Design and analysis of Bar-seq experiments

David G. Robinson[1], Wei Chen[2], John D. Storey[1,4] and David Gresham[3,4]

[1]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton. [2]Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, 13125 Berlin, Germany. [3]Center for Genomics and Systems Biology, Department of Biology, New York University, New York.

[4]correspondence: dgresham@nyu.edu and jstorey@princeton.edu

Running title: High-throughput genetic screens

Key words
- yeast
- Bar-Seq
- galactose
- functional genomics
- Sacchromyces cerevisiae

David Gresham
12 Waverly Place, Rm 203
New York Univeristy
New York NY 10003
USA
Ph: 212-998-3879
E-mail: dgresham@nyu.edu

John Storey
Carl Icahn Labs
Princeton University
Princeton, NJ 08544
USA
Email: jstorey@princeton.edu

High-throughput quantitative DNA sequencing enables the parallel phenotyping of pools of thousands of mutants. However, the appropriate analytical methods and experimental design that maximize the efficiency of these methods while maintaining statistical power are currently unknown. Here, we have used Bar-seq analysis of the *Saccharomyces cerevisiae* yeast deletion library to systematically test the effect of experimental design parameters and sequence read depth on experimental results. We present computational methods that efficiently and accurately estimate effect sizes and their statistical significance by adapting existing methods for RNA-seq analysis. Using simulated variation of experimental designs we find that biological replicates are critical for statistical analysis of Bar-seq data whereas technical replicates are of less value. By sub-sampling sequence reads we find that when using four-fold biological replication, 6 million reads per condition achieves 96% power to detect $\geq$2-fold change at a 5% false discovery rate (FDR). Our guidelines for experimental design and computational analysis enables the study of the yeast deletion collection in up to 30 different conditions in a single sequencing lane. These findings are relevant to a variety of pooled genetic screening methods that use high-throughput quantitative DNA sequencing including Tn-seq.

INTRODUCTION

Uncovering the connection between genotype and phenotype remains one of the central challenges of modern genetics. At the same time, the rate at which new genomes are sequenced currently outpaces our capacity to functionally annotate those genomes. Addressing these challenges requires efficient means of quantifying phenotypes associated with defined genetic perturbations. Methods for uniquely identifying and quantifying phenotypic effects of mutant alleles in complex mixtures enables the parallel analysis of hundreds to thousands of genotypes. Pooled mutant analysis entails the use of either libraries of defined mutants tagged with unique DNA sequences (molecular barcodes) [28, 8] or complex libraries of tens of thousands of unique mutants generated by random insertional mutagenesis. Analogously, comprehensive libraries of short hairpin RNAs (shRNAs) enables parallel analysis of perturbations of mammalian genes in cell culture [17, 18, 19].

Recently, methods for estimating mutant abundances in complex mixtures have been introduced that capitalize on advances in high-throughput quantitative DNA sequencing. Barcode Analysis by Sequencing (Bar-seq) was first developed to analyze libraries of thousands of *Saccharomyces cerevisiae* gene deletion mutants [20] and has subsequently been used to analyze a library of deletion mutants in *Schizzosaccharomyces pombe* [10]. The use of Bar-seq enables efficient, accurate and comprehensive genetic screens for addressing a variety of questions such as defining the genetic requirements for initiation and maintenance of cell quiescence in response to distinct starvation signals [9]. In organisms for which barcoded mutant libraries are not available, high-throughput DNA sequencing of pools of transposon insertion mutants (Tn-seq) enables multiplexed mutant analysis. Tn-seq was initially applied in studies of *Streptococcus pneumonia* [26] and *Haemophilus influenzae* [7] and has subsequently been adapted for use in diverse organisms [3, 6]. Similarly, PhiTSeq facilitates simultaneous analysis of thousands of transposon-mutagenized haploid human cells [4]. The wide-spread adoption of pooled mutant screens using high-throughput quantitative DNA sequencing attests to the power of these methods for efficient genetic analysis.

In contrast to the rapid technological advances in pooled mutant analysis, there has not yet been a statistical treatment of the experimental design and analysis of data generated by high-throughput DNA sequence analysis of these complex libraries. Thus, major methodological and analytical questions remain unanswered. What is the appropriate statistical framework for analyzing DNA sequence count data? What are the sources of variation? What is the appropriate study design for maximizing the power and accuracy to detect differences in mutant abundances? What sequence read depth maximizes the precision of these methods while minimizing the cost and resources required?

