Scotty: An Interactive Web Tool for Designing RNA-Seq Experiments to Measure Differential Gene Expression: Supplemental Methods and Figures

Michele A. Busby, Chip Stewart, Chase Miller, Krzysztof R. Grzeda, Gabor T. Marth

Table of Contents

Sections

1. Sources of variance in expression measurements
2. Reduction of Poisson uncertainty by doubling the sequencing depth versus doubling the number of replicates
3. Prototype data
4. Determination of sequencing depth
5. Determination of variance
6. Calculation of statistical power and rationale for choice of the t-test
7. Datasets used, alignment and quantification strategy, and comparison metrics

Figures

S1. Effects of doubling read depth versus doubling the number of replicates
S2. Power comparison between t-test and DESeq
S3. Scotty workflow
S4. Performance of Scotty predictions
S5. Comparison of saturation rate among different human liver datasets
S6. Extra-Poisson variance differs among experiments
S7. Differential expression at high versus low read counts
S8. Contribution of each source of variance to total measurement uncertainty
S9. Distribution of reads per gene in yeast and human samples
S10. The compound Poisson lognormal distribution
S11. Performance of rarefaction estimates in different datasets
S12. Distribution of variance in Cheung samples
S13. Performance of estimates of the distribution of variance in various datasets
S14. Consistency of t-test p-value with measured false positive rate
Supplement 1. Sources of variance in expression measurements

In order to identify a difference in expression between a test and control condition as statistically significant, it is necessary for the difference in the expression between the two conditions to be larger than the uncertainty in the estimate of the mean expression within a single condition (Figure S6A). The sources of uncertainty in RNA-Seq measurements stem from the following types of variance:

**Poisson variance** describes a type of variance that occurs in count data when something is sampled according to a Poisson process. We use it here to describe the variance in RNA-Seq data that occurs because each particular RNA is selected at random from among the RNA and counted. Poisson variance is higher relative to the total count for genes measured with low counts versus high counts. For example, the difference in expression of a gene measured with one read versus two reads is inherently less certain than the differences in expression of a gene measured with 100 reads versus 200 reads, even though both differences are nominally a 2X fold change (Figure S6B).

The Poisson variance in measurements ($\sigma_p^2$) for any count can be approximated by the count itself Equation 1:

$$\sigma_p^2 \approx R$$

where R is the read count for the gene. The relative standard deviation due to Poisson counting noise ($\sigma_p$) is thus calculated as follows:

Equation 2:

$$RSD_p = \frac{\sqrt{R}}{R}$$

Because of this relationship, Poisson uncertainty is highest as a proportion of the gene’s measurement when genes are measured with a low number of reads. Poisson noise adds relatively high uncertainty to low read counts, particularly those below ~10 (Figure S7A).

**Non-Poisson Technical Variance** is measurement imprecision beyond Poisson noise that stems from the inability of RNA-Seq measurements to measure expression perfectly. This imprecision is seen when expression from the same biological material is measured twice. Sources of measurement imprecision may include PCR amplification errors during library preparation or machine errors.

While non-Poisson technical variance is low in RNA-Seq experiments, in general it appears to be high enough that any differential expression calls based on a Poisson model will overestimate the significance of differences in expression (Busby, et al., 2011). Therefore, Scotty does not support power analysis based on Poisson models, despite the fact that they have been widely used in the literature (reviewed in (Anders and Huber, 2010)).
Biological variance is variance that naturally occurs within the samples under investigation. This variance stems from the fact that the expression of any given gene is likely to naturally fluctuate within the cells themselves, and between samples of the same condition. Sources of biological variance include genetic differences among samples and gene expression responses to the environment. Experimentalists should expect that any difference in the environment will lead to some differences in gene expression.

The relative contribution of each source of variance is shown in Figure S7A and S7B. A version of Figure S7A was originally published in the supplement to Busby et al. 2011 and we have reproduced it here for the convenience of our readers. We note that the proportions of technical and biological variance differ between experiments depending on how much variance is present in technical and biological replicates. Because the Poisson variance is based on the value of the count itself, it is not experiment-specific and the degree of Poisson variance at any given sequencing level is the same across experiments.

**Supplement 2. Reduction of Poisson uncertainty by increasing the sequencing depth versus increasing the number of replicates**

If one wanted to reduce Poisson uncertainty in an experiment by doubling the number of reads, one could do this either by doubling the number of replicates, or by retaining the same number of replicates but doubling the sequencing depth of each replicate. Here we demonstrate that both approaches theoretically have the same effect. This effect is easy to see if one imagines an experiment where Poisson uncertainty was the only source of variance. Adding counts from another replicate would be no different than doubling the sequencing depth.

Mathematically, this can be proven by observing the effects on the error in the estimate of the mean (σ_μ^2). The error will be reduced by either approach, and the resulting relative error in the mean is the same.

