Smooth Quantile Normalization

Stephanie C. Hicks[1,2], Kwame Okrah[3], Joseph N. Paulson[1,2], John Quackenbush[1,2], Rafael A. Irizarry[1,2], and Héctor Corrada Bravo[4,5,*]

[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute
[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[3]Genentech
[4]Department of Computer Science, University of Maryland, College Park
[5]Center for Bioinformatics and Computational Biology, Institute of Advanced Computer Studies, University of Maryland, College Park
[*]Corresponding Author

Emails:
Stephanie C. Hicks, shicks@jimmy.harvard.edu
Kwame Okrah, kwame.okrah@gene.com
Joseph Paulson, jpaulson@jimmy.harvard.edu
John Quackenbush, johnq@jimmy.harvard.edu
Rafael A. Irizarry, rafa@jimmy.harvard.edu
Héctor Corrada Bravo, hcorrada@umiacs.umd.edu

Abstract

Between-sample normalization is a critical step in genomic data analysis to remove systematic bias and unwanted technical variation in high-throughput data. Global normalization methods are based on the assumption that observed variability in global properties is due to technical reasons and are unrelated to the biology of interest. For example, some methods correct for differences in sequencing read counts by scaling features to have similar median values across samples, but these fail to reduce other forms of unwanted technical variation. Methods such as quantile normalization transform the statistical distributions across samples to be the same and assume global differences in the distribution are induced by only technical variation. However, it remains unclear how to proceed with normalization if these assumptions are violated, for example if there are global differences in the statistical distributions between biological conditions or groups, and external information, such as negative or control features, is not available. Here we introduce a generalization of quantile normalization, referred to as *smooth quantile normalization* (qsmooth), which is based on the assumption that the statistical distribution of each sample should be the same (or have the same distributional shape) within biological groups or conditions, but allowing that they may differ between groups. We illustrate the advantages of our method on several high-throughput datasets with global differences in distributions corresponding to different biological conditions. We also perform a Monte Carlo simulation study to illustrate the bias-variance tradeoff of qsmooth compared to other global normalization methods. A software implementation is available from https://github.com/stephaniehicks/qsmooth.

Keywords

global normalization, quantile normalization

Introduction

Multi-sample normalization methods are an important part of any data analysis pipeline to remove systematic bias and unwanted technical variation, particularly in high-throughput data, where systematic effects can cause perceived differences between samples irrespective of biological variation. Many *global adjustment* normalization methods (Gagnon-Bartsch and Speed 2012; Hicks and Irizarry 2015) have been developed based on the assumption that observed variability in global properties is due to technical reasons and are unrelated to the biology of the system under study (Bolstad et al. 2003; Reimers 2010). Examples of global properties include differences in the total, upper quartile (Bullard et al. 2010) or median gene expression, proportion of differentially expressed genes (Anders and Huber 2010; Robinson and Oshlack 2010; Love, Huber, and Anders 2014), observed variance across expression levels (Durbin et al. 2002) and statistical distribution across samples.

Quantile normalization is a global adjustment normalization method that transforms the statistical distributions across samples to be the same and assumes global differences in the distribution are induced by technical variation (Amaratunga and Cabrera 2001; Bolstad et al. 2003). The observed distributions are forced to be the same to achieve normalization and the average distribution (average of each quantile across samples) is used as the reference.

Several studies have evaluated quantile normalization and other global adjustment normalization methods (Robinson and Oshlack 2010; Bullard et al. 2010; Dillies et al. 2013; Aanes et al. 2014). Under the assumptions of global adjustment normalization methods, quantile normalization has been shown to reduce the variance in observed gene expression data with a tradeoff of inducing a small amount of bias (due to the bias-variance tradeoff) (Bolstad et al. 2003; Qiu, Hu, and Wu 2014). However, when the assumptions of global adjustment normalization methods are violated (for example, if the majority of genes are up-regulated in one biological condition relative to another (Lovén et al. 2012; Aanes et al. 2014; Hu et al. 2014; Evans, Hardin, and Stoebel 2016), forcing the distributions to be the same can lead to errors in downstream analyses. Graphical and quantitative assessments (Hicks and Irizarry 2015) have been developed to assess the assumptions of global normalization methods.

If global adjustment methods are found not to be appropriate, another class of normalization methods can be applied (*application-specific* methods), but these often rely on external information such as positive and negative control features or experimentally measured data (Lovén et al. 2012; Aanes et al. 2014). However, it is unclear how to proceed with normalization if the assumptions about the observed variability in global properties are violated, such as they may occur when there are global differences in the statistical distributions between tissues (Figure 1), and external information is not available.

Here we introduce a generalization of quantile normalization, referred to as *smooth quantile normalization* (qsmooth), which is based on the assumption that the statistical distribution of each sample should be the same (or have the same distributional shape) within a biological group (or condition), but that the distribution may differ between groups. At each quantile, a weight is computed comparing the variability between groups relative to the total variability between and within groups (Equation 1). In one extreme with a weight of zero, qsmooth is quantile normalization within each biological group when there are global differences in distributions corresponding to differences in biological groups. As the variability between groups decreases, the weight increases towards one and the quantile is shrunk towards the overall reference quantile (Equation 2) and is equivalent to standard *quantile normalization*. In certain portions of the distributions, the quantiles from different biological groups may be more or less similar to each other depending on the biological variability, which is reflected in the weight varying between 0 and 1 across the quantiles.

Using several high-throughput datasets, we demonstrate the advantages of qsmooth, which include (1) preservation of global differences in distributions corresponding to different biological conditions, (2) non-reliance on external information, (3) applicability to many different high-throughput technologies, and (4) the return of normalized data that can be used for many types of downstream analyses including finding differences in features (genes, CpGs, etc), clustering and dimensionality reduction. We also perform a Monte Carlo simulation study to illustrate the bias-variance tradeoff when using qsmooth.

Results

qsmooth: smooth quantile normalization.

Consider a set of high-dimensional vectors $Y_1, Y_2, ..., Y_n$ each of length J representing samples from a high-throughput experiment and each associated with a covariate Z_i representing the biological group or condition. We define $F_i^{-1}(u)$ as the empirical quantile function for the i^{th} sample and the u^{th} quantile where $u \in [0,1]$. Quantile normalization begins by calculating a reference distribution, which is the average at each quantile across the samples, $\overline{F}_i^{-1}(u) = \frac{1}{n}\sum_{i=1}^{n} F_i^{-1}(u)$. Our method begins by assuming that following form $F_i^{-1}(u) = Z_i\beta(u) + \varepsilon_i$. This model is similar to the model described in the functional normalization method proposed by Fortin et al. (Fortin et al. 2014), which relates the quantile functions of a set of high-dimensional vectors to a set of known covariates Z_i that are not associated with biological group or condition. Functional normalization attempts to remove the influence of unwanted technical variation using control features leaving the biological variation in the data. We take a different approach that does not depend on the use of control features and uses a covariate Z_i that is associated with the biological group or condition. In addition, our model extends the model of Fortin et al. by adaptively weighting group information in the normalization transformation applied. Here, $\hat{\beta}(u)$ are the estimated regression coefficients representing the reference distributions within each biological group at each quantile and the predicted values, $\hat{F}_i^{-1}(u) = Z_i\hat{\beta}(u)$ correspond to quantile normalized data within biological groups. We partition the total sum of squares ($SST_{(u)}$) into the residual sum of squares ($SSE_{(u)}$) and the explained sum of squares ($SSB_{(u)}$),

$$\sum_{i=1}^{n}(F_i^{-1}(u) - \overline{F}_i^{-1}(u))^2 = \sum_{i=1}^{n}(F_i^{-1}(u) - \hat{F}_i^{-1}(u))^2 + \sum_{i=1}^{n}(\hat{F}_i^{-1}(u) - \overline{F}_i^{-1}(u))^2$$

For each quantile u, we calculate the weight ($w_{(u)}$),

$$w_{(u)} = median\left\{1 - \frac{SSB_{(j)}}{SST_{(j)}}\right\} \quad \text{for } j = u-k, ..., u, ..., u+k \qquad (1)$$

where we use a rolling median across $j = u-k, ..., u, ..., u+k$ quantiles with a width of $\pm k$ where $k = floor(N * 0.05)$ to smooth the weights at quantiles with a high variance. The number 0.05 is a flexible parameter than can be altered to change the window of the number of quantiles considered. The smooth quantile normalized data is a weighted average,

$$F_i^{qsmooth}(u) = w_{(u)}\overline{F}_i^{-1}(u) + (1 - w_{(u)})\hat{F}_i^{-1}(u) \qquad (2)$$

The raw feature values are substituted with the $F_i^{qsmooth}(u)$ values and then the transformed values are placed in the original order similar to quantile normalization.