We undertook a study that aimed to address these questions with the goal of providing guidance for the design and analysis of pooled mutant screens using high-throughput DNA sequencing. Using experimental analysis of the *S. cerevisiae* gene deletion collection in two different conditions we studied the contribution of treatment and biological and technical variation to Bar-seq data (**Figure 1**). We demonstrate that the negative binomial model used to analyze RNA-seq data is also directly applicable to Bar-seq data. Using computational subsampling of our experimental data, we studied the effect of different experimental designs on the results from Bar-seq analysis. We find that biological replicates substantially improve statistical power whereas technical replicates provide only moderate additional statistical power. We also find that increasing sequencing depth beyond 6 million reads per condition provides limited improvement to the experimental results regardless of experimental design.

Our results provide information directly relevant to designing future high-throughput quantitative DNA sequencing experiments of pooled mutants. For example, using an experimental design of four-fold biological replication and no technical replication, we show that detection of mutants in the 4295 strain yeast deletion collection with $\geq$2-fold change between conditions can be achieved with 96% power at a 5% false discovery rate (FDR) using as few as 6 million reads per condition. This corresponds to a requirement of 1397 sequence reads per mutant per condition or 349 reads per biological replicate library. Using our experimental and analytical methods for Bar-seq analysis it is possible to analyze the yeast deletion collection in up to 30 different conditions using a single 200 million read lane without sacrificing statistical power. Our findings should be informative for other methods of pooled mutant analysis such as Tn-seq.

MATERIALS AND METHODS

Strains, Media and Sampling Procedures: We used a haploid prototrophic gene deletion collection constructed using the synthetic genetic array method [24]. The library contains the identical gene deletion alleles as the standard yeast knockout collection [28] excluding gene deletions that result in auxotrophies. Gene deletion alleles are marked with the kanMX4 cassette conferring G418 resistance, which is flanked by a unique 5' molecular barcode (the UPTAG) and a unique 3' molecular barcode (the DNTAG). Each MAT**a** mutant contains a functional copy of the *URA3*, *LYS2*, *LEU2*, *MET15* genes and the $can1\Delta::STE2_{pr}$-*SpHIS5*, *lyp1*Δ*0* and *his3*Δ*1* alleles. We used standard YPD and YPGal media containing either 2% glucose or 2% galactose respectively [1].

Following growth of individual mutants on YPD agar plates, all mutants were pooled to a final density of 1.5 x 10^9 cells/mL. Each agar plate contained single colonies of individual genotypes and replicated colonies of the control *HO*Δ*0* strain. To define the replicated time-zero (t_0) samples we obtained two independent samples of 0.5mL (i.e. 7.5 x 10^8 cells) from the pooled library. We inoculated 5μL from the pooled library (i.e. 7.5 x 10^6 cells) into four-fold replicated cultures of either 5mL YPD or YPGal. Cells were grown for 24 hours (t_0) to a final density of 3.3 x 10^8 cells/mL in both conditions. We removed 2mL (i.e. 6.6 x 10^8 cells) samples from each of the four YPD cultures and four YPGal cultures and purified genomic DNA using Qiagen Genomic-Tip 100 columns.

Library Preparation and Sequencing: We designed a two-step PCR protocol for efficient multiplexing of Bar-seq libraries. In the first PCR step UPTAGs from a single sample were amplified with the primers *Illumina UPTAG Index #* (5'-

ACG CTC TTC CGA TCT NNNNN GTC CAC GAG GTC TCT-3') and *Illumina UPkanMX* (CAA GCA GAA GAC GGC ATA CGA GAT GTC GAC CTG CAG CGT ACG-3') and DNTAGs from the same sample were amplified with the primers *Illumina DNTAG Index #* (5'-ACG CTC TTC CGA TCT NNNNN GTG TCG GTC TCG TAG-3') and *IlluminaDNkanMX* (5'-CAA GCA GAA GAC GGC ATA CGA GAT ACG AGC TCG AAT TCA TCG-3') in separate PCR reactions. Illumina UPTAG and Illumina DNTAG primers contain a 5 base pair sequence (denoted NNNNN in the primer sequence) that uniquely identifies the sample. We designed 120 unique sample indices that differ by at least two nucleotides. A complete list of primer sequences is provided in **Table S1**. We normalized genomic DNA concentrations to 10ng/μL and using 100ng of template amplified barcodes using the following PCR program: 2 minutes at 98 °C followed by 20 cycles of 10 seconds at 98 °C, 10 seconds at 50 °C and 10 seconds at 72 °C and a final extension step of 2 minutes at 72 °C. PCR products were confirmed on 2% agarose gels and purified using QIAquick PCR purification columns.

