Quantitative phenotyping via deep barcode sequencing

Andrew M. Smith,[1,2,3] Lawrence E. Heisler,[3,4] Joseph Mellor,[5,6] Fiona Kaper,[7] Michael J. Thompson,[7] Mark Chee,[7] Frederick P. Roth,[5,6] Guri Giaever,[1,3,4,8] and Corey Nislow[1,2,3,8]

[1]*Department of Molecular Genetics, University of Toronto, Toronto, Ontario M5S 1A8, Canada;* [2]*Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5G 1L6, Canada;* [3]*Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario M5S 3E1, Canada;* [4]*Department of Pharmaceutical Sciences, University of Toronto, Toronto, Ontario M5S 3M2, Canada;* [5]*Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts 02115, USA;* [6]*Center for Cancer Systems Biology, Dana Farber Cancer Institute, Boston, Massachusetts 02115, USA;* [7]*Prognosys Biosciences, Inc., La Jolla, California 92037, USA*

Next-generation DNA sequencing technologies have revolutionized diverse genomics applications, including de novo genome sequencing, SNP detection, chromatin immunoprecipitation, and transcriptome analysis. Here we apply deep sequencing to genome-scale fitness profiling to evaluate yeast strain collections in parallel. This method, Barcode analysis by Sequencing, or "Bar-seq," outperforms the current benchmark barcode microarray assay in terms of both dynamic range and throughput. When applied to a complex chemogenomic assay, Bar-seq quantitatively identifies drug targets, with performance superior to the benchmark microarray assay. We also show that Bar-seq is well-suited for a multiplex format. We completely re-sequenced and re-annotated the yeast deletion collection using deep sequencing, found that ~20% of the barcodes and common priming sequences varied from expectation, and used this revised list of barcode sequences to improve data quality. Together, this new assay and analysis routine provide a deep-sequencing-based toolkit for identifying gene–environment interactions on a genome-wide scale.

[Supplemental material is available online at http://www.genome.org. All data and analysis tools are available at http://chemogenomics.med.utoronto.ca/supplemental/barseq/.]

Genomics has benefited from continued innovations and advances in automation and information management. New technologies will continue to increase the rate of discovery; however, the future requires tools to analyze the vast amount of data collected in highly multiplexed assays that are capable of interrogating biological systems en masse. To date, high-density barcode microarray platforms have been used for the comprehensive analysis of transcription factor binding sites (ChIP-chip), gene expression, nucleosome occupancy, and fitness profiling, to name a few examples. More recently, next-generation sequencing (NGS) technologies have been applied to tackle these same applications with promising results, including RNA-seq (Nagalakshmi et al. 2008), ChIP-seq (Robertson et al. 2007), genome analysis (Bentley et al. 2008), nucleosome occupancy (Ozsolak et al. 2007), and many other applications (e.g., de novo sequencing, SNP detection). For a more detailed review of next-generation sequence applications, we refer you to MacLean et al. (2009).

We previously established a genome-wide chemogenomic assay (Giaever et al. 2004; Hillenmeyer et al. 2008; Hoon et al. 2008) that uses barcoded yeast deletion strains in a competitive growth assay (combined with a barcode microarray readout) to identify the genes important for growth in the presence of compound, e.g., haploinsufficiency profiling (HIP) or homozygous profiling (HOP). Although high-density barcode microarrays are well-suited for such assays, the assay platform requires re-tooling to investigate other organisms or strain collections. For example, one may have to either design a new barcode microarray for each organism or cell type, or re-engineer strains such that they carry specific barcodes. Either case will carry significant up-front costs. Furthermore, a new array design may require a priori sequence information, whereas an NGS approach does not.

We adapted a validated barcode microarray-based chemogenomic assay and directly compared the barcode microarray data to that of high-throughput sequencing. This protocol (Barcode Analysis by Sequencing, or "Bar-seq") directly "counts" each barcode in a complex sample via sequencing. For this comparison, we used the well-characterized yeast deletion strain library and assessed its ability to identify the known targets for several well-characterized drugs. Together, Bar-seq, combined with a reannotation of the yeast deletion collection and development of methods to analyze the data, promise to make Bar-seq a powerful tool for understanding gene function.

Results

Bar-seq outperformed barcode microarray hybridization, based on several performance metrics; including (1) sensitivity, (2) dynamic range, and (3) limits of detection (based on the number of sequencing reads we could reliably detect vs. the hybridization level we could reliably detect). Bar-seq was also able to assess and "rescue" those barcodes having sequence errors that made them undetectable by barcode microarray hybridization. Accordingly, we

[8]**Corresponding authors.**
E-mail corey.nislow@gmail.com; fax (416) 978-4842.
E-mail ggiaever@gmail.com; fax (416) 978-4842.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.093955.109.

19:1836–1842 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

characterized all the barcodes and common primer sites in the yeast knockout collection by sequencing and were able to reassign 2000 barcodes, compiling a high-confidence list in the process for future analysis of Bar-seq screens.