Global differences in distributions between tissues in gene expression and DNA methylation data. We compared qsmooth to other normalization methods using publicly available gene expression and DNA methylation datasets with global differences in distributions. We assessed how global normalization methods impact control features, namely the External RNA Control consortium (ERCC) spike-ins (Jiang et al. 2011), in samples comparing the gene expression from brain and liver tissue in rats (see bodymapRat data set in Methods).

We found that global normalization methods remove the global differences in distribution between brain and liver tissues and induce artificial differences in the spike-in controls compared to using the raw data, including quantile normalization (p <2.2e^{-16}), Relative Log Expression (RLE) normalization (Anders and Huber 2010) (p <2.2e^{-16}), and median normalization (p <2.2e^{-16}) (Figure 2; Supplementary Figures 1-2). In contrast, our method, qsmooth, greatly reduces artificial differences induced between the distributions of the spike-in control genes (p = 9.2e^{-05}).

Using the data from the Genotype-Tissue Expression project (GTEx) (GTEx Consortium 2015), we compared qsmooth to a number of scaling normalization methods including, RLE, Trimmed Mean of M-Values (TMM) (Robinson and Oshlack 2010), and upper quartile scaling (Bullard et al. 2010). We observed that scaling methods did not sufficiently control for variability between distributions within tissues; in particular, we observed stark differences in global distribution for a number of body regions, most pronounced between testis, whole blood and other tissues such as artery tibial (Figure 3; Supplemental Figure 3). Normalizing tissues with global differences (in distribution) using a tissue-specific reference distribution, such as in qsmooth, can reduce the root mean squared errors (RMSE) of the overall variability across distributions compared to quantile normalization (Paulson et al. 2016). This occurs because qsmooth is based on the assumption that the statistical distribution of each sample should be similar within a biological group, but not necessarily across biological groups.
To demonstrate the importance of preserving tissue-specific differences, we assessed the impact of normalization using quantile normalization and qsmooth using two genes, ENSG00000160882 (*CYP11B1*) and ENSG00000164532 (*TBX20*). These two genes are known to be highly expressed in specific tissues (Figure 4; Supplementary Table 1). The *CYP11B1* gene has been shown to play a critical role in congenital adrenal hyperplasia (Zachmann, Tassinari, and Prader 1983; Curnow et al. 1993; Joehrer et al. 1997) and the *TBX20* gene plays an important role in cardiac chamber differentiation in adults (Cai et al. 2005; Singh et al. 2005; Stennard et al. 2005; Takeuchi et al. 2005; Qian et al. 2008). In both genes, we found that quantile normalization removes the biologically known tissue-specific expression. In contrast, qsmooth preserves the tissue-specificity, which is also observed just using the raw data. In particular, the *CYP11B1* gene is highly expressed in the testis tissue using both qsmooth normalized and raw data, but it is reported as lowly expressed in the testis tissue after applying quantile normalization. Using qsmooth normalized data and raw data, we observe the tissue-specific gene *TBX20* as highly expressed in heart atrial appendage and heart left ventricle tissues, but lowly expressed in the same tissues after applying quantile normalization. Furthermore, quantile normalization results in this gene being spuriously inflated in other tissues.

We also tested qsmooth using publicly available DNA methylation (DNAm) data from six purified cell types in whole blood that are known to exhibit global differences in DNAm (Hicks and Irizarry 2015). Using qsmooth, the global differences in distributions are preserved across purified cell types (Figure 5).

Furthermore, the cell types cluster more closely along the first two principal components compared to using the raw data or quantile normalized data, because qsmooth accounts for cell type-specific differences in DNAm and removes technical variability across samples within each cell-type.

The bias-variance tradeoff of qsmooth.

We performed a Monte Carlo simulation study to evaluate the performance of qsmooth when the assumptions related to the observed variability in global properties are violated with the detection of differentially expressed genes as a measure of overall performance. We generated gene-level RNA-Seq counts and varied the proportion of differentially expressed genes between biological groups.

As others have noted, when testing for differential expression between groups, quantile normalization results in increased bias with a tradeoff of a reduction in variance compared to using the raw data. Under the assumptions of global normalization methods, qsmooth improves upon this tradeoff, resulting in lower bias compared to quantile normalization, but also less variance compared to using the raw data, and better overall detection of differential expression. As the number of differentially expressed genes increases, quantile normalization and qsmooth both reduce the variance compared to using the raw data, but qsmooth also reduces the bias compared to using the raw data by accounting for global differences between the biological groups, particularly when the assumptions of global normalization methods are violated (Supplementary Figure 4).

Conclusions

Global normalization methods are useful for removing unwanted technical variation from high-throughput data. However, they are based on the assumption that observed variability in global properties is due only to technical factors and is unrelated to the biology of the system under study. While these assumptions are usually fine when comparing closely related samples, large-scale studies are increasingly generating data where those assumptions do not hold. In cases where these global assumptions are violated, more robust forms of normalization are needed to allow for different distributions in different classes of samples.

Application-specific normalization methods can be applied, but these methods rely on the use of external information such as positive or negative control features or experimentally measured information, which are often not available. Furthermore, these methods are also based on assumptions about the nature of the measured distributions, and these have been shown to be violated in many situations (Dillies et al. 2013; Risso et al. 2014).

The new method we describe here, *smooth quantile normalization* (qsmooth), is based on the assumption that the statistical distribution of each sample should be the same (or have the same distributional shape) within a biological group or condition, but it does not require that different groups or conditions have the same distribution. Our method also does not require any external information other than sample group assignment, it is not specific to one type of high-throughput data, and it returns normalized data that can be used for many types of downstream analyses including finding differences in features (genes, CpGs, etc), clustering and dimensionality reduction.

We demonstrated the advantages of qsmooth using several high-throughput datasets that exhibit global differences in distributions between biological conditions, such as the global changes in gene expression profiles in brain and liver. We illustrated the bias-variance tradeoff when using qsmooth, which preserves global differences in distributions corresponding to different biological conditions. We have implemented our normalization method into the qsmooth R-package, which is available on GitHub (https://github.com/stephaniehicks/qsmooth).

Methods

Datasets with global differences in distributions.
We downloaded Affymetrix GeneChip gene expression data for alveolar macrophages (GSE2125), brain (GSE17612, GSE21935), and liver (GSE29721, GSE14668, GSE6764) samples in human as reported by a number of studies archived in the Gene Expression Omnibus (GEO) (Edgar, Domrachev, and Lash 2002). We extracted the raw Perfect Match (PM) values from the CEL files using the *affy* (Gautier et al. 2004) R/Bioconductor package for gene expression.

We downloaded raw RNA-Seq gene counts from the *T. cruzi* life cycle (Li et al. 2016). We also downloaded and mapped raw sequencing reads to obtain raw RNA-Seq gene counts for multiple tissues from the Rat BodyMap project (Yu et al. 2014) (GSE53960). This data is also available as an R data package on GitHub, (https://github.com/stephaniehicks/bodymapRat) (see Supplementary Material for more details). Counts have an added pseudocount of 1 and then are $\log_2$ transformed.We used the Kolmogorov–Smirnov test for global differences in distributions in spike-ins from the bodymapRat gene expression data.

Gene expression data from the Genotype-Tissue Expression (GTEx) consortium was downloaded from the GTEx portal (http://www.gtexportal.org/) and processed using YARN (Paulson et al. 2016) (bioconductor.org/packages/yarn) (see Supplementary Materials for more details).

The sorted whole blood cell populations measured on Illumina 450K DNA methylation arrays were obtained from *FlowSorted.Blood.450k* R/Bioconductor data package (Jaffe 2015) and the raw beta values were extracted using the *minfi* R/Bioconductor package (Aryee et al. 2014).

Monte Carlo simulation study.
We used the polyester R/Bioconductor package (Frazee et al. 2015) to simulate gene-level RNA-Seq counts while varying the proportion of differentially expressed genes (pDiff) to obtain samples with global differences in the distributions between biological conditions. Each simulation study considered ten samples from two groups (total of 20 samples). We added additional non-linear sample-specific noise by splitting the sample into four quartiles and scaling each quartile within the sample with a draw from a uniform distribution ranging from 0.5 to 3. This is more realistic than linearly scaling each sample.