The estimate of the error in the mean decreases with each additional sampling as follows:

Equation 3

\[ \sigma_μ^2 = \frac{\sigma^2}{N} \]

Where \( \sigma_μ \) is the error in the estimate of the mean, \( \sigma \) is the standard deviation of the distribution from which the sample was drawn and \( N \) is the number of samples (See (Bevington and Robinson, 2003) for derivation).
The Poisson variance of a count can be approximated as the count itself, in this case designated by \( R \) for “reads”. The Poisson uncertainty in the mean is thus:

Equation 4

\[ \sigma_{\mu}^2 \approx \frac{R}{N} \]

To compare the total variance in measurements at different scales it is convenient to use the relative standard deviation (RSD), which requires taking the square root of the variance and dividing it by the value of the mean:

Equation 5

\[ RSD \approx \frac{1}{R} \sqrt{\frac{R}{N}} \]

An equivalent number of reads can be used to decrease the uncertainty in the gene expression measurement that is due to Poisson uncertainty. If we use these reads to increase the number of replicates by the multiple \( X \) we will have:

Equation 6

\[ RSD \approx \frac{1}{R} \sqrt{\frac{R}{XN}} \]

If we use these same reads to increase the sequencing depth of the existing sample it will increase the mean of each measurement by a factor \( X \), as follows:

Equation 7

\[ RSD \approx \frac{1}{RX} \sqrt{\frac{RX}{N}} \]

Which is algebraically equivalent to Equation 6. These relationships are also shown in Figure S1.

Supplement 3. Pilot Data

Scotty uses prototype data to assess the how quickly new genes are quantified as additional reads are sequenced and how much variance is present between replicates. The prototype data can be one of the pre-loaded datasets, or it can be pilot data generated by the user. If it is pilot data it will ideally contain sample data from two pairs of replicates from the control and two pairs from the test condition. However, Scotty’s analysis can also be done with two replicates from a single condition, e.g. the control condition. In this case, Scotty will analyze the data with
the assumption that the variance and saturation rate of the test condition is identical to that of
the control condition.

As we point out in our main manuscript, pilot data should be generated in as close a manner to
the experimental data as possible. For example, growing culture samples in two different
batches can be expected to introduce more variance than growing samples in a single batch,
due to greater fluctuations in gene expression due to slight differences in the environment. If
the samples in the experiment will be grown in two batches, the most accurate power estimates
will be obtained if the pilot data is also grown in two batches. Samples should also be
sequenced according to the same protocols that will be used in the actual experiment.

Once the sequencing runs for the pilot data are completed, data should be aligned to a
reference genome or transcriptome sequence, and a file should be created that counts the
number of reads that align uniquely to each gene or transcript, based on an annotation file.

Pilot data is input into Scotty as a tab delimited text file. The file should be formatted in the
following format:

<table>
<thead>
<tr>
<th>Gene_Name</th>
<th>Control_Rep_1</th>
<th>Control_Rep_2</th>
<th>Test_Rep_1</th>
<th>Test_Rep_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene_A</td>
<td>123</td>
<td>154</td>
<td>223</td>
<td>102</td>
</tr>
<tr>
<td>Gene_B</td>
<td>12</td>
<td>7</td>
<td>8</td>
<td>4</td>
</tr>
</tbody>
</table>

Under the Scotty statistical model, the count data represents the number of reads that map
uniquely to the gene in question. Reads that do not map uniquely in the genome should not be
included in count data.

However, several strategies have been published that consider reads that map to
multiple isoforms and call transcript abundance based on both the uniquely and
multiply-mapped. Scotty provides limited support for counts generated by other programs as
follows:

We will first consider an example using the basic reallocation method which is a
simplified version of the method described in (Mortazavi, et al., 2008). We can consider
a simple example of a gene with two alternative transcripts (A and B). Reads can
align uniquely to one transcript or can map to both transcript A and transcript B. In
this example we have:

Transcript A  6 unique reads
Transcript B  3 unique reads
Both Transcripts:  90 reads
Under the Mortazavi approach, the 90 reads which map to both transcripts will be divided proportionally with 60 reads assigned to transcript A and 30 reads going to transcript B. While this will increase the nominal count of the reads for both Transcript A and Transcript B there will not be a simple corresponding reduction in Poisson variance that would be expected from higher, more precise counts because the re-allocation itself is based on counts which have high Poisson noise and is thus itself noisy. Thus, the accuracy of the quantification of the transcript abundances are fundamentally limited by the error observed in the uniquely mapped reads.

While the reallocation of multiply-mapping reads has been demonstrated to improve the absolute quantification of transcript abundance, as measured for instance by the correlation with microarray expression values, an accurate absolute quantification is not required for differential expression calls. Differential expression instead requires an accurate relative expression measurement when one compares the expression of the same transcript in two conditions, and an accurate relative measurement can be attained from the unique reads alone.

