise. Zheng et al., (2011) proposed an approach based on generalized additive model to simultaneously correct different sources of biases, such as GC-content, gene length and dinucleotide frequencies. Under this approach, a principal component analysis on GC content and dinucleotide frequencies is performed first, then the resulting principal components and gene length are used as covariates to fit a generalized additive model with smoothing spline on gene expression levels. This strategy is implemented in the R package *RNASeqBias*. Risso et al., (2014) proposed a normalization strategy called *remove unwanted variation* (*RUV*) that adjusts for technical effects by performing factor analysis on sets of control genes (e.g. ERCC spike-ins) or samples (e.g. replicate libraries). This is a general approach for removing technical effects and it does not aim at specifically modeling GC and length bias. Correcting the estimates of expression fold-changes, this approach promises to be valuable for large collaborative projects involving multiple laboratories, technicians and sequencing platforms. However, this general approach is limited in the case where the bias correlates with experimental factors. *svaseq* is another method that targets batch effects or another unwanted variation in RNA-seq data using factor analysis (Leek, 2014). In a recent work, Love et al., (2015) introduced *alpine*, a method for estimation of bias-corrected transcript abundance based on expectation-maximization algorithm. This method addresses the linear combination of factors and it improves the identification of the major isoforms, however it does not consider the interaction between factors.

In this article we propose a new strategy to simultaneously correct for GC-content and gene length bias in RNA-seq data considering their interaction. We model effects as a parameter-free continuous function of the two independent variables GC and gene length. Different from existing approaches we do not aim at modeling the absolute bias, but model and correct only in individual samples the deviation from the common bias in the data set.

2 Methods

3.1 Library Bias Correction Algorithm

Our library bias correction (*LBC*) approach assumes that the RNA-seq sample and library preparation protocols introduce a sample specific bias that depends on the GC content and length of the gene. Specifically the model assumes that the observed expression counts for gene g in sample i are $Y_{g,i}$ with

$$Y_{g,i}|\mu_{g,i} \sim \text{Poisson}(\mu_{g,i})$$

$$\mu_{g,i} = M_i \times \exp\{\theta_{g,i} + \alpha_{g,i} + \beta_g + f_i(\text{GC}(g), \text{LogLength}(g))\}.$$

Where the second equation indicates that the expected counts $\mu_{g,i}$ for g in sample i depend on the sequencing depth M_i , the log expression levels $\theta_{g,i}$, experimental factors of interest $\alpha_{g,i}$, and a gene dependent offset β_g . This gene dependent and sample independent offset represents the bias that is a result of all protocol steps starting from the tissue extraction and that is common to all samples. It includes the common GC effect but may also include other biases like e.g. gene dependent read mappability *etc*. The last function, the bias function f_i finally represents the portion of the bias that may exhibit sample-to-sample differences. Here we assume that it solely depends on the genes through the GC content and the gene length. With gene length, we refer to the length of the transcribed isoform. In practice this may not be available and can be approximated by the length of the primary isoform or by the average of all isoforms (weighted by their relative expression). The bias function is modeled as a continuous function of the GC content and gene length that we approximate with a bilinear function on a 2D grid in the GC-LogLength space of the genes. The relative expression counts within the samples are

$$\rho_{gi} = \log(\mu_{g,i}/M_i) = \theta_{g,i} + \alpha_{g,i} + \beta_g + f_{ig}$$

and if we average across all samples we have

$$\overline{\rho_g} = \overline{\theta_g} + \overline{\alpha_g} + \beta_g$$

where we make use of our definition that implies $\overline{f_g} = 0$ and the fact that β_g is sample independent. The difference of each sample to the average is subsequently

$$\rho_{gi} - \overline{\rho_g} = \theta_{g,i} - \overline{\theta_g} + \alpha_{g,i} - \overline{\alpha_g} + f_{ig}$$

which is the equation that we use to estimate the bias function. We define R bins for the GC content and S bins for the logarithmic length that define the gene sets Γ_{rs} consisting of the genes in the r-th GC bin and the s-th length bin. We now assume that the log-expression levels and the effects of the experimental factors are not correlated with GC content and gene length, i.e.