As our measure of performance in the detection of differentially expressed genes, we compared the output of qsmooth to both quantile normalized data and raw (unnormalized) gene counts. We assessed the bias-variance tradeoff of the $\log_2$ fold change using these three methods while varying the proportion of differentially expressed genes between two groups. The plots were created with the ggplot2 R package (Wickham 2009).

Software
The R-package *qsmooth* implementing our method is available on GitHub (https://github.com/stephaniehicks/qsmooth).

Supplementary Materials

Supplementary materials are available in a single pdf, which contain supplemental figures and a detailed description of the bodymapRat and GTEx datasets. All scripts containing the code for these analyses are available on Github (https://github.com/stephaniehicks/qsmoothPaper).

Acknowledgments

S.C.H and R.A.I. were supported by NIH R01 grants GM083084 and RR021967/GM103552. J.N.P and J.Q. were supported in part by grants 5P01HL105339, 5R01HL111759, 5P01HL114501, 5P50CA127003, 1R35CA197449, 1U01CA190234, 5P30CA006516, and 5R01AI099204. H.C.B and K.O. were supported by NIH R01 grants HG005220 and GM114267.

Conflict of Interest

None declared.

References

Aanes, Håvard, Cecilia Winata, Lars F. Moen, Olga Østrup, Sinnakaruppan Mathavan, Philippe Collas, Torbjørn Rognes, and Peter Aleström. 2014. "Normalization of RNA-Sequencing Data from Samples with Varying mRNA Levels." *PloS One* 9 (2): e89158.

Amaratunga, Dhammika, and Javier Cabrera. 2001. "Outlier Resistance, Standardization, and Modeling Issues for DNA Microarray Data." In *Statistics in Genetics and in the Environmental Sciences*, edited by Luisa Turrin Fernholz, Stephan Morgenthaler, and Werner Stahel, 17–26. Trends in Mathematics. Birkhäuser Basel.

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10): R106.

Aryee, Martin J., Andrew E. Jaffe, Hector Corrada-Bravo, Christine Ladd-Acosta, Andrew P. Feinberg, Kasper D. Hansen, and Rafael A. Irizarry. 2014. "Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays." *Bioinformatics* 30 (10): 1363–69.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. 2003. "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias." *Bioinformatics* 19 (2): 185–93.

Bullard, James H., Elizabeth Purdom, Kasper D. Hansen, and Sandrine Dudoit. 2010. "Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments." *BMC Bioinformatics* 11 (February): 94.

Cai, Chen-Leng, Wenlai Zhou, Lei Yang, Lei Bu, Yibing Qyang, Xiaoxue Zhang, Xiaodong Li, Michael G. Rosenfeld, Ju Chen, and Sylvia Evans. 2005. "T-Box Genes Coordinate Regional Rates of Proliferation and Regional Specification during Cardiogenesis." *Development* 132 (10): 2475–87.

Curnow, K. M., L. Slutsker, J. Vitek, T. Cole, P. W. Speiser, M. I. New, P. C. White, and L. Pascoe. 1993. "Mutations in the CYP11B1 Gene Causing Congenital Adrenal Hyperplasia and Hypertension Cluster in Exons 6, 7, and 8." *Proceedings of the National Academy of Sciences of the United States of America* 90 (10): 4552–56.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. "A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis." *Briefings in Bioinformatics* 14 (6): 671–83.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Durbin, B. P., J. S. Hardin, D. M. Hawkins, and D. M. Rocke. 2002. "A Variance-Stabilizing Transformation for Gene-Expression Microarray Data." *Bioinformatics* 18 Suppl 1: S105–10.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10.

Evans, Ciaran, Johanna Hardin, and Daniel Stoebel. 2016. "Selecting between-Sample RNA-Seq Normalization Methods from the Perspective of Their Assumptions." *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1609.00959.

Fortin, Jean-Philippe, Aurélie Labbe, Mathieu Lemire, Brent W. Zanke, Thomas J. Hudson, Elana J. Fertig, Celia Mt Greenwood, and Kasper D. Hansen. 2014. "Functional Normalization of 450k Methylation Array Data Improves Replication in Large Cancer Studies." *Genome Biology* 15 (12): 503.

Frazee, Alyssa C., Andrew E. Jaffe, Ben Langmead, and Jeffrey T. Leek. 2015. "Polyester: Simulating

RNA-Seq Datasets with Differential Transcript Expression." *Bioinformatics* 31 (17): 2778–84.

Gagnon-Bartsch, Johann A., and Terence P. Speed. 2012. "Using Control Genes to Correct for Unwanted Variation in Microarray Data." *Biostatistics* 13 (3): 539–52.

Gautier, Laurent, Leslie Cope, Benjamin M. Bolstad, and Rafael A. Irizarry. 2004. "Affy--Analysis of Affymetrix GeneChip Data at the Probe Level." *Bioinformatics* 20 (3): 307–15.

GTEx Consortium. 2015. "Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." *Science* 348 (6235): 648–60.

Hicks, Stephanie C., and Rafael A. Irizarry. 2015. "Quantro: A Data-Driven Approach to Guide the Choice of an Appropriate Normalization Method." *Genome Biology* 16 (June): 117.

Hu, Zheng, Kaifu Chen, Zheng Xia, Myrriah Chavez, Sangita Pal, Ja-Hwan Seol, Chin-Chuan Chen, Wei Li, and Jessica K. Tyler. 2014. "Nucleosome Loss Leads to Global Transcriptional up-Regulation and Genomic Instability during Yeast Aging." *Genes & Development* 28 (4): 396–408.

Jaffe, A. E. 2015. "FlowSorted. Blood. 450k: Illumina HumanMethylation Data on Sorted Blood Cell Populations." *R Package Version* 1 (0).

Jiang, Lichun, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R. Gingeras, and Brian Oliver. 2011. "Synthetic Spike-in Standards for RNA-Seq Experiments." *Genome Research* 21 (9): 1543–51.

Joehrer, K., S. Geley, E. M. Strasser-Wozak, R. Azziz, H. A. Wollmann, K. Schmitt, R. Kofler, and P. C. White. 1997. "CYP11B1 Mutations Causing Non-Classic Adrenal Hyperplasia due to 11 Beta-Hydroxylase Deficiency." *Human Molecular Genetics* 6 (11): 1829–34.

Li, Yuan, Sheena Shah-Simpson, Kwame Okrah, A. Trey Belew, Jungmin Choi, Kacey L. Caradonna, Prasad Padmanabhan, et al. 2016. "Transcriptome Remodeling in Trypanosoma Cruzi and Human Cells during Intracellular Infection." *PLoS Pathogens* 12 (4): e1005511.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.

Lovén, Jakob, David A. Orlando, Alla A. Sigova, Charles Y. Lin, Peter B. Rahl, Christopher B. Burge, David L. Levens, Tong Ihn Lee, and Richard A. Young. 2012. "Revisiting Global Gene Expression Analysis." *Cell* 151 (3): 476–82.

Paulson, Joseph N., Cho-Yi Chen, Camila M. Lopes-Ramos, Marieke L. Kuijjer, John Platig, Abhijeet R. Sonawane, Maud Fagny, Kimberly Glass, and John Quackenbush. 2016. "Tissue-Aware RNA-Seq Processing and Normalization for Heterogeneous and Sparse Data." *bioRxiv*. doi:10.1101/081802.

Qian, Li, Bhagyalaxmi Mohapatra, Takeshi Akasaka, Jiandong Liu, Karen Ocorr, Jeffrey A. Towbin, and Rolf Bodmer. 2008. "Transcription Factor neuromancer/TBX20 Is Required for Cardiac Function in Drosophila with Implications for Human Heart Disease." *Proceedings of the National Academy of Sciences of the United States of America* 105 (50): 19833–38.

Qiu, Xing, Rui Hu, and Zhixin Wu. 2014. "Evaluation of Bias-Variance Trade-off for Commonly Used Post-Summarizing Normalization Procedures in Large-Scale Gene Expression Studies." *PloS One* 9 (6): e99380.

Reimers, Mark. 2010. "Making Informed Choices about Microarray Data Analysis." *PLoS Computational Biology* 6 (5): e1000786.

Risso, Davide, John Ngai, Terence P. Speed, and Sandrine Dudoit. 2014. "Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples." *Nature Biotechnology* 32 (9): 896–902.

Robinson, Mark D., and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3): R25.