We quantified purified PCR products using a Qubit fluorimeter and combined 60ng from each of the 20 different UPTAG libraries and, in a separate tube, 60ng from each of the 20 different DNTAG libraries. The multiplexed UPTAG libraries were then amplified using the primers *P5* (5'-A ATG ATA CGG CGA CCA CCG AGA TCT ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT-3') and *Illumina UPkanMX* and the combined DNTAG libraries were amplified using the *P5* and *IlluminaDNkanMX* primers using the identical PCR program as the first step with 20ng of template. The 140 base pair UPTAG and DNTAG libraries were purified using QIAquick PCR purification columns, quantified using a Qubit fluorometer, combined in equimolar amounts and adjusted to a final concentration of 10nM (i.e. 0.924 ng/μL). In total, the sequencing library contained 20 UPTAG and 20 DNTAG libraries from 20 different samples (**Table S2**). The library was sequenced on a single lane of an Illumina HiSeq 2000 using standard methods including the use of the standard Illumina sequencing primer (5'-ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT-3'). The qseq files for each of the 20 samples are available from the NCBI Short Read Archive with the accession number SRA101498.

Read Matching and Statistical Analysis: Sequence reads were matched to the yeast deletion collection barcodes reannotated by Smith et al 2009. Inexact matching was performed by identifying barcode sequences that are within a Levenshtein distance of 2 from each read [13]. Reads matching equally to multiple barcodes were discarded. Sample indices were similarly matched using a maximum Levenshtein distance of 1. The final matrix of counts matching the UPTAG and DNTAG of each of the 20 samples is provided as **Table S3**. A set of 359 outliers was identified that had fewer than 100 total reads across all 20 samples (**Figure S1**). These low-count matches are likely due to sequencing error, and were removed. In addition, our pooled yeast gene deletion library included a highly abundant strain (the HO gene deletion mutant, which was present on each of the 96-well plates containing individual mutants prior to pooling). The HO deletion mutant represented 19% of all reads and was removed prior to computational analyses leaving a total of 139.8 million reads mapped to 4295 mutants.

Eigen-R^2 was used to determine the percent of variance explained by the different factors in our experimental design for the t_{24} samples [5]. Barcode counts were normalized using the TMM method [16] after adding 1 to each value, and then were log-transformed, to avoid including differences in per-library read depth as a source of variation. The bottom 10% of mutants were filtered out because lower counts have a disproportionate effect on the technical variation. Eigen-R^2 was used to compute the percent of variance explained by the treatment factor (R^2_T) and the biological replicate factor (R^2_B). As the treatment factor is contained within the biological factor, we report the biological percent of variation as $R^2_B - R^2_T$, and the technical variation as $1 - R^2_B - R^2_T$.

For differential abundance analysis, we first summed UPTAGs and DNTAGs for technical replicates within each biological replicate. The `edgeR` package (version 3.2.4) was used to perform dispersion estimation and to perform an exact negative binomial test to calculate a p-value and log fold change for each mutant using the `exactTest` function, using the default parameters [15]. The `qvalue` package was used to compute q-values [22].

Gene set enrichment analysis was performed using the Biological Process ontology from SGD. Gene sets that had fewer than four genes among the detected deletions were discarded in advance. We used the Wilcoxon rank-sum test to compare the distribution of the estimated log fold changes within each gene set to those outside of the set [9]. We used the `qvalue` package to set a FDR=5% threshold above which gene sets were declared significantly enriched.

Read Subsampling: Separate subsamplings were performed for each combination of replicates in each design. This requires 1 combination for the full $2 \times 4 \times 2$ design, 2 combinations for the $2 \times 4 \times 1$ design, $\binom{4}{3} = 4$ combinations for the $2 \times 3 \times 2$ design, $\binom{4}{3} \times 2 = 8$ combinations for the $2 \times 3 \times 2$ design, $\binom{4}{2} = 6$ combinations for the $2 \times 2 \times 2$ design, and $\binom{4}{2} \times 2 = 12$ combinations for the $2 \times 2 \times 1$ design. For each combination, we performed subsampling over a sequence of 400 evenly spaced fractions of reads corresponding to $0.25\%, 0.50\%, \ldots, 99.75\%, 100\%$.

For each fraction p, a subsampled count matrix S was generated based on the full experiment matrix, as $S_{i,j} \sim \text{Binom}(X_{i,j}, p)$. This is equivalent to choosing a random sample of the sequenced reads and then mapping them. The same analysis steps used for the full data set were used to analyze each subsample and the same metrics were applied to assess the results as used for the full experiment.

As the results for each experimental design depend on which of the replicates was chosen for subsampling, the results were smoothed for each experimental design using a natural cubic spline with 20 degrees of freedom for estimates of the power, accuracy and FDR. For estimates of the informativeness of each experimental design, we used 15 degrees of freedom as the number of significant gene sets identified in each subsample showed greater variance than the other three metrics.

