Chapter 6

Microarray-Based Genetic Footprinting Strategy to Identify Strain Improvement Genes after Competitive Selection of Transposon Libraries

Alison K. Hottes and Saeed Tavazoie

Abstract

Successful strain engineering involves perturbing key nodes within the cellular network. How the network's connectivity affects the phenotype of interest and the ideal nodes to modulate, however, are frequently not readily apparent. To guide the generation of a list of candidate nodes for detailed investigation, designers often examine the behavior of a representative set of strains, such as a library of transposon insertion mutants, in the environment of interest. Here, we first present design principles for creating a maximally informative competitive selection. Then, we describe how to globally quantify the change in distribution of strains within a transposon library in response to a competitive selection by amplifying the DNA adjacent to the transposons and hybridizing it to a microarray. Finally, we detail strategies for analyzing the resulting hybridization data to identify genes and pathways that contribute both negatively and positively to fitness in the desired environment.

Key words: Genetic footprinting, *Escherichia coli*, Strain engineering, Transposon, Bacterial genetics, Microarray analysis, Statistics

1. Introduction

Strain engineering starts with an existing cellular network and determines how best to modify that network to optimize a phenotype of interest, such as production of a metabolite. Complicating the design process, however, is the biological reality that multiple cellular pathways affect many phenotypes of commercial and medical importance, such as ethanol tolerance and antibiotic susceptibility (1, 2). While mutations can be directed to regions of interest (3), exploring all possible cellular networks that are within even a few mutational steps of the original network is not currently feasible. Fortunately, although not all

James A. Williams (ed.), *Strain Engineering: Methods and Protocols*, Methods in Molecular Biology, vol. 765,
DOI 10.1007/978-1-61779-197-0_6, © Springer Science+Business Media, LLC 2011

mutations are additive, many are (1, 2). Thus, discovering single perturbations that influence a phenotype is a productive first step toward identifying combinations of mutations likely to further enhance a phenotype.

Transposon mutagenesis is a convenient way to generate a collection of strains each with a single mutation in a readily identifiable location (4–8). Transposon insertions can produce a wide range of phenotypes from null alleles caused by insertions in coding regions, to overexpression phenotypes resulting from insertions in intergenic regions that increase the expression of neighboring genes, to hypomorphs produced by insertions in the extreme 3′ end of genes. Furthermore, many commercial companies, such as Epicentre Biotechnologies, Finnzymes, and New England Biolabs offer transposons and transposases with desirable properties such as high transposition efficiency and low insertion site sequence bias.

Although some studies have tested large numbers of transposon insertion mutants individually (7), working with a library en masse is frequently more convenient and cost-effective (9). Subjecting a transposon library to a competitive selection enriches for strains with insertions that increase fitness and depletes the library of insertions that decrease fitness. Insertions that enhance fitness are obviously relevant to strain engineering. Since many insertions that decrease fitness are in genes essential to the behavior of interest, such genes are good candidates for targeted upregulation. Thus, strain engineering requires knowledge of both beneficial and deleterious insertion locations. Strongly beneficial insertion locations can often be identified by individually mapping the location of the transposon insertions in a number of cells isolated from a population after competitive selection. Individual colony methods, however, are not suitable for identifying insertion locations that decrease fitness or that increase fitness only moderately. More global methods, however, can characterize the full distribution of transposon insertion locations in a population before and after a selection and provide quantitative information about the contribution of each gene to a phenotype.

Here, we first discuss key considerations for designing an informative competitive transposon library selection. We then describe how to selectively amplify the DNA adjacent to transposons and hybridize it to a microarray to quantify the distribution of transposon insertion locations in a population. Finally, we address the main issues in data analysis: array normalization, identification of transposon insertion sites that cause fitness effects significant at a chosen false discovery rate (FDR), and discovery of pathways underlying the phenotype of interest.

The protocols presented were developed by Badarinarayana et al. (9) and Girgis et al. (10) using *Escherichia coli*, but should be readily adaptable to other organisms. A wide variety of related protocols are available (e.g., see refs. 11, 12).

2. Materials

2.1. Competitive Library Enrichment

1. Transposon insertion library, preferably frozen at –80°C in single-use aliquots. Figure 1 shows a typical transposon and the specific elements needed for this protocol.
2. Enrichment-specific materials.
3. LB + 30% glycerol: 0.5% yeast extract (w/v), 0.5% NaCl (w/v), 1% tryptone (w/v), and 30% glycerol (v/v). Autoclave to sterilize. Store at room temperature (see Note 1).
4. Dry ice.
5. Ethanol.