While more complex algorithms exist than this approach, any method can only impute the origin of multiply-mapped reads from the observed uniquely mapped reads and will all be fundamentally limited by the information content in the uniquely mapped reads. The question then becomes whether including the multiply-mapped reads decreases the variance across replicates (and increased power) or increases the variance (and decreases power). Thus, we considered both the strengths and the limitations of Scotty in estimating power based on the output of these programs:

We do not believe that Scotty will have any inaccuracies when extrapolating the power from two pilot replicates to many replicates when the sequencing depth remains unchanged. The increase or decrease of variance that is introduced by the reallocation method will be observed in the total variance that is measured across the replicates. Thus, when the sequencing depth in the projected experiment is the same as the sequencing depth of the pilot data then the estimation of power using an increased number of replicates will be correct.

However, the amount of error that is introduced by the reallocation of multiply-mapped reads is dependent upon the Poisson uncertainty present in the original read counts. Thus, it is not possible for Scotty to model the change in variance that will occur with increased sequencing depth from the data that we collect because the error in the reallocation is both correlated with sequencing depth and specific to the statistical model that is used to reallocate the reads.

While we cannot completely address this inherent characteristic of the data we offer two options to our users who wish to use transcripts quantified by other programs: The first is to use Scotty with counts generated by the transcript-quantifying program with the caveat that the extrapolation of power to higher sequencing depth will be subject to a certain degree of error. We have added a description of the problem to the Scotty help section. The second option is to simply quantify the transcript abundances using uniquely mapped reads. We believe this is the better
option as most of the information that is useful for differential expression calls is contained in the uniquely-mapped reads, and these results will provide a reasonable approximation of the power that will be attained if you also include the multiply-mapped reads, which contain relatively little information.

This assumption is easy for the user to test if they process their pilot data using both the transcript caller and the uniquely mapped reads. If the power observed at the sequencing depth of the pilot data is the same for both approaches then the extrapolation based on the uniquely aligned reads will provide a good estimate of the power attained by the transcript caller at higher read depths.

Supplement 4. Determination of sequencing depth

The general approach to all of Scotty’s analyses of pilot data is the assumption that with a limited amount of data it is unlikely that the experimental parameters for an individual gene will be estimated accurately. However, it is possible to use a large number of genes to fit distributions that will give reasonable approximations of power for the data in aggregate.

Pilot data is used to assess the rate at which RNAs are quantified. In RNA-Seq data, the count of reads per gene will span many orders of magnitude (Figure S8). A few rare genes will produce very high read counts, while the majority of genes will produce read counts several orders of magnitude lower. The expected number of reads per gene can be estimated at any sequencing depth by calculating the probability of a read from each gene being observed as a function of the number of reads observed for the gene ($R_g$) divided by the total number of reads sequenced for all genes ($R$). This probability ($p_g$) can then be multiplied by any total number of reads at a given sequencing depth.

Equation 8

$$p_g = \frac{R_g}{\sum R}$$

Genes measured with low counts, $p_g$ will be measured with high Poisson noise and introduce a binning effect where many genes are measured with one read, which causes a multiplicative extrapolation of gene counts at higher read depths to give erratic results. To address this issue, we smoothed $p_g$ values that were based on measurements of only one read by fitting the observed data to a lognormal Poisson distribution (Bulmer, 1974).

To justify this fit we examined datasets sequenced to saturation. When datasets approach saturation there are a greater number of genes observed with counts of 2 than counts of one. The yeast datasets which we examined appear to be saturated, with the majority of genes having read counts greater than 10 and the distribution of read counts appears reasonably lognormal (Busby, et al., 2011; Nagalakshmi, et al., 2008 Figure S8A). None of the human datasets that we used in this analysis are sequenced deeply enough to observe a complete distribution (Cheung, et al., 2010; Kasowski, et al., 2010; Marioni, et al., 2008). However, in the human datasets, we observed that the distribution of logged read counts appears to be
reasonably approximated by a truncated normal distribution (Figure S8B). These observations together suggest that it is reasonable to model the true distribution of reads per gene as a lognormal distribution, which is consistent with observations previously reported using microarray data (Hoyle, et al., 2002).

While we assume that the true distribution of gene expression is lognormal, in sample data read abundances are measured with discrete measurements and Poisson sampling noise. The distribution of observed counts is therefore better approximated by a Poisson-lognormal distribution (Figure S9) which is defined in (Engen, 2007) as:

Equation 9:

$$p_i(\mu, \sigma^2) = \int_{-\infty}^{\infty} \frac{e^{i(\mu+\sigma u)}}{i!} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

Where $\ln(\alpha)$ is normally distributed with mean $\mu$ and variance $\sigma^2$.

In the calculations of power, Scotty only relies on the fit of the data to the lognormal distribution to calculate the true probabilities of genes observed with one read. In other calculations, the observed probabilities are used. Performance of estimates is shown in Figure S10.