The Bar-seq assay differs from the barcode microarray-based assay at the analytical readout step. For our analysis, several different pools of yeast mutants were grown competitively in diverse conditions, and following growth, genomic DNA was extracted, molecular barcodes were amplified by PCR, and barcode amplicons were either labeled and hybridized to a barcode microarray as described (Pierce et al. 2007) or sequenced using an Illumina Genome Analyzer (Bennett 2004). For the barcode microarray samples, barcode abundance was inferred based on the normalized fluorescence intensity (Pierce et al. 2006, 2007) following detection with an Affymetrix confocal laser scanner. For Bar-seq, barcode abundance was determined by counting the number of times each unique barcode was sequenced (see Methods). A significant difference between Bar-seq and hybridization is that the entire barcode is not necessarily required for unambiguous determination of each barcode sequence. Theoretically, not all 20 bases of sequence are necessary to discriminate between the yeast barcodes; in practice, most barcodes can be uniquely identified with as few as eight to nine sequenced bases (Supplemental Fig. 1), but to avoid losing any barcodes, all 20 bases were sequenced. For a read of length 20 bases (the entire length of each barcode), the calculated sequencing error rate is currently <5%, which represents the sum of errors for the first, second, to the 20th base. We expect that future improvements in chemistry and software will lower this rate. Therefore, sequence error rates will have minimal impact on the results of Bar-seq screens.

Sensitivity of high-throughput sequencing vs. barcode microarray hybridization

We constructed and compared two pools of yeast strains, one containing 953 strains included at approximately equal representation ("Pool-constant") and a second ("Pool-variable") containing the same strains as Pool-constant but with each strain represented at one of four different levels of representation (0.25×, 0.5×, 1.0×, and 2.0×) relative to Pool-constant (see Supplemental Methods for a description of each pool). By comparing Pool-constant with Pool-variable, we could quantify and compare the abundance of each strain/barcode either by microarray hybridization or sequencing. Microarray signals were transformed to account for the observation that they saturate at high probe/signal levels (Pierce et al. 2007). We found that both platforms clearly distinguished all four levels of strain abundance between the two pools (Fig. 1). Others have reported that NGS can increase the dynamic range of "counting assays" such as Bar-seq. Indeed, the separation of strains present in the $\log_2$ ratio is significantly improved using Bar-seq as compared with barcode microarrays. In Figure 1B, we used an ANOVA test to show that the four box-plots representing the four different subpools are significantly different. Furthermore, we found that the separation of these subpools was significantly improved for sequencing vs. microarrays, although the results were not perfectly linear. Comparing Bar-seq to microarray, we found that the correlation between replicates was extremely high (r = 0.999) for Bar-seq and (r = 0.993) for barcode microarrays.

Application of Bar-seq to drug target identification

To assess Bar-seq in a practical assay context, we performed a chemogenomic assay to identify the targets of several well-characterized drugs. Pool-constant was challenged with two drugs, cerivastatin and tunicamycin, for 20 generations of growth, and barcodes were quantified by barcode microarray or Bar-seq. The $\log_2$ ratios of each treated pool (relative to DMSO controls) are shown in Figure 2. The known targets of cerivastatin and tunicamycin (Hmg1 and Alg7, respectively) are clearly identified by both platforms, supporting the practical utility of both approaches. We next screened doxorubicin (an anticancer antibiotic) vs. a pool of 1100 essential heterozygous deletion strains (see Supplemental Methods; Hoon et al. 2008). Our previous results suggest that Ssl2,