2.2. Genetic Footprinting

2.2.1. Restriction Digestion

1. Lysis buffer: Prepare just before use, per sample, combine 96 μl water, 12 μl 10× NEBuffer 2 [500 mM NaCl, 100 mM Tris–HCl, 100 mM $MgCl_2$, 10 mM dithiothreitol pH 7.9 (New England Biolabs, Ipswich, MA)], and 6 μl Triton X-100.
2. Alkaline phosphatase, 1 U/μl (Roche). Store at 4°C.
3. HinP1I, 10 U/μl (New England Biolabs or equivalent). Store at –20°C.
4. MspI, 20 U/μl (New England Biolabs or equivalent). Store at –20°C.

2.2.2. Y-Linker Ligation

1. 3 M sodium acetate pH 5.2: Use acetic acid for pH adjustment. Autoclave to sterilize; store at room temperature.
2. Ethanol chilled at –20°C.
3. 70% ethanol chilled at –20°C.

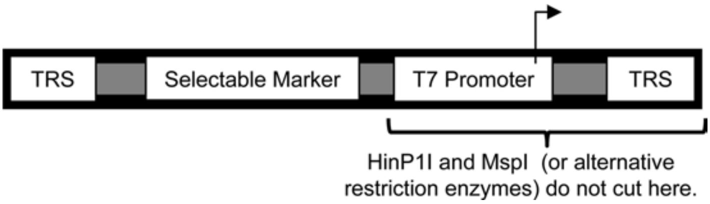
TRS
Selectable Marker
T7 Promoter
TRS
HinP1I and MspI (or alternative restriction enzymes) do not cut here.

Fig. 1. Transposon structure and required components. Transposon ends contain transposase recognition sequences (TRS) that are recognized by the corresponding transposase. Transposons typically also contain a selectable marker that can facilitate selecting for strains that contain the transposon. The protocol presented here requires the presence of an outward-reading T7 promoter near one of the transposon's ends. Additionally, the protocol assumes that the HinPI1 and MspI restriction enzymes do not cut between the T7 promoter and the end of the transposon. Otherwise, alternative restriction enzymes must be substituted (see Note 7). The modified Tn5 transposon described in ref. (10) that meets these criteria and was used to develop the methods described herein is available upon request.

4. Y-linker (40 pmol/μl): Purchase the following HPLC-purified primers:

   5′ – ACTACGCACGCGACGAGACGTAGCGTC – 3′ (YCG5) and

   5′ – P-CGGACGCTACGTCCGTGTTGTCGGTCCTG – 3′ (YCG3).

   Note that YCG3 is phosphorylated on the 5′ end. Dissolve each in water at a concentration of 100 pmol/μl. In a PCR tube, combine 30 μl primer YCG5, 30 μl primer YCG3, 7.5 μl 10× annealing buffer [1 M NaCl, 100 mM Tris–HCl (pH 8.0), 10 mM EDTA (pH 8.0)], and 7.5 μl water. Using a thermocycler, heat the mixture at 94°C for 1 min and then drop the temperature in 2°C increments every 30 s until reaching 26°C. The reaction may be scaled up as needed. Y-linker should be frozen at −20°C in single use aliquots (e.g., 25 μl aliquots are ideal for processing samples in batches of eight).
5. T4 DNA ligase (400 U/μl) and 10× buffer [500 mM Tris–HCl, 100 mM $MgCl_2$, 10 mM ATP, 100 mM dithiothreitol, pH 7.5] (New England Biolabs or equivalent). Store at −20°C.
6. QIAquick PCR Purification Kit (Qiagen, Valencia CA)

2.2.3. Repair Nicks

1. 10× NEBuffer 2 [500 mM NaCl, 100 mM Tris–HCl, 100 mM $MgCl_2$, 10 mM dithiothreitol pH 7.9 (New England Biolabs)]. Store at −20°C.
2. dNTP mix: 2.5 mM each of dATP, dCTP, dGTP, and dTTP. Store at −20°C.
3. *E. coli* DNA polymerase I (10 U/μl) (New England Biolabs or equivalent). Store at −20°C.