While it is also possible that some genes will be truly expressed but measured with zero reads due to insufficient sequencing, Scotty does not attempt to estimate power for these genes due to the fact that there is often insufficient information to accurately assess the number of genes that are expressed but not observed. Power calculations are therefore based only on observed genes. However, power calculations for genes expressed at lower levels can be observed in the final output charts of the Scotty application.

**Supplement 5. Determination of variance**

Because each gene’s variance is in part determined by the gene’s biological function, we expect that each gene will have its own unique variance, which will be poorly measured in pilot data due to the low number of replicates. However, when all gene-specific variances are combined we expect that they will form a distribution that is approximately lognormal. To estimate the parameters of this distribution, we fit the measured variances to an underlying lognormal distribution using a procedure that accounts the expectation that the distribution will be skewed somewhat by the low number of replicates in pilot data.

We examined the Busby, Cheung, Karowski, and Marioni datasets (see Section 7) and found that measured gene-specific relative non-Poisson standard deviation ($\sigma_{mv}$, Equation 10) present in the gene’s measurements between replicates is uncorrelated with the gene’s read count.
(Pearson correlation < 0.05 in all four cases). Therefore, this estimated distribution of non-Poisson variance is used in downstream power calculations to estimate the power of the genes using the assumption that the distribution of variances would be representative of the variance present at all sequencing depths.

As described in S1, the total variance for a gene’s expression, as measured in replicates ($\sigma_{\text{reps}}^2$), will consist of both Poisson and non-Poisson variance ($\sigma_{np}$). The Poisson variance of a gene’s measurement is determined by the sequencing depth of the experiment and will vary between the pilot and experimental data if the two datasets are sequenced to different depths. Therefore, it is useful to separate out the Poisson and non-Poisson variance for the purpose of power calculations. To do this, we used same approach we used in (Busby, et al., 2011). This approach is based on the fact that measurements of uncorrelated variances can be added to form the total variance (Bevington and Robinson, 2003).

The mean of the number of reads that were used to make the measurement of variance in replicates (R) is used as an estimate of Poisson variance, and separated from non-Poisson variance as follows:

Equation 9:

$$\sigma_{np}^2 = \sigma_{\text{reps}}^2 - R$$

The measured gene-specific relative non-Poisson standard deviation ($\sigma_m$) is found by dividing the non-Poisson variance by the read count:

Equation 10:

$$\sigma_m = \frac{\sigma_{np}}{R}$$

This value ($\sigma_m$) can then be used to observe the variability in all genes expressed on the same scale, independent of sequencing depth. It can thus be calculated as:

Equation 11:

$$\sigma_m = \sqrt{\frac{\sigma_{\text{reps}}^2 - R}{R}}$$

By Cochran’s theorem (Cochran, 1934) we expect that if N measurements for a single gene are drawn from a population a large number of times, the measured variances for that gene will themselves be well approximated by a scaled $\chi^2$ distribution with N-1 degrees of freedom. The distribution of measured variances are therefore expected to be skewed relative to the distribution of true variances, with more high-variance outliers measured than exists in the true distribution.
We then assume that the true distribution of the combined relative measured non-Poisson variances for each individual gene ($\sigma_{m}^{2}$) will be well approximated by a lognormal distribution. We chose a lognormal rather than a normal distribution based on observation that the distribution of normalized variances that is observed within the 40 biological replicates of the Cheung dataset contain a long right hand tail (Figure S1). Because a $\chi^2$ distribution with 39 degrees of freedom has only minimal skew and no pronounced tail, the tail is unlikely to have been caused by the number of measurements, suggesting that the true variances will be well modeled with a lognormal distribution. Additionally, variance cannot be less than zero as would be allowed by a normal distribution.

Therefore, we expect that the distribution of the relative non-Poisson variances measured from pilot data ($\sigma_{m}^{2}$) will be well described by a compound distribution that consists of a lognormal distribution convoluted by a scaled $\chi^2$ distribution. The parameters that are of interest are the parameters ($\mu$ and $\sigma$) of the underlying lognormal distribution. These are estimated using the following iterative process:

1. Wide boundaries of the parameters are established. The boundaries of the variance are therefore set to 0.001 and the maximum measured variance.
2. The values within the boundaries are divided into 5 steps. The distribution created by each possible pair of steps is tested against stimulation data. For the simulation data, N values are generated for each gene that is in the measured data, where N is the number of replicates in the pilot data (we expect this will usually be two). For each gene, a lognormal distribution is simulated. The mean is taken from the gene’s expression level and the pilot data. The relative non-Poisson variance is taken from a value chosen at random from the lognormal distribution generated by the parameters being tested. We assume that the expression levels and the variance are uncorrelated, so these values are simulated independently. A final value for each simulated gene measurement is found by simulating Poisson variance for the measurement.
3. The relative non-Poisson variances of the N simulated values for each gene are measured.
4. The fit of the measured variances to the simulated variances is measured using a Kolmogorov-Smirnov test.
5. The boundaries are then narrowed to values on either side of the best fit. This process is repeated until a fit is found that does not differ significantly from the measured distribution ($p$ of differing $>$ 0.5) or for five rounds.
6. If a fit is not found that does not differ from the measured distribution ($p$ $>$ 0.10) after five rounds, then calculations proceed with the best fit found, but Scotty will give a warning message that the model variance is poorly fit to the distribution in question.

