RESEARCH ARTICLE

A Mechanistic Beta-Binomial Probability Model for mRNA Sequencing Data

Gregory R. Smith, Marc R. Birtwistle*

Department of Pharmacology and Systems Therapeutics, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America

* marc.birtwistle@mssm.edu

Abstract

A main application for mRNA sequencing (mRNAseq) is determining lists of differentially-expressed genes (DEGs) between two or more conditions. Several software packages exist to produce DEGs from mRNAseq data, but they typically yield different DEGs, sometimes markedly so. The underlying probability model used to describe mRNAseq data is central to deriving DEGs, and not surprisingly most softwares use different models and assumptions to analyze mRNAseq data. Here, we propose a mechanistic justification to model mRNAseq as a binomial process, with data from technical replicates given by a binomial distribution, and data from biological replicates well-described by a beta-binomial distribution. We demonstrate good agreement of this model with two large datasets. We show that an emergent feature of the beta-binomial distribution, given parameter regimes typical for mRNAseq experiments, is the well-known quadratic polynomial scaling of variance with the mean. The so-called dispersion parameter controls this scaling, and our analysis suggests that the dispersion parameter is a continually decreasing function of the mean, as opposed to current approaches that impose an asymptotic value to the dispersion parameter at moderate mean read counts. We show how this leads to current approaches overestimating variance for moderately to highly expressed genes, which inflates false negative rates. Describing mRNAseq data with a beta-binomial distribution thus may be preferred since its parameters are relatable to the mechanistic underpinnings of the technique and may improve the consistency of DEG analysis across softwares, particularly for moderately to highly expressed genes.

OPEN ACCESS

Citation: Smith GR, Birtwistle MR (2016) A Mechanistic Beta-Binomial Probability Model for mRNA Sequencing Data. PLoS ONE 11(6): e0157828. doi:10.1371/journal.pone.0157828

Editor: Marinus F.W. te Pas, Wageningen UR Livestock Research, NETHERLANDS

Received: July 9, 2015

Accepted: June 6, 2016

Published: June 21, 2016

Copyright: © 2016 Smith, Birtwistle. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was funded by the Icahn School of Medicine at Mount Sinai, an IBM Faculty Award, and the NIH Grants P50 GM071558 (Systems Biology Center New York), R01GM104184, R21CA196418 and U54HG008098 (LINCS Center). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Since the advent of the microarray around the turn of the 20th century, whole transcriptome profiling has been of great importance to systems biology [1–8]. The ability to observe how every transcript in a cell population responds to, for example, treatment with a drug or a change in the expression of a gene-of-interest, gives insight into the wiring and function of biological systems. A common method for deriving biological knowledge from such perturbation experiments is to identify lists of differentially expressed transcripts or genes (DEGs) between two (or more) conditions. By analyzing the genes which show up on such lists, one can identify larger functional units such as biological processes, pathways, networks, and organelles that are involved in the response, giving clear hypotheses for further targeted experiments [9–13]. The centralized collection of most transcriptome experiments in databases such as the gene expression omnibus (GEO) and the connectivity map (CMAP) has given further insight by enabling the use of big data methods to identify general trends and connections that do not emerge from a single experiment (or even a handful) [14–16].

While the microarray was the transcriptomic workhorse in the 2000s, the advent of massively parallel sequencing has given rise to deep mRNA sequencing (mRNAseq) [17,18], an alternative way to measure the transcriptome. Like most new technologies, mRNAseq was originally much more expensive than microarrays; however, it has now become quite competitive, and in many ways a superior technical method for transcriptome profiling [19–22]. The basic premise is to isolate mRNA from a sample, PCR amplify it, and then subject it to tens-of-millions of "short" (~50–100 bp typically) sequencing reads. By aligning the resulting sequence reads with the known genome, and then counting the number of reads that align to a particular gene or transcript, one obtains a measurement of expression. One caveat of this traditional form of quantification is the inherent PCR bias that can distort the original number of transcripts in the sample. A recent method based on incorporating a short unique molecular identifier (UMI) sequence into every transcript molecule provides a new method of quantification that reduces PCR bias and thus improves linearity and precision [23–25].

Several open source software suites with associated probability models have been developed to analyze mRNAseq data and identify DEGs. The first was Cufflinks / Cuffdiff [17], which has an elegant underlying mathematical model to estimate the "fragments per kilobase of transcript length per million mapped reads" (FPKM) metric of gene expression, and a t-test based on approximate normality of the resulting FPKM estimate. Cuffdiff2 [26] more accurately estimates false discovery rates for DEGs. Using this FPKM metric, Cuffdiff2 is specialized to a transcript-resolution of gene expression, and comparison across different transcripts, but not to count based data, which we focus on here. Other widely used software suites are EdgeR [27], DESeq2 [28] and BaySeq [29], which, as opposed to the FPKM metric of Cufflinks/Cuffdiff, retains the count-based nature of mRNAseq data and describes it with a negative binomial model (also called Poisson-gamma). This probability model describes mRNAseq count data well, and was predominantly used because it is the common choice to describe count-based data that are "overdispersed" (i.e. variance that is greater than the mean) relative to the Poisson distribution (variance = mean); it is well established that mRNAseq data are overdispersed [30,31]. A recent meta-analysis found that each of these softwares can produce quite different DEGs from the same dataset, a result that is common and not entirely surprising given the different modeling and assumptions used. Further, it was shown that the intersection of DEGs from these softwares are preferred to reduce false positives, which indicates that each might benefit from improvements to the underlying probabilistic treatment of the mRNAseq data [32].

To that end, other probabilistic distributions have been examined. The beta-binomial distribution has also been explored, and it also reflects the overdispersion of the data [33,34]. DEG analysis based upon a beta-binomial distribution is now available as an option for BaySeq solely for paired data (distinct from traditional DEG analyses) [35] and in the software BBSeq [36]; however, a derivation of the mean-variance relationship inherent in the beta-binomial distribution has yet to be undertaken. Furthermore, each software, as with negative-binomial or Poisson methods, has its own specific interpretation of the probabilistic models utilized resulting in often very different selections of DEGs following analysis. This suggests the necessity of a theoretical derivation of an appropriate probabilistic distribution: a ground-up, first-principles approach to modeling the mean-variance relationship and overdispersion which, to date, has not been deeply investigated.

Here, we propose that the basic mRNAseq experimental process is mechanistically a binomial experiment: a series of *N* trials (reads) with an essentially constant probability of success for a particular transcript/gene in each trial. This gives rise to a binomial distribution for counts from technical mRNAseq replicates, with parameters that have physical interpretation. We highlight how this binomial model agrees well with literature data for technical replicates. For biological replicates, we propose that a beta-binomial distribution, where the probability of success follows a beta distribution, can describe the data, and demonstrate its fit to two large literature datasets. Given ranges of beta-binomial parameter values typical for mRNAseq experiments, a quadratic polynomial scaling between variance and mean emerges, as is consistently experimentally observed. The dispersion parameter is the quadratic coefficient that controls this scaling, and our analysis suggests that the dispersion parameter is a continually decreasing function of the mean. Surprisingly, this is different from current approaches that impose an asymptotic value on the dispersion parameter at moderate and high mean read counts. We show how this leads to overestimating variance for moderately to highly expressed genes, which inflates false negative rates in downstream DEG analysis. Because the beta-binomial model emerges from the mechanism of the mRNAseq technique, it may be preferred, and its use might not only help improve consistency in deriving DEGs, but also variance estimation for moderately to highly expressed genes.