2.2.4. Amplify Transposon-Adjacent DNA by PCR

1. Water.
2. dNTP mix: 2.5 mM each of dATP, dCTP, dGTP, and dTTP. Store at −20°C.
3. Primer Y-COMP (5′-ACTACGCACGCGACGAGACG-3′), 10 μM. This primer anneals to the complement of the single-stranded part of the Y-linker (see Fig. 2). Store at −20°C.
4. Primer T7-UPSTRM, 10 μM. This primer, in conjunction with primer Y-COMP should amplify the end of the transposon, including the T7 promoter (see Fig. 2). Store at −20°C.
5. Ex Taq polymerase and 10× Ex Taq buffer (Takara). Store at −20°C.
6. QIAquick PCR Purification Kit (Qiagen).
7. Nuclease-free water.

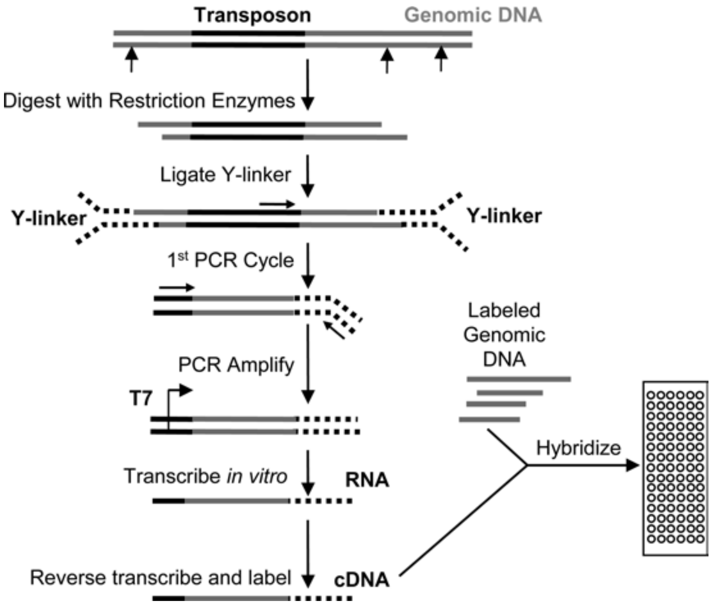

Fig. 2. Genetic footprinting protocol overview. First, genomic DNA from the transposon insertion library is digested with restriction enzymes; the DNA adjacent to a transposon insertion serves as the marker for the insertion site. Then, a Y-linker with an overhang compatible with the restriction digestion is ligated to the DNA. Next, PCR is used to amplify the ends of the transposons and the adjacent DNA. During the first PCR cycle, the primer from the transposon primes the synthesis of DNA complementary to one strand of the Y-linker. The second PCR primer then anneals to the newly synthesized DNA and participates in subsequent rounds of amplification. To reduce the nonlinearities introduced by PCR, the number of cycles is limited as much as possible. To obtain sufficient product for hybridization, the DNA adjacent to the transposon is further amplified by in vitro transcription using a T7 promoter located on the transposon. The resulting RNA is then typically converted into cDNA and labeled in a way suitable for the chosen microarray hybridization technology. Finally, a microarray is used to quantify the fraction of the library population with transposon insertions near each array probe (modified from ref. (10), which was published by Public Library of Science as an open-access article under a Creative Commons Attribution License).

2.2.5. Further Amplify Transposon-Adjacent DNA Using In Vitro Transcription

1. MEGAscript T7 Kit (Ambion Inc., Austin, TX).
2. RNeasy Mini Kit (Qiagen).

2.2.6. Microarray Hybridization

1. Genomic DNA from the transposon library's parental strain: DNA should be fragmented to an appropriate size and suitably labeled for hybridization using the chosen microarray platform (see Note 2).
2. Reagents needed to synthesize cDNA suitably labeled for the chosen microarray platform from RNA.
3. Reagents needed for a microarray hybridization.

3. Methods

3.1. Competitive Library Enrichment

1. Subject the transposon insertion library to the experimental conditions of interest (see Note 3 and Fig. 3).
2. Preserve samples of the population throughout the course of the experiment by mixing equal volumes of culture and LB + 30% glycerol, snap-freezing in dry ice and ethanol, and storing at –80°C. Archival samples allow for detailed studies of the progression of the selection and can also be searched for mutants with transposon insertions in sites of interest.
3. At times of interest, collect samples for genetic footprinting (see Note 4). For each sample, pellet ~10^7 cells by centrifugation, remove all supernatant possible using a pipette, and store the pellet at –80°C until needed.