Results of the procedure are shown in Figure S12.

**Supplement 6. Calculation of statistical power and rationale for choice of the t-test**
Power calculations for the t-test are based on equations as published in Harrison and Brady (Harrison and Brady, 2004) which are based on the work of Chow et al. (Chow, et al., 2002). These equations make use of a non-central t-distribution to approximate the power of an experiment (1-the false negative rate (β)) of samples with means (m), standard deviations (σ) and replicate number (n):

Equation 12 (reproduced from (Harrison and Brady, 2004))

\[ 1 - \beta = T_v\left(\pm t_{\alpha/2,v} \left| \frac{m_1 - m_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right) - T_v\left(-t_{\alpha/2,v} \left| \frac{m_1 - m_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \right) \right) \]

Where \( T_v(\pm t_{\alpha/2,v} | \theta) \) is the cumulative distribution function of a non-central t-distribution with non-central parameter \( \theta \) and degrees of freedom \( v \), and \( t_{\alpha/2,v} \) represents the critical value of a t-distribution at the false positive rate \( \alpha \).

We note that it is more common to find power calculations formulas for a t-test in which the t-distribution is approximated by a normal distribution than those that use the non-central t-distribution. However, at the low number of samples commonly available in RNA-Seq experiments this approximation will lead to an overestimation of power (Chao, Shao, and Wang, 2002).

**Estimating Total Power**

Total power is calculated for each configuration of a given number of replicates and total sequencing depth. Within each configuration there will be a mixture of genes of different variance and different sequencing depths. We have previously reported and confirmed with the datasets used in this analysis that gene-specific read depth and variance are uncorrelated (Busby, et al., 2011). To calculate the total power for each configuration, we build a matrix representative of the dataset. In the first dimension of the matrix, an evenly spaced selection of 50 variances is drawn from the lognormal distribution of gene-specific variances, having parameters estimated as described in Supplement 5. The second dimension of the matrix represents the sequencing depth. For this dimension we rely primarily on the measured observed probabilities of a gene being selected, with a small number of additional values added to represent genes that are expressed but not observed, as described in Supplement 4. While we could use this entire table of probabilities, to improve performance we pseudo-randomly select 200 values from this table to provide a representative sample of the read depths expected. This then provides a matrix of powers representative of the whole gene dataset. The total power at each configuration is the mean of that matrix.

When calculating total power, we only included the genes that are expressed at a level that is observable in the prototype data. The reason that we excluded genes with lower expression levels is as follows: It is easy to imagine that with an infinite number of samples and infinite sequencing all of the annotated genes would eventually be detected (sometimes incorrectly: through misaligned reads, DNA contamination, or other artifacts) as expressed. However, many
of these genes will be expressed at very low levels. These genes are not likely to either be of interest to the user or quantifiable using current RNA-Seq technology at practical costs. Therefore, Scotty assumes that only the genes that are present in the prototype are of interest to the user.

Rationale for Choosing the T-Test

We had two motivations for choosing the t-test as the test used for power calculations in our analysis. The first motivation is that it produces unbiased calls of differential expression, which provides for less complicated downstream analyses. The second is that it has readily available formulas for calculating power.

Unbiased Calls

Before choosing the t-test, we also considered two other well-performing and closely-related statistical packages for calling differential gene expression: DESeq and edgeR (Anders and Huber, 2010; Robinson, et al., 2009). Both of these packages call differential gene expression by performing an exact test based on a negative binomial distribution. There are two large differences between these tests and a t-test. 1) Both packages rely on the discrete negative binomial distribution as their test distribution. 2) Both packages under their default settings share information between genes to provide a more accurate estimate of each gene’s variance.

We chose not to use a statistical metric which relies on information sharing. With a very low number of replicates, both the mean and the variance measured for each gene will deviate from the gene’s true variance (Figure S15). DESeq and EdgeR increase power by improving variance estimates on aggregate by smoothing all variances towards a central value. However, this procedure introduces a bias into which genes are called differentially expressed (Busby et al. 2011). Genes that truly have large variances are more likely to be called differentially expressed because their variances are systematically underestimated. This bias can make it difficult to interpret the biological significance of findings because different biological classes of genes have different variances. For example, we corroborated microarray-based findings that TATA controlled genes in yeast have higher variance in biological replicates, and we found that snoRNA had higher technical variance in our data (Busby et al. 2011). It is therefore our opinion that, as sequencing costs have dropped to a point where experiments with a larger number of replicates is becoming the norm, during the planning stages of an experiment it is better to plan to use a large enough number of replicates to allow variance to be more accurately estimated for each individual gene than to rely on information sharing.