Singh, Manvendra K., Vincent M. Christoffels, José M. Dias, Mark-Oliver Trowe, Marianne Petry, Karin Schuster-Gossler, Antje Bürger, Johan Ericson, and Andreas Kispert. 2005. "Tbx20 Is Essential for Cardiac Chamber Differentiation and Repression of Tbx2." *Development* 132 (12): 2697–2707.

Stennard, Fiona A., Mauro W. Costa, Donna Lai, Christine Biben, Milena B. Furtado, Mark J. Solloway, David J. McCulley, et al. 2005. “Murine T-Box Transcription Factor Tbx20 Acts as a Repressor during Heart Development, and Is Essential for Adult Heart Integrity, Function and Adaptation.” *Development* 132 (10): 2451–62.

Takeuchi, Jun K., Maria Mileikovskaia, Kazuko Koshiba-Takeuchi, Analeah B. Heidt, Alessandro D. Mori, Eric P. Arruda, Marina Gertsenstein, et al. 2005. “Tbx20 Dose-Dependently Regulates Transcription Factor Networks Required for Mouse Heart and Motoneuron Development.” *Development* 132 (10): 2463–74.

Wickham, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer New York.

Yu, Ying, James C. Fuscoe, Chen Zhao, Chao Guo, Meiwen Jia, Tao Qing, Desmond I. Bannon, et al. 2014. “A Rat RNA-Seq Transcriptomic BodyMap across 11 Organs and 4 Developmental Stages.” *Nature Communications* 5: 3230.

Zachmann, M., D. Tassinari, and A. Prader. 1983. “Clinical and Biochemical Variability of Congenital Adrenal Hyperplasia due to 11 Beta-Hydroxylase Deficiency. A Study of 25 Patients.” *The Journal of Clinical Endocrinology and Metabolism* 56 (2): 222–29.

Figures

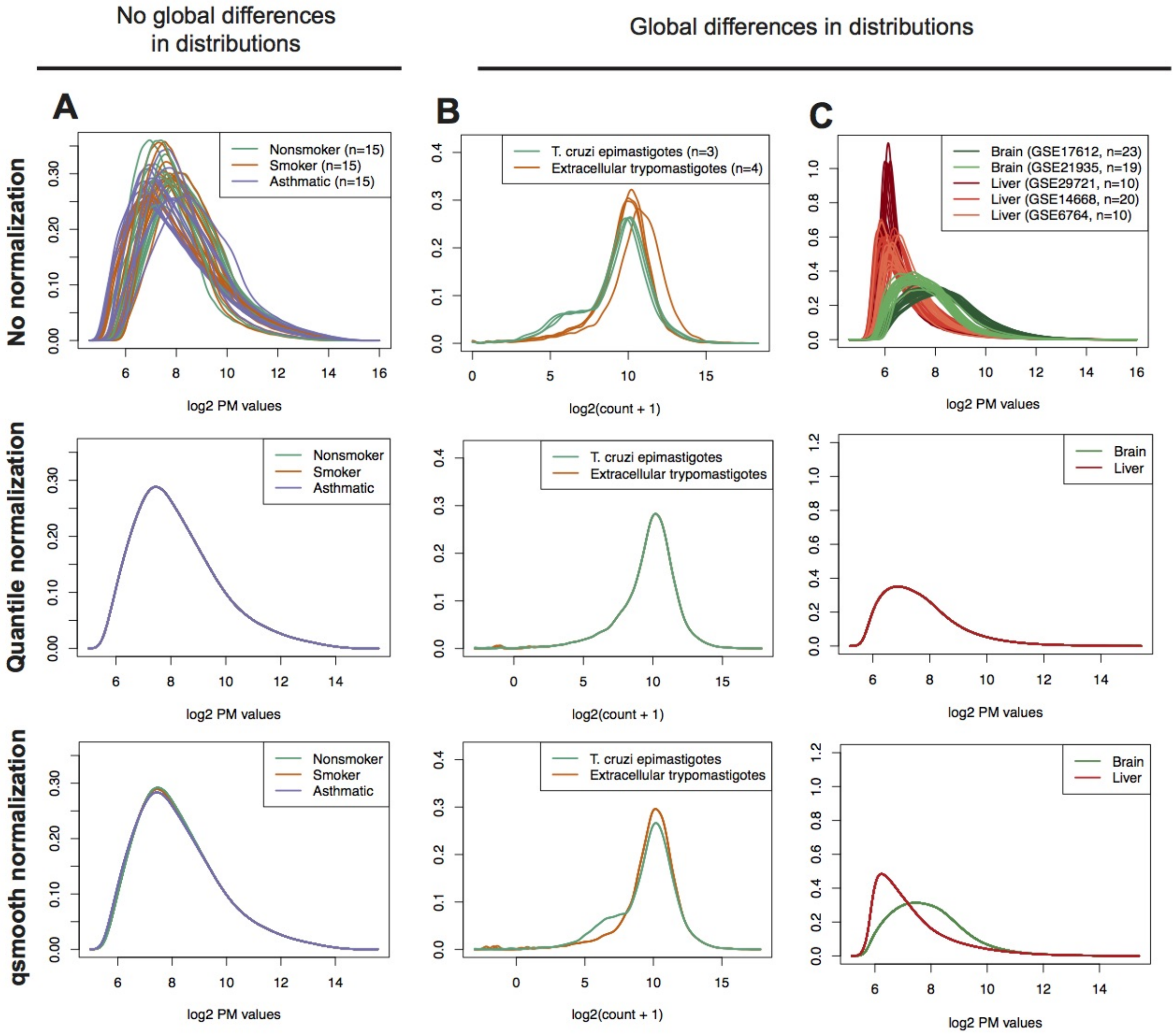


Figure 1: Using biological information to preserve global differences in distributions. Under the conditions of no global differences in distributions (A), qsmooth is similar to standard quantile normalization. Under the conditions of global differences in distributions (B) and (C), quantile normalization removes the global differences by making the distributions the same, but qsmooth preserves global differences in distributions. Examples of gene expression data with (A) Perfect match (PM) values from n = 45 arrays comparing the gene expression of alveolar macrophages from nonsmokers (green), smokers (red) and patients with asthma (blue). (B) Gene counts from n = 7 from RNA-Seq samples comparing the *T. cruzi* life cycle at the epimastigote (insect vector) stage and extracellular trypomastigotes. Counts have an added pseudocount of 1 and then are $\log_2$ transformed. (C) PM values from n = 82 arrays comparing brain and liver tissue samples colored by tissue (brain [green] and liver [orange]). The shades represent different Gene Expression Omibus (GEO) IDs.