$$\overline{\sum_{g\in\Gamma_{rs}} \theta_{g,\imath} - \overline{\theta_g}} = 0$$

$$\overline{\sum_{g\in\Gamma_{rs}} \alpha_{g,\imath} - \overline{\alpha_g}} = 0$$

which means the interpolation points for the bias function can be computed as

$$\widehat{f_{\imath,rs}} = 1/|\Gamma_{rs}| \sum_{g\in\Gamma_{rs}} \rho_{g,i} - \overline{\rho_g} \, .$$

The bias function of a sample is defined as the bias relative to the average of all samples in the dataset; it is not a global bias function. We typically observe that a majority of the samples in a dataset has the same library characteristics with respect to GC and length. Only a minority of the samples deviates from that. Therefore, in practice we use the median in the above equation since that leaves the majority of the samples uncorrected and corrects only a minority of the samples.

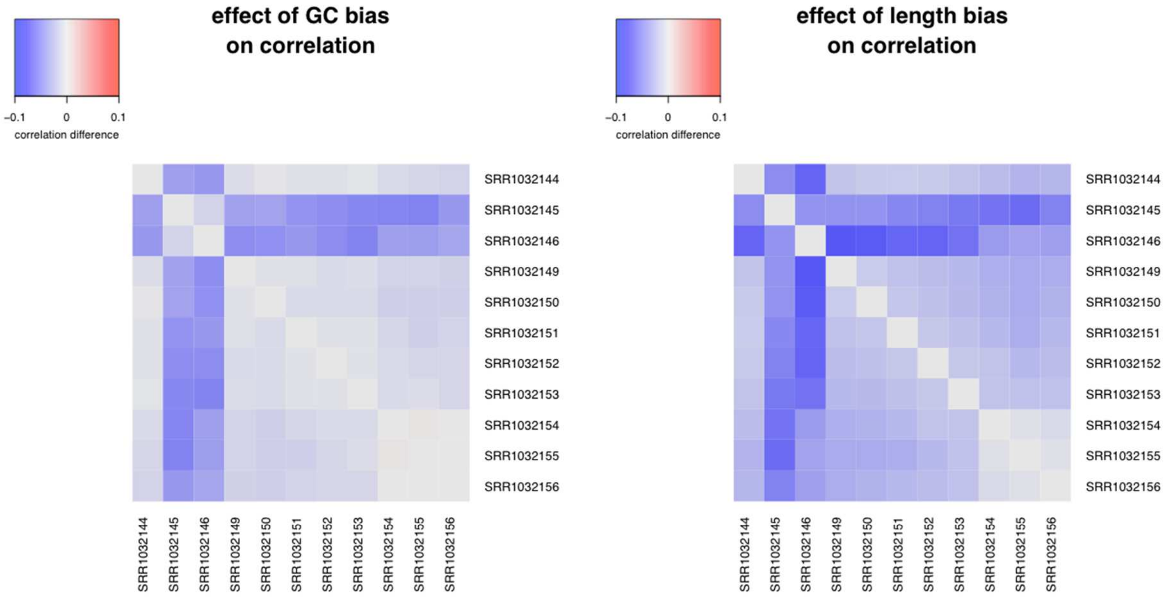

Fig. 1. Effect of GC and length bias on sample-sample correlation. The plot shows the difference in sample-sample correlation when using the genes with extreme (large and small) GC and gene length properties relative to a control gene set with well defined, medium GC content and gene length. In both cases, when looking at genes with extreme GC content (left) and extreme gene length (right), the second and third sample show an effect. This effect is due to the poly-A selection protocol that is used in those samples.

Our implementation returns a matrix of corrected expression counts as well as a matrix of offset terms that can be fed into differential count analysis algorithms, like e.g. *edgeR* (McCarthy, et al., 2012)..

We put special emphasis on correctly handling low or zero counts that cannot be well represented after log transform. Therefore we don't generate additive offsets to expression counts. We do make use of the low expressed genes for estimating the bias function as long as the median expression is finite, *i.e.* a gene must have a non-zero count in at least 50% of the samples. For genes with an initial read count of zero, the corrected count will again be zero.

