Marth Lab

Help

Introduction

Welcome to Scotty!

Scotty is used to plan RNA-Seq experiments that measure differential gene expression.

At the start of every experiment, someone must ask the question, "How many reads do we need to sequence?"

If you are new to RNA Seq you may find it helpful to read our tutorial, Thinking About RNA Seq Experimental Design. Our intended audience for this tutorial was people who do not have an extensive background in statistics and computational biology, for example, a first year graduate student in wet lab biology.

Input Data

A prototype dataset is required for power calculations. The prototype dataset is used to measure how much variance is present between replicates. This largely determines how many replicates are required. This data is also used to quantify how many genes are detected as more reads are sequenced. This largely determines how deeply each replicate needs to be sequenced.

There are two options for obtaining prototype data. The first is to use an existing datasets from other people's experiments. Several publically available datasets are preloaded. Additional datasets will be added upon user request. Please send an email to busbym at bc dot edu with a link to the dataset you would like loaded. We can only load datasets that include biological replicates and may not be able to load datasets that are subject to embargo.

The second option is to use pilot data generated within your own lab. Pilot data is preferable. We have found that there is a great deal of inter-experiment variability in both the variation between replicates and the rate of gene discovery, even between samples using the same species and tissue type.

However, when it is not feasible to generate pilot existing experiments with a very close protocol may serve as an acceptable substitute, and will of course be better than no data at all.

Generating Pilot Data

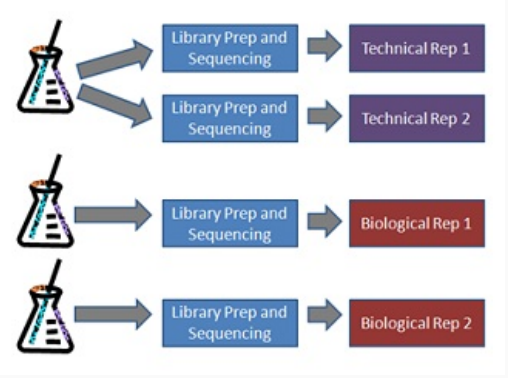
The most accurate approach

Scotty requires, at a minimum, two replicates of the same condition sequenced to a read depth where a reasonable portion of genes are measured. This does not have to be as deep as the final expected sequencing run, and reasonable results may be found with fairly low depths.

However, more data will lead to improved predictions. Scotty will generate the most accurate results if you have two replicates of both the control and the test condition. This allows Scotty and the user to see how variance differs between conditions. However, we recognize that this is not always feasible. If replicates are not available for one of the conditions Scotty assumes that both conditions have the same variance.

What type of replicates should I use?

There are two types of replicates: technical and biological.
In RNA-Seq, technical replicates are replicates where the biological material is the same in each replicate but the technical steps used to measure gene expression are performed separately. For example, a single culture of S. cerevisiae is grown and two samples are drawn from it and sequenced separately. In true technical replicates, all steps are replicated, including library preparation. Biological replicates consist of different biological samples that are processed through the sequencing process separately, for example, two cultures of S. cerevisiae are grown separately and sequenced separately.



Deciding whether to use biological or technical replicates is fairly straightforward. You simply have to decide whether or not you have biological variance in your system. For example, if you want to indentify the genes that are differentially expressed between two strains of yeast then you will most likely grow each of the two strains in different flasks. Growing the strains in different flasks of course introduces some type of biological variance. This can be seen because if you grew two flasks of the same yeast strain then the expression would be different. These differences are caused by biological variance. Thus, biological replicates are required for this experiment.

There may be cases where there is not biological variance. For example, if you are interested in the differences between tumor tissue and matched normal tissue in a single patient then it could be argued that you only have technical variance. In this case there may be only one tissue in each control and test condition. Thus, technical replicates could be used. However, you cannot say anything about cancer in general from this experiment because it only looked at one patient. You can only limit your findings to *that one patient*, and specifically about the two sections of tissue that were involved. However, we expect that in most differential expression experiments there will be some biological variance and biological replicates will be required.

If this explanation unclear to you then you may find it helpful to read our tutorial, Thinking About RNA Seq Experimental Design. If you are still confused then you should probably e-mail us at busbym at bc dot edu because a proper design cannot be devised without understanding these principles.

Once pilot samples are sequenced, the reads will need to be aligned to a reference genome or transcriptome. Then, the number of reads which are aligned to each gene must be tabulated. The alignment and quantification process is currently outside the scope of Scotty, but we can recommend our lab's gkno Project tools. If you have difficulty getting started please contact us for assistance at busbym at bc dot edu and we will be happy to help.

How deeply should I sequence pilot data?

Pilot data does not have to be sequenced as deeply as the final sequencing run. However, Scotty will calculate the total power based on the number of genes that are observed in the pilot data with at least a read count of 1. That is, if 10,000 genes are quantified in the pilot data and Scotty expects you to be able to find 5,000 of these genes differentially expressed the total power is reported as 50%. The reason Scotty calculates the power this way is that it is difficult to accurately estimate the number of genes that are expressed but unobserved in the data. These unobserved genes are expected to have as substantially lower power than the observed genes and an incorrect number of them can skew the results. In general, it is reasonable to assume that there will be very fairly low power to detect differentially expressed genes that were not observed in either replicate of the pilot samples. To see this, the power of unobserved genes is shown in the final output chart. Scotty does not attempt to predict the number of genes in that class. The rarefaction output chart in Scotty gives an idea of how quickly new genes are quantified. For some species it is practical to sequence even pilot data to saturation. That is, all of the genes that can be observed have been observed.