Assumption of a Normal Distribution

The t-test assumes that the true underlying expression values are normally distributed. As with any parametric test, the empirical data must be approximated by the theoretical distribution well enough to identify genes which are differentially expressed. We tested whether this approximation was appropriate for RNA-Seq data using two approaches. We first tested the
number of false positives called at p<0.01 using simulated data to see if they conformed to the expectation that the p-value would reflect the false positive rate. We then compared the performance of the t-test with another metric to examine how accurately it can identify differentially expressed genes.

An initial concern in using the t-test is that a normal distribution is continuous. The true data of the expression per gene across replicates forms an underlying distribution which is measured using count data. Count data yields discrete measurements. To examine the false positive rate at these low count levels we simulated human datasets using the simulation procedures we described in (Busby, et al., 2011). The data was simulated to reflect moderate coverage where approximately 30% of the genes were measured with fewer than 10 reads. To test the robustness of the t-test, we performed simulations twice: the first time with each gene’s expression modeled as being drawn from a normal distribution and the second time it was modeled as being drawn from a lognormal distribution. Each simulation was performed 50 times. We found that at mean counts greater than or equal to 1 the number of false positives appeared consistent with a correct statistical model, and robust to the true nature of the underlying distribution. By contrast, while the number of false positives called by DESeq was in total consistent with the false positive rate, the values increased with read depth, further biasing the call set in favor of calling more highly expressed genes differentially expressed (Figure S13). We concluded that the normal approximation was reasonable for RNA-Seq data, and that the test itself is fairly robust to the true underlying distribution of the data.
Power Comparison Between a T-Test and DESeq

We also compared power of the t-test to the power of DESeq using the simulation data (Figure S14). In this simulation the distribution we modeled the data as lognormal. We modeled differential expression as a change in biological expression with an effect size (Cohen’s D) of 5 in random directions. We used the standard t-test formula with pooled variance.

We found that the t-test performs worse than DESeq when there were only two replicates present. This is an expected result because the t-test is a very conservative test when there are a low number of replicates because it measures the variance for each gene independently and accounts for the fact that each variance is poorly measured with two data points. The performance between the two tests becomes comparable with an increased number of replicates. For example, at 5 and 6 replicates the t-test performed slightly better than DESeq in this simulation, and at all replicate numbers produces an unbiased call set (Figure S14).

In general, based on our previous experience using only two replicates in (Busby, et al., 2011), we recommend against using only two replicates because the power will be low regardless of what statistical test is used and because it is difficult to identify outlier replicates when there are only two samples to observe.

Supplement 7. Datasets used, alignment strategy, and comparison metrics

We obtained publically available datasets from the Gene Expression Omnibus. We named the datasets based on the first author of their accompanying publications. The datasets included liver datasets (Blekhman, et al., 2010; Marioni, et al., 2008; Pan, et al., 2008; Xu, et al., 2011), Lymphocyte cell line datasets (Cheung, et al., 2010; Kasowski, et al., 2010), and yeast (Busby, et al., 2011; Nagalakshmi, et al., 2008). All datasets use Illumina sequencing of mRNA with the exception of our Busby dataset which used whole transcriptome sequencing with rRNA removal on AB sequencing technology.

We chose these samples for inclusion because they had specific characteristics that we wanted to examine, for example a large number of replicates (Cheung dataset) or multiple experiments using the same tissue type. Technical qualities of the samples were not a factor for inclusion. Therefore, we assume that these samples are likely to be representative of available datasets. However, we note that because of the rapid adoption of next-generation sequencing technology the datasets currently available to the public were in many cases prepared at an earlier stage in protocol development than the protocols currently being used.

Samples were aligned to reference genomes using Mosaik (http://bioinformatics.bc.edu/marthlab/Mosaik) allowing for two mismatches. The human genome used was build 36, downloaded from the NCBI. Annotations were downloaded from Gencode. The S. cerevisiae genome and annotation were downloaded from the "Saccharomyces Genome Database" http://downloads.yeastgenome.org on 1/28/2010. Alignments were processed using the same scripts we developed for (Busby, et al., 2011) which utilize the
BamTools API (Barnett, et al., 2011). The read counts per gene are defined as the number of reads uniquely aligning to the region in the genome where the gene is annotated.

In our comparison of liver samples (Figure S4), all reads were trimmed to 32 bases before alignment. This was done to demonstrate that the differences in saturation rate were due to factors other than read length. Thirty-two bases is the length of the shortest reads, in the Pan dataset.