The *LBC* algorithm is implemented in an *R* script and is available in the supplementary material.

3.2 Datasets Used

To evaluate the *LBC* algorithm we processed in total 390 datasets from the Short Read Archive (SRA). We obtained the data from the Digital Expression Explorer (DEE) (Ziemann M, et al., 2015) that provides expression values from SRA data after uniform preprocessing. Reads have been mapped with the *STAR* RNA-seq aligner (Dobin, et al., 2013) and expression counts were generated with *featureCounts* (Liao, et al., 2014) (for details see http://dee.bakeridi.edu.au/).

3.3 Library Bias Measures

We detect the presence of a sample specific bias that is related to GC-content and gene length by a correlation measure. For pairs of samples, we compute the Spearman correlation coefficient using once the genes at the extreme ends of the GC-range (GC > 65% and GC < 40%) and then compute the difference to the correlation computed using the set of genes that have a GC value well confined around the mode of the GC distribution (45% < GC < 50%) and well defined around the mode of the gene length distribution (1024nt < length < 1448nt). The latter gene set has a very low variation in GC content and gene length and the computed correlation is therefore unaffected from any bias that is related to these gene properties. The same procedure is applied for genes at the extremes of the length distribution.

This correlation-based measure detects biases in their entirety and is independent of the expression level of the change and does not need any between-sample normalization.

3 Results

3.1 Bias Characterization

We demonstrate the *LBC* approach with the expression data in the SRA project *SRP033117*. RNA-seq profiles in this data set have been generated using a poly-A selection protocol (2 samples) and a ribo-depletion protocol (9 samples). When computing and visualizing the bias measures in Figure 1, we see that for two samples, the GC and gene length have an effect on the correlation with the remaining samples. The figure shows also that these two samples have a different gene length but the same GC bias.

While other approaches like *CQN* do consider GC and gene length bias as independent additive effects without interaction, we do model these effects as a true 2D function. Figure 2 shows in the left plot the estimated interpolation points of *LBC*'s bilinear bias function. The right plot shows the position of all genes in the GC-gene length space and with the color code the applied correction. The 2D-heatmap of correction functions shows that this sample exhibits a GC and length bias with an interaction and justifies the use of a true two-dimensional correction function.