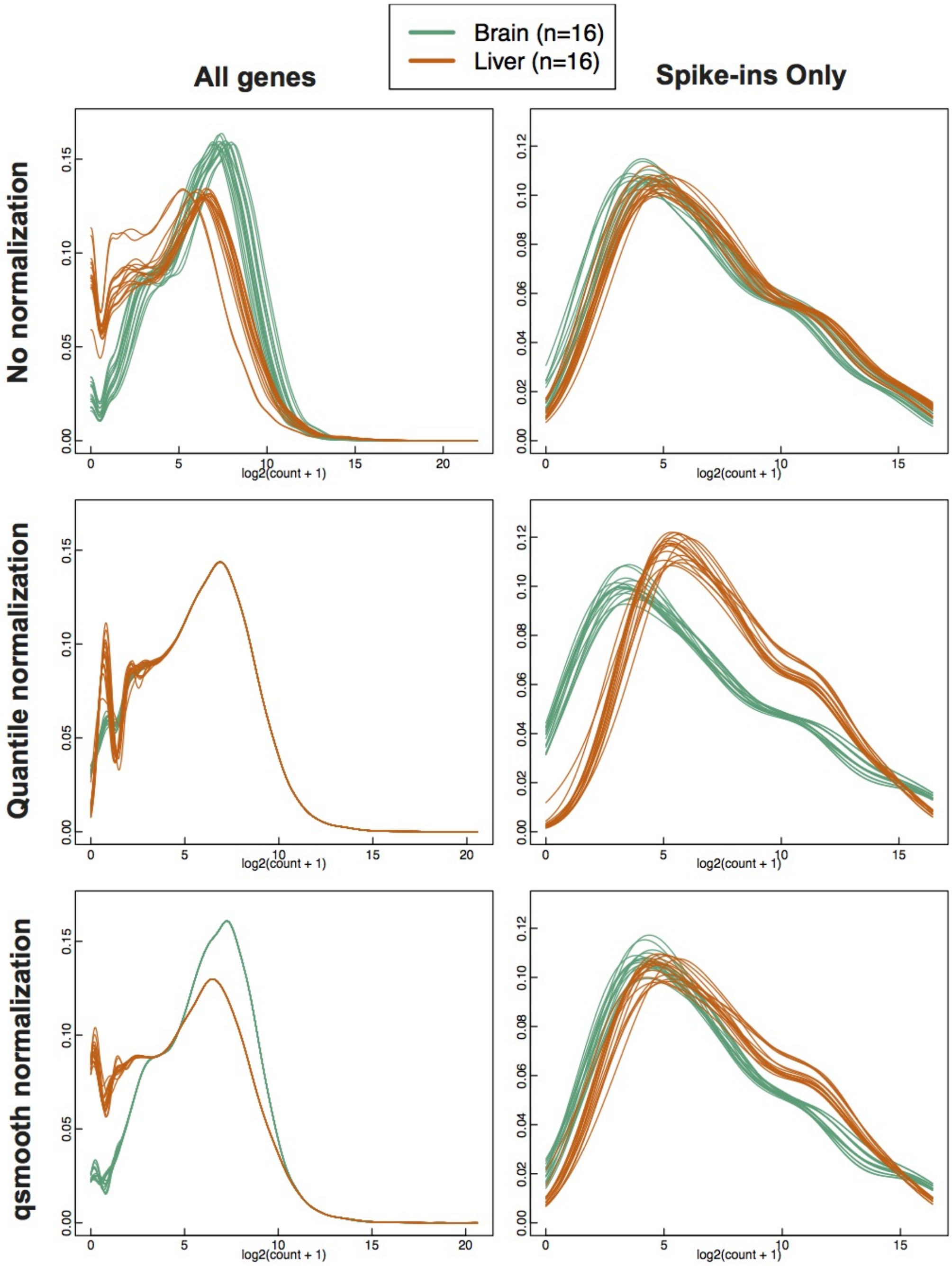


Figure 2: Quantile normalization induces artificial differences in spike-in control genes using data with global differences in distributions. Comparing no normalization (row 1), quantile normalization (row 2) and qsmooth (row 3) applied RNA-Seq gene counts from brain (green) and liver (orange) tissues in the bodymapRat dataset. Column 2 contains the density plots for only the spike-in control genes. Counts have an added pseudocount of 1 and then are $\log_2$ transformed.

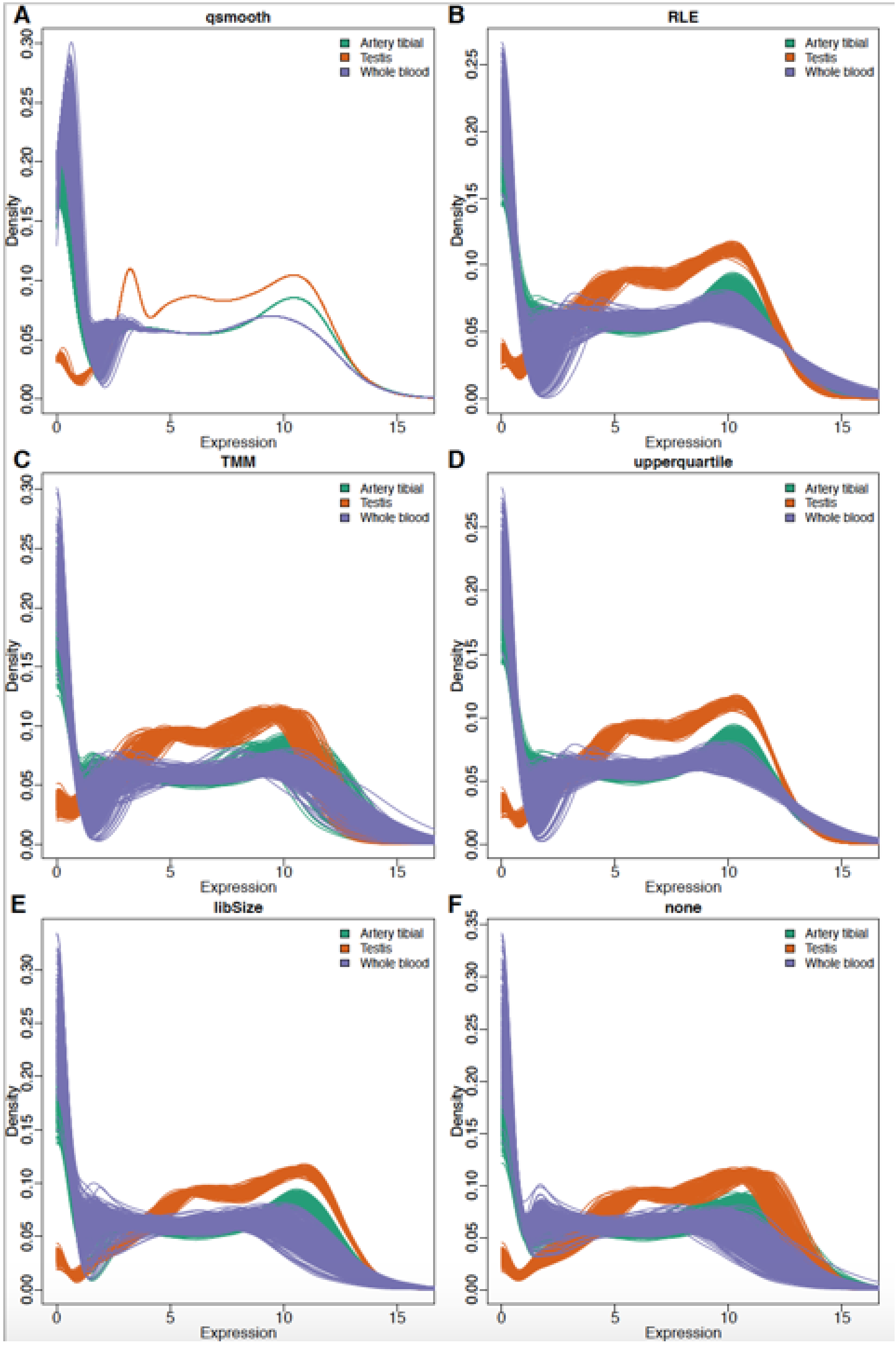


Figure 3: Scaling normalization methods do not adequately control within-group variability. Comparing density plots following either qsmooth (A), Relative Log Expression (RLE) (B), Trimmed Mean of M-Values (TMM) (C), upper quartile scaling (upperquartile) (D), library size (libSize) (E), or no (none) (F) normalization. Plotted are the artery tibial (green) and the testis (orange) tissues from the GTEx consortium. All counts have an added pseudocount of 1 and then are log2 transformed.