How do I make sure my pilot data looks like my experimental data?

Control the environment in the same way that you will control the environment in the actual experiment. Environmental effects, such as growing samples in separate batches can significantly alter the power of experiments. Further, we recommend using the same protocol that you will use in the actual experiment. We observed a lot of variability in how many genes were required to observe a fixed number of genes, and expect the results to vary by the library preparation protocol used.

Upload File Format

Pilot data should be uploaded as a tab-delimited text file in the following format:

Gene	Control_1	Control_2	Control_3	Test_1	Test_2
Gene1_ID	153	225	16	7	0
Gene2_ID	20	15	16	13	12
Gene3_ID	1	0	2	3	2
Gene4_ID	15	4	10	3	11

...
To ensure a reliable upload, the gene names should contain only letters and numbers (no spaces). If you have both technical replicates (lanes or libraries from same biological sample) and biological replicates (lanes or libraries from different biological samples) all of the technical replicates from each biological sample should be added together.

Thus, if you have samples from four different cultures sequenced with two lanes each your pilot data should have four columns, not eight. For paired end data count each pair once.

Selecting Preloaded Data

A less accurate approach
CAUTION In the Scotty manuscript we speak at length about the potential shortcomings of using existing data to plan future experiments.

In short, we looked at several different experiments and found that the two factors that determine the power of an experiment (how many reads it took to measure new genes and how much variance there was between replicates) were not repeating very well between experiments, even when the experiments were nominally "the same", that is using measuring the same type of sample. We emphasize that we do not believe that this is not a shortcoming of the studies we examined. Rather, this happens for a variety of reasons and is to be expected in the normal course of laboratory work. For example, it would be expected to be more difficult to control the biological in an experiment with a large number of replicates and thus these samples would be expected to have higher variance compared to samples in smaller experiments.

Further, the number of reads that is required to measure new genes will vary between libraries based on several factors that can influence library complexity. For examples, if there is a low quantity of input RNA available in one of the experiments you may see a higher rate of read duplicates (an artifact) and the loss of certain RNA species. Thus, there will be a requirement for deeper sequencing to observe the same number of genes, and some genes may not be observed at all. Certain library preparation protocols may be better at producing complex libraries than others. Because protocols are always improving, we extend the caveat that the data in existing experiments may not be an ideal model for your experiment.

Thus, choosing a good model of pilot data for your experiment is tricky and the results of this analysis will certainly be less accurate than using your own pilot data. However, we believe that pre-loaded data will provide a better informed design than no data, and it is useful when it is not possible to generate pilot data, at the early stages of experiment planning, and after pilot data has been run to serve as a benchmark for measuring how well your own libraries are replicating.

Using Quantifications From Other Programs

The Scotty statistical model assumes that the read counts will represent the number of reads that align uniquely to a single spot in the reference. The reference can be the genome, in which case Scotty will estimate the power to quantify gene expression, or the transcriptome, in which case Scotty will estimate the power to quantify whether each transcript is differentially expressed.

In the case of a transcriptome alignment, users can expect that for complex species such as human the majority of reads that align to a transcript will not align uniquely. That is, the reads are likely to align to exons that are present in more than one isoform originating from the same gene. Therefore, substantially deeper sequencing is required to quantify differential expression at the transcript level.

Several strategies have been published that consider reads mapping to multiple isoforms

and call transcript abundance based on both the uniquely and multiply-mapped reads. Scotty can be used with the output of these programs with the following considerations and cautions:

To illustrate how Scotty works with the output of these programs, we will first consider an example using a very simple reallocation method. We will consider an idealized example of a gene with two alternative transcripts of the same length (A and B) and completely even read coverage. 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 a basic reallocation 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 counting noise that you would see if the measurements were actually higher due to deeper sequencing. This occurs because the re-allocation itself is based on counts which a have level of noise and is thus itself noisy.

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 two conditions, and an accurate relative measurement can be attained from the unique reads alone. While more complex algorithms exist than this simple approach, these methods 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 across replicates (and decreases power).

We considered both the strengths and the limitations of Scotty in estimating power based on the output of these programs and offer this guidance to our users:

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 level of counting noise 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. Further, it is likely to be experiment-specific in regards to factors such as how well the data conforms to the model's assumptions of parameters such as evenness of coverage.

Thus, Scotty cannot completely address this inherent characteristic of the data, but we do offer this guidance to users who wish to analyze transcripts quantified by programs that consider multiply-mapped read counts. We offer suggestions: 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. 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 are likely to provide a reasonable approximation of the power that will be attained if you also include the multiply-mapped reads, which contain relatively less information.

This assumption is easy for the user to test if they process their pilot data using both the transcript caller and only 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.

Cost Data

Cost data should be entered as follows:

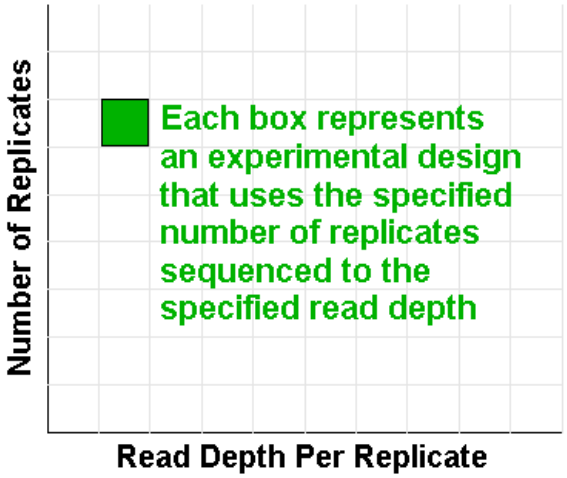
Cost per replicate, excluding reads - This is the cost of producing each replicate, for instance the cost of obtaining the biological material and preparing the library.