RESULTS

Experimental Results: We aimed to dissect the sources of variation in pooled mutant screens and determine the appropriate analytical framework and experimental design that maximizes the value of the assay while minimizing cost and resources. All pooled genetic screens using mixtures of mutants require 1) generation of a library of mutants, 2) experimental treatment of the pooled mutants, and 3) identification and quantification of DNA sequences that uniquely identify each mutant using high-throughput DNA sequencing. We designed an experiment to compare growth of haploid yeast non-essential gene deletion mutants in two different carbon sources: glucose (YPD) and galactose (YPGal), using Bar-seq analysis of the molecular barcodes that uniquely identify each mutant. To address the goals of our study we prepared four biological replicates grown for 24 hours in each condition and two technical replicates (i.e. independent sequencing library preparation of the same DNA sample) of each biological replicate (**Figure 1A** and **Table S2**). We also obtained two independent samples from the unselected library (time point 0) from which we prepared technical replicates.

To generate libraries for sequencing with an Illumina HiSeq, we designed a simple two stage PCR protocol (**methods**). Each gene deletion is marked by two different molecular barcodes: one $5'$ (the UPTAG) and one $3'$ (the DNTAG) of the drug resistance cassette. To multiplex sequencing of different Bar-seq libraries we developed a PCR-based method for library preparation that incorporates a unique sequence index for each library (**methods**). We sequenced 40 libraries (20 UPTAG and 20 DNTAG) from 20 samples in a single lane of an Illumina HiSeq 2000. We obtained 185.2 million reads that passed quality filters, and matched them to the molecular barcodes by identifying sequences within a Levenshtein edit distance of 2, which resulted in mapping 93.3% of reads. Using a Levenshtein distance cutoff of 0 (i.e. an exact match) or 1 results in successful mapping of 62.6% and 84.6% of the reads respectively.

For the majority of mutants, the number of reads per barcode across all experiments follows an approximately log-normal distribution, and ranges between 1,000 - 100,000 (**Figure S1**). Low-count outliers that likely result from sequencing errors were removed (**methods**). We found that UPTAGs and DNTAGs for each mutant had similar counts in the majority of samples, with 2574 mutants within a 2-fold difference of each other (**Figure S2**). However, many mutants have highly divergent counts: 1264 have more than a ten-fold difference and 1052 have more than a 100-fold difference. These discrepancies are likely due to one of the barcodes being lost due to sequencing error in either the barcode or PCR priming site.

Correlation analysis of barcode counts shows that the lowest correlations are between mutant abundance in the unselected library (t_0) and mutant abundance following 24 hours of growth in either glucose- or galactose-containing media, indicating that differences in cell growth rates results in substantial changes in the relative abundance of mutants (**Figure 1B**). Growth in YPD yields higher correlation with the t_0 sample than growth in YPGal, indicating that growth in galactose led to a greater shift in the mutants' relative abundances than did growth in glucose. To identify differential effects of mutants during growth in glucose and galactose, we restricted our analysis to the t_{24} samples. We used eigen-R^2 [5] to partition the variance among these samples, and found that 63.5% of the variance is explained by the treatment, 20.3% by biological variation, and 16.1% by technical variation (**methods**). The apportionment of variance is consistent across a wide range of percentile thresholds and using a variety of normalization methods (**Figure S3**).

Computational Analysis of Differential Mutant Abundance: The goal of pooled mutant screens is to identify mutants that exhibit differences in abundance as a result of a defined treatment. The appropriate statistical methods depend on the nature of the data, which in the case of quantitative DNA sequencing of molecular barcodes are discrete count data. As we observed in [9] the data are best described by an overdispersed Poisson distribution (i.e. the variance of biological replicates is greater than the mean) (**Figure S4**). The problem of comparing count data between samples with different read depths while assuming overdispersed Poisson variation is related to that presented by differential expression analysis of RNA-seq data, for which a negative binomial test is used. In addition to the fact that Bar-seq data present some characteristics problematic for t-tests (i.e. lack of normality and a strong mean-variance relationship), there is an important motivation for utilizing models specifically designed for count data. For example, consider two mutants in two different conditions where one's data is simply 1000× the other in read depth (e.g., counts 8 and 9 versus 13 and 14 for mutant A and 8000 and 9000 versus 13000 and 14000 for mutant B). Whereas a t-test results in the same p-value for both mutants, a negative binomial model directly takes the difference in read depth into account resulting in drastically different p-values. Because the difference between the mutants with the lowest total number of reads to the highest number of reads is ~2600-fold in our experiment (**Figure S1**), this is a valid issue to address. Therefore, we used a negative binomial model to test for mutants that are differentially abundant as a result of the treatment.

We utilized the `edgeR` software package [15], which has an efficient implementation of the negative binomial test that accounts for differing read depth and uses shrinkage to help estimate dispersion parameters. We observed that dispersion estimates undergo considerable shrinkage even when four biological replicates are used (**Figure S4**). We found RNA-seq analysis methods that also fit a negative binomial model, such as that implemented in DESeq [2], produce qualitatively comparable results (**Figure S5**). Alternative methods, including DEGSeq [27] and Myrna [12] make overdispersion assumptions less consistent with our data, whereas other methods, including Cuffdiff, use an implementation specific to RNA-seq [25].