Methods

Solving for the Dispersion Parameter

For each gene *i*, we assume $\sigma_{ij}^2 = \mu_{ij} + \varphi_i\mu_{ij}^2$ and solve for φ_i as follows. First, we expand the right hand side of the equation:

$$\mu_{ij} + \varphi_i\mu_{ij}^2 = \frac{N_j\alpha_i}{\alpha_i + \beta_i} + \varphi_i\frac{N_j^2\alpha_i^2}{(\alpha_i + \beta_i)^2} = \frac{N_j\alpha_i(\alpha_i + \beta_i) + \varphi_iN_j^2\alpha_i^2}{(\alpha_i + \beta_i)^2}$$

$$= \frac{N_j\alpha_i^2 + N_j\alpha_i\beta_i + \varphi_iN_j^2\alpha_i^2}{(\alpha_i + \beta_i)^2}$$

Including the left hand side provides the following equation:

$$\sigma_{ij}^2 = \frac{N_j\alpha_i\beta_i(N_j + \alpha_i + \beta_i)}{(\alpha_i + \beta_i)^2(1 + \alpha_i + \beta_i)} = \frac{N_j\alpha_i(\alpha_i + \beta_i + \varphi_iN_j\alpha_i)}{(\alpha_i + \beta_i)^2}$$

After simplifying:

$$\frac{\beta_i(N_j + \alpha_i + \beta_i)}{(1 + \alpha_i + \beta_i)} = \alpha_i + \beta_i + \varphi_iN_j\alpha_i$$

Writing in terms of φ_i:

$$\varphi_i N_j \alpha_i = \frac{\beta_i(N_j + \alpha_i + \beta_i)}{(1 + \alpha_i + \beta_i)} - \alpha_i - \beta_i$$

$$\varphi_i = \frac{\beta_i(N_j + \alpha_i + \beta_i)}{(1 + \alpha_i + \beta_i)N_j\alpha_i} - \frac{\alpha_i + \beta_i}{N_j\alpha_i}$$

$$\varphi_i = \frac{\beta_i(N_j + \alpha_i + \beta_i) - (\alpha_i + \beta_i)(1 + \alpha_i + \beta_i)}{(1 + \alpha_i + \beta_i)N_j\alpha_i}$$

$$\varphi_i = \frac{\beta_i N_j + \beta_i\alpha_i + \beta_i^2 - \alpha_i - \beta_i - \alpha_i^2 - 2\alpha_i\beta_i - \beta_i^2}{(1 + \alpha_i + \beta_i)N_j\alpha_i}$$

After some cancellation, this can be broken into two terms:

$$\varphi_i = \frac{\beta_i N_j - \alpha_i - \beta_i - \alpha_i^2 - \alpha_i\beta_i}{(1 + \alpha_i + \beta_i)N_j\alpha_i} = \frac{\beta_i(N_j - 1)}{(1 + \alpha_i + \beta_i)N_j\alpha_i} - \frac{(1 + \alpha_i + \beta_i)\alpha_i}{(1 + \alpha_i + \beta_i)N_j\alpha_i}$$

$$= \frac{\beta_i}{(1 + \alpha_i + \beta_i)\alpha_i}\frac{(N_j - 1)}{N_j} - \frac{1}{N_j}$$

Since N_j is very large, $\frac{N_j-1}{N_j} \approx 1$ and $\frac{1}{N_j} \approx 0$. Therefore, we find that:

$$\varphi_i \approx \frac{\beta_i}{\alpha_i(1 + \alpha_i + \beta_i)}$$

This corroborates well with our original estimate. For $\beta_i >>> \alpha_i$, $\varphi_i \approx \frac{\beta_i}{\alpha_i(1+\alpha_i+\beta_i)} \approx \frac{\beta_i}{\beta_i\alpha_i} = \frac{1}{\alpha_i}$.

Downloading mRNAseq Data

UMI count data were obtained from the DToXS LINCS website (http://research.mssm.edu/pst/DToxS) on July 1st, 2015, from DToXS LINCS ID Raw-Data-R2015-06-30. Raw (Level 1) transcriptomic data released June 30th, 2015 were downloaded, and data from batch identifier SR-1 were used in this study. There were 15 control samples (with sample name prefix CTRL), but the sample CTRL.1.C1 was excluded because it showed poor correlation with the remaining 14 samples. There were six samples treated with the kinase-inhibitor Sorafenib, (SOR), but samples 1 and 3 were excluded as they had poor correlation compared to the remaining four. Gierlinski yeast data were acquired from the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/data/view/ERP004763) consisting of 672 fastq files: 2 cell lines each with 48 biological replicates each with 7 technical replicates. Raw reads from the fastq files were then aligned using Bowtie [37] against the Saccharomyces cerevisiae genome removing reads with multiple alignments to the genome. Aligned reads were then sorted using Samtools [38] and converted into files of gene read counts using Bedtools [39]. We followed the author's method for removing "bad replicates" that did not satisfy a quality score based upon median correlation coefficient, outlier fraction and median reduced χ^2 of pileup depth. We corroborated their calculations and removed six WT biological replicates (21, 22, 25, 28, 34, 36) and four Δsnf2 biological replicates (6, 13, 25, 35) just as they had done. All raw data are given in S1–S4 Tables.

Estimating Beta-Binomial Distribution Parameters

First, the integer count data in S1–S4 Tables were divided by their respective sequencing depth, which was calculated by summing the counts along a single column (sample). The resulting probability of success estimates for each gene were fit to a beta distribution using method of moments estimates for α and β as follows:

$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right)$$

$$\hat{\beta} = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1\right)$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the sample mean and $\bar{v} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{x})^2$ is the sample variance. These α and β parameter estimates for each gene are also given in S1–S4 Tables.

Data Normalization

We normalize the data by scaling each sample to have an equivalent sequencing depth as the sample with the maximum sequencing depth. That is, we take $\bar{N} = \max(N_j)$ and for each sample j, the normalized read counts are:

$$\bar{k}_{ij} = k_{ij}\frac{\bar{N}}{N_{ij}}$$

Estimation of Dispersion

To obtain a smooth trend of dispersion that follows the data as implied by our beta-binomial formulation, we fit an empirical quadratic polynomial to the plot of log(mean) vs log(dispersion) using the MATLAB fit tool ($y = p1^*x^2+p2^*x+p3$). The parameter values for each data set, in order of (p1,p2,p3) are Gierlinski WT (0.06, -0.90, 0.236), Gierlinski Δsnf2 (0.04, -0.93, 0.26), LINCS Mapped Reads (0.007, -0.83, 0.57), and LINCS UMI (0.020, -0.96, 0.26).

To compare our approach of modeling dispersion with previous methods, we downloaded the R packages DESeq2, Version 1.12.2 [28], and EdgeR, Version 3.14.0 [27]. For DESeq2, we uploaded each data set and used the *estimateDispersions* command which generates three separates formulations of dispersion for each gene: *dispGeneEst* reflects the raw dispersion estimate from the data, *dispFit* represents a curve fit to the dispGeneEst data following the distribution expected by DESeq2 and lastly *dispersion* which is a modified version of *dispGeneEst* with outliers corrected to reflect the trend of *dispFit* values. For the purposes of our work, we use the *dispersion* value for each gene in each data set as that is the recommended setting by DESeq2. For EdgeR, we use the *estimateDisp* command which also generates three dispersion estimates for each gene: *common.dispersion* is a single value over all genes as a best estimate of global dispersion, t*rended.dispersion* represents a curve fit to genewise dispersion similar to DESeq2's *dispFit*, and *tagwise.dispersion* is a gene-specific estimate of dispersion that is modified to reflect the value in *trended.dispersion* again similar to DESeq2's *dispersion* value. For our work, we chose the *tagwise.dispersion* value for each gene.