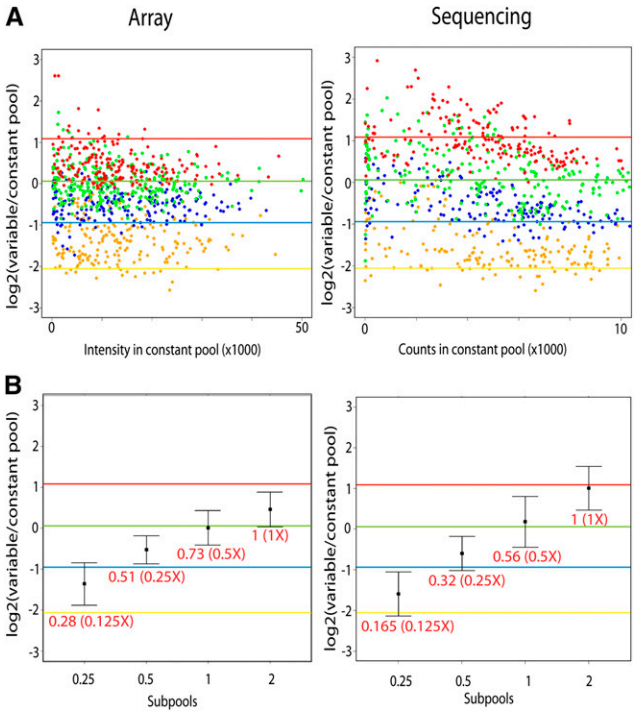

Figure 1. Comparison of barcode microarray hybridization and Bar-seq data on identical samples. A pool of 953 strains was created that contains four subpools of approximately 250 yeast deletion strains each. The strains in this pool were selected to contain two well-characterized drug targets and an additional 951 control heterozygote strains. These were mixed together in a constant pool (Pool-constant) at a ratio of 1:1:1:1 and in a variable pool (Pool-variable) at a ratio of 0.25:0.5:1.0:2.0. Log_2 signals for each strain were determined, and the relative abundance across subpools was assessed. For tag-array analysis, the signal refers to the raw intensities corrected for saturation effects as described previously (Pierce et al. 2007), whereas for sequencing analysis, the signal refers to the sequencing counts. Data were filtered to remove strains with signal below an arbitrary background level (signal of 40 for sequencing data, 200 for hybridization data). (*A*) Scatterplot of the $\log_2$ ratio of the signal for each strain in the variable pool (0.25:0.5:1.0:2.0) over the signal in the constant pool (1:1:1:1). The subpools are shown in different colors: red, green, blue, and yellow correspond to ratios within Pool-variable of 0.25:0.5:1.0:2.0, respectively. The red, green, blue, and yellow lines indicate the expected $\log_2$ ratios. The data for this panel were scale-normalized using the green group, which is at equal concentration in both pools. (*B*) The distribution of the $\log_2$ ratios between variable (0.25:0.5:1.0:2.0) and constant (1:1:1:1) pools is shown for each subpool. The mean of each distribution is shown, with error bars representing one standard deviation. The *y*-axis is the $\log_2$ intensity or counts for each subpool present in the variable pool over the constant pool. The red numbers are the ratio of each subpool's mean over the mean of the 2 subgroup; in brackets is the expected ratio. All subgroups are statistically different in both the barcode microarray and Bar-seq data sets with *P*-values $<10^{-6}$.

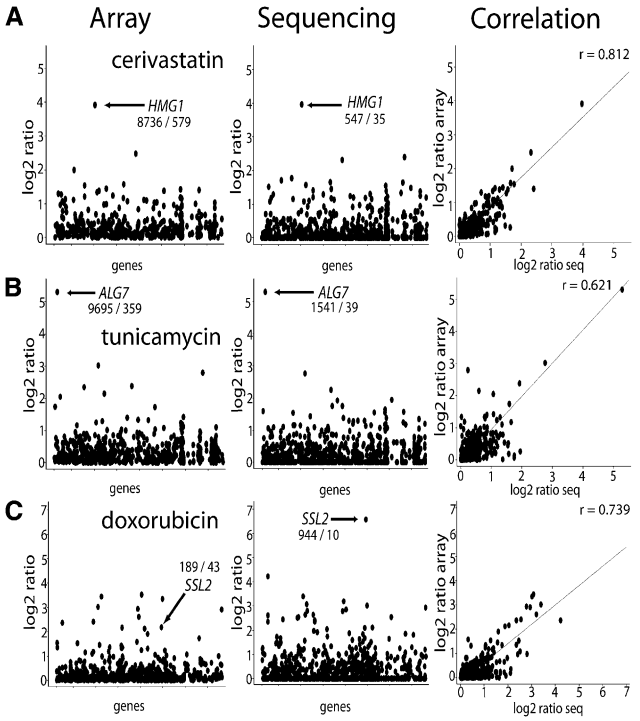

Figure 2. Results of the yeast deletion pools assayed by array and Bar-seq. Log_2 results for both TAG4 barcode microarray hybridization and Illumina sequencing are presented. All axes represent log_2 ratios of control over treatment vs. genes (alphabetically ordered). (*A*,*B*) Results for the downtags for the drug treatments of the constant pool for (*A*) cerivastatin and (*B*) tunicamycin. (*C*) Results for the heterozygote essential pool treated with doxorubicin. The *r*-value in the *righthand* column indicates the correlation of the log_2 ratios between the array vs. sequencing data. (Arrow) Known drug targets are labeled. The sequencing data were collected using a single sequencing reaction for four independent samples (four-plex). The correlation data were filtered based on greater than 10 counts in the Bar-seq DMSO control and an intensity of more than 200 in the DMSO array control, prior to correlation calculation. These data were collected in four-plex sequencing reactions. For details, see Methods.

a component of the RNA polymerase transcription machinery, is a likely target of doxorubicin. Both barcode microarray and Bar-seq platforms identify Ssl2 (S Hoon, RP St Onge, G Giaever, and C Nislow, unpubl.) as a doxorubicin-sensitive strain, yet barcode microarray analysis scored Ssl2 as one of the 10 most-sensitive strains, while Bar-seq identified it as the most-sensitive deletion strain (Fig. 2). Bar-seq does not always outperform the microarray; for example, when we analyzed the tunicamycin assay, we found that arrays outperformed Bar-seq for the down barcode, but not for the up barcode (Supplemental Fig. 2). We note that the *ALG7* up barcode does have a mutation in its U1 primer, which could adversely effect the Bar-seq PCR.