Cost per million reads that are aligned to genes - A certain portion of the reads that are sequenced will contain too many errors to align to the reference. The portion that is not alignable will vary by experiment. Scotty only considers the reads that are aligned to genes. Therefore, the cost per read should reflect how much each aligned read cost. For example, if one million reads cost $1,000 but only half of those reads align to genes then you should enter $2,000 as this cost.

The currency does not matter. We just put $ because that is what is on our keyboard.

Constraints for Power Optimization

Scotty will calculate the power, cost, and measurement bias for a number of experimental configurations having different read depths and replicate number:



The first two inputs specify which experimental configurations to test:

Maximum number of replicates per condition - Scotty will test experimental configurations that use between two and the maximum number of biological replicates specified. Each experiment that is tested contains the same number of replicates in the control and test condition. While we display the power observed for only two replicates, we generally recommend that the minimum number of replicates used is 3. Gene variance is very poorly measured with only two replicates and it is difficult to detect an outlier (bad) replicate if there are only two samples.

Assess the power of sequencing depths between (entries) - For each replicate number, Scotty will test 10 different read depths between the minimum and maximum values that are specified. The read depth is *not* the number of reads sequenced. It is the number of reads that align to genes. Therefore, if you expect that about half your reads will align to reads, you will need to sequence twice as many reads as Scotty calculates to get to the correct read depth.

The next inputs specify the constraints that are on the experimental design:

Detect at least (a certain percentage) of genes that are differentially expressed by (a fold change) at a (p-value) - Differential expression is defined as a fold change relative to the control condition. For example, in a 2X fold change the test condition will have expression that is twice as high as the control condition.

The percentage of genes detected is defined as the percentage of genes with a true fold change of the magnitude specified in the test condition relative to the control condition. Genes that are expressed at levels that are too low to be

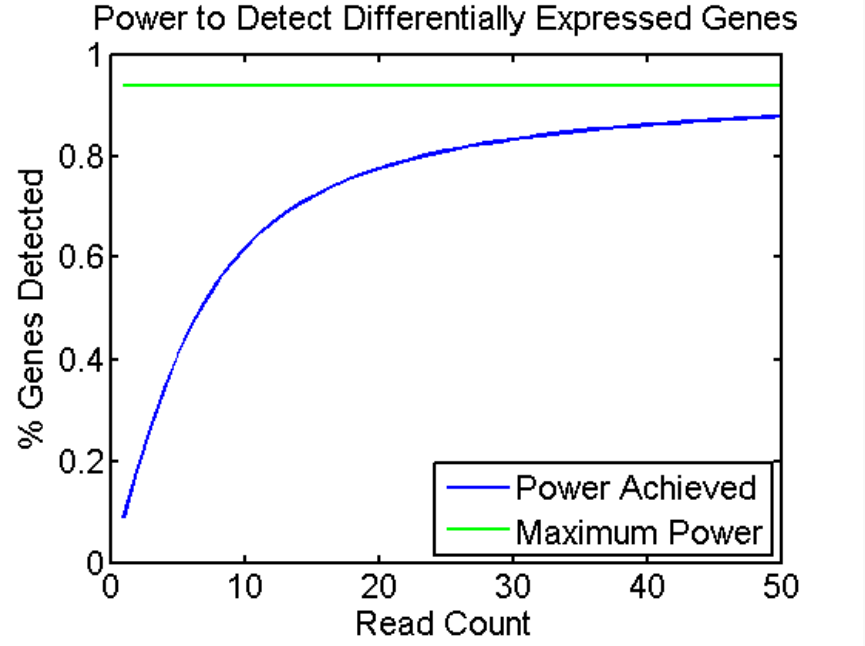
observed in the protoype data are not included in this calculation. Use the full percentage, e.g. 80 not 0.8 which we will interpret at 0.8%.

The p-value is based on the percentage that will be calculated in a t-test. See Scotty's accompanying paper for details of the statistical test and comparisons of power between the t-test and a negative binomial test.

Experiment will cost no more than (budget) - Scotty adds up the cost of the replicates and the cost of the reads and tells the user which experiments are feasible. This does not include data analysis costs. Processing a large number of reads will require computational power beyond what is available on normal desktop computers. These resources may be available at, for example, university computer clusters. Alternatively, processing power can be obtained through cloud services but should be budgeted for.

Limit measurement bias by measuring at least (percentage) of genes with at least (percentage) of maximum power - This is the most complicated of Scotty's inputs. RNA-Seq measurements quantify data with read counts. Because of the count nature of the data, genes which are measured with low read counts (particularly below 10) have imprecise measurements. The imprecision occurs because you can be fairly certain that counts of 100 and 200 represent true differences than you can be about counts of 1 and 2. (See our tutorial Thinking About RNA Seq Experimental Design for a longer explanation of this effect, which is called Poisson variance). Because of this effect, there will be substantially lower power to detect differentially expressed genes measured with low counts than there would have been if the expression of these genes had been measured on a continuous scale (e.g. a continuous measurement of 3.5 reads versus the forced binning of measurements into 3 or 4 reads).