Estimation of p-values

For each method of acquiring a dispersion estimate, we calculate an estimated variance dependent upon the mean by solving the formula $\sigma_{ij}^2 = \mu_{ij} + \varphi_i\mu_{ij}^2$ given the normalized mean read

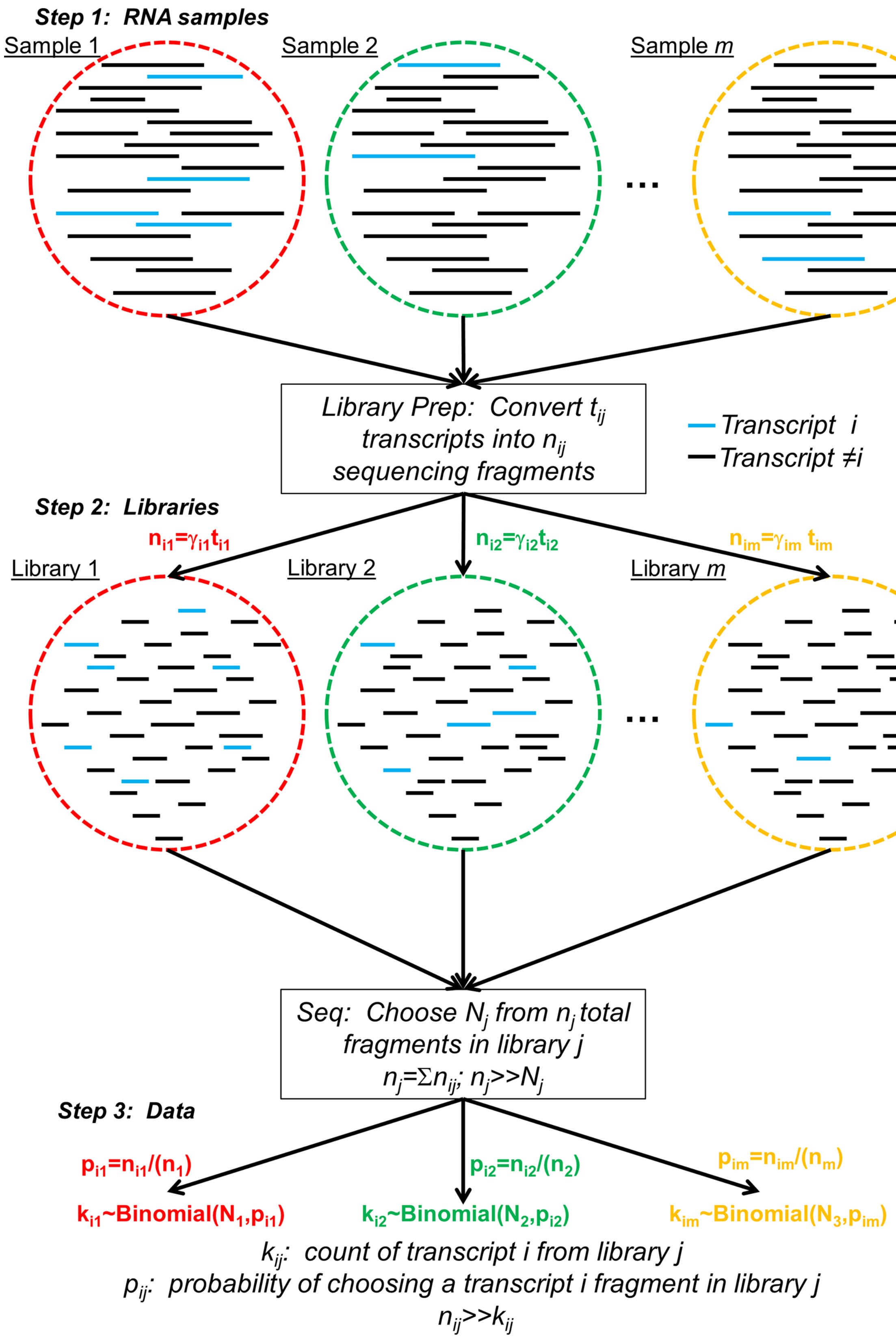
Step 1: RNA samples
Sample 1
Sample 2
Sample *m*
...
Library Prep: Convert t_{ij} transcripts into n_{ij} sequencing fragments
Transcript *i*
Transcript ≠*i*
Step 2: Libraries
$n_{i1}=\gamma_{i1}t_{i1}$
$n_{i2}=\gamma_{i2}t_{i2}$
$n_{im}=\gamma_{im}\ t_{im}$
Library 1
Library 2
Library *m*
...
Seq: Choose N_j from n_j total fragments in library *j*
$n_j=\Sigma n_{ij}$; $n_j>>N_j$
Step 3: Data
$p_{i1}=n_{i1}/(n_1)$
$p_{i2}=n_{i2}/(n_2)$
$p_{im}=n_{im}/(n_m)$
k_{i1}~Binomial(N_1,p_{i1})
k_{i2}~Binomial(N_2,p_{i2})
k_{im}~Binomial(N_3,p_{im})
k_{ij}: count of transcript *i* from library *j*
p_{ij}: probability of choosing a transcript *i* fragment in library *j*
$n_{ij}>>k_{ij}$

Fig 1. Schematic of the General mRNAseq Process. There are three main steps depicted here, from top to bottom. First is obtaining RNA samples, which contain full length transcripts. Different samples are denoted by different color circles, and transcripts by straight lines within those circles. We highlight one transcript blue to enable following it through the process. Next, library preparation converts the transcripts in each sample to a library of fragments that can be sequenced. Finally, the libraries are sequenced by choosing fragments from the library, and the number of reads that align to particular transcripts are counted for the readout of expression.

doi:10.1371/journal.pone.0157828.g001

counts μ_{ij} and dispersion estimate φ_i for each gene i in each dataset. Then we conduct a Welch's t test for the hypotheses that the UMI CTRL and SOR samples have the same mean for a given gene and that the Gierlinski WT and Δsnf2 mutant samples have the same mean for a given gene. To do this, we modified the Matlab method ttest2 to accept as input parameters an estimate for the mean and variance for each sample as opposed to the normalized read counts themselves generating a p value for each gene in each dataset. This is to show how different estimates of dispersion, and thus different estimates of variance, affect the resulting p values for each gene tested in each dataset.

Results and Discussion

mRNA Sequencing as a Binomial Experiment

An mRNA sequencing (mRNAseq) experiment consists of three main steps (Fig 1). First is isolating mRNA from biological samples (sample index $j \in \{1,2,\ldots,m\}$). Second, the mRNA samples are converted into a library that is compatible with the sequencing platform. This often includes fragmenting the original mRNA molecules, along with one or more PCR steps, into n_j total fragments (sometimes isolation of mRNA from total RNA is part of the library preparation). Let the number of molecules from a particular transcript i in the library j be $n_{ij} = \gamma_{ij}t_{ij}$, where i is the transcript index, t_{ij} is the original number of transcript i molecules in library j, $\gamma_j \geq 0$ is the amplification factor, and $n_j = \sum_i n_{ij}$.

The library is then subjected to the sequencing process, where N_j of the n_j library molecules are randomly chosen for sequencing. The number of trials N_j is often called the sequencing depth.

The probability of choosing a molecule for sequencing from library j that maps to transcript i is (except in relatively rare cases of capture bias)

$$p_{ij} = \frac{n_{ij}}{n_j} \tag{1}$$

Denote p_{ij} as the probability of success for transcript i in library j. If the total number of molecules in the library far exceeds the total number of reads ($n_j >> N_j$), then "taking" a fragment from the library for sequencing has negligible effect on this probability, making it essentially constant throughout the selection process. For the common Illumina platform, $n_j \sim 10^9$ library molecules are loaded onto the instrument (e.g. ~75 μL of a 20 pM library), and a typical sequencing depth for an mRNAseq experiment is $N_j \sim 10^7$ reads, giving $n_j >> N_j$ and essentially constant p_{ij} for all but the few most lowly expressed transcripts.