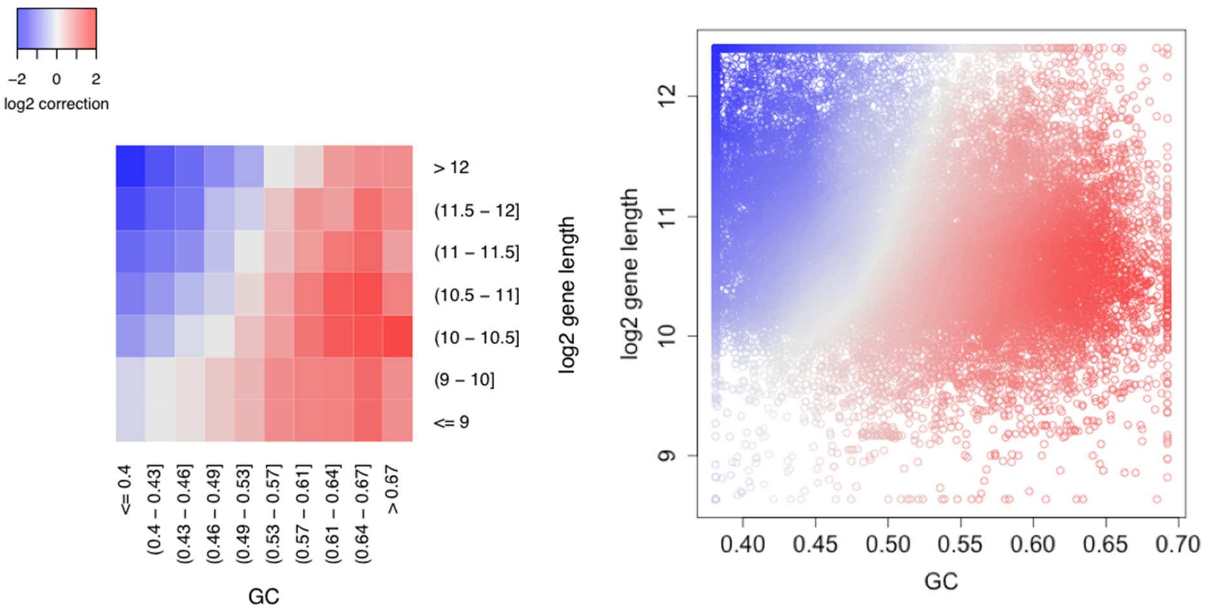

Fig. 2. ***LBC* estimates at the interpolation points and applied offsets.** The left plot shows the estimates on the 2D GC-log length grid and the right plot shows the correction factors applied to the genes. The correction factors are obtained by bilinear interpolation. For visualization, the range of the GC content and gene length have been truncated

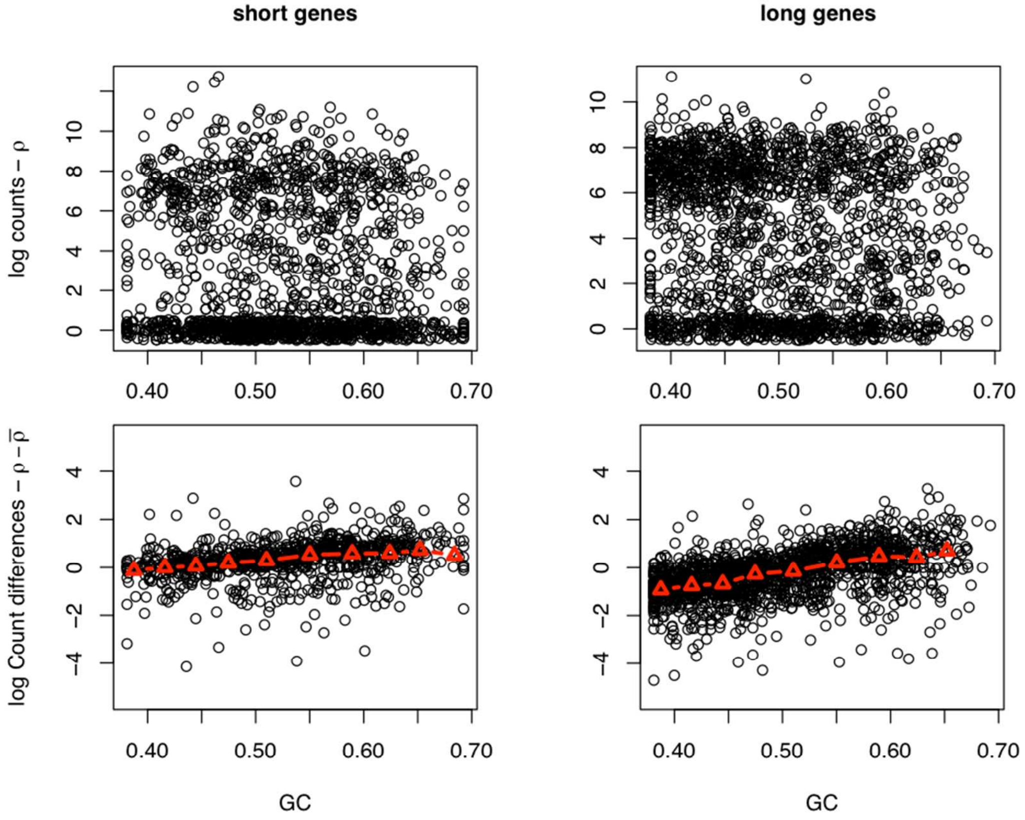

Fig. 3. **Log count measures and their dependency on GC content and gene length.** Upper row: The GC dependency of the log count measures is masked by the large dynamic range of the gene expression. This holds for short genes (left) as well as long genes (right). Lower row: Differences of the log counts relative to the gene-wise median in the dataset that was normalized for sequencing depth. The GC dependency becomes visible and the fitted values (red) show that short genes show a different GC-trend than long genes. This plot shows data from the sample *SRR1032145* in the short read archive (SRA) project *SRP033117*.