However, the amount of uncertainty that is introduced into the measurements because of their count nature can be easily calculated. Thus, it is possible to calculate how much power the measurement would have had if the gene expression had been measured with a continuous measurement. We define the power that would have been calculated with continuous measurements as the **maximum power**. As the read count increases, bins between the counts become smaller relative to the count, and the power that is achieved using counts asymptotically approaches the maximum power:



In some experiments, it is undesirable to have a large number of genes measured with less power than other genes because it adds complexity to downstream analyses. This is particularly true in cases where findings of functional enrichment, for example, are in and of themselves experimental results. In such cases, it is important to account for measurement biases in the analysis to make sure that enrichment for individual biological categories is observed because gene in these categories are more likely to be differentially expressed rather than they are more likely to be detected if they are differentially expressed. This might be the case for genes which are measured with a high number of reads. While this analysis can be done, it is more difficult and will add some expense to

downstream analyses. Thus, it may be worth considering whether spending more to obtained less biased data may be paid for by the cost saving achieved by a simpler analysis, particularly in species with smaller transcriptomes such as yeast where sequencing samples beyond Poisson noise is economically feasible. In other experiments, such as when the primary findings are target identification, measurement biases will matter less.

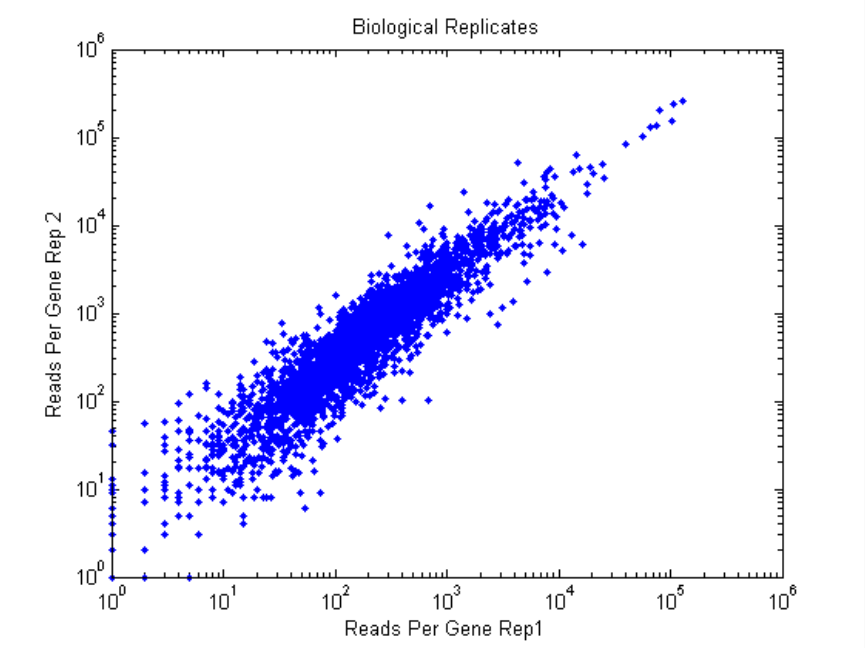
Outputs

Here we show some examples of output data. Most of the data shown is taken from the human liver data samples from Bleckhman et al. 2010.

Replication

We charted the reads per gene for each replicate against the first replicate in the condition (i.e. Control Replicate 2 versus Control Replicate 1).

This is an example of a fairly typical pair of biologcial replicates:



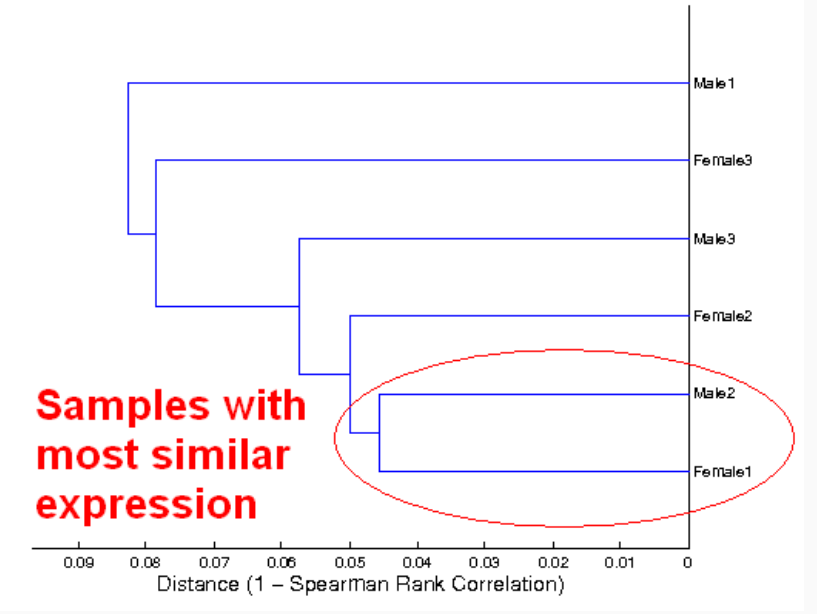
If your pilot data is more dispersed than this then it could indicate a problem. If it is truely a mess then we would recommend first checking whether your data file is corrupt. If not, then there might be a problem with the library prep or the samples may not be replicating properly.

Control Versus Test Data

If you entered both control and test data, Scotty charted the reads per gene for the control data versus the test data. You can use this chart to show how different the control and test sample are. The black line going through the center of the data points is the slope line. The slope line is used to normalize the test and control samples. This line should be going through the center of the data. If the line seems far off, this could mean that you have a lot of differential expression in your samples, or that the library complexity between the control and the test data is very different. If this is the case, you may want to investigate including spike ins in your library prep to get more accurate sample normalization.