Multiplexing Bar-seq samples

The minimal unit of output for the Illumina Genome Analyzer is a single lane of a single flow cell, delivering, at the time of this report, 5–10 million reads. Because this number of reads is in vast excess of the number of reads required to determine changes in strain abundance, we designed multiplex experiments to be combined within a single lane of one flow cell. Multiplexing has been reported for other applications, for example, for ChIP-seq (Craig et al. 2008; Lefrancois et al. 2009); however, this is the first example of a multiplexed barcode-sequencing application. In fact, all the drug treatments described in Figure 2 were performed on multiplexed samples. To explore the theoretical limits of multiplexing, we analyzed two replicate single-plex Bar-seq runs, then randomly undersampled one replicate and calculated the correlation between the undersampled and the constant replicates. The correlation between these simulations exceeded $r = 0.95$ when the number of counts was greater than 50,000 ($k > 50{,}000$). This test suggests that, for a pool of 1000 yeast strains, satisfactory data are achieved with 50,000 total counts (approximately 50 counts per strain), and extrapolation of these data suggests that Bar-seq can be multiplexed by a factor of 200 per sequencing lane without significant loss in data quality. Supplemental Figure 3A also shows that once 50,000 counts are achieved, >98% of the barcodes within the pool are sequenced. We performed a second simulation on the data from Figure 2, A and B, by undersampling the number of Bar-seq counts and asked at what level of undersampling identification of the known drug target was affected. We found that, in agreement with Supplemental Figure 3A, ~50,000 counts/experiment or 50 counts/strain are needed for clear target identification (Supplemental Fig. 3B,C). Additional constraints, for example, errors introduced by liquid handling and other preparative steps, will, of course, limit the upper level of multiplexing. Nonetheless, the sampling error for Bar-seq (assuming 10 million reads/lane) is 0.03%, well below the bottleneck sampling error introduced by cell-harvesting and liquid-handling steps of the assay, which can be as high as ~5%–10% (Pierce et al. 2007).

Re-characterization of yeast deletion collection

To complement the Bar-seq assay and to support its routine use, we assayed the commercially available version of the heterozygote yeast deletion collection pool (the "Invitrogen 6000" pool) (Giaever et al. 2002) by Bar-seq (see Supplemental Methods). To validate this pool, we screened the drugs alverine citrate and clotrimazole and found that Bar-seq identified the putative target of alverine citrate, Erg24, and the well-characterized target of clotrimazole, Erg11 (Supplemental Fig. 4; Hillenmeyer et al. 2008), respectively.

While the deletion strains have been previously assessed by Sanger and pyrosequencing, we speculated that the increased sampling afforded by deep sequencing would improve the quality of the data. We therefore re-sequenced each 20-mer barcode associated with each gene deletion, sequenced every common amplification primer, and also prepared libraries of genomic DNA fragments to unambiguously associate each particular barcode with its genomic location using paired-end Illumina sequencing (Fig. 3). To generate the fragments that were sequenced, we first isolated genomic DNA from the Invitrogen 6000 pool. We selected this pool because it is commercially available, and, during the yeast knockout project, all strains were initially constructed as heterozygous diploids; therefore, this collection should contain the most complete representation of all barcodes and common primers. For the paired-end sequencing reactions, we fragmented the DNA and ligated adaptors on the ends. Using PCR, we enriched for genomic DNA that possesses the KanMX junction. We used one primer that was homologous to the KanMX cassette and one adaptor primer. This allowed us to selectively amplify the paired-end fragments seen in Figure 3. For the single end sequence reactions, we used U1 or D1 common priming sites and a primer immediately downstream from U2 or immediately upstream of D2, respectively.

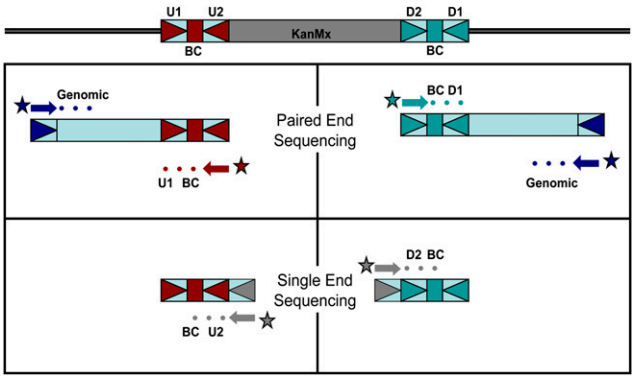

Figure 3. Schematic showing sequencing strategy for re-characterization of barcode and common priming sequences. (U1, U2/D1, D2) Common priming sites for uptag/downtag barcodes. (BC) Barcode. (*Top* panels) We used a paired-end sequencing reaction to identify both genomic position (from one read) and the barcodes and U1/D1 sequences (from the second read). (*Bottom* panels) In an additional sequencing reaction, we identified the barcodes and U2/D2 sequences in a single Illumina sequencing read by using a primer with homology with the KanMX4 cassette and flanking the U2/D2 sequences (shown in gray). (Colored circles) The bases that are being sequenced; (colored arrows) the primers used in the sequencing reaction; (square) the uptag barcode; (light-blue square) the downtag barcode. (Triangles flanking the colored boxes) The common primers; (dark blue triangle) the ligated adaptor sequence used to sequence the genomic DNA flanking the cassette.

Our analysis confirmed the identity and genomic location of the majority of the yeast barcodes. Surprisingly, we found that 2042 barcodes deviated from previously reported sequences (Supplemental Fig. 5; Eason et al. 2004). This is likely due to the increased sampling using NGS vs. Sanger sequencing. We found that ~90% of our data agree with the Sanger sequencing of the barcoded yeast deletion collection (Supplemental Fig. 5). Altogether, ~82% of all barcode tags are correct (i.e., an exact match to expectation), and ~18% had mutations (Fig. 4; Supplemental Table 1). Of these mutations, ~28% were single substitutions, ~40% single deletions, ~1.5% single insertions, and ~31% were "other," that is, multiple basepair substitution/deletions (Fig. 4; Supplemental Table 1). We found that ~18% of all universal priming sequences had errors, which likely explains those strains that are not detected in Figure 4 and Supplemental Table 2. In addition, we found that 99.5% of the yeast barcodes mapped to their correct loci in the genome. As shown in Supplemental Figure 6, the small number of strains that are missing in Bar-seq are statistically enriched for strains with errors in their common priming sites. Our precise characterization of barcode and primer sequences at this high level of coverage will improve Bar-seq data analysis and may also help interpretation of array data.