An mRNAseq experiment with library j can thus be cast as a series of N_j trials, with each trial selecting one library fragment for sequencing. We define a trial to be a success for transcript i if a fragment subsequently aligned to it is chosen for sequencing; the probability of success is p_{ij}. This scenario, as described, is analogous to a binomial experiment [40]. Therefore, the probability of selecting k_{ij} fragments from library j that map to transcript i should follow a binomial distribution,

$$k_{ij} \sim Binomial(N_j, p_{ij}). \tag{2}$$

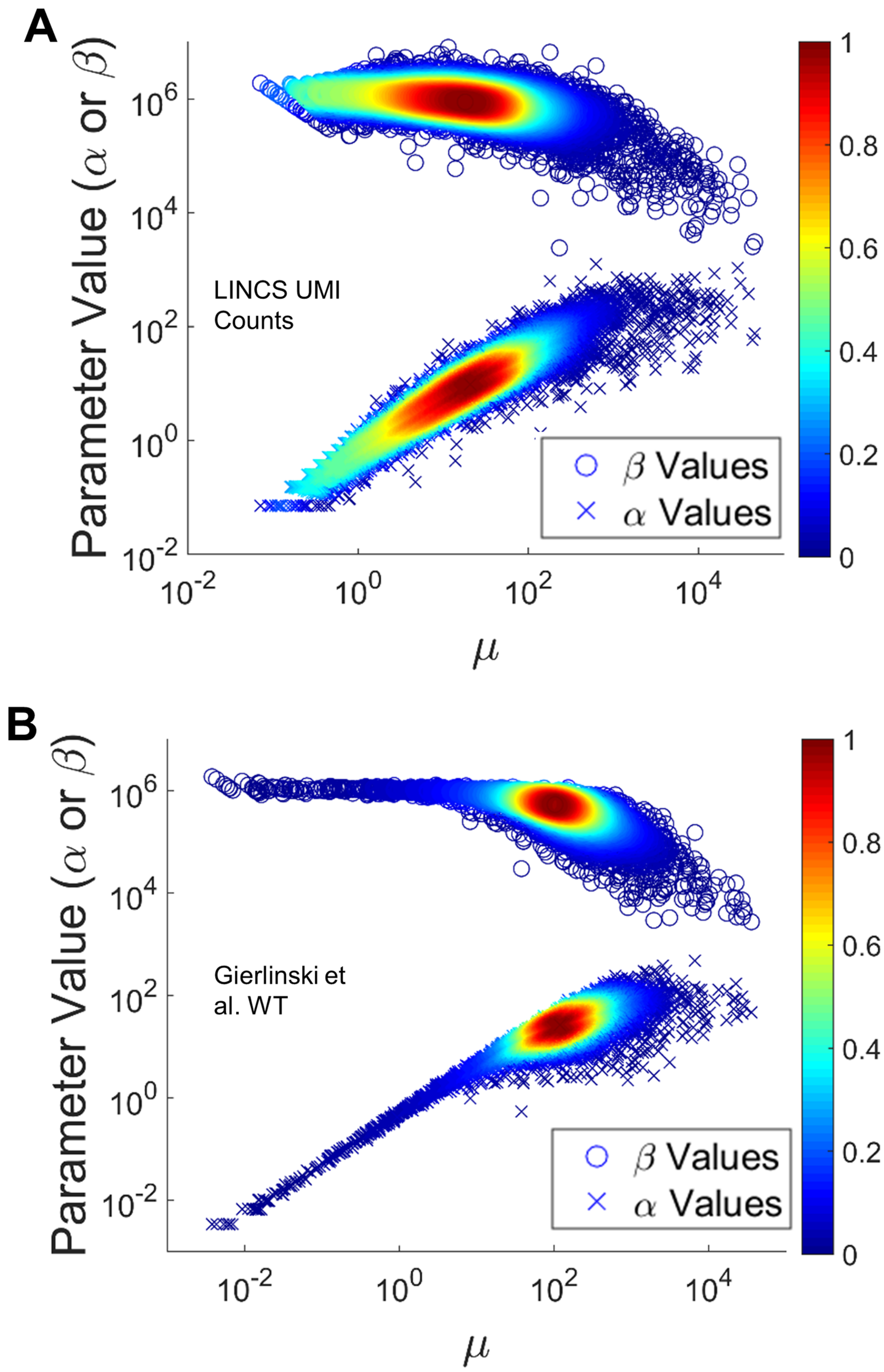

Fig 2. Estimated α and β values Plotted Against the Mean for each Gene. Each panel is a log-scale scatter plot of mean vs α and β over all genes for one of the following datasets tested: LINCS UMI (A) and Gierlinski WT (B). The results for the two remaining datasets are shown in S1 Fig. The x's reflect α values and the circles reflect β values with color dependent upon the density of points in the scatter plot.

doi:10.1371/journal.pone.0157828.g002

The random variable k_{ij} is often referred to as the number of uniquely mapped reads to transcript i, and has mean $\mu = N_j \cdot p_{ij}$ and variance $\sigma^2 = N_j \cdot p_{ij} \cdot (1-p_{ij})$. In general, $p_{ij} << 1$ due to the large number of different expressed transcripts in a cell (typically ~10,000 [41,42] and see non-zero entries in S1 and S2 Tables). This gives $\mu = \sigma^2$ for most transcripts, as one expects from a Poisson distribution. This is in excellent agreement with data from technical replicates sequenced from the same library [22], giving direct experimental support for the notion that the mRNAseq process can be cast as a binomial experiment.

Describing Inter-Library Variability with a Beta-Binomial Distribution

When mRNAseq experiments are performed across biological replicates which have different libraries, the probability of success for a transcript varies. Dividing the number of mapped reads for a transcript by the sequencing depth N_j gives an estimate of the true (inter-library) probability of success, p_i. Because p_i is continuous on the unit interval ($0 \le p_i \le 1$), a potentially suitable model is a beta random variable [40], with density

$$f(p_i) = \frac{p_i^{(\alpha_i-1)}(1-p_i)^{(\beta_j-1)}}{B(\alpha_i,\beta_i)} \tag{3}$$

where B denotes a Beta function of the first kind and α_{i} and β_i are parameters to be estimated from biological replicates. The expected value of p_i is

$$E[p_i] = \frac{\alpha_i}{\alpha_i+\beta_i} = \frac{E[n_{ij}]}{n_j}. \tag{4}$$

We have also used Eq 1 and the fact that the total number of library molecules is essentially constant across libraries, due to concentration normalization during loading.

When the probability of success for a binomial random variable follows a beta distribution, the resulting random variable is said to follow a beta-binomial distribution. The mean and variance of a beta-binomial distribution are, respectively [43]

$$\mu_{ij} = \frac{N_j\alpha_i}{\alpha_i+\beta_i} \tag{5}$$

$$\sigma_{ij}^2 = \frac{N_j\alpha_i\beta_i(N_j+\alpha_i+\beta_i)}{(\alpha_i+\beta_i)^2(1+\alpha_i+\beta_i)} \tag{6}$$

Table 1. CV^2 and LS fits for the dispersion parameter φ for each dataset under raw and normalized conditions. R^2 values are also included for the quality of the corresponding fit to the raw data.