Clustering

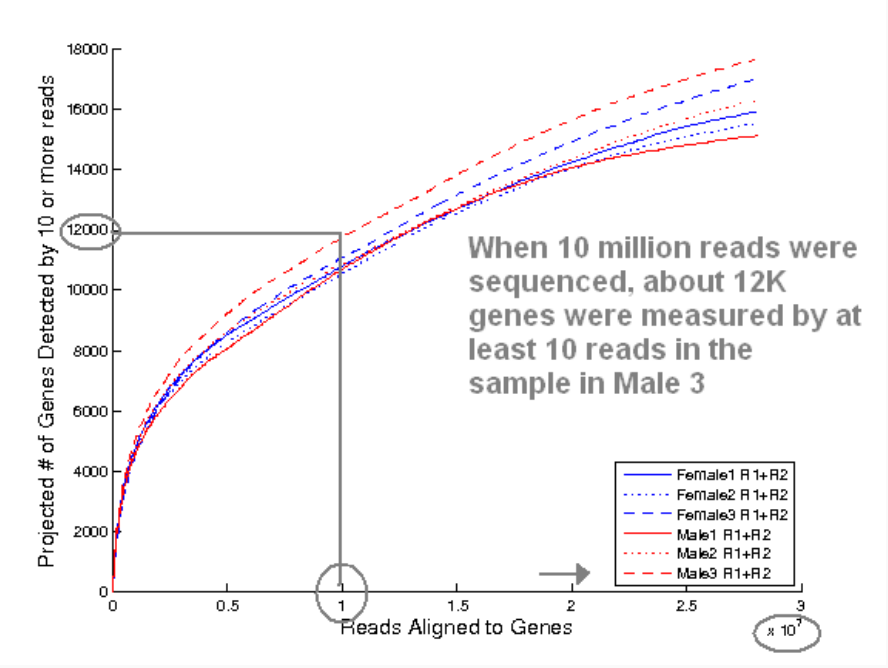
Scotty clusters the data to identify any problems, such as outlier samples and sample swaps. In general, you should expect samples from the control condition to cluster in one branch, and samples from the test condition to cluster in another branch. This, however, will not be the case if the groups are similar. For instance, if you are comparing gene expression between two HapMap groups then we expect the groups to share branches because, while expression individual genes may be different, the global expression is close enough that there is no clear delineation between the groups using this clustering (based on what we have observed in existing datasets). However, this clustering has proven useful for us in the past to identify an instance where one of our samples was mislabelled. For this cluster we used a Spearman rank correlation as the distance metric. We excluded genes expressed with fewer than 10 reads to avoid the findings of the chart being dominated by Poisson noise.



In the example above, we see that the samples labeled Female1 and Male2 had the most closely related expression. If there were large differences between male and female expression you would have expected the male samples to cluster together, and the female samples to cluster together separately. However, that effect was not observed.

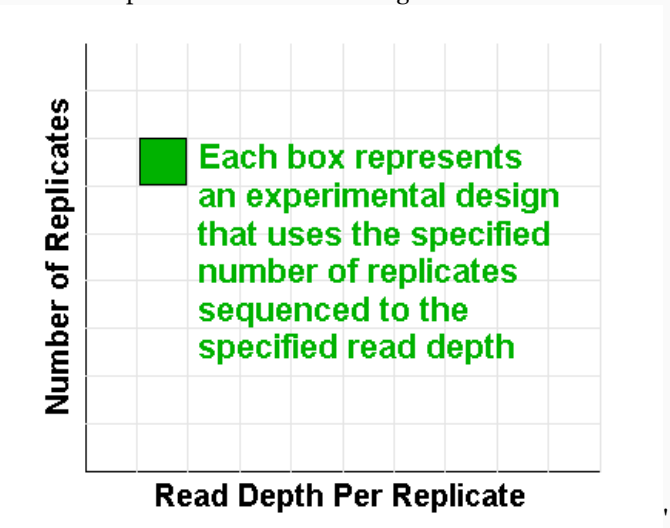
Genes Detected as a Function of Sequencing Depth

As sequencing depth increases, a greater number of genes will be measured. This chart shows how many genes are measured by at least 10 reads as a function of sequencing depth. We chose 10 as the cutoff because measurements with fewer than 10 reads tend to be imprecise due to the effects of Poisson counting noise. Samples that are nominally measuring the same thing within the same experiment (i.e. all human liver samples) would be expected to have similar rarefaction curves. This example shows human liver data taken from six different individuals.