Discussion

Bar-seq has the potential to perform 200-plex experiments of 1000 strain pools or 33-plex assays of 6000 strain pools in a single Illumina lane at current sequencing densities. Although the data presented here used Illumina sequencing, Bar-seq should be readily transferable to other NGS technologies (e.g., ABI SOLiD, Polonator, or 454 Life Sciences [Roche] technologies). This level of multiplexing will increase directly as the number of reads per lane increases and greater multiplexing reduces the cost per experiment. With the current average cost per Illumina run being approximately $1500 USD, a 20-plex experiment defines the break-even point for Bar-seq vs. microarrays, assuming microarray costs of $150–$300 USD.

Here we defined the parameters for practicing Bar-seq, and used NGS to re-annotate the commercial yeast deletion collection. We characterized 2042 barcodes such that they can now be definitively included in genome-wide screens. We did, however, still fail to identify some of the "common" priming sites. For example, as shown in Figure 4, we detect fewer of the total U2 and D2 common primer sequences compared to U1 and D1, respectively. This is likely due to the difference in read length (i.e., 12 bases rather than 18 bases) obtained for these barcodes. To verify our re-annotation, we selected 35 uptags for which we had both paired-end sequence and single end read data that disagreed with Eason et al. (2004). We found that 96% (29/30) of the sequences we called correct were confirmed by Sanger sequence as correct, and 80% (4/5) were verified as having a mutation in the barcode (Supplemental Table 3). This indicates that our re-characterization of the yeast deletion collection is a high-quality annotation of the yeast barcodes and common primer sites.

As an example of the Bar-seq assay in practice, Figure 2C illustrates the performance advantages of sequencing vs. barcode microarray-based methods. Specifically, by barcode microarray, the tag intensity of the *SSL2* deletion strain in the control treatment is quite low (189 fluorescence units) and declines only fourfold (to 43) after treatment. In contrast, when the same samples were subjected to NGS, this tag was counted 940 times in the control sample and diminished to 10 counts following drug treatment, a significant reduction. One benefit of having re-sequenced the

	U1	Up Tag	U2	KanMX4 Cassette	D2	Down Tag	D1
Total	6004	6004	6004		5811	5811	5811
# Not Found	684	289	1124		1169	197	586
# Found	5320	5715	4880		4642	5614	5552
% Correct	80.15%	82.2%	82.9%		84.08%	81.7%	80.96%
% Mutated	19.85%	17.8%	17.1%		15.92%	18.3%	19.04%
% Substitutions	34.8%	29%	10%		59.2%	27%	32.1%
% Deletions	39.8%	43%	66.9%		18.3%	37%	37.4%
% Insertions	0.2%	1%	0.6%		0.3%	2%	0.6%
% Other	25.2%	28%	22.5%		22.1%	34%	29.9%

Figure 4. Yeast knockout collection characterization. (*Top*) An illustration of the yeast deletion cassette; (*bottom*) the table represents the total number of barcodes found, the percent correct (i.e., sequences found to exactly match the designed sequence), and the percent incorrect (i.e., sequences found to deviate from the expected sequence). Also shown is a breakdown of the incorrect sequences that were identified. This breakdown includes the percentage of single substitutions, single deletions, single insertions, and other mutations (i.e., multiple deletions). These data were collected in two paired-end sequencing reactions and two single sequencing reactions. For details, see Supplemental Methods.

yeast deletion collection is that we found a previously undetected mutation in the *SSL2* deletion cassettes uptag (designed barcode, AGATTGACTACACGCTCTTC; actual barcode, AGATTGACTACA CCTCTTC), which likely explains the low intensity in the barcode microarray data. Both microarray and Bar-seq can successfully identify drug targets, and while Bar-seq does have several potential advantages (e.g., greater sensitivity and dynamic range), the barcode microarray assay is the assay benchmark against which improvements of Bar-seq are and will be judged. We compared Bar-seq counts vs. barcode microarray intensities, for all barcodes present in Pool-constant (Supplemental Fig. 7). We see that the barcodes that have a correct common priming site score well in both assays, while barcodes that contain errors score much lower in both assays. Even if all barcodes were perfect, there are inherent limitations in the microarray readout, including saturation effects, limited dynamic range, and issues of cross-hybridization, which do not similarly effect sequencing. Furthermore, barcodes for many loci in parallel may be constructed using degenerate oligonucleotide synthesis, since there are no hybridization-related sequence constraints that require an independent oligonucleotide synthesis for each barcode. As mentioned above, it is important to underscore that the barcodes in this study were specifically designed for microarray hybridization and, as a result, are much more similar to one another than is desirable for a sequencing application. If we repeated the same comparison with a set of barcodes optimized for sequencing (e.g., for maximal diversity), the improvement in performance of sequencing over microarrays would be expected to be even greater, and the use of even shorter barcodes might be equally effective.