Previous studies have used measurements of the UPTAG and DNTAG for each deletion mutant in different ways including selection of the barcode for each mutant with the highest count [21] and independent analysis of each barcode [9]. As the UPTAG and DNTAG are measurements of the same mutant, summing the counts within each sample provides a means of combining the information from both barcodes while remaining robust to cases where one barcode is lost. Furthermore, with count data, summing across technical replicates provides a superior method for minimizing technical variation compared with calculating an average value. Therefore, we summed UPTAGs and DNTAGs for each mutant over technical replicates, such that each condition has four biological replicates, and applied tests using a negative binomial model to identify mutants that are significantly different in abundance in YPGal compared with YPD after 24 hours of growth. The sixteen samples comprising this dataset include a total of 112 million reads.

Analysis of our dataset identified 1992 mutants that are differentially abundant between the two conditions at a 5% FDR. The effect sizes of individual gene deletions are widely distributed (**Table S4** and **Figure 2A**). Notably, the gene deletion mutants for 8 of the 11 genes required for galactose metabolism [23] are significantly decreased in abundance in YPGal and mutants deleted for two genes known to repress galactose metabolism are significantly increased in abundance in YPGal (**Figure 2A**). Gene set enrichment analysis using a Wilcoxon rank-sum test found 192 enriched gene sets at FDR=5%, and the top sets are related to respiration and mitochondrial processes, consistent with the increased importance of respirative metabolism when yeast cells grow in galactose (**Table S5**). Mutants identified as significantly differing in abundance between YPGal and YPD are identified across a range of sequence read depths, although smaller effect sizes tend to be called statistically significant as read depth increases (**Figure 2B**). The ability to detect significant changes in mutant abundance is not greatly affected when total read counts are greater than 1000, and two-fold differences are still detected as statistically significant with total read depths as low as 100. These observations suggest that we oversampled in our study and that similar results would be obtained with approximately an order of magnitude fewer reads.

Effect of Experimental Design on Statistical Results: We aimed to identify the experimental design features that have the greatest effect on the results of a Bar-seq experiment. In practice, the experimental considerations that are most easily controlled are the extent of biological and technical replication and the depth to which each library is sequenced. We computationally simulated variation in each of these experiment design parameters using random subsampling of sequence reads from our complete experiment (**methods**). For the purpose of assessing results from these subsamples we compared them to results obtained from analysis of the complete dataset, which we define as the gold standard. The negative binomial model we fit requires at least two biological replicates. Therefore, to study the effect of biological replication we simulated the use of experimental designs using 3 or 2 biological replicates while retaining two technical replicates for each sample. To study the effect of technical replicates we simulated the use of experimental designs using one technical replicate for each of the biological replicates. For each simulated experimental design we sampled a subset of the reads to simulate varying read depths. We considered four metrics that assess the quality of each simulated experimental dataset: statistical power, accuracy, informativeness, and false discovery rate.

We assessed the power of each experimental design for different sequence read depths by determining the number of mutants identified as differentially abundant at FDR=5%. In all cases, the statistical power of each experimental design increases with read depth; however, it rapidly saturates (**Figure 3A**). Considering our full experimental design, it takes just 1.7 million reads per condition to detect half of the significant mutants that are detected using the complete dataset and 75% are detected with 4.3 million reads per condition. Mutants that are most differentially abundant can be detected at very low read depths: the 13 most significant mutants identified using the complete dataset are all identified as significant even at the lowest depth tested, 140,000 mapped reads (i.e. a 400-fold lower sequence read depth than the total), and are ranked among the 15 most significant mutants in all but the lowest read depth. Table S6 shows the effect size, significance and rank of the 7 most significant galactose-related genes at each level of subsampling, demonstrating that they remain highly significant even at very low read depths.

Reducing the number of biological replicates results in reduced statistical power for a given read depth. Using three biological replicates rather than four decreases the statistical power by about 16%, and using only two biological replicates decreases it by 38%. In practice, this effect is far more relevant than the read depth: 10 million mapped reads using two biological replicates achieves approximately the same power as 2 million total reads across four biological replicates, and the difference cannot be compensated by increasing sequence read depth. Technical replicates only marginally increase the power of the experimental design. This improvement is because pooling multiple replicates decreases the noise added by the library preparation, and therefore decreases the within-treatment variation, analogous to previously-studied strategies of pooling multiple replicates on a single microarray [14, 11].

Although the maximum power possible with each experimental design differs, it is interesting to note that the point at which statistical power begins to asymptote is very similar across experimental design: at around 6 million reads per condition (**Figure 3A**). This suggests that at this point, experimental noise attributable to the sequencing machine itself no longer decreases and additional variation is due to noise introduced by biological variability and library preparation. Statistical power varies within each subset of the designs depending on which replicates were selected (**Figure S6**) indicating that different replicates added different amounts of variance to the experiment, which cannot be predicted *a priori*.