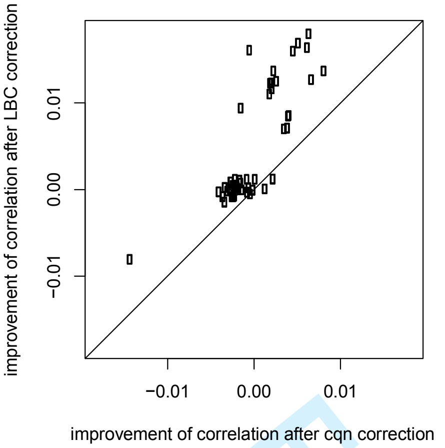

Fig. 4. Improvement of sample correlation after bias correction. The *LBC* correction leads to higher sample correlation as compared to CQN correction. Black circles represent correlations for samples with the same protocol. Blue circles represent the correlation of the two biased samples relative to the rest. The red circle is the correlation of the two biased samples relative to each other.

With the *LBC* approach, the bias function is estimated for each sample from the differences to the median expression in the dataset while other approaches use directly the log expression counts. The different characteristics of these two approaches are illustrated with Figure 2. The top row shows the normalized log expression counts plotted against the gene GC-content. For the display, genes with zero counts (negative log count) are substituted with a random log count between -0.5 and 0.5. No strong GC trend is apparent in this sample. The plots in the lower row of Figure 2 show the log count differences relative to the gene-wise median in the dataset (genes where the log ratio imply zero counts are not shown). From these differences it becomes apparent that the GC trend in this specific sample deviates from the majority of the samples in the data set. When comparing the figures in the left and right column one can also observe that the GC trend differs for the short and long genes. This difference indicates a general interaction of the GC and gene length effects on the gene expression counts.

3.2 Bias correction performance

We measure the performance of the bias correction by the increase of the sample-sample correlation of the expression values. For the study *SRP033117* this is shown in Figure 4. Generally, *LBC* leads to a positive change of the sample-sample correlation and the change is higher as compared to the corresponding *CQN* correction. Especially the correlation of the two poly-A samples with the ribo-depletion samples (blue circles) increases. The correlation of the ribo-depletion samples among each other stays the same with *LBC* and slightly decreases with *CQN* correction. Only the correlation of the two biased samples relative to each other decreases after correction with both methods (red circle).

3.3 Prevalence of bias in SRA studies

We globally screened SRA data sets for the presence of samples with deviating biases within the data sets. Out of all human data sets available at the Digital Expression Explorer we selected data sets according to the following criteria

- Samples must have at least 5 Mio reads aligned
- Samples must have a QC rating of PASS in the meta-information provided by DEE.
- Data sets must have at least 8 samples and at most 100 samples

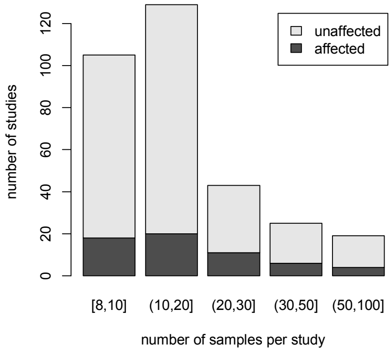

Fig. 5. Overview of human SRA studies affected by GC and gene length bias. The plot shows the studies stratified by sample size and for each strata the fraction of studies where more than 10% of the samples have a diverging GC and gene length bias.

We restricted ourselves to these criteria in order to have a good basis for judging whether the *LBC* algorithm performs well. Altogether this left us with 329 data sets where we applied bias correction. We considered a sample as affected if for at least 500 genes a correction factor greater or equal to a two-fold change was applied. Furthermore, we considered a study itself as affected if at least 10% of the samples were affected. Figure 5 plots the number of affected studies stratified by study size. We observe that around 15% of the studies can be considered as affected. This shows that variations in GC and length bias are not a rare event. but do happen in many situations.