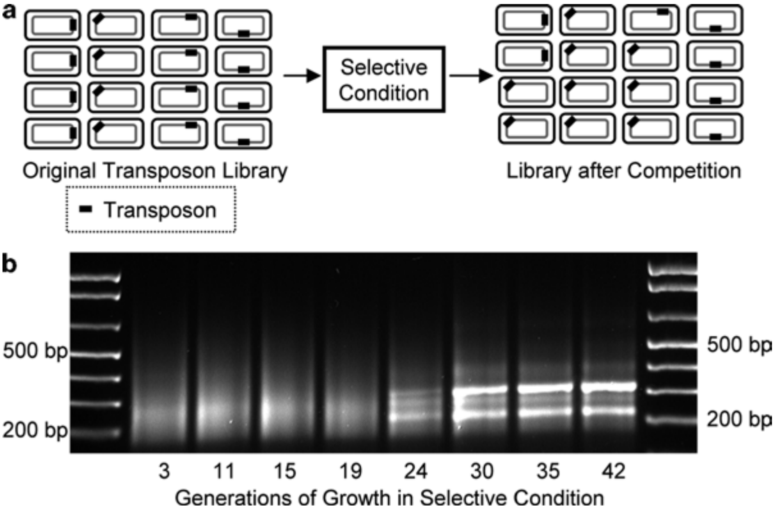
a
Selective Condition
Original Transposon Library
Library after Competition
▬ Transposon
b
500 bp
200 bp
500 bp
200 bp
3
11
15
19
24
30
35
42
Generations of Growth in Selective Condition

Fig. 3. Library diversity as a function of generations of competitive selection. (**a**) The original, high-diversity transposon library is subjected to a competitive selection that increases the abundance of strains with beneficial transposon insertions and reduces the abundance of strains with deleterious transposon insertions. Ideally, a selection should span enough generations to detectably magnify the abundance of strains with small fitness increases over the wild-type strain, but not so many generations that both strains of average and below-average fitness drop out of the population completely. (**b**) Samples of a transposon library propagated in defined media with aspartic acid as the sole carbon source for the indicated number of generations were subjected to genetic footprinting. The resulting PCR products (the output of Subheading 3.2.4) were then run on a 2% agarose gel. DNA band sizes are indicated in the far left and right lanes. The presence of discrete bands indicates that a clone reached high density in the population. The clone either contained a highly beneficial transposon insertion or, as happens more commonly, a beneficial spontaneous mutation that allowed the endogenous transposon insertion to hitchhike to prominence. In our experience, spontaneous mutations typically become problematic after about 20 generations.

3.2. Genetic Footprinting

See Fig. 2 for an overview of the procedure.

3.2.1. Restriction Digestion

1. Thaw the sample pellet briefly at room temperature and suspend it in 114 μl lysis buffer.
2. Transfer 48 μl of cells to each of two PCR tubes (see Note 5).
3. Incubate the tubes at 99°C for 40 s in a thermocycler to lyse the cells, and then cool to room temperature.
4. Add 1 μl alkaline phosphate to both tubes (see Note 6).
5. Add 1 μl HinP1I to one tube and 1 μl MspI to the other (see Note 7). Mix.
6. Incubate at 37°C for 3 h.
7. Heat at 65°C for 20 min to deactivate the restriction enzymes (see Note 8).

3.2.2. Ligate Y-Linker

1. Combine the two restriction digests.
2. Add 10 μl 3 M sodium acetate (pH 5.2) and transfer the mixture to a microfuge tube.
3. Add 0.3 ml of −20°C ethanol and mix.
4. Freeze at −20°C for at least 1 h.
5. Centrifuge at >13,000 × *g* for 10 min at 4°C (maximum RPM in microfuge).
6. Pour off the supernatant without disturbing the pellet.
7. Add 0.5 ml −20°C 70% ethanol.
8. Centrifuge at >13,000 × *g* for 10 min at 4°C.
9. Pour off the supernatant without disturbing the pellet.
10. Centrifuge the tube briefly to collect the remaining liquid in the bottom of the tube.
11. Pipet out the residual liquid.
12. Allow the pellet to dry to remove the remaining ethanol. This can either be done in a speed-vac for ~1 min or in a fume hood for ~30 min. Do not over-dry.
13. Resuspend the pellet in 23 μl water, 3 μl 10× T4 DNA ligase buffer, and 3 μl Y-linker. Keep on ice.
14. Add 1 μl T4 DNA ligase.
15. Place the sample in a floating microfuge tube rack in a container with 2 l of room temperature water. Place the container with water in a 4°C room overnight. Alternatively, the sample can be ligated at 16°C overnight.
16. Clean up the sample using a Qiaquick PCR purification kit according to the manufacturer's directions. In the last step, elute in 26 μl of water; approximately 24 μl will flow through.