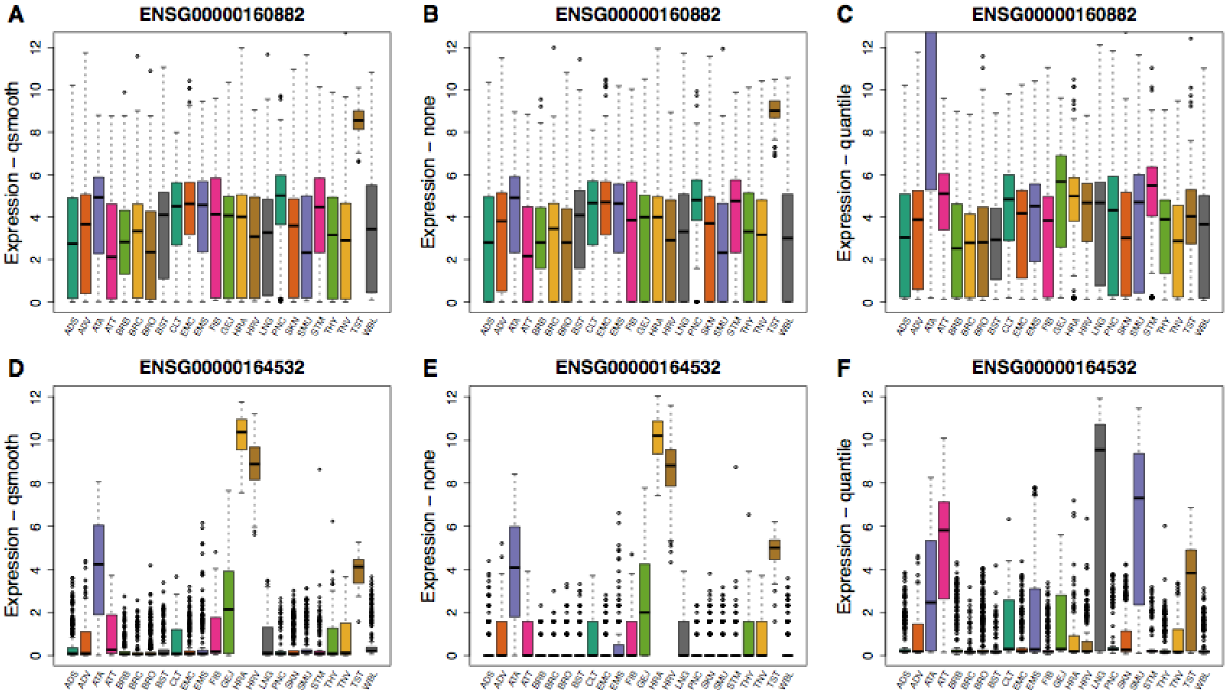


Figure 4: Gene-specific effects induced from quantile normalization. Boxplots of the normalized expression for ENSG00000160882 (*CYP11B1*) and ENSG00000164532 (*TBX20*) are shown for 24 tissues profiled by GTEx. Top, we see *CYP11B1* is more highly expressed in testis (TST) and more lowly expressed in other tissues in both (A) qsmooth and (B) raw expression profiles. However, following quantile normalization (C) *CYP11B1* is relatively lowly expressed in TST but now more variably and highly expressed in the artery aorta (ATA). *CYP11B1* produces 11 beta-hydroxylase, a final step necessary to convert 11-deoxycortisol into cortisol. Steroid 11 beta-hydroxylase deficiency is the second most common cause (5-8%) of congenital adrenal hyperplasia (Zachmann, Tassinari, and Prader 1983; Curnow et al. 1993; Joehrer et al. 1997). Bottom (D, E), *TBX20* is a member of the T-box family and encodes the TBX20 transcription factor and helps dictate cardiac chamber differentiation and in adults regulates integrity, function and adaptation (Cai et al. 2005; Singh et al. 2005; Stennard et al. 2005; Takeuchi et al. 2005; Qian et al. 2008). We see *TBX20* highly expressed in both raw and qsmooth normalized heart atrial appendage and left ventricle tissues (HRA, HRV). However, following (F) quantile normalization, expression of the gene in both heart tissues is almost zero and several other tissues are more highly or variably expressed.

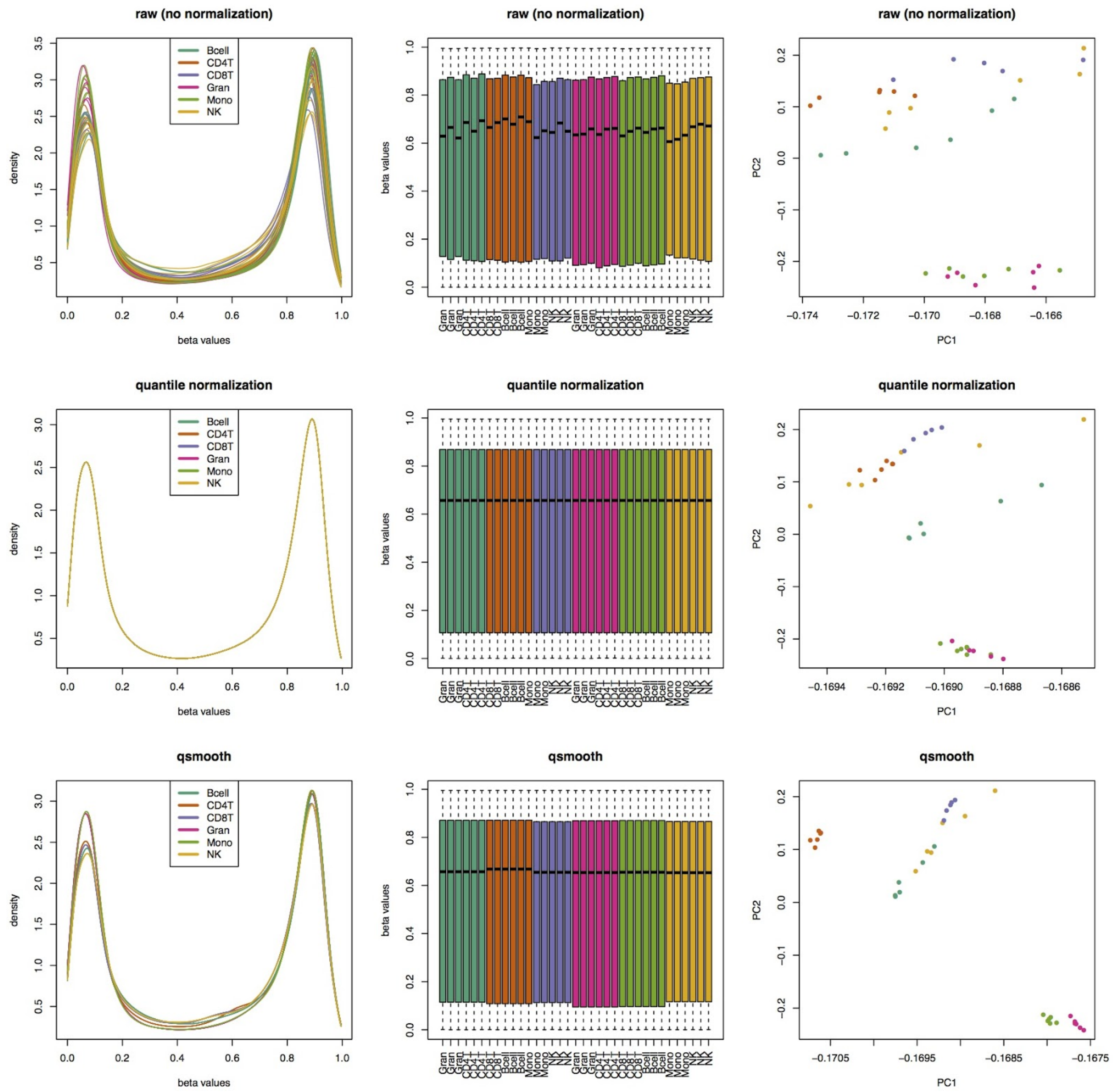


Figure 5: Density plots (column 1) and boxplots (column 2) with global changes in distributions of beta values from n = 35 Illumina 450K DNA methylation arrays comparing raw data (row 1), quantile normalized data (row 2) and qsmooth data (row 3) on six purified cell types from whole blood: CD14+ Monocytes (Mono), CD19+ B-cells (Bcell), CD4+ T-cells (CD4T), CD56+ NK-cells (NK), CD8+ T-cells (CD8T), and Granulocytes (Gran). Column 3 shows first two principal components using three normalization methods.

Supplementary Material to: Smooth Quantile Normalization

bodymapRat data.
Gene expression data from brain and liver tissues in rat measured using RNA-Seq was obtained from the rat RNA-Seq transcriptomic BodyMap (Yu et al. 2014) (GSE53960), which performed the rat BodyMap across 11 organs and 4 developmental stages. We download the raw FASTQ files and mapped the reads to a custom genome reference made up of the ENSEMBL rat genome (ftp://ftp.ensembl.org/pub/release-80/fasta/rattus_norvegicus/dna), ENSEMBL annotation files (ftp://ftp.ensembl.org/pub/release-80/gtf/rattus_norvegicus/) and the ERCC RNA spike-ins (Jiang et al. 2011) (https://tools.lifetechnologies.com/content/sfs/manuals/ERCC92.zip).

The reads were mapped to the Rat genome using STAR (Dobin et al. 2013) version 2.3.1 with default parameters. Reads mapping to annotated exons were counted using the summarizeOverlaps function in the *GenomicAlignments* R/Bioconductor package. The raw gene counts are available as an ExpressionSet in an R data package on GitHub (https://github.com/stephaniehicks/bodymapRat), which includes a complete description of processing the raw sequencing reads to obtain gene counts. We filtered out the genes with the sum of counts (across rows) less than the number of samples (columns).

GTEx data.
Gene expression data from the Genotype-Tissue Expression (GTEx) consortium (GTEx Consortium 2015) was downloaded from the GTEx portal website (www.gtexportal.org) and preprocessed using YARN (Paulson et al. 2016). In preprocessing the data we removed samples with very few samples, including the bladder, cervix - ectocervix, cervix - endocervix, fallopian tube, and the leukemia cell line samples. Following the YARN pipeline we used tissues as the biology of the system under study. For most analyses we display only tissues with at least 150 samples. See Supplementary Table 1 for a list of the abbreviations for the tissues.