Dataset	Processing	CV^2 φ Fit	CV^2 R^2	LS φ Fit	LS R^2
LINCS MR	Raw	.0700	.9218	.0898	.9687
	Normalized	.0101	.7084	.0118	.7229
LINCS UMI	Raw	.0785	.9703	.0867	.9791
	Normalized	.0099	.7205	.0098	.7205
Gier WT	Raw	.0815	.9909	.0799	.9913
	Normalized	.0271	.8083	.0173	.8974
Gier SNF2	Raw	.0684	.9794	.0606	.9961
	Normalized	.0159	.7982	.0118	.9078

doi:10.1371/journal.pone.0157828.t001

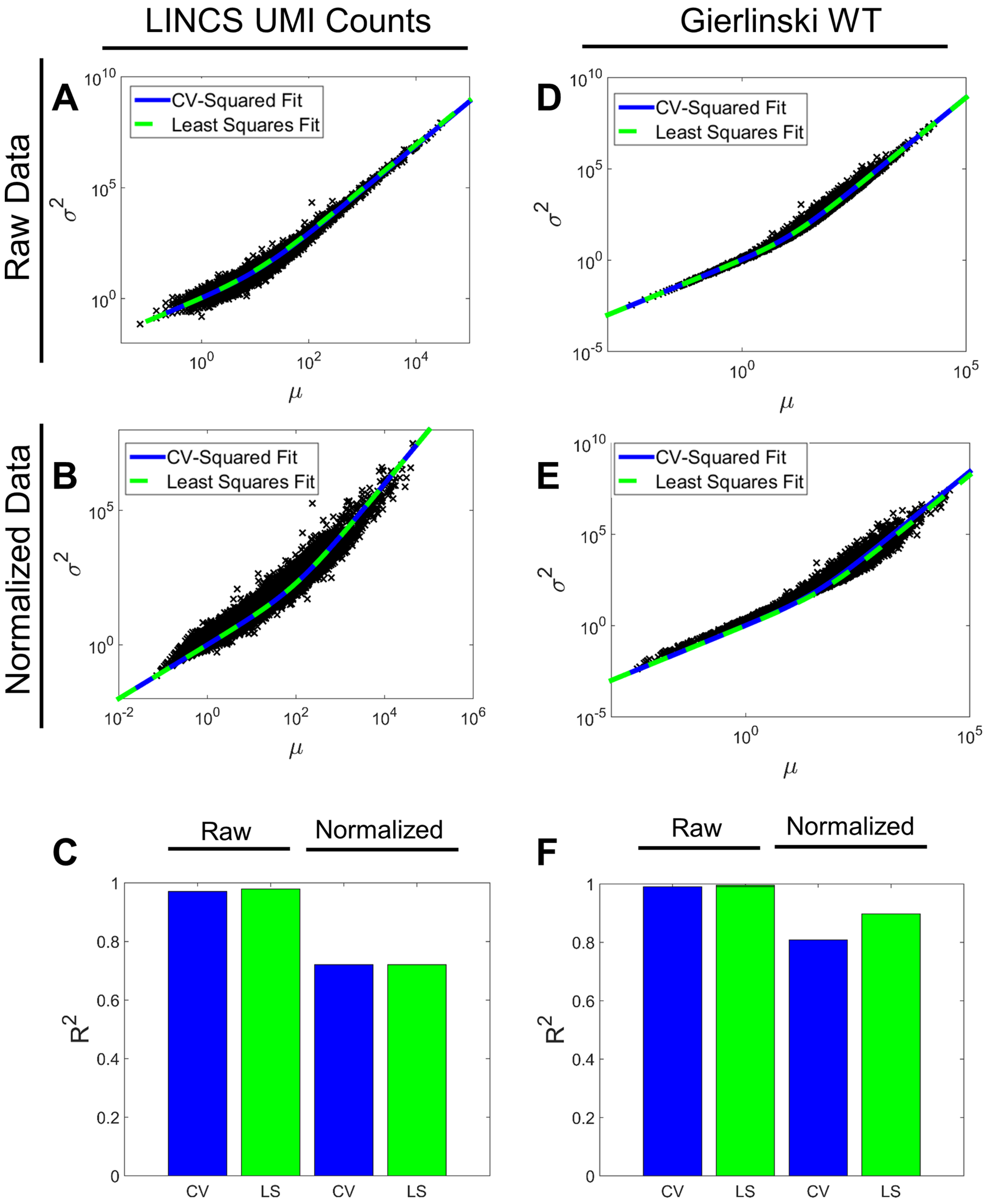
LINCS UMI Counts
Gierlinski WT
Raw Data
Normalized Data
A
10^{10}
10^{5}
10^{0}
σ^2
CV-Squared Fit
Least Squares Fit
10^{0}
10^{2}
10^{4}
μ
D
10^{10}
10^{5}
10^{0}
10^{-5}
σ^2
CV-Squared Fit
Least Squares Fit
10^{0}
10^{5}
μ
B
10^{5}
10^{0}
σ^2
CV-Squared Fit
Least Squares Fit
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}
μ
E
10^{10}
10^{5}
10^{0}
10^{-5}
σ^2
CV-Squared Fit
Least Squares Fit
10^{0}
10^{5}
μ
C
Raw
Normalized
R^2
1
0.8
0.6
0.4
0.2
0
CV
LS
CV
LS
F
Raw
Normalized
R^2
1
0.8
0.6
0.4
0.2
0
CV
LS
CV
LS

Fig 3. Mean-Variance Relationship for Raw and Normalized mRNAseq Data. Each column of three panels reflects one of the following datasets tested: LINCS UMI (A-B) and Gierlinski WT (D-F). The two remaining datasets are shown in S2 Fig. For each column of three panels, the first panel (A,D) shows the CV^2 fit (solid blue line) and Least Squares fit (dashed green line) to the raw data points plotting mean vs variance (black x's). The second panel (B,E) shows the same fits for the normalized data. The third panel (C,F) shows the respective R^2 values for the CV^2 and Least Squares (LS) fits for the raw and normalized data.

doi:10.1371/journal.pone.0157828.g003

As described above, predominantly, $p_i << 1$. Given Eq 4, this implies that $\beta_i >> \alpha_i$ for the majority of transcripts. Moreover, since the number of molecules in the library n_j is much greater than 1, it is likely that $\beta_i >> 1$. Given these considerations, the mean and variance reduce to

$$\mu_{ij} \approx \frac{N_j \alpha_i}{\beta_i} \tag{7}$$

$$\sigma_{ij}^2 \approx \frac{N_j \alpha_i}{\beta_i} + \frac{N_j^2 \alpha_i}{\beta_i^2} = \mu_{ij} + \frac{1}{\alpha_i} \mu_{ij}^2 \tag{8}$$

This reveals a characteristic scaling prediction between the mean and the variance via a "dispersion parameter" $1/\alpha_i$. Such scaling has indeed been well-described for mRNAseq experiments [27,28,30,31]. The full functional form for the dispersion parameter given a beta-binomial distribution is given in the Methods section.

Evaluating the Beta-Binomial Model with Data from Multiple Biological Replicates

Two large mRNAseq datasets were utilized to evaluate the beta-binomial model proposed above. The first is available via the Library of Integrated Network-Based Cellular Signatures (LINCS) (see Methods—DToXS LINCS ID Raw-Data-R2015-06-30). The dataset consisted of 14 biological replicate samples (RNA isolated from independent cell batches) of PromoCell cardiomyocyte-like cells treated under control (DMSO/vehicle) conditions (S1 and S2 Tables). The sequencing libraries were prepared using unique molecular identifiers (UMI) [23–25], which allows removal of PCR biases (by experimentally estimating the γ_{ij} factor—see Fig 1) via quantification by UMI counts, on the level of genes. We refer to this metric as "Unique UMI Counts". It is also possible to retain quantification by the traditional means of counting the number of reads that uniquely align to a gene. We refer to this metric as "Unique Mapped Read Counts". The beta distribution parameters for each gene were estimated as described in Methods from the 14 biological replicates.

A second mRNAseq dataset developed by Gierlinski et al is available on the ENA archive (see Methods - project ID PRJEB5348), consisting of 48 biological replicate samples in two *S. cerevisiae* lines: WT and snf2 knock-out mutant [44]. The replicates underwent standard Illumina multiplexed TruSeq library preparation. Each biological replicate consists of seven technical replicates producing 336 datasets in each cell line resulting in "Unique Mapped Read Counts" (S3 and S4 Tables). As with the LINCS data, the beta distribution parameters for each gene were then estimated for each cell line as described in the Methods.