3.2.3. Repair Nicks (see Note 9)

1. To the 24 µl sample, add 3 µl 10× NEBuffer 2, 2 µl dNTP mix, and 1 µl *E. coli* DNA polymerase I.
2. Incubate at 25°C for 2 h.
3. Inactive the enzyme by heating at 75°C for 20 min.

3.2.4. Amplify Transposon-Adjacent DNA by PCR

1. Combine the following in order: 25.8 µl water, 5 µl 10× Ex Taq buffer, 4 µl dNTP mix, 5 µl T7-UPSTRM primer, 5 µl Y-COMP primer, 5 µl of nick-repaired ligation product, and 0.2 µl Ex Taq polymerase.
2. Heat in a thermocycler at 94°C for 2 min. Then, cycle at 94°C for 30 s, 68°C for 30 s, and 72°C for 3 min 30 times (see Note 10). Finally, heat at 72°C for 10 min.
3. Clean up the sample using a Qiaquick PCR purification kit according to the manufacturer's directions. In the last step, elute in 30 µl nuclease-free water.
4. If desired, visualize the sample on a 2% agarose gel as in Fig. 3b.

3.2.5. Further Amplify Transposon-Adjacent DNA Using In Vitro Transcription

1. Combine the following components (from the MEGAscript T7 kit) in a PCR tube at room temperature: 2 µl each of ATP, CTP, GTP, and UTP solutions (8 µl total), 2 µl of 10× reaction buffer, 1 µg of PCR product from the reaction above, and enough nuclease-free water to bring the total volume to 18 µl (see Note 11).
2. Add 2 µl T7 enzyme mix (from kit).
3. Incubate for 4 h at 37°C.
4. Add 1 µl TURBO DNase (2 U/µl) from the MEGAscript T7 kit and incubate for 15 min at 37°C.
5. Purify the RNA using the RNeasy Mini Kit according to the manufacturer's directions. In the final step, elute in 40 µl RNase-free water.

3.2.6. Microarray Hybridization

1. Select a microarray platform (see Note 2).
2. For two-color, comparative platforms, prepare a labeled, genomic DNA reference (see Note 12). See Girgis et al. (10) or the array manufacturer's instructions.
3. Synthesize cDNA suitably labeled for the chosen microarray platform from the in vitro transcribed RNA.
4. Hybridize the sample to the chosen array.

3.3. Data Analysis

This section focuses on the analysis of samples either hybridized to single channel platforms (e.g., Affymetrix arrays) or hybridized to two-channel platforms (e.g., Agilent arrays) using genomic DNA as a common reference. Data sets from competitive selections, similar to expression data sets, are large and for reasons of expense typically contain few repetitions. The large number of genes per array necessitates an awareness of the number of false positives expected due to multiple hypothesis testing (see Note 13). The small number of repetitions favors the use, at least initially, of simple analysis techniques with few parameters to fit. Here, we describe basic analysis techniques that work well with most data sets; numerous alternative algorithms are described in the literature that may be helpful in special situations (see refs. 13–15 for a sampling of reviews).

3.3.1. Obtain Data Describing the Composition of the Transposon Library Prior to Competitive Selection

1. For comparative purposes, process and hybridize *at least* three samples of the original, unselected library. Five samples are commonly used (10). To make the null distribution as accurate as possible, each sample should be processed independently starting with the genetic footprinting step (Subheading 3.2).

3.3.2. Perform Suitable Within-Array Normalization (see Note 14)

1. Compensate for background and off-target hybridization as dictated by the technology.
2. Additionally, for two-channel arrays, scale the signals so that the contribution of each channel is equal (10). In other words, the sum of the signal from the first channel over all of the probes should be equal to the sum of the signal from the second channel over all the probes.
3. Combine data from all probes representing each gene as appropriate for the array.