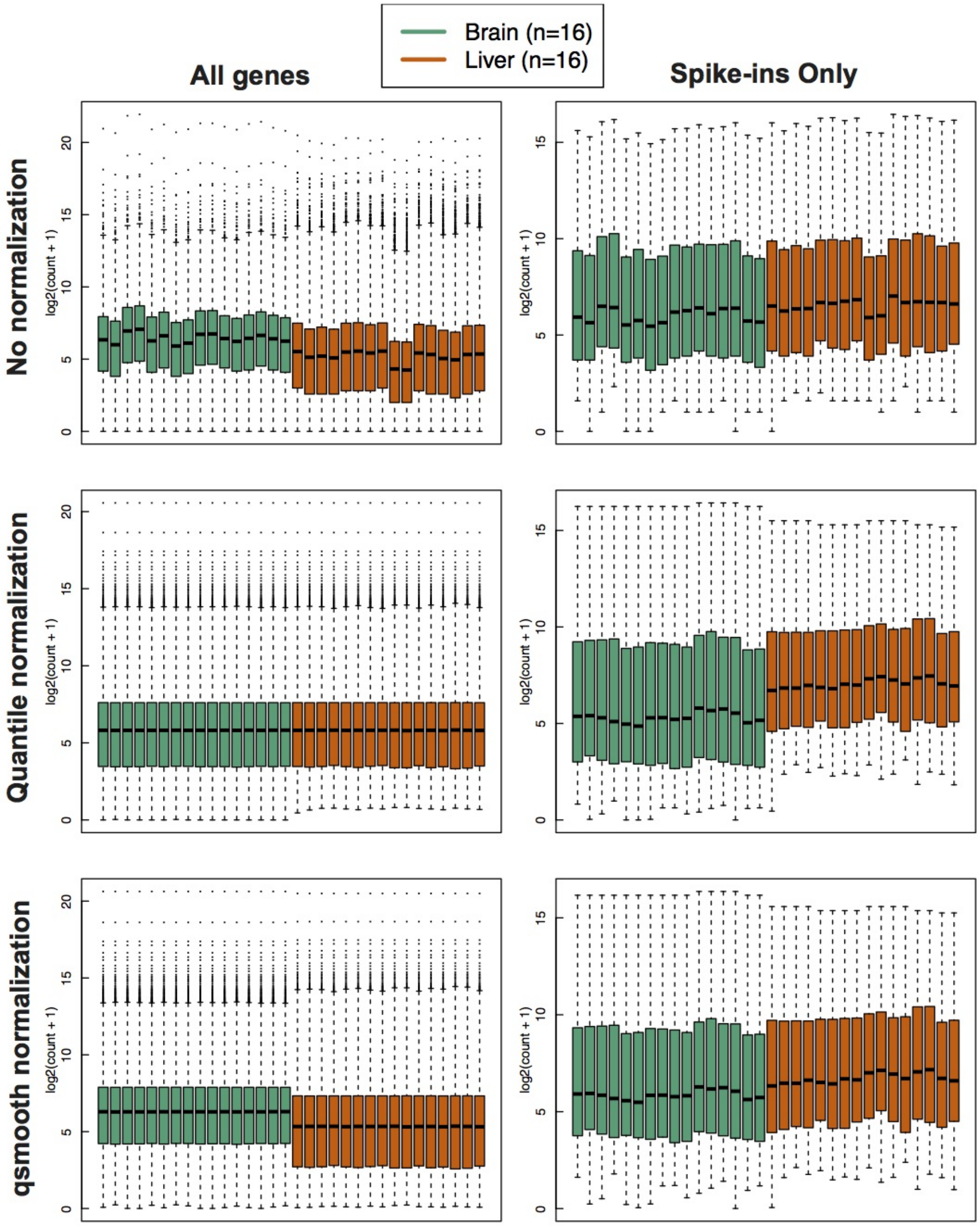
Supplemental Tables

Supplemental Table 1. Abbreviations for GTEx data

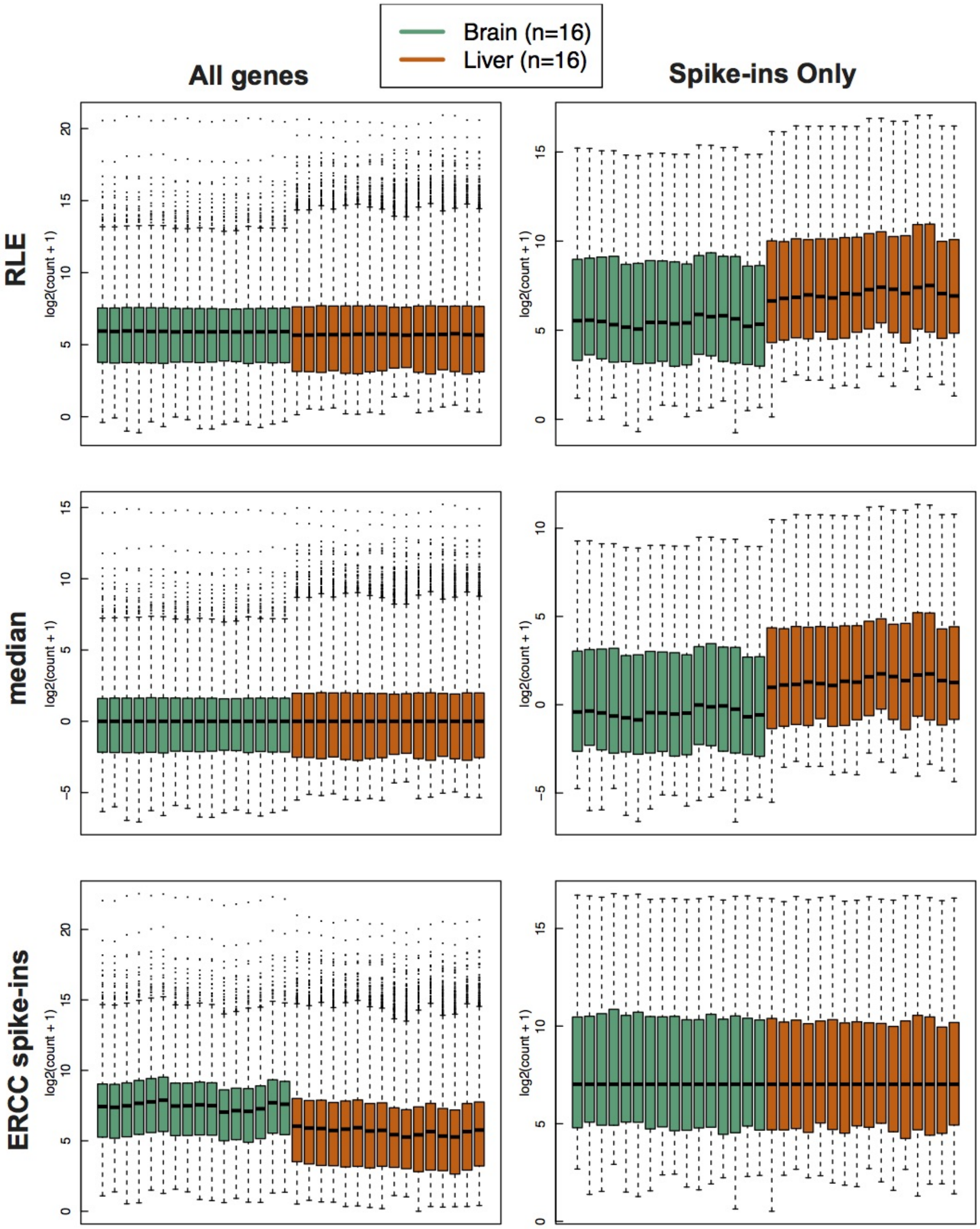
Tissue	Abbreviation	Subtissue
Adipose subcutaneous	ADS	Adipose - Subcutaneous
Adipose visceral	ADV	Adipose - Visceral (Omentum)
Artery aorta	ATA	Artery - Aorta
Artery tibial	ATT	Artery - Tibial
Brain other	BRO	Brain - Amygdala
		Brain - Anterior cingulate cortex (BA24)
		Brain - Cortex
		Brain - Frontal Cortex (BA9)
		Brain - Hippocampus
		Brain - Hypothalamus
		Brain - Spinal cord (cervical c-1)
		Brain - Substantia nigra
Brain cerebellum	BRC	Brain - Cerebellar Hemisphere
		Brain - Cerebellum
Brain basal ganglia	BRB	Brain - Caudate (basal ganglia)
		Brain - Nucleus accumbens (basal ganglia)
		Brain - Putamen (basal ganglia)
Breast	BST	Breast - Mammary Tissue
Fibroblast cell line	FIB	Cells - Transformed fibroblasts
Colon transverse	CLT	Colon - Transverse
Gastroesophageal junction	GEJ	Esophagus - Gastroesophageal Junction
Esophagus mucosa	EMC	Esophagus - Mucosa