We first sought to understand the space of estimated α and β parameters for the datasets studied. Given the relationship between the beta distribution parameters and expected value for the probability of success in Eq 4, one would predict that β_i should remain relatively constant across genes, since most transcript types are a very small fraction of the total number of transcripts in a cell. Furthermore, we would like to evaluate the assumption above that $\beta_i >> \alpha_i$. Fig 2 shows log scale plots of α and β values plotted against the mean for two sets of count

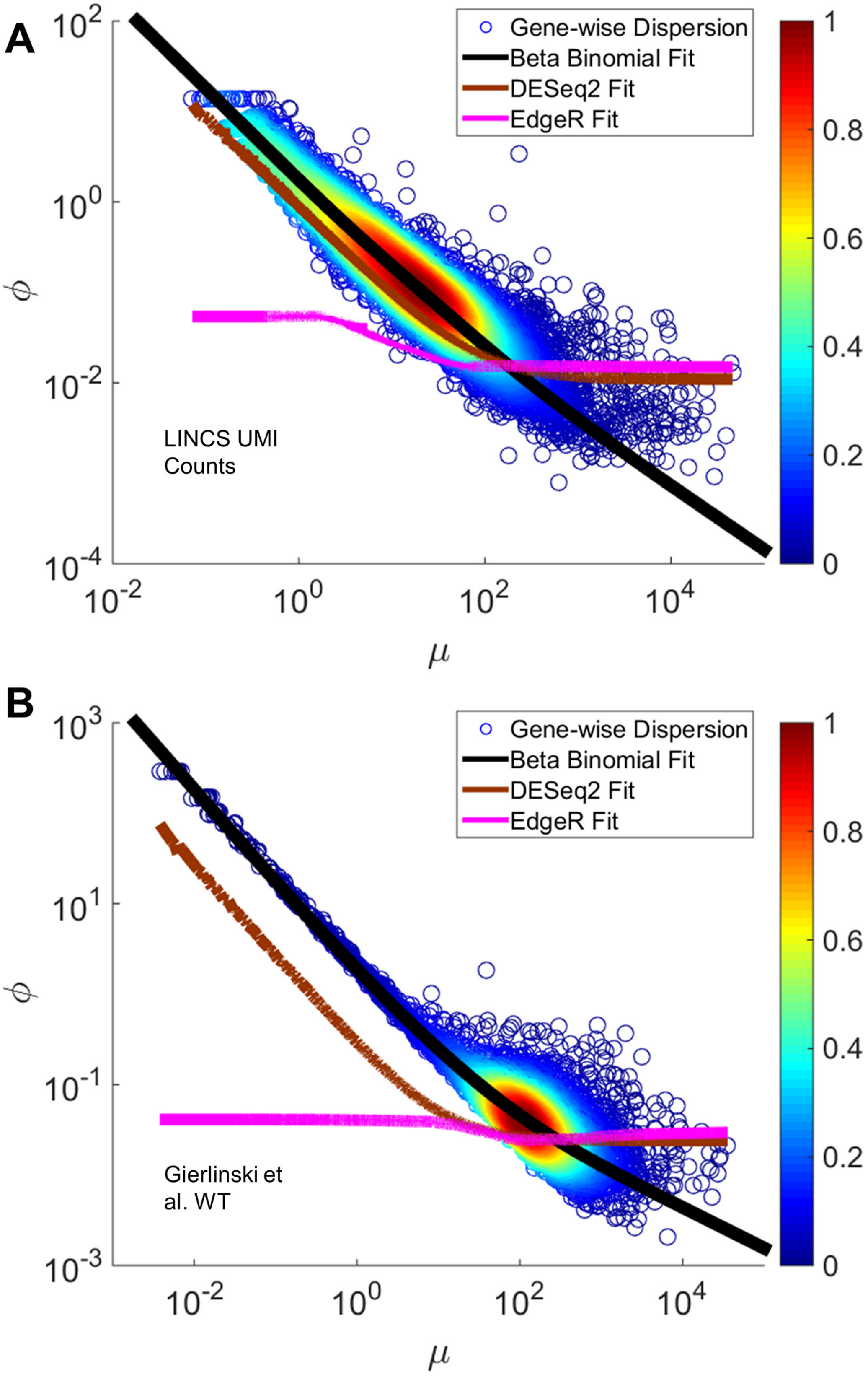
A
Gene-wise Dispersion
Beta Binomial Fit
DESeq2 Fit
EdgeR Fit
LINCS UMI Counts
ϕ
μ
10^{2}
10^{0}
10^{-2}
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
1
0.8
0.6
0.4
0.2
0
B
Gene-wise Dispersion
Beta Binomial Fit
DESeq2 Fit
EdgeR Fit
Gierlinski et al. WT
ϕ
μ
10^{3}
10^{1}
10^{-1}
10^{-3}
10^{-2}
10^{0}
10^{2}
10^{4}
1
0.8
0.6
0.4
0.2
0

Fig 4. Comparing Beta-Binomial Dispersion with DESeq2 and EdgeR Dispersion Estimates. Each panel reflects one of the following datasets tested: LINCS UMI (A) and Gierlinski WT (B). The remaining two datasets are shown in S3 Fig. Each panel shows a density scatter plot of mean versus dispersion values for each gene in each sample. The black line represents our fit showing the non-asymptotic relationship between mean and variance (see Methods). The brown line shows the DESeq2 dispersion fit while the magenta line shows the EdgeR dispersion fit (see Methods).

doi:10.1371/journal.pone.0157828.g004

data: the LINCS UMI Counts (Fig 2A) and the Gierlinski Yeast WT Mapped Read Counts (Fig 2B). Two further sets of count data are shown in S1 Fig: the LINCS Mapped Read Counts (S1A Fig) and the Gierlinski Yeast Δsnf2 Mapped Read Counts (S1B Fig). In each panel, α values are represented by x's and β values are represented by circles. First, we observe that β values are indeed significantly larger than α values for all genes tested. Second, β is largely invariant across the transcriptome, consistent with expectations, only slightly decreasing for genes at higher counts (relative to changes in α values). With more typical mRNAseq datasets where one might expect to have three or even fewer replicates, this result implies that a global fit of β across genes may be quite appropriate, similar to "information sharing" approaches of current softwares [27,28]. This might allow improved estimation of the dispersion parameter for each gene, particularly for those with low abundance, which is critical for estimation of variance and downstream differential expression testing [27,28,30,31]. Lastly, it is clear that the mean is largely determined by α, implying that dispersion is strongly linked to the mean.

We next evaluated whether the beta-binomial model captured the mean-variance structure of the mRNAseq data, which is critical for determining differential expression. Here, we focus on a global gene-independent dispersion parameter, and explore gene-specific dispersion parameters subsequently. We calculated the mean and variance for each gene in each of the datasets studied and compared this to the Eq 8 prediction given a beta-binomial model and one of two global estimates for the dispersion parameter. The first estimate for dispersion is based on previous approaches: CV^2 [27]. The second estimate utilizes least squares (LS) regression. We made this comparison for each dataset both before and after a simple scaling normalization procedure (see Methods) to account for differences in sequencing depth between samples. Table 1, Fig 3 and S2 Fig show the dispersion estimates based upon the two procedures and their respective R^2 values. Genome-wide estimated dispersion values are very close for the LS and CV^2 fits. However, R^2 values are only high when fitting to the raw and not read-depth normalized data. This observation, along with Eq 8, suggests that the dispersion parameter strongly depends on the mean.

Relationship Between Dispersion and Mean

Previous work allows for gene-specific estimation of dispersion [27,28], which imposes a relationship where the gene-specific dispersion parameter asymptotes to a lower bound as mean increases. This relationship derives from the widely accepted quadratic function between variance and mean. This fixed lower bound of dispersion is sometimes called the biological squared coefficient of variation [27], and typically reaches this lower limit at moderate read counts.