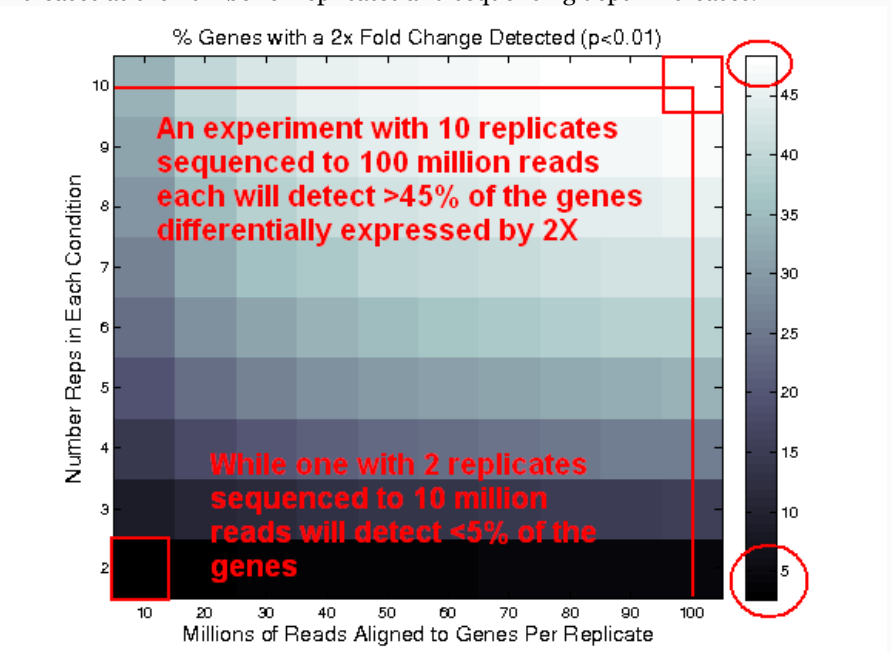
Optimization

Results are presented in the following format:



Power Optimization

The power optimization chart shows how the power to detect differentially expressed genes increases as the number of replicates and sequencing depth increases.



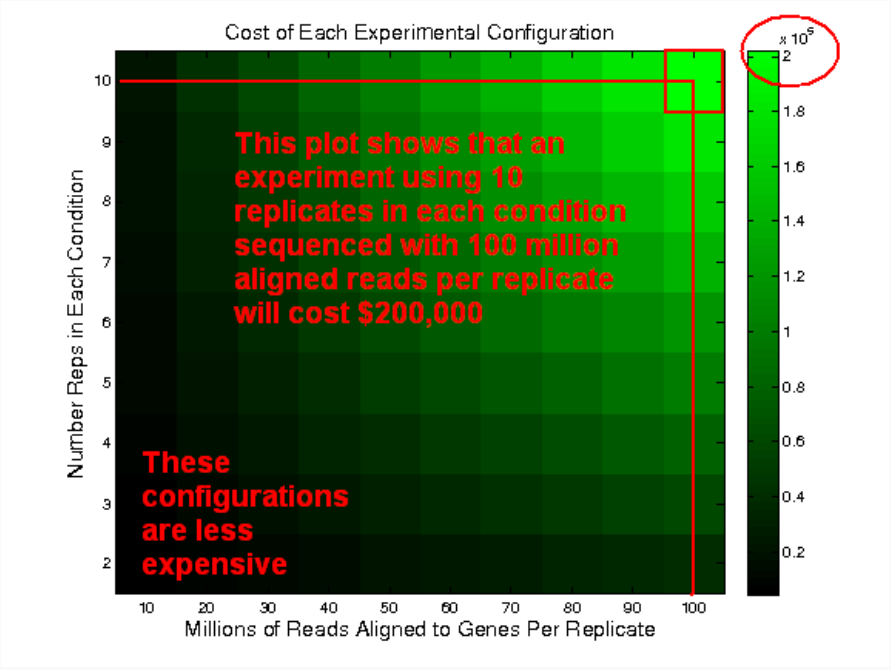
Genes that are expressed at levels that are too low to be observed in the protoype data are not included in this calculation. The reason that we excluded these genes is as follows: It is easy to imagine that with an infinite number of samples and infinite sequencing all of the annotated genes would eventually be detected as expressed. However, as a practical matter, many of these genes appear to be expressed at very low levels (below one transcript per cell). 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 the genes that are present in the protoype data are the genes that are of interest to the user.

It is worth noting that in datasets where many genes are measured with fewer than ~10 reads, as is the case in most human datasets, the calculation of total power may not be the most informative metric. In these cases, when there is a sufficiently high replicate number, the total power will be driven primarily by the mixture of read depths. For genes measured with high counts, almost all of the differentially expressed genes will be detected, while for genes measured with low counts almost none of the differentially expressed genes will detected. The total power, therefore, becomes primarily a function of how many genes have high versus low counts. The total power metric is therefore best used in conjunction with the final output chart in Scotty, which shows the power by read depth for allowed experiments. (Note that this chart does not appear if there are no allowed configurations.)

Cost Optimization

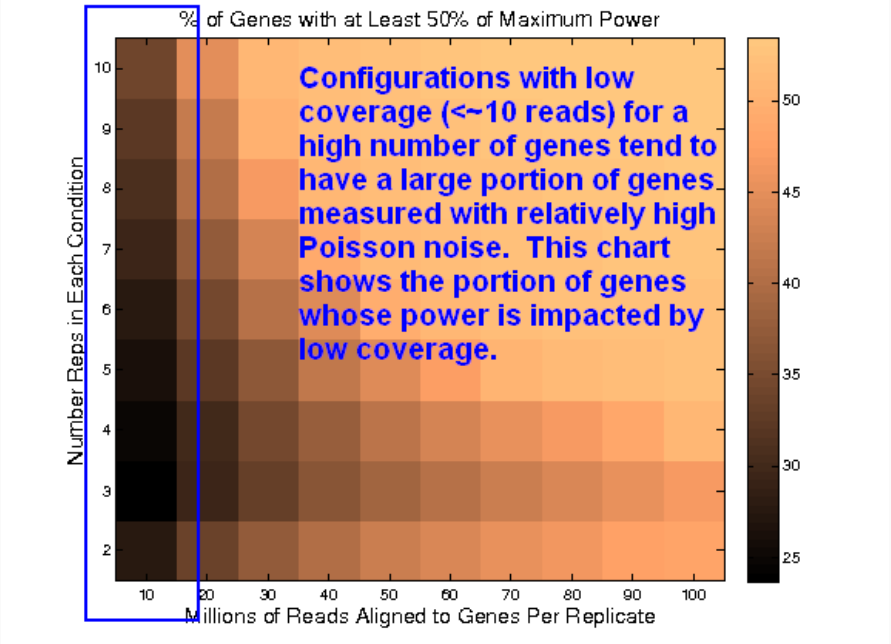
The cost optimization chart shows how the cost of the experiment will increase as the reads sequenced and the number of replicates used will increase.