The utility of an experimental design can also be assessed in terms of the accuracy with which effect sizes are estimated, as quantified by the mean square error, the informativeness of the analysis, as quantified by the number of significant gene sets identified by gene set enrichment analysis, or the false discovery rate, as quantified by the proportion of genes found significant that are not significant in the full experiment. Assessment of the quality of each experimental design considering accuracy (**Figure 3B**), informativeness (**Figure 3C**) and false discovery rate (**Figure 3D**) shows that the greatest improvements are found with addition of biological replicates and that improvements beyond 6 million reads per condition are minimal regardless of experimental design. Although there is some variation in the point at which each metric ultimately saturates, the points at which each metric begins to asymptote are highly concordant. Thus, beyond a surprisingly low threshold of 6 million reads per condition, additional sequencing depth provides little additional value.

Although pooled mutant screens enable simultaneous sensitive measurement of each mutants' effect, they are frequently employed as a means of identifying those mutants of greatest effect. We analyzed the statistical power of an experimental design using four biological replicates and no technical replication for different effect sizes **(Figure 4)**. As few as 2.5 million sequence reads per condition (625,000 reads per sample) is sufficient to detect 90% of mutants that change more than two-fold in the full experiment. Increasing the read depth to 6 million reads per condition detects 96% of all mutants that change more than two-fold, 91% of all mutants that change more than 1.5-fold and 72% of all mutants that are significant in the full experiment.

DISCUSSION

High-throughput quantitative DNA sequencing has resulted in rapid advances in a range of problems from the analysis of genome variation to the three dimensional organization of genomes. The coupling of high-throughput quantitative sequencing with large-scale mutagenesis (either systematic or random) enables the pooled analysis of mutant phenotypes with broad applications including the study of gene function, drug targets and genetic interactions. Here, we have studied one realization of pooled mutant analysis - Bar-seq - with the goal of determining experimental design and analytical methods that provide excellent levels of sensitivity, specificity, and efficiency.

We have shown that statistical models employed for RNA-seq analysis are directly applicable to the analysis of Bar-seq data. Tools for RNA-seq analysis, such as those used here, are therefore readily adapted to Bar-seq analysis providing estimates of effect sizes and statistical significance for each mutant. For Bar-seq analysis, UPTAGs and DNTAGs represent additive measurements of the same genotype and therefore should be summed for each sample. Similarly, technical replicates should be combined by addition of barcode counts.

Biological replication is essential for rigorous assessment of statistical significance. At least two biological replicates should always be performed in order to use the within-treatment variation for determining statistical significance. Some software packages have the option of guessing the dispersion in advance, but this is not recommended as an incorrect estimate would make subsequent tests for statistical significance either too conservative or too generous. Moreover, we have found that different experiments can contribute different amounts of variation. Therefore, we recommend performing at least four biological replicates in order to maximize statistical power and accuracy of effect size. The use of technical replicates of biological replicates results in marginal improvements and is likely unnecessary.

Importantly, we found that Bar-seq does not require a high read depth to accurately detect differential abundances of mutants and that additional reads add little to the results. In our study using nearly 60 million mapped reads per condition to analyze 4295 mutants, we demonstrate that the quality of our dataset is maintained with approximately 10-fold fewer reads. Our experimental method for Bar-seq includes 120 uniquely indexed adaptors (**Table S2**), meaning that on a 200 million read sequencing lane one can analyze four biological replicates of 30 different conditions, resulting in approximately 6 million reads per condition. Based on our analysis, that read depth would be expected to identify 96% of genes with a 2-fold change, 91% of mutants with a 1.5-fold change, and 72% of all mutants that would be detected with ten times greater read depth and two technical replicates (**Figure 4**). These findings can be extended to other methods for pooled genetic screens by noting that it corresponds to ∼1400 reads per genomic target per condition. Increasing sequence read depth beyond this value provides only an incremental increase. Thus, our analysis provides guidelines about the tradeoff between per-condition read depth and statistical power that can be used for the design of future experiments.

Figure Legends

Figure 1. **Experimental design and results.** a) Our experimental design entailed two treatments (twenty-four hours of growth in glucose/YPD or galactose/YPGal), four biological replicates and two technical replicates, along with four samples at time point 0 (not shown in panel a). b) Heatmap of the Spearman correlation matrix of mutant counts by sample. Samples cluster according to time point, and also by treatment (YPD vs YPGal) and biological replicate.

Figure 2. **Bar-seq quantifies mutant effects across a range of sequence read depths.** a) Volcano plot showing the relationship between the p-value (log-scale) and log fold change. Genes known to be involved in activation or repression of the galactose utilization pathway are highlighted. The p-value of the rightmost red point is computationally indistinguishable from 0. b) Plot of reads per mutant in the entire experiment compared with the estimated fold change following treatment.