We have shown that Bar-seq is a highly robust assay, with the potential of multiplexing to high levels. This assay is not necessarily limited to gene–environment interactions and can be extended to other contexts. For example, with the creation of the yeast "barcoder" strains (Yan et al. 2008), barcoding any yeast strain is possible, which can then be pooled and analyzed using Bar-seq, including the analysis of DAmP (Yan et al. 2008) or temperature-sensitive allele collections (Ben-Aroya et al. 2008), barcoded open-reading frame (ORF) assays (Ho et al. 2009), or for the analysis of double-mutant pools created by SGA (Tong et al. 2001). Bar-seq is also not limited to yeast barcodes and could be applied to diverse other organisms such as the barcoded *Escherichia coli* collection (Kitagawa et al. 2005) or RNAi collections.

In summary, Bar-seq represents a novel deep-sequencing-based assay for quantitatively characterizing complex pools. The data gathered from either platform are highly reproducible; however, Bar-seq is more sensitive at discriminating between the sample groups tested (Fig. 1). We demonstrated its successful multiplexed application to quantitative genome-scale fitness profiling of yeast deletion pools to characterize gene–environment interactions and drug mechanisms of action. In addition, we used NGS to re-sequence the barcodes and common priming sites and showed that it improved the results obtained from Sanger sequencing and the data collected from pooled assays.

Methods

Construction of contrived pools with fixed numbers of barcoded strains

A pool of 953 different heterozygous mutants was selected to contain two well-known drug targets. Pool-constant was constructed by growing each strain in 100 μL of YPD to saturation in 96-well plates, then pooling 20 μL from each well. Pool-variable consisted of the same 953 strains, but the number of cells of each strain was varied systematically with approximately one-quarter of the 953 strains added at a ratio of 0.25:0.5:1.0:2.0 when compared to the same strain abundance in Pool-constant.

Pooled growth assays

Pool-constant was thawed and diluted in YPD containing 2% DMSO or drug to a final OD_{600} of 0.062. Drug was applied at a dose that produced a 10%–20% wild-type growth inhibition. Using an automated pipetting liquid-handler robot that pipettes every five generations, 600 μL of the pools was harvested robotically at an OD_{600} of 0.76 after 20 generations of growth. In some experiments, we used a pool of 1100 essential heterozygous deletion mutants that contained a deletion strain of a putative target of doxorubicin, Ssl2 (S Hoon, RP St Onge, G Giaever, and C Nislow, unpubl.).

Assessing fitness of barcoded yeast strains by barcode microarray

Except where indicated, pooled assays were performed as described by Pierce et al. (2007). Genomic DNA was isolated from cells grown for 20 generations, and barcodes were amplified and hybridized to barcode microarrays, where each barcode deletion mutant is represented by 10 hybridization signals (the uptag and downtag for each strain are represented on the array five times). Array measurements were quantile-normalized such that all tags hybridized with the sample pool had similar distributions. Following normalization, a correction factor was applied to correct for feature saturation (Pierce et al. 2007), and the fitness of each barcoded deletion strain was then determined. Positive fitness defect scores signify a decrease in strain abundance after drug treatment.

Assessing fitness of barcoded yeast strains by Illumina sequencing

DNA was isolated from the deletion pools at time 0 and after 20 generations of growth as described (Pierce et al. 2007). Each 20-mer barcode was amplified with primers that were comprised of the common barcode primers and the sequences required for cluster formation on the Illumina flow cell (underlined). For the Uptags: 5′-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGATGTC CACGAGGTCTCT-3′ (F) and 5′-AATGATACGGCGACCACCGACA CTCTTTCCCTACACGACGCTCTTCCGATCTGTCGACCTGCAG CGTACG-3′ (R) were used. For the Dntags: 5′-CAAGCAGAAGA CGGCATACGAGCTCTTCCGATCTGAAAACGAGCTCGAATTCA TCG-3′ (F) and 5′-AATGATACGGCGACCACCGACACTCTTTCC CTACACGACGCTCTTCCGATCTCGGTGTCGGTCTCGTAG-3′(R) were used. PCR reactions had a final volume of 100 μL, which contained PCR buffer [50 mM Tris/HCl, 10 mM KCl, 5 mM $(NH_4)_2SO_4$, 2 mM $MgCl_2$, pH 8.3], 200 μM each dNTP, 4 U of FastStart Taq DNA polymerase (Roche Applied Science), 100 ng of genomic DNA, and 400 nM each primer (IDT). PCR amplification was conducted in a GeneAmp PCR System 9700 thermocycler (Applied Biosystems) with the following conditions: 4 min at 95°C; 25 cycles of 30 sec at 95°C, 30 sec at 55°C, and 30 sec at 72°C; followed by 7 min at 72°C. The 150-bp PCR products were gel-purified using 20% TBE gels (Invitrogen), and the crush and soak method was followed by ethanol precipitation; samples were used directly for cluster formation on the Illumina flow cell. For multiplexed Illumina sequencing, 5-mer tag sequences were incorporated into each primer between the Illumina and barcode primer sequence. This multiplexing tag allowed postsequencing assignment of each amplicon to a particular experiment. For