Note that Scotty's models are based on the cost per read that is aligned to a gene. We use this model to simplify user inputs and because alignment rates differ between experiments and between sequencing centers. A fraction of sequenced reads will not align to genes due to errors in the reads and alignment to non-gene regions. Therefore, the cost per aligned read is higher than the cost per sequenced read. For the pre-loaded datasets, the percentage of sequenced reads that were aligned to genes is available in the documentation for the pre-loaded datasets.



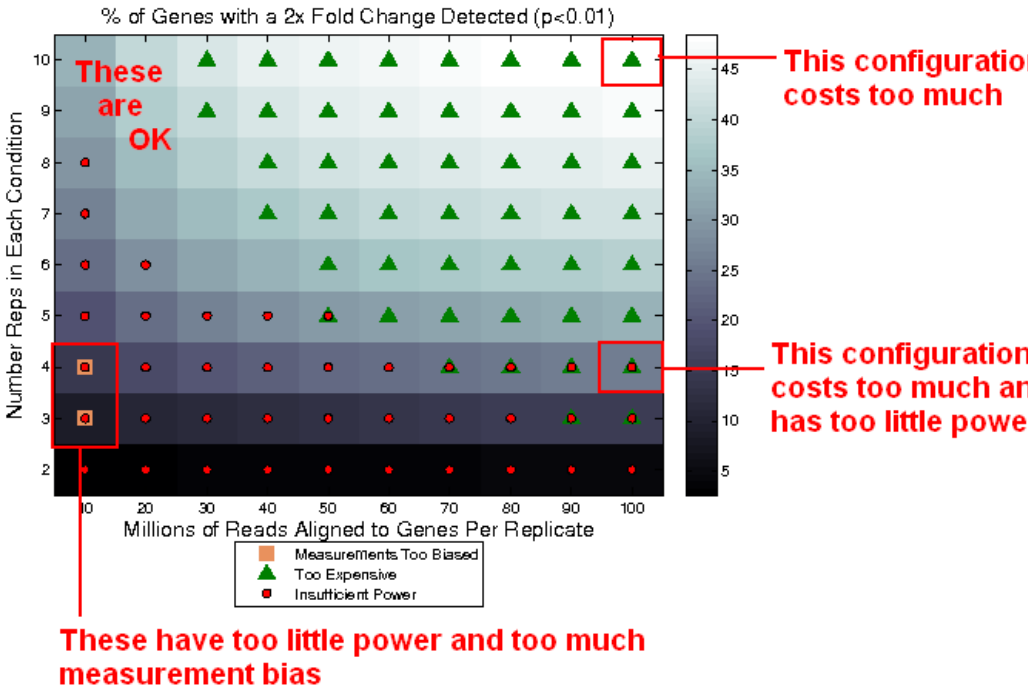
Detection Differences Due to Poisson Noise

The measurement bias metric is described in detail above.



Excluded Experimental Configurations

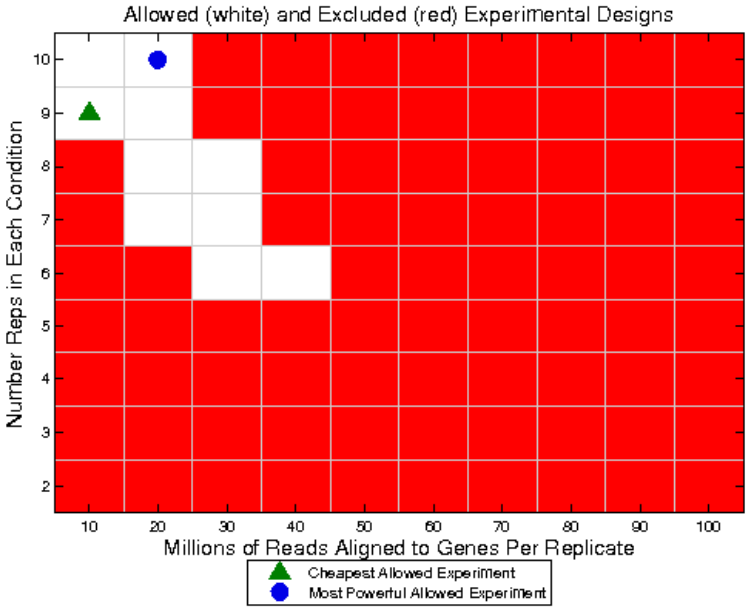
This plot shows which experimental configurations meet the user's requirements.



In this experiment, we see that the experiments in the top right hand corner are too expensive. Those on the bottom have insufficient power. This leaves ten of the original 90 configurations that meet the user's requirements. The reasons for the exclusions are given in the key.