Figure 3. **Simulation analysis of variation in experimental design.** The effect of read depth on a) the number of mutants found significant at FDR=5%, b) the mean squared error between the estimate of the log fold change and the value for the full experiment, c) the number of significant GO terms identified using a Wilcoxon rank sum test at FDR=5%, and d) the percentage of significant genes that were *not* found as significant in the full experiment. Curves are shown for the full experiment, 2 treatments x 4 biological replicates x 2 technical replicates, as well as for subsampled $2 \times 3 \times 2$, $2 \times 2 \times 2$ experimental designs (solid lines). Subsamplings were also performed to simulate each experimental design using a single technical replicate (dashed lines). Each curve was smoothed using a natural cubic spline.

Figure 4. **Statistical power varies with effect size and sequence read depth.** The effect of read depth on the proportion of genes identified as significant (FDR=5%) at different fold-change thresholds using 4 biological replicates x 1 technical replicate for each condition. The fold change for each mutant determined from the full 4 biological replicate x 2 technical replicate experiment is defined as the gold standard. The solid curve shows the proportion of genes found significant relative to the total experiment, while the dotted and dashed curves show the proportion of mutants that had at least a 1.5- or a 2-fold change, respectively. The horizontal dashed line indicates the 90% power level.

Figure S1. **Distribution of the number of reads for all identified mutants.**

Most mutants follow an approximately log-normal distribution in terms of their abundance, with an additional group of mutants that had fewer than 100 counts across all 20 samples, probably due to sequencing error.

Figure S2. **Comparison of UPTAG and DNTAG counts for each mutant.** While the counts were closely correlated for many mutants, a large proportion of mutants had unusually low counts for one barcode, with some missing either an UPTAG or DNTAG entirely, probably due to a mutation in the barcode or the primer.

Figure S3. The percent of variance explained by treatment and biological and technical replication as determined by eigen-R^2. The results are qualitatively identical regardless of the normalization method and the percentile threshold for the minimum number of required reads for inclusion of a mutant.

Figure S4. **Comparison of mean barcode count with the associated variance.** Grey points are the raw measurements for each strain, the red X's are the average variance in each bin, and the blue points are the estimated variance of each strain after dispersion shrinkage has been performed. Variance tends to be substantially greater than the mean suggesting that a overdispersed Poisson or negative binomial model is appropriate.

Figure S5. **P-values for the YPD/YPGal comparison for each strain, calculated using the negative binomial models with `edgeR` and `DESeq`.** The methods show a Spearman correlation of 0.99, indicating only slight differences in their approach.

Figure S6. **The number of significant mutants at different read depths for different subsets of subsampling experiments**. A spline is fit to the results for each of comparison.

Supplementary Table Legends

Table S1. 120 UPTAG and DNTAG indexed primer sequences for multiplexed Barseq analysis of the yeast deletion collection.

Table S2. The UPTAG and DOWNTAG primer and index used for each of the 20 samples analyzed in the current study.

Table S3. The matrix of raw read counts that matched to each tag in each replicate. The first three columns give the systematic and gene name of the deletion and an indication as to whether the mutant was among the 4295 included in the analysis.

Table S4. The p-value and q-value for the test for differential abundance using both DESeq and edgeR for each mutant. Also shown are the estimated $\log_2$ fold changes, the total number of reads matching the gene across both conditions, and the annotation of the biological process indicated in Figure 2.

Table S5. The p-values for gene set enrichment analysis using the Wilcoxon rank-sum test on the estimated $\log_2$ fold changes. The gene sets shown are those in the Biological Process ontology that had at least four genes in the set of analyzed deletions.

Table S6. The estimated log fold change, q-value, and significance rank for the 7 most significant GAL genes at each of the 400 levels of read subsampling.

Acknowledgements

We thank members of the Gresham and Storey labs for helpful discussions. We thank Amy Caudy and David Hess for construction of the prototrophic deletion collection and Zachary Kurtz for design of sample index primers. This work was supported by the National Science Foundation (MCB-1244219 to D.G.), the National Institutes of Health (1R01GM107466 to D.G., R01HG002913 and R21HG006769 to J.D.S.) and a Dupont Young Professor Award to D.G.

References

[1] D. C. Amberg, D. Burke, and J. N. Strathern. *Methods in Yeast Genetics*. A Cold Spring Harbor Laboratory Course Manual. CSHL Press, 2005.

[2] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[3] E. D. Brutinel and J. A. Gralnick. Anomalies of the anaerobic tricarboxylic acid cycle in Shewanella oneidensis revealed by Tn-seq. *Molecular microbiology*, 86(2):273–283, Oct. 2012.