Where necessary, samples were normalized to the median value of the samples. In this process a normalization factor was calculated by taking median of the read count for each gene for the first sample divided by the count for the second sample. This method is the same as was used in (Anders and Huber, 2010) and (Busby, et al., 2011). We will note that we found previously that this method works very well when we used samples prepared by the same protocol, within the same experiment, but we observed that it failed to correctly normalize the data when we compared samples from different experiments. The failure appeared to be caused by the differences in library complexity that we observed in Figure S4.
Figure S1: Effects of doubling read depth versus doubling the number of replicates

S1A: Theoretical samples with only Poisson variance

S1B: Realistic samples with only Poisson variance and biological plus technical variance

Figure S1: These figures show the difference in the reduction of total measurement uncertainty that is obtained by doubling the number of reads versus doubling the number of samples. The measurement uncertainty is defined as the relative standard deviation of the uncertainty in the mean as described in Supplement Section 2. Figure S1A shows theoretical samples where the only source of measurement uncertainty is Poisson noise. Figure S1B shows a more realistic scenario where the variance in gene expression is over dispersed from Poisson by 30% by a combination of technical and biological variance. The degree of improvement of adding replicates over deeper sequencing is greatest at higher sequencing depths, and proportional to the degree of non-Poisson variance in the samples, i.e. the improvement is greater when samples have a higher degree of technical and biological noise.
Figure S13: Consistency of t-test p-value with measured false positive rate

Here we use data simulated to model a human dataset with moderate coverage to show that at low read counts the p-value calculated by a t-test (blue) accurately represents the true false positive rate (green). Measurements across replicates are simulated as being drawn from normal (left) or lognormal (right) distribution for each gene. Corresponding values for DESeq are plotted in red.
Figure S14: Performance of t-test versus DESeq

Continued...
Figure S14: Performance of t-test versus DESeq (continued)

Figure S14: This shows the performance of the t-test versus DESeq in simulated data modeled to represent moderately covered human data. The change in expression is modeled as an effect size change of 5 relative to biological expression in random directions.
Figure S2: The Scotty workflow represented schematically. Blue boxes represent user inputs.
Figure S3: Data were simulated to model human data using the assumption that each gene’s expression level will form a normal distribution across replicates. Data were simulated at ten different coverage depths and power was calculated at each depth using 2-10 replicates (top panel, left). Power was then estimated by Scotty’s algorithm using the two lowest coverage replicates for each condition (top panel, right). Correspondence between the estimated and measured value is shown in the lower panel. Only genes observed in the simulated prototype data were considered in the calculations of the true power.
Figure S4: This figure shows how many genes are detected with 10 or more reads at each sequencing depth. The rate at which genes are quantified can differ dramatically between experiments, even when they are nominally using the same species and tissue type. Sequencing depths are based on the number of reads that are uniquely aligned to genes. Each line represents the saturation rate observed in a single sample of an experiment measuring human liver expression. Gene counts at different read depths were generated using Monte Carlo sampling of the existing data. If we compare samples at the final read depth we find that the number of genes quantified with 10 reads differs significantly between samples with the lowest Pan sample having 4845 genes while the highest (one of the Bleckhman Male samples) has 7525 genes. Therefore, if one had predicted the this Bleckhman Male’s saturation rate based on the Pan dataset you would find 55% more genes than you were expecting at this read depth. The Bleckhman dataset contains samples from 6 different individuals with two technical replicates of each. The other datasets consist of technical replicates of the same sample. The sources of the datasets are listed in Supplement 7. The Bleckhman and Marioni dataset were produced by the same group. For the purposes of this analysis, reads in all samples were trimmed to a common length and aligned to the same human genome version using the same settings to eliminate these factors as a source of error.
Figure S5: Extra-Poisson variance differs among experiments

For example, the whisker plot for the Cheung data shows the median non-Poisson over dispersion ($\sigma_{np}^2$) between each pairwise comparison of the 40 replicates within the experiment, calculated as described in Equation 9. To eliminate differences in sequencing depth as a source of variance only genes with at least 10 reads per replicate were included. Overall, the Cheung experiment had more variability than the replicates in the Kasowski dataset, though both were of Human B Cells. We note that the Cheung experiment contained 40 replicates while the Kasowski dataset only contained 14, and we expect it to be more technically challenging to control variance in an experiment with a larger number of replicates than in a smaller experiment. Correspondingly, our experiment (Busby yeast) also contained more replicates than the Nagalakshmi yeast experiment (8 biological replicates versus 2) and had higher variance between replicates, likely because we split our samples into two batches. A comparison of a greater number of experiments using the same type of biological material would be necessary to see if the trend for experiments with higher replicate number to have higher variance can be generalized. This finding emphasizes the need for pilot data to be predictive of the conditions that will be present during the actual experiment. In instances within the Cheung and the Blekhman experiment where there were multiple biological replicates from the same individual only the first replicate was used.
Figure S6: Differential expression at high versus low read counts