3.3.3. Employ Between-Array Normalization to Correct for Signal Strength Variations Between Arrays

1. Identify the genes that are present on all of the unselected library hybridizations and all of the experimental samples of current interest. This step does not distinguish between experimental and reference samples; all of the arrays should be processed together.
2. For a one-channel technology, let $s_{i,j}$ be the signal from the ith gene on the jth array; for a two-channel technology, let $s_{i,j}$ be the ratio of the competitive enrichment signal and the genomic DNA signal for the ith gene on the jth array.
3. For each array, compute t_j, the total signal from array j for all genes present on all arrays. That is, find $t_j = \sum_{i=1}^{N} s_{i,j}$, where the index, i, runs over the N genes with

   signal present on all arrays.
4. For each array, j, replace $s_{i,j}$ with $s_{i,j}C/t_j$ where C is an arbitrary constant chosen to put the numbers on a convenient scale. Make the replacement for all genes, not just those with valid signals on all arrays.

3.3.4. Calculate z-scores

A *z*-score, $z_{i,j}$, should be calculated for each gene, *i*, and each hybridization, *j*, of the competitively selected library.

1. Let μ_i be the average of the normalized signal for gene *i* from the hybridizations of the unselected library.
2. Let σ_i be the standard deviation of the normalized signal for gene *i* from the hybridizations of the unselected library.
3. Define $z_{i,j} = (s_{i,j} - \mu_i)/\sigma_i$ where $s_{i,j}$ is the normalized signal calculated above (see Note 15). A positive *z*-score indicates that the fraction of strains with insertions in or near gene *i* increased during the selection; a negative *z*-score indicates that the fraction of strains with insertions in or near gene *i* decreased during the selection. The normalization by σ_i accounts for the expected variability of each gene.

3.3.5. Identify Genes that Changed Compared to the Unselected Library

1. Let *z* be the significance threshold.
2. Consider a gene *i* to have caused a significant effect in competitive selection *j* if $|z_{i,j}| > z$ where $z_{i,j}$ is the *z*-score calculated in Subheading 3.3.4 (see Note 16).

3.3.6. Estimate the False Discovery Rate

The FDR is the fraction of the set deemed significant using a particular *z*-score, *z*, that is expected to consist of false positives (16).

1. Let *S* be the number of significant genes from Subheading 3.3.5.
2. Use the hybridizations of the unselected library as a model for the null distribution. For each gene, randomly remove one of the measurements from the set of unselected library hybridizations and designate it as "signal." Then, calculate *z*-scores as in Subheading 3.3.4. Take care *not* to use the data designated as "signal" in calculating the per-gene means and standard deviations. See Fig. 4a.
3. Calculate *FP*, the expected number of false positives. FP is the number of samples in the null distribution with *z*-scores of greater magnitude than *z*, the significance threshold used in Subheading 3.3.5.
4. The FDR is FP/*S*. See Fig. 4b, c.

3.3.7. Combine Data from Multiple Competitive Selections, if Available

1. If three or more samples are available for each gene, use the median.
2. If only two samples are available, and both have *z*-scores of the same sign, use the one closest to zero; otherwise assign a *z*-score of zero (see Note 17).
3. If desired, reestimate the FDR by generating a null distribution that reflects how multiple samples were combined.