The beta-binomial model makes a different prediction about the dependence of dispersion with the mean. Namely, because increases in mean are predominantly driven by increases in α (β is mostly constant across genes), and the dispersion parameter is essentially inversely proportional to α (Eq 8), then we expected the dispersion parameter to be smaller than that imposed by the currently used formalisms in DESeq2 and EdgeR. We compared the beta-binomial dispersion trends with those calculated by DESeq2 and EdgeR (Fig 4 and S3 Fig) for both datasets analyzed above, along with direct estimates of dispersion based on the data themselves. The results indeed displayed evidence that current estimation methods were overestimating

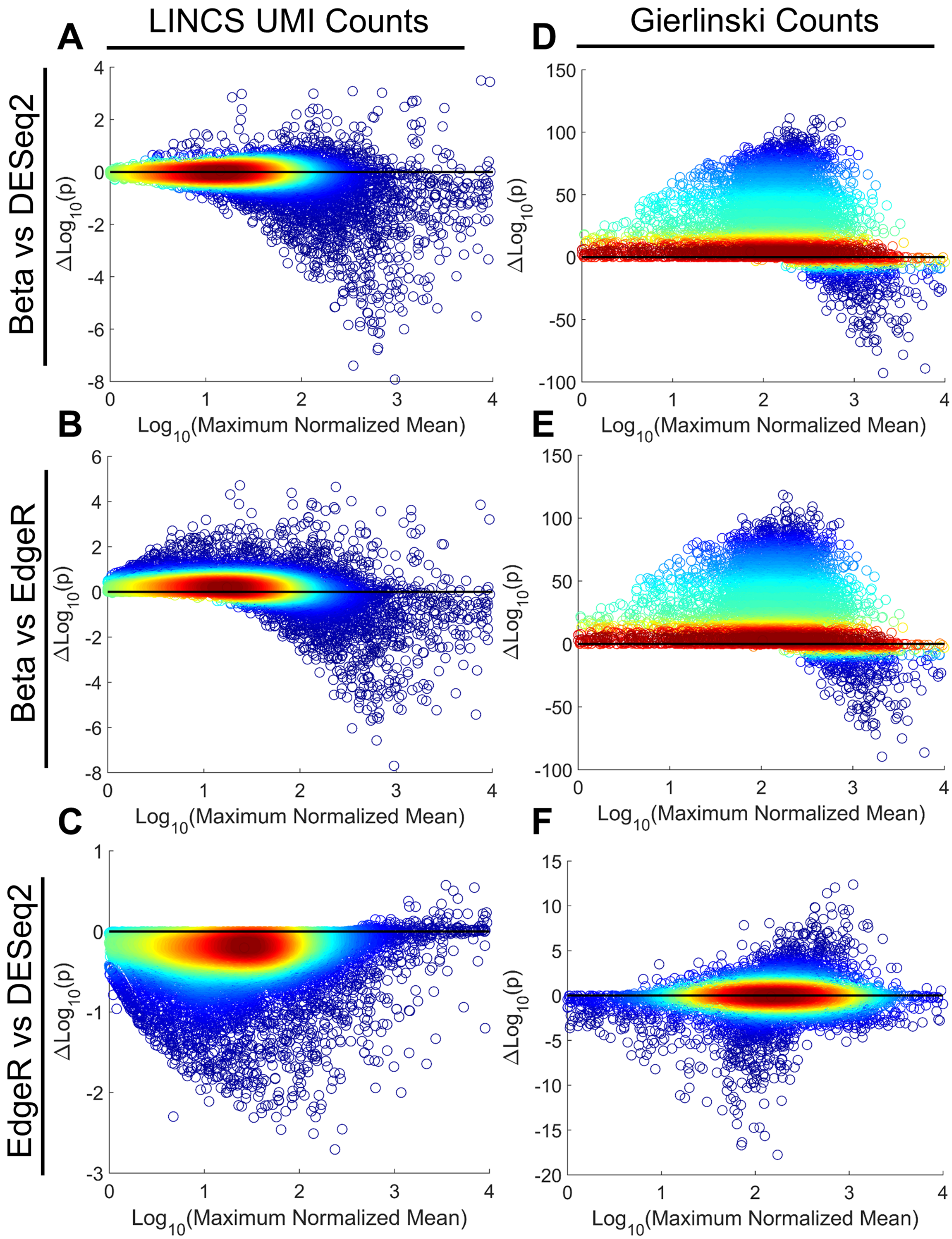
LINCS UMI Counts
Gierlinski Counts
A
B
C
D
E
F
Beta vs DESeq2
Beta vs EdgeR
EdgeR vs DESeq2
ΔLog10(p)
Log10(Maximum Normalized Mean)
4
2
0
-2
-4
-6
-8
6
1
-1
-2
-3
150
100
50
-50
-100
15
10
5
-5
-10
-15
-20
0.2
0.4
0.6
0.8

Fig 5. Differential p-values for Negative Binomial vs. Beta-Binomial Dispersion Methods. Each panel reflects a comparison of p-values for beta binomial-based dispersion or negative binomial-based dispersion generated from the UMI count data, CTRL vs SOR (A-C), or the Gierlinski data, WT vs Δsnf2 (D-F). Each panel is a scatter plot of the base-10 logarithm of the maximum normalized mean (maximum of the CTRL mean or SOR mean for UMI or the WT mean or Δsnf2 mean for Gierlinski) against the difference in base-10 logarithm of the corresponding p-values being compared for each gene. Color indicates density of points. The top row compares the beta binomial formulation versus DESeq2 (A,D). The second row compares beta binomial versus EdgeR (B,E). The third row compares EdgeR and DESeq2 (C,F).

doi:10.1371/journal.pone.0157828.g005

dispersion at read counts starting at ~100 (5–10% of the genes). We conclude that a beta-binomial representation of mRNAseq data might allow for more precise estimation of gene-specific dispersion, and further that current methods might overestimate dispersion and therefore variance for moderately to highly expressed genes. This may have implications for downstream DEG analysis, since a larger variance would lead to a higher false negative rate.

Statistical Significance of Moderately to Highly Expressed Genes

To demonstrate explicitly how overestimating dispersion could lead to identification of new DEGs, we explored a comparison of treated vs. control data for the UMI data set (DMSO vs. sorafenib) and the Gierlinski dataset (WT vs. Δsnf2). We expected that for genes with moderate to high mean read counts, we would have on average higher statistical significance than current negative binomial based methods. As representative of negative binomial methods we used DESeq2 and EdgeR. Fig 5 shows precisely this prediction; as mean read counts increase, the p-values calculated for dispersion estimates of a beta-binomial model are much lower than that from typical negative binomial models. This is evidenced by a preponderance of data below zero on the difference of p-value scatter plots above 100 counts for UMI, and 200 for Gierlinski (Fig 5). This leads to several new genes being called as DEGs, which gives rise to potential new biology being uncovered. Specifically, 597 genes from the Gierlinski dataset and 1023 genes from the LINCS dataset (S5 and S6 Tables). Thus, not only does the beta binomial distribution better capture the statistical dispersion properties of mRNAseq data, but it also has biologically meaningful implications.

Conclusions

Use of mRNAseq to measure transcriptomes is expected to increase, and derivation of DEGs is essential for extracting knowledge from such data. There is no uniform agreement on what probabilistic assumptions and models to use and as such various mRNAseq analysis softwares produce different (sometimes markedly) DEGs. This paper proposes that the mRNAseq process is inherently a binomial process, and a beta-binomial model is an appropriate choice for describing mRNAseq data. We found that current methods may be overestimating dispersion and therefore variance for moderately to highly genes, and that the beta-binomial description can correct this to achieve better sensitivity for medium to highly expressed genes. Standardizing modeling approaches can help to harmonize the DEG outputs from different softwares and thus help to increase knowledge extracted from these increasing amounts of data.