[4] J. E. Carette, C. P. Guimaraes, I. Wuethrich, V. A. Blomen, M. Varadarajan, C. Sun, G. Bell, B. Yuan, M. K. Muellner, S. M. Nijman, H. L. Ploegh, and T. R. Brummelkamp. Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. *Nature Biotechnology*, 29(6):542–546, June 2011.

[5] L. Chen and J. Storey. Eigen-R2 for dissecting variation in high-dimensional studies. *Bioinformatics*, 2008.

[6] L. A. Gallagher, J. Shendure, and C. Manoil. Genome-scale identification of resistance functions in Pseudomonas aeruginosa using Tn-seq. *mBio*, 2(1):e00315–10, 2011.

[7] J. D. Gawronski, S. M. S. Wong, G. Giannoukos, D. V. Ward, and B. J. Akerley. Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. *Proceedings of the National Academy of Sciences of the United States of America*, 106(38):16422–16427, Sept. 2009.

[8] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Véronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. André, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Güldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kötter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C.-y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896):387–391, July 2002.

[9] D. Gresham, V. M. Boer, A. Caudy, N. Ziv, N. J. Brandt, J. D. Storey, and D. Botstein. System-level analysis of genes and functions affecting survival during nutrient starvation in Saccharomyces cerevisiae. *Genetics*, 187(1):299–317, Jan. 2011.

[10] T. X. Han, X.-Y. Xu, M.-J. Zhang, X. Peng, and L.-L. Du. Global fitness profiling of fission yeast deletion strains by barcode sequencing. *Genome biology*, 11(6):R60, 2010.

[11] C. Kendziorski, R. A. Irizarry, K. S. Chen, J. D. Haag, and M. N. Gould. On the utility of pooling biological samples in microarray experiments. *PNAS*, 2005.

[12] B. Langmead, K. D. Hansen, and J. T. Leek. Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome biology*, 11(8):R83, 2010.

[13] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707710, 1966.

[14] X. Peng, C. L. Wood, E. M. Blalock, K. Chen, P. W. Landfield, and A. J. Stromberg. Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics*, 4(1):26, 2003.

[15] M. Robinson and D. McCarthy. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, 2010.

[16] M. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3):R25, 2010.

[17] M. R. Schlabach, J. Luo, N. L. Solimini, G. Hu, Q. Xu, M. Z. Li, Z. Zhao, A. Smogorzewska, M. E. Sowa, X. L. Ang, T. F. Westbrook, A. C. Liang, K. Chang, J. A. Hackett, J. W. Harper, G. J. Hannon, and S. J. Elledge. Cancer proliferation gene discovery through functional genomics. *Science (New York, NY)*, 319(5863):620–624, Feb. 2008.

[18] J. M. Silva, K. Marran, J. S. Parker, J. Silva, M. Golding, M. R. Schlabach, S. J. Elledge, G. J. Hannon, and K. Chang. Profiling Essential Genes in Human Mammary Cells by Multiplex RNAi Screening. *Science (New York, NY)*, 319(5863):617–620, Feb. 2008.

[19] D. Sims, A. M. Mendes-Pereira, J. Frankum, D. Burgess, M.-A. Cerone, C. Lombardelli, C. Mitsopoulos, J. Hakas, N. Murugaesu, C. M. Isacke, K. Fenwick, I. Assiotis, I. Kozarewa, M. Zvelebil, A. Ashworth, and C. J. Lord. High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome biology*, 12(10):R104, 2011.

[20] A. M. Smith, L. E. Heisler, J. Mellor, F. Kaper, M. J. Thompson, M. Chee, F. P. Roth, G. Giaever, and C. Nislow. Quantitative phenotyping via deep barcode sequencing. *Genome research*, 19(10):1836–1842, Oct. 2009.

[21] A. M. Smith, L. E. Heisler, R. P. St Onge, E. Farias-Hesson, I. M. Wallace, J. Bodeau, A. N. Harris, K. M. Perry, G. Giaever, N. Pourmand, and C. Nislow. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Research*, 38(13):e142, July 2010.

[22] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci*, 100(16):9440–5, Aug 2003.

[23] D. J. Timson. Galactose metabolism in Saccharomyces cerevisiae. *Dynamic Biochemistry*, 2007.

[24] A. H. Tong, M. Evangelista, A. B. Parsons, H. Xu, G. D. Bader, N. Pagé, M. Robinson, S. Raghibizadeh, C. W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294(5550):2364–2368, Dec. 2001.

[25] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1):46–53, Jan. 2013.

[26] T. van Opijnen, K. L. Bodi, and A. Camilli. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nature methods*, 6(10):767–772, Oct. 2009.

[27] L. Wang, Z. Feng, X. Wang, and X. Wang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 2010.

[28] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Véronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, Aug. 1999.