multiplexing, the location of each barcode was recorded, and each cluster on the slide was hybridized twice, first to a mixture of custom sequencing primers corresponding to the multiplexing tags in use (5′-CGCTCTTCCGATCTNNNNNGTCGACCTGCAGC GTACG-3′ for uptags or 5′-CGCTCTTCCGATCTNNNNNCGGTG TCGGTCTCGTAG-3′ for downtags, where NNNNN indicates the 5-mer multiplexing tag) to sequence the barcode and then to the common sequencing primer (5′-CACTCTTTCCCTACACGACGCT CTTCCGATCT-3′) to sequence the multiplexing tag. This two-step sequencing reaction was performed to allow us to sequence fewer bases and to obtain the full barcode sequence for barcode identification. To analyze the Bar-seq data, all counts were quantile-normalized between experiments such that each experiment had the same count distribution. We added 10 pseudo-counts to all sequence tag tallies to prevent division by zero during data analysis. By analogy with barcode microarray fitness experiments, fitness defect ratios were calculated and expressed as the $\log_2$ ratio of control counts over counts after drug treatment for each strain.

Characterizing the commercially available Invitrogen 6000 deletion pool

Yeast genomic DNA from the Invitrogen 6000 deletion pool (CAT #95401.H4POOL) was isolated and fragmented using Adaptive Focused Acoustics technology with an S-series instrument (Covaris) to a size of 150–700 bp. The standard Illumina adaptors (5′-P-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG-3′ and 5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′) were ligated to the ends of the fragments following the Illumina genomic DNA library preparation protocol. Using a directed PCR with a primer targeted to the adaptor sequence (5′-CAAGCAG AAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCG CTCTTCCGATCT-3′) and a primer homologous to the KanMX cassette (5′-AAGCTAAACAGATCTGGCGCGCC-3′ for the uptag or 5′-TAACGCCGCCATCCAGTGTCG-3′ for the downtag), genomic DNA fragments containing the KanMX4 junction were amplified. PCR reactions had a final volume of 50 μL, which contained Phusion HF buffer, 250 μM each dNTP, 1 U of Phusion Hot Start DNA polymerase (Finnzymes), and 250 nM each primer (IDT). PCR amplification was conducted in a GeneAmp PCR System 9700 thermocycler (Applied Biosystems) with the following conditions: 30 sec at 98°C; 20 cycles of 10 sec at 98°C, 30 sec at 65°C, and 30 sec at 72°C; followed by 5 min at 72°C. PCR fragments of 500 bp were gel-purified as described and served as template in a subsequent nested PCR amplification with the adaptor primer used in the first PCR and primers that were comprised of the common yeast barcode primers and the sequences required for paired-end cluster formation on the Illumina flow cell (5′-AATGATACGGCGACCACCGAGATCTAC ACTCTTTCCCTACACGACGCTCTTCCGATCTGTCGACCTGCA GCGTACG-3′ for uptags and 5′-AATGATACGGCGACCACCGAG ATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGAAAAC GAGCTCGAATTCATCGAT-3′ for downtags). PCR conditions were the same as the first PCR reaction. Paired-end sequencing with 5′-CTCTTCCGATCTGTCGACCTGCAGCGTACG-3′ (uptags) or 5′-CTCTTCCGATCTGAAAACGAGCTCGAATTCATCGAT-3′ (downtags) in the first read captured the complete barcode sequence and the U1/D1 common priming sites. The second read with the standard sequencing primer (5′-CGGTCTCGGCATTCCTGCTGAA CCGCTCTTCCGATCT-3′) sequenced short sequences within the genomic DNA ~300 bp upstream (uptags) or downstream (downtags) from the barcoded KanMX4 cassette.

To sequence the U2/D2 common primer sites, a second single read-sequencing run was performed. The molecular barcodes were amplified with the universal U1/D1 primers extended with the Illumina cluster formation sequences (5′-CAAGCAGAAGACGG CATACGAGCTCTTCCGATCTGATGTCCACGAGGTCTCT-3′ for uptags or 5′-CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT CGGTGTCGGTCTCGTAG-3′ for downtags) and a primer complementary to the region of the KanMX4 cassette flanking the U2/D2 common priming site extended with the Illumina cluster formation sequences (5′-AATGATACGGCGACCACCGACACTCTTTCC CTACACGACGCTCTTCCGATCTAAGCTAAACAGATCTGGCGC GCC-3′ for uptags or 5′- AATGATACGGCGACCACCGACACTCTT TCCCTACACGACGCTCTTCCGATCTTAACGCCGCCATCCAGT GTC-3′ for downtags). PCR conditions were as described. Amplified DNA was sequenced on the Illumina platform from the KanMX4 site across the U2/D2 site and partially into the barcodes assigned to each strain with 5′-TCTGGCGCGCCTTAATTA ACCCGGGGATCC-3′ for uptags and 5′-CTTCCGATCTTAACGC CGCCATCCAGTGTC-3′ for downtags. The resulting sequences were identified by alignment to a database of anticipated sequences consisting of the barcodes, the common priming sites, and flanking genomic regions using maq software (Li et al. 2008) and BLAST. The genomic portion of the paired-end sequence was used to verify positioning within the correct locus for the associated barcode portion. Expected common priming sites and barcodes were aligned to reads to characterize any sequence alterations.