Esophagus muscularis	EMS	Esophagus - Muscularis
Heart atrial appendage	HRA	Heart - Atrial Appendage
Heart left ventricle	HRV	Heart - Left Ventricle
Lung	LNG	Lung
Skeletal muscle	SMU	Muscle - Skeletal
Tibial nerve	TNV	Nerve - Tibial
Pancreas	PNC	Pancreas
Skin	SKN	Skin - Not Sun Exposed (Suprapubic)
		Skin - Sun Exposed (Lower leg)
Stomach	STM	Stomach
Testis	TST	Testis
Thyroid	THY	Thyroid
Whole blood	WBL	Whole Blood

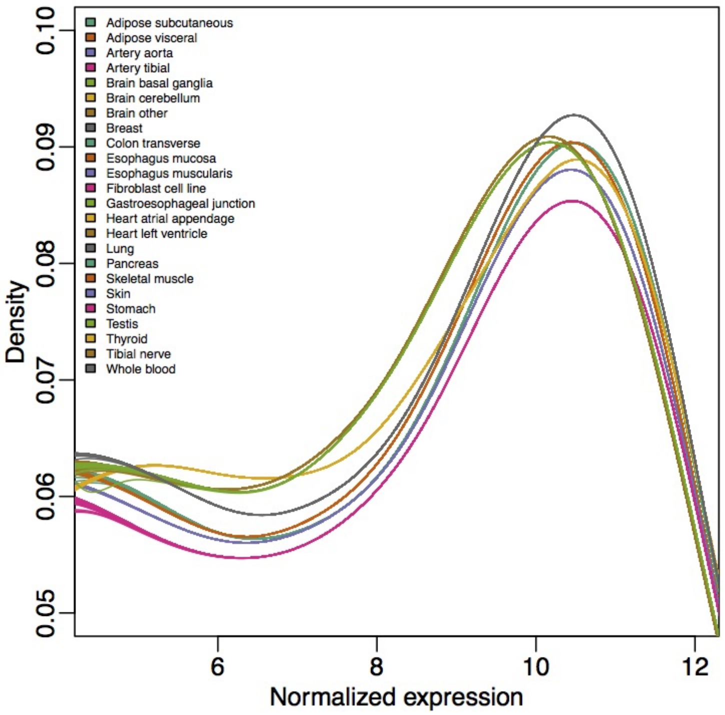
Supplemental Figures



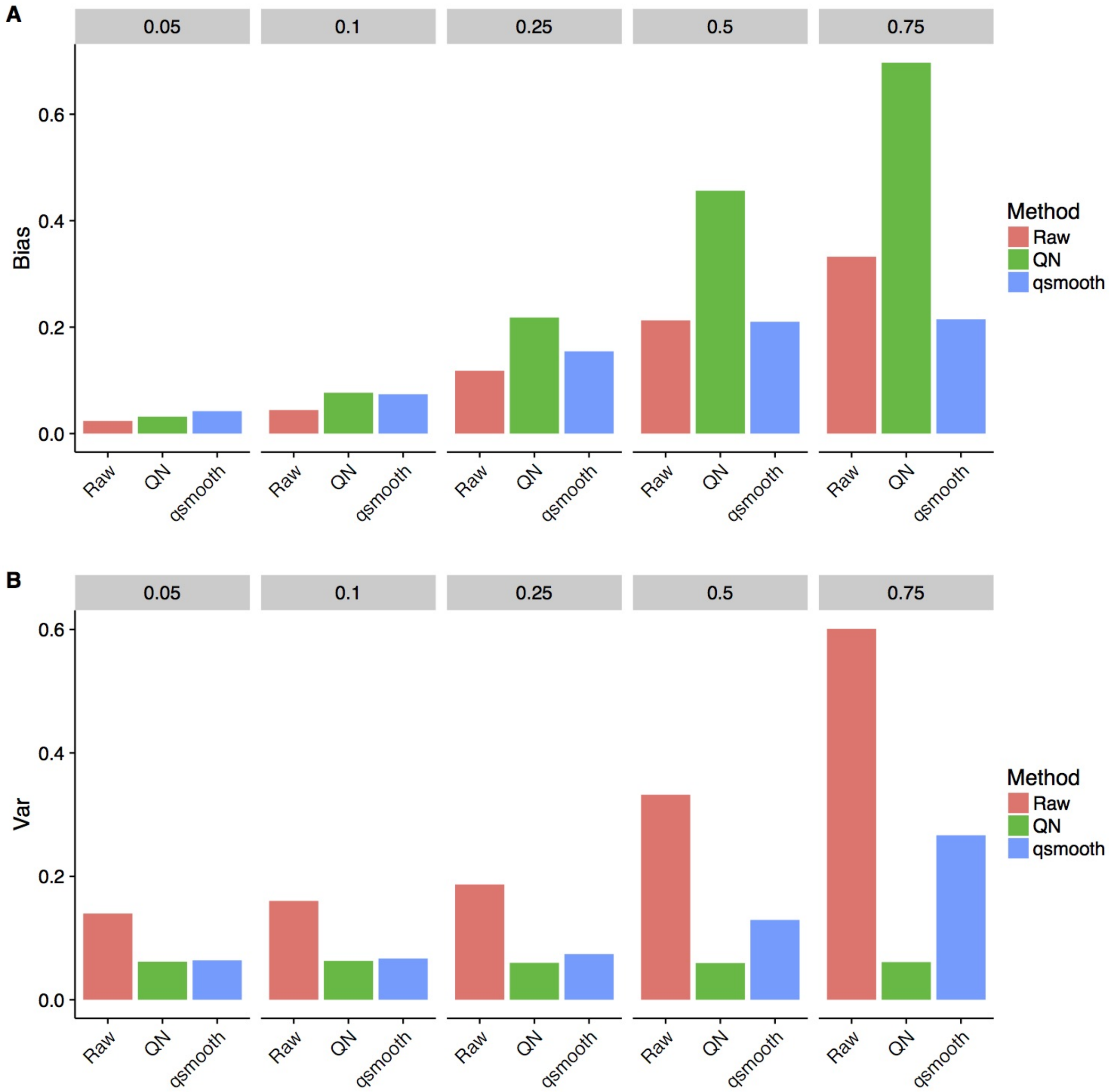
Supplemental Figure 1: Comparing no normalization (row 1), quantile normalization (row 2) and qsmooth (row 3) applied RNA-Seq gene counts from brain (green) and liver (orange) tissues in the bodymapRat dataset. Column 2 contains the boxplots for only the spike-in control genes. Counts have an added pseudocount of 1 and then are $\log_2$ transformed.



Supplemental Figure 2: Comparing scaling normalization methods including Relative Log Expression (RLE) normalization (row 1), median normalization (row 2) and scaling by the ERCC spike-ins (row 3) applied RNA-Seq gene counts from brain (green) and liver (orange) tissues in the bodymapRat dataset. Column 2 contains the boxplots for only the spike-in control genes. Counts have an added pseudocount of 1 and then are $\log_2$ transformed.



Supplemental Figure 3: Densities of the gene expression from the GTEx RNA-sequencing samples colored by tissue. Tissue-specific differences in the gene expression distribution can be seen using qsmooth normalization. Only tissues with at least 150 samples are displayed.



Supplemental Figure 4: Bias and variance (Var) trade-off of the raw high-throughput data (Raw), quantile normalized data (QN), and smooth quantile normalized data (qsmooth). In each column, we simulated data 10 samples from two biological groups while sampling the proportion of differentially expressed genes (pDiff) from a Uniform[0, X] distribution, where X is listed as the column heading in the figure. Under the assumptions of global normalization methods, qsmooth results in less bias compared to quantile normalization, but also less variance compared to using the raw data. As the number of differentially expressed genes increases, quantile normalization and qsmooth both reduce the variance compared to using the raw data, but qsmooth also reduces the bias compared to using the raw data by accounting for global differences between the biological groups when the assumptions of global normalization methods are violated.