S6A) Identifying differences in gene expression as statistically significant

![Graph showing probability distributions for identified and not identified as differentially expressed](image)

S6B) Uncertainty in a 2X fold change at different read depths

![Graphs showing uncertainty at different read depths](image)

Figure S6A: The expression for any one gene will vary with repeated measurements. Observed measurements can be used to estimate the distribution of a gene’s true expression level, shown as lognormal. Here we show the overlaps of two normal distributions that could theoretically be estimated from observed counts. In the left panel, the two distributions show differential expression. In the second panel, while the mean expression levels are different, there is a high degree of overlap between two distributions, and therefore a low degree of statistical certainty that the measured differences represent true differences.  

S2B: Here we show the overlap of Poisson distributions of single measurements at different read counts. Because relative Poisson uncertainty is high at low read counts, a count of 1 versus 2 has very little power to discriminate a true 2X fold change, though at higher counts a 2X fold change becomes significant. In an actual experiment, the width of the distribution would be greater due to additional biological and technical uncertainty, but the uncertainty to the mean expression would narrow with each additional replicate.
Figure S7: Contribution of each source of variance to total measurement uncertainty

S7A.

Contribution of Each Source of Variance to Total Measurement Uncertainty

S7B.

Measurement Uncertainty In Different Types of Replicates

Figure S7: The contribution of each source of variance to the total uncertainty in a gene’s expression measurement at a given read depth. S7A. This shows the contribution of each source of uncertainty. We use relative standard deviation as the metric of uncertainty. We represent these as constants because we found that in our data the non-Poisson variance was uncorrelated with read depth. The uniform over dispersion is an aggregate value and the true relative standard deviation will be higher or lower than this value for individual genes. The over dispersion for the biological replicates includes both biological and non-Poisson technical variance. In interpreting this chart, note that uncorrelated variances (not the standard deviations) can be added to achieve the total variance. The data shown is S. cerevisiae data published in Busby et al. 2011. A version of this chart appeared in that paper’s Supplementary Material. S7B. By overlaying normalized biological and technical replicates with simulated replicates showing only Poisson noise we show the relative contribution of each sources of variance.
Figure S8: Distribution of reads per gene in yeast and human samples

S8A: *S. cerevisiae* samples
S8B: Human samples

Marioni Human Liver Combined Technical Reps

Cheung Single Human B-Cell Sample
Figure S8: For each sample the chart on the left shows the distribution of reads per gene in log space. The chart on the right shows the distribution of reads per gene in ordinal space for counts ranging from 1 to 10. While the yeast samples (A) appear saturated, in all human samples (B) the number of genes with higher read counts decreases at each count (i.e. there are more 1’s than 2’s, etc.) suggesting none of the human samples are saturated.
Figure S9: The compound Poisson lognormal distribution

The compound Poisson lognormal distribution is a compound distribution of a Poisson and lognormal distribution that arises when items sampled from a lognormal distribution are measured with discrete measurements with Poisson counting noise.
Figure S10: Performance of estimates of sequencing depth predictions in different datasets

S10: Prediction of Genes Covered by 5 Reads

Human

Figure S10: For each dataset, we took a subsample of the data (black). This dataset was used to project the coverage at higher read depths as described in Supplement 4 (red). A comparison with the actual data is shown in blue.

Yeast

Figure S10: For each dataset, we took a subsample of the data (black). This dataset was used to project the coverage at higher read depths as described in Supplement 4 (red). A comparison with the actual data is shown in blue.
Figure S11: Distribution of variance in Cheung samples

Figure S11: The top chart shows the distribution of non-Poisson variances \((\sigma_n^2)\) present in 40 replicates from the Cheung data. We excluded genes with mean read counts lower than 20 to avoid possible confounding effects of Poisson noise. The bottom chart shows a Chi Square distribution with 39 degrees of freedom. By Cochrane’s theorem, measured variance will be approximately drawn from a \(\chi^2\) distribution with N-1 degrees of freedom. The Cheung data has a larger right hand tail than what could be expected from measurement effects alone as this pronounced tail is not apparent in the corresponding \(\chi^2\) distribution. For this reason, we modeled gene-specific biological variance using a lognormal distribution.
Figure S12: Performance of estimates of the distribution of variance in various datasets

A. Fit of variance estimates to actual measurements in Marioni human liver dataset (3 technical replicates)

B. Fit of variance estimates to actual measurements in Cheung human cell line data (2 biological replicates, different cell lines)
Figure S12: Comparisons between measured and model variance to show the fit between the actual data and the modeled data. The panels on the left of each chart show the correspondence between the measured distributions and simulated distributions using the parameters estimated by the variance estimation procedure described in Supplement Section 5. The degree of correspondence shows that variance can be accurately modeled on aggregate using only a low number of replicates from pilot data.
References