Statistical analysis

Differences in the distributions represented by the box-plots in Figure 1 were determined by analysis of variance between the four groups. All distributions were distinct as determined by Tukey's HSD post-hoc analysis that indicated a P-value of $<10^{-6}$ between any two groups.

We tested the yeast deletion strains that were not detected in either or both platforms based on Bar-seq or barcode microarray data, for an enrichment or depletion of barcode errors from the designed barcode sequence using a hypergeometric test (Supplemental Fig. 6). Tags or primers not identified in the re-characterization sequencing reaction were excluded from both the test sets as well as the entire population of yeast deletion mutants used to generate the distribution.

Acknowledgments

This work was supported by US National Institutes of Health (NIH) grants R01 HG003224 and R21 HG004756 to F.P.R. and HG003788 to M.C. J.M. was supported by an individual NRSA Fellowship from the NIH/NHGRI. G.G. and C.N. were supported by the CIHR (MOP-81340 to G.G., MOP-84305 to C.N.) and the NHGRI (HG00317-05). We thank Angela Chu, Bob P. St. Onge, and all members of the HIP-HOP lab for advice; and Lixin Zhou for developing NGS analysis methods at Prognosys Biosciences. We thank Ron Davis and all members of the YKO consortium.

Authors' contributions: C.N., A.M.S., G.G., and M.C. conceived of the project and designed experiments. M.C. provided sequencing capabilities and discussion. A.M.S. and F.K. developed methods for sample multiplexing and performed the experiments. A.M.S., L.E.H., J.M., F.P.R, C.N., G.G., and M.J.T. analyzed the data. A.M.S., C.N., and G.G. wrote the paper.

References

Ben-Aroya S, Coombes C, Kwok T, O'Donnell KA, Boeke JD, Hieter P. 2008. Toward a comprehensive temperature-sensitive mutant repository of the essential genes of *Saccharomyces cerevisiae*. *Mol Cell* **30:** 248–258.
Bennett S. 2004. Solexa Ltd. *Pharmacogenomics* **5:** 433–438.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53–59.

Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, Pawlowski TL, Laub T, Nunn G, Stephan DA, et al. 2008. Identification of genetic variants using barcoded multiplexed sequencing. *Nat Methods* **5:** 887–893.

Eason RG, Pourmand N, Tongprasit W, Herman ZS, Anthony K, Jejelowo O, Davis RW, Stolc V. 2004. Characterization of synthetic DNA bar codes in *Saccharomyces cerevisiae* gene-deletion strains. *Proc Natl Acad Sci* **101:** 11046–11051.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418:** 387–391.

Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW. 2004. Chemogenomic profiling: Identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci* **101:** 793–798.

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St Onge RP, Tyers M, Koller D, et al. 2008. The chemical genomic portrait of yeast: Uncovering a phenotype for all genes. *Science* **320:** 362–365.

Ho CH, Magtanong L, Barker SL, Gresham D, Nishimura S, Natarajan P, Koh JL, Porter J, Gray CA, Andersen RJ, et al. 2009. A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat Biotechnol* **27:** 369–377.

Hoon S, Smith AM, Wallace IM, Suresh S, Miranda M, Fung E, Proctor M, Shokat KM, Zhang C, Davis RW, et al. 2008. An integrated platform of genomic assays reveals small-molecule bioactivities. *Nat Chem Biol* **4:** 498–506.

Kitagawa M, Ara T, Arifuzzaman M, Ioka-Nakamichi T, Inamoto E, Toyonaga H, Mori H. 2005. Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): Unique resources for biological research. *DNA Res* **12:** 291–299.

Lefrancois P, Euskirchen GM, Auerbach RK, Rozowsky J, Gibson T, Yellman CM, Gerstein M, Snyder M. 2009. Efficient yeast ChIP-seq using multiplex short-read DNA sequencing. *BMC Genomics* **10:** 37. doi: 10.1186/1471-2164-10-37.

Li H, Ruan J, Durbin R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18:** 1851–1858.

MacLean D, Jones JD, Studholme DJ. 2009. Application of "next-generation" sequencing technologies to microbial genetics. *Nat Rev Microbiol* **7:** 287–296.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.

Ozsolak F, Song JS, Liu XS, Fisher DE. 2007. High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol* **25:** 244–248.

Pierce SE, Fung EL, Jaramillo DF, Chu AM, Davis RW, Nislow C, Giaever G. 2006. A unique and universal molecular barcode array. *Nat Methods* **3:** 601–603.

Pierce SE, Davis RW, Nislow C, Giaever G. 2007. Genome-wide analysis of barcoded *Saccharomyces cerevisiae* gene-deletion mutants in pooled cultures. *Nat Protocols* **2:** 2958–2974.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* **4:** 651–657.

Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294:** 2364–2368.

Yan Z, Costanzo M, Heisler LE, Paw J, Kaper F, Andrews BJ, Boone C, Giaever G, Nislow C. 2008. Yeast barcoders: A chemogenomic application of a universal donor-strain collection carrying bar-code identifiers. *Nat Methods* **5:** 719–725.

Received March 19, 2009; accepted in revised form July 9, 2009.