Allowed Experimental Configurations

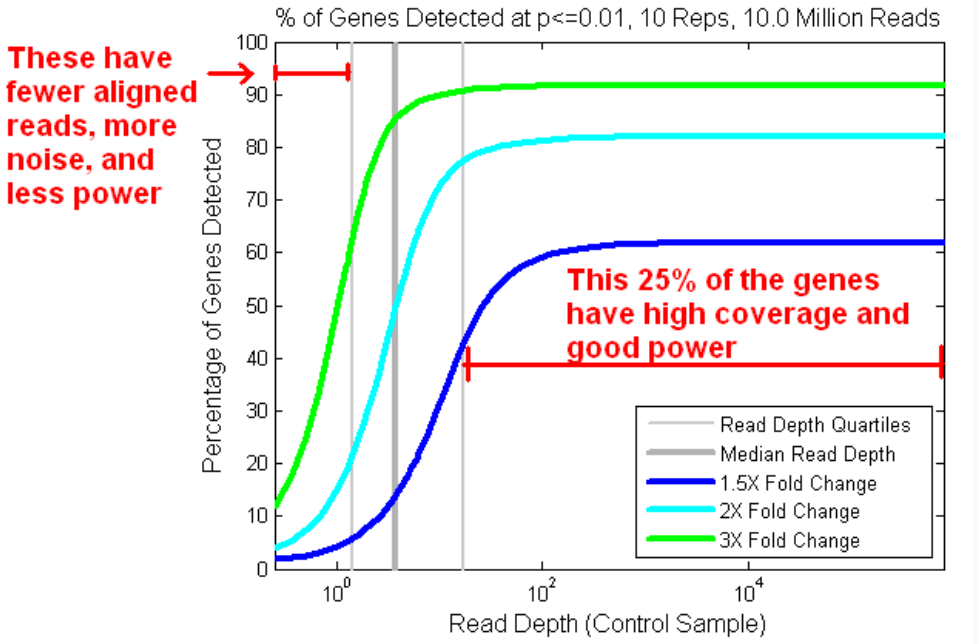
This plot shows the optimal experimenetal configurations in a simpler format.



All experimental configurations plotted with white boxes meet the user-defined criteria. In this experiment, the cheapest of these allowed configurations uses 9 replicates of each condition, sequenced with 10 million reads each. The most powerful has 10 replicates with 20 million reads per replicate.

Predicted Statistical Power

In these plots we show the statistical power to identify a 3X, 2X, and a 1.5X change in expression.



In the example above, notice how much more cheerful this plot is than the previous power plots. While the overall power of the experiment is poor (around 45%) in this case this is due mostly because of the existence of a large class of genes which are measured with few reads, and have low power. The number of these genes can be seen in the lines defining the quartiles of expression. Among the genes that are measured with counts that are high enough not to be affected very much by Poisson noise (~10 or more) the power is much higher. This chart indicates that under this experimental configuration most changes in expression in genes with low read counts will be missed, but most changes in expression among higher read count genes will be found. The total power of the experiment then becomes dependent on what this coverage mix is.

Why Does Scotty Use a T-Test?

Our selection of a t-test as our statistical methods was made after considerable deliberation. The primary reason we used a t-test is that we think it is a very good test.

For example, in the Scotty manuscript we used simulations to compare the t-test's ability to identify differentially expressed genes to a widely-used benchmark method that is based on the negative binomial distribution and uses an information-sharing strategy to calculate variance. We found that the t-test does identify fewer differentially expressed genes when the experiment contains are only two replicates. This is to be expected because variance measured with only two replicates is inaccurate. However, when there were at least three replicates the t-test performed nearly as well as the benchmark test and with 4 to 6 replicates the t-test comparably. At all read depths above one read per gene the false positive rate was closely matched by the p-value, despite the fact that RNA Seq data is count-based. We found that this performance and the accuracy of the p-values was maintained whether expression was modeled as being normally or lognormally distributed. Additionally, the t-test does not introduce any biases into the findings which can occur when information-sharing approaches are used to calculate the variance. This, we think makes downstream analysis easier because findings of, for example, enrichment in biological categories are less likely to be influenced by biases in the statistical test. Formulas are also readily available, which improves Scotty's performance. We use the formuas decribed in Chow, Shao, and Wang 2002 which are more accurate at low sample sizes than the ones commonly found in textbooks.

We would generally recommend using at least three replicates. An experiment that uses only two it is likely to be underpowered and it is also difficult to tell if one of your replicates is an outlier.

Beyond Differential Expression

The power calculated by Scotty is only predictive of the power to detect differential gene expression between control and test conditions. As a general rule, other purposes, such as variant detection, require much deeper coverage. Some tools for splice detection, for example, only perform well when there are hundreds of reads aligned to the spliced gene. Accurate single nucleotide variant detection require several reads to align across each individual variant.

For comparison, most of the genes detected as expressed in the earlier Solexa sequencing runs of human contained fewer than 10 reads across the entire gene, which may stretch for thousands of base pairs. These measurements, therefore, do not contain enough information to detect much besides expression level for the majority of the genes that were expressed in the sample.

To determine how deeply samples need to be sequenced for other purposes, we recommend reading the papers that describe tools for processing the data and seeing how many reads are used in the performance benchmarks. The figures under the optimization charts can be used to give you an idea of how many genes will be quantified at a given read depth.

Citations Keep the Money Rolling

If you use Scotty to design your experiment, we hope you will have with awesome, publishable results! If you do, please don't forget us! We can be referenced in our paper in Bioinformatics:

Scotty: A Web Tool For Designing RNA-Seq Experiments to Measure Differential Gene Expression. M.A. Busby; C. Stewart; C. Miller; K. Grzeda; G. Marth Bioinformatics 2013; doi: 10.1093/bioinformatics/btt015

Thank you!

Contact Us

We love hearing from our users! Contact Michele Busby with any questions at busbym at bc dot edu.

Good Luck!

Scotty is dedicated to Hillary St. Pierre. May all your experiments be as powerful as Hillary.