Supporting Information

S1 Fig. Estimated α and β values Plotted Against the Mean for each Gene. Continuation of Fig 2 on the two remaining datasets: LINCS Mapped Reads (A) and and Gierlinski Δsnf2 (B). The x's reflect α values and the circles reflect β values with color dependent upon the density of points in the scatter plot.
(TIF)

S2 Fig. Measuring Quality of Fit for the Beta-Binomial Model to Raw and Normalized mRNAseq Data. Continuation of Fig 3 on the two remaining datasets: LINCS Mapped Reads (A-C) and Gierlinski Δsnf2 (D-F). For each column of three panels, the first panel (A,D) shows the CV^2 fit (solid blue line) and Least Squares fit (dashed green line) to the raw data points plotting mean vs variance (black x's). The second panel (B,E) shows the same fits for the normalized data. The third panel (C,F) shows the respective R^2 values for the CV^2 and Least Squares (LS) fits for the raw and normalized data.
(TIF)

S3 Fig. Comparing Beta-Binomial dispersion derivation with DESeq2 and EdgeR dispersion estimates. Each panel reflects one of the following datasets tested: LINCS Mapped Reads (A) and Gierlinski Δsnf2 (B). The black line represents our fit showing the non-asymptotic relationship between mean and variance. The brown line shows the DESeq2 dispersion fit while the magenta line shows the EdgeR dispersion fit.
(TIF)

S1 Table. Raw Data and Beta Distribution Parameter Estimates for LINCS Mapped Read Data.
(XLSX)

S2 Table. Raw Data and Beta Distribution Parameter Estimates for LINCS UMI Data.
(XLSX)

S3 Table. Raw Data and Beta Distribution Parameter Estimates for Gierlinski WT Data.
(XLSX)

S4 Table. Raw Data and Beta Distribution Parameter Estimates for Gierlinski Δsnf2 Data.
(XLSX)

S5 Table. Base-10 Logarithm p-value Differences for Predicting Differential Gene Expression in Gierlinski Count Data.
(XLSX)

S6 Table. Base-10 Logarithm p-value Differences for Predicting Differential Gene Expression in LINCS UMI Count Data.
(XLSX)

Acknowledgments

We acknowledge funding from the Icahn School of Medicine at Mount Sinai, an IBM Faculty Award, and the NIH Grants P50 GM071558 (Systems Biology Center New York), R01GM104184, R21CA196418 and U54HG008098 (LINCS Center). We thank Alan Stern and Mehdi Bouhaddou for critical reading of the manuscript, and the DToXS LINCS Center for making the UMI mRNAseq dataset available.

Author Contributions

Conceived and designed the experiments: GRS MRB. Performed the experiments: GRS MRB. Analyzed the data: GRS MRB. Contributed reagents/materials/analysis tools: GRS MRB. Wrote the paper: GRS MRB.

References

1. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JC, et al. (1999) The transcriptional program in the response of human fibroblasts to serum. Science 283: 83–87. PMID: 9872747

2. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G. (1999) Making and reading microarrays. Nat Genet 21: 15–19. PMID: 9915495

3. Bowtell DD (1999) Options available—from start to finish—for obtaining expression data by microarray. Nat Genet 21: 25–32. PMID: 9915497

4. Cole KA, Krizman DB, Emmert-Buck MR (1999) The genetics of cancer—a 3D model. Nat Genet 21: 38–41. PMID: 9915499

5. Hacia JG (1999) Resequencing and mutational analysis using oligonucleotide microarrays. Nat Genet 21: 42–47. PMID: 9915500

6. Debouck C, Goodfellow PN (1999) DNA microarrays in drug discovery and development. Nat Genet 21: 48–50. PMID: 9915501

7. Bubendorf L, Kononen J, Koivisto P, Schraml P, Moch H, Gasser TC, et al. (1999) Survey of gene amplifications during prostate cancer progression by high-throughout fluorescence in situ hybridization on tissue microarrays. Cancer Res 59: 803–806. PMID: 10029066

8. Vente A, Korn B, Zehetner G, Poustka A, Lehrach H (1999) Distribution and early development of microarray technology in Europe. Nat Genet 22: 22. PMID: 10319856

9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. PMID: 10802651

10. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929–934. PMID: 11340206

11. Smith JJ, Marelli M, Christmas RH, Vizeacoumar FJ, Dilworth DJ, Ideker T, et al. (2002) Transcriptome profiling to identify genes involved in peroxisome assembly and function. J Cell Biol 158: 259–271. PMID: 12135984

12. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, et al. (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science 309: 1078–1083. PMID: 16099987

13. Bromberg KD, Ma'ayan A, Neves SR, Iyengar R (2008) Design logic of a cannabinoid receptor signaling network that triggers neurite outgrowth. Science 320: 903–909. doi: 10.1126/science.1152662 PMID: 18487186

14. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 313: 1929–1935. PMID: 17008526

15. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210. PMID: 11752295

16. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. (2011) NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res 39: D1005–1010. doi: 10.1093/nar/gkq1184 PMID: 21097893

17. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515. doi: 10.1038/nbt.1621 PMID: 20436464

18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621–628. doi: 10.1038/nmeth.1226 PMID: 18516045

19. Zhang W, Yu Y, Hertwig F, Thierry-Mieg J, Thierry-Mieg D, et al. (2015) Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol 16: 133. doi: 10.1186/s13059-015-0694-1 PMID: 26109056

20. Shendure J (2008) The beginning of the end for microarrays? Nat Methods 5: 585–587. doi: 10.1038/nmeth0708-585 PMID: 18587314

21. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63. doi: 10.1038/nrg2484 PMID: 19015660

22. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18: 1509–1517. doi: 10.1101/gr.079558.108 PMID: 18550803

23. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. (2012) Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 9: 72–74.

24. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 11: 163–166. doi: 10.1038/nmeth.2772 PMID: 24363023

25. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS (2014) Characterization of directed differentiation by high-throughput single-cell RNA-Seq.

26. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31(1): 46–53. doi: 10.1038/nbt.2450 PMID: 23222703

27. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. doi: 10.1093/bioinformatics/btp616 PMID: 19910308

28. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550. PMID: 25516281

29. Hardcastle TJ, Kelly KA (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11: 422. doi: 10.1186/1471-2105-11-422 PMID: 20698981

30. Yu D, Huber W, Vitek O (2013) Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. Bioinformatics 29: 1275–1282. doi: 10.1093/bioinformatics/btt143 PMID: 23589650

31. McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40: 4288–4297. doi: 10.1093/nar/gks042 PMID: 22287627

32. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, et al. (2014) A comparative study of techniques for differential expression analysis on RNA-Seq data. PLoS One 9: e103207. doi: 10.1371/journal.pone.0103207 PMID: 25119138

33. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768–772. doi: 10.1038/nature08872 PMID: 20220758

34. Cai G, Li H, Lu Y, Huang X, Lee J, Müller P, et al. (2012) Accuracy of RNA-Seq and its dependence on sequencing depth. BMC Bioinformatics 13(Suppl 13): S5. doi: 10.1186/1471-2105-13-S13-S5 PMID: 23320920

35. Hardcastle TJ, Kelly KA (2013) Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. BMC Bioinformatics 14: 135. doi: 10.1186/1471-2105-14-135 PMID: 23617841

36. Zhou Y, Xia K, Wright FA (2011) A powerful and flexible approach to the analysis of RNA sequence count data. BMC Bioinformatics 27(19):2672–2678.

37. Langmead B, Trapnell C, Pop M, Salzberg SL (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 PMID: 19261174

38. Li H, Handsaker B, Wysoker A, Fennel T, Ruan J, Homer N, et al. (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–9. doi: 10.1093/bioinformatics/btp352 PMID: 19505943

39. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. doi: 10.1093/bioinformatics/btq033 PMID: 20110278

40. Ogunnaike BA (2010) Random phenomena: fundamentals of probability and statistics for engineers. Boca Raton, FL: CRC Press. xli, 1015 p. p.

41. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. (2011) Global quantification of mammalian gene expression control. Nature 473: 337–342. doi: 10.1038/nature10098 PMID: 21593866

42. Alberts B (2002) Molecular biology of the cell. New York: Garland Science. xxxiv, 1548 p. p.

43. Weisstein EW Beta Binomial Distribution. MathWorld—A Wolfram Web Resource: Wolfram.

44. Gierlinski M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, et al. (2015) Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics 31(22):3625–3630. doi: 10.1093/bioinformatics/btv425 PMID: 26206307