4 Discussion

Current RNA-seq protocols do not provide an unbiased view on the transcriptome in the sense that the generated expression estimates do represent the molar concentrations of the transcripts up to a single global scaling factor. The individual steps of the protocol do lead to biases that depend on physical properties of the transcripts but may also depend on other features like read mappability. Biases per se do not render the data unusable. As long as they are constant across samples, the data still allows differential analyses without trade-off. Biases are, however, a concern in situations where they lead to a strong underrepresentation of a gene and consequently to an unreliable low or zero expression estimate which represents an irreversible information loss. Additionally, biases are a concern when they vary between samples processed with the same protocol. And finally, biases that differ between protocols are a concern when data generated by different protocols are to be integrated. As reported by others, we have observed that bias in RNA-seq data is subject to sample-to-sample differences, and we propose an algorithm to correct for probably the most common biases encountered, GC and gene length dependency.

In contrast to other approaches, we do not aim at modeling the biases in a data set in their entirety, but we rather focus on modeling the deviations that individual samples have from the dominant bias in a dataset and we apply only correction factors that mitigate these deviations. After our *LBC* correction, an RNA-seq data set is not bias free. The LBC approach rather makes the bias consistent such that it no longer differs between samples.

Deliberately, we restricted ourselves to the biases explained by the independent variables GC content and gene length but we are well aware that additionally, general amplification biases might be present. Such type of biases do e.g. originate from different amounts of starting material.

As a limitation, *LBC* will not work for studies where the effect of experimental factors does not primarily depend on the biological function of the gene, but rather on the physical properties of the genes, GC content and length. In those studies the *LBC* algorithm would remove the effects of the experimental factor.

Acknowledgements

We sincerely acknowledge all members of the Functional Genomics Center Zurich (FGCZ) for the constructive feedback.

Conflict of Interest: none declared.

References

Aird, D., *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries, *Genome Biol*, **12**, R18.

Dobin, A., *et al.* (2013) STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, **29**, 15-21.

Filloux, C., *et al.* (2014) An integrative method to normalize RNA-Seq data, *BMC Bioinformatics*, **15**, 188.

Finotello, F., *et al.* (2014) Reducing bias in RNA sequencing data: a novel approach to compute counts, *BMC Bioinformatics*, **15 Suppl 1**, S7.

Hansen, K.D., Irizarry, R.A. and Wu, Z. (2012) Removing technical variability in RNA-seq data using conditional quantile normalization, *Biostatistics*, **13**, 204-216.

Leek, J.T. (2014) svaseq: removing batch effects and other unwanted noise from sequencing data, *Nucleic Acids Res*, **42**.

Li, S., *et al.* (2014) Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study, *Nat Biotechnol*, **32**, 915-925.

Li, S., *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data, *Nat Biotechnol*, **32**, 888-895.

Liao, Y., Smyth, G.K. and Shi, W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*, **30**, 923-930.

Love, M.I., Hogenesch , J.B. and Irizarry, R.A. (2015) Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. bioRxiv.

McCarthy, D.J., Chen, Y. and Smyth, G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation, *Nucleic Acids Res*, **40**, 4288-4297.

Nagalakshmi, U., Waern, K. and Snyder, M. (2010) RNA-Seq: a method for comprehensive transcriptome analysis, *Curr Protoc Mol Biol*, **Chapter 4**, Unit 4.11.11-13.

Pickrell, J.K., *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing, *Nature*, **464**, 768-772.

Risso, D., *et al.* (2014) Normalization of RNA-seq data using factor analysis of control genes or samples, *Nat Biotechnol*, **32**, 896-902.

Risso, D., *et al.* (2011) GC-content normalization for RNA-Seq data, *BMC Bioinformatics*, **12**, 480.

Stegle, O., Teichmann, S.A. and Marioni, J.C. (2015) Computational and analytical challenges in single-cell transcriptomics, *Nat Rev Genet*, **16**, 133-145.

Zheng, W., Chung, L.M. and Zhao, H. (2011) Bias detection and correction in RNA-Sequencing data, *BMC Bioinformatics*, **12**, 290.

Ziemann M, *et al.* (2015) Digital Expression Explorer: A user-friendly repository of uniformly processed RNA-seq data. ComBio2015.