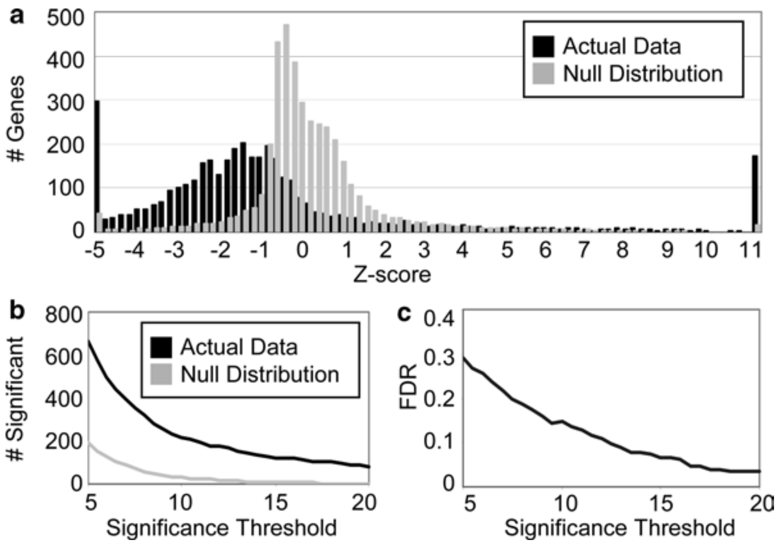

Fig. 4. Calculating the false discovery rate (FDR) as a function of the significance threshold. *Z*-scores relative to five hybridizations of the original, unselected library were calculated for data from a competitive selection to find *E. coli* mutants that remain motile in high salt concentrations. A null distribution was simulated by treating one of the five reference samples for each gene as data as described in Subheading 3.3.6. A global component equal to one-tenth of the average standard deviation was added to the standard deviation of each gene (see Note 15). (**a**) The histogram displays the *z*-scores for the real data and the null distribution. The real data has a larger spread and heavier tails than the null distribution indicating that the library contained some mutants of above- and below-average fitness. During the course of the selection, several strains became a substantial part of the population and reduced the prevalence of the average mutant. As a result, the mean *z*-score for the real data is lower than the mean *z*-score of the null distribution. (**b**) A gene was considered significant if the absolute value of its *z*-score was greater than the indicated threshold. (**c**) As the significance threshold decreases, both the estimated FDR and the number of true positives increase. The FDR will not necessarily increase monotonically as the number of true positive increases, but it usually does. All data were published in Girgis et al. (10).

3.3.8. Search for Pathways that Contributed to Fitness in the Competition

1. Pathway analysis looks for commonalities among the genes with similar *z*-scores. By examining the data set as a whole, *z*-scores that are individually too small to be considered significant can still contribute to the identification of large-scale patterns.
2. Many pathway analysis tools are available. In particular, the Tavazoie lab has developed iPAGE (17), which identifies pathways and gene ontology (GO) terms (18) that are enriched or depleted for each range of *z*-scores. See Fig. 5 for an example.

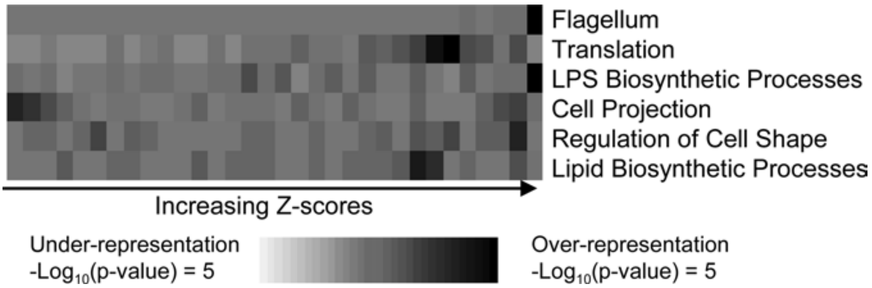

Fig. 5. Using iPAGE (17) to identify pathways involved in χ-phage susceptibility. *Z*-scores from a competitive selection to find *E. coli* mutants with reduced sensitivity to χ-phage were calculated relative to five hybridizations of the original, unselected library. A global component equal to one half of the average standard deviation was added to each gene's standard deviation (see Note 15). Data from two independent repetitions were combined by taking the value closest to zero when the repetitions had the same sign and using a value of zero otherwise (see Subheading 3.3.7). Columns, from left to right, correspond to equally populated bins of increasing *z*-scores; values of zero are present in the second through fifth columns from the right. The darker (lighter) the rectangle, the more the range of *z*-score was enriched (depleted) for the indicated functional category; no significant regions of depletion were identified in this data set. The results suggest that LPS or flagella defects increase χ-phage resistance while defects in cell projection processes (e.g., fimbrial-like proteins) increase susceptibility (10). iPAGE can detect functional enrichments in middle ranges of *z*-scores as well as in the most extreme ranges. For example, *z*-scores just below zero are enriched for genes with products involved in translation; members of the set, which consists mainly of genes encoding essential ribosomal proteins, were largely absent in the library both before and after the selection. Data came from Girgis et al. (10). *LPS* lipopolysaccharides.

4. Notes

1. Unless stated otherwise, solutions and media should be made with deionized water.
2. We have successfully used Affymetrix tiling arrays (unpublished), Agilent oligo arrays (unpublished), and in-house arrays containing a PCR product from each open reading frame (ORF) (1, 2, 10, 19). The size of the features on an array (i.e., 25 mers, 60 mers, or ~1 kb ORFs) determines the precision with which the technology will be able to resolve transposon insertion locations. The combination of the density of the features and the size of the transposon-adjacent DNA amplified, which is set by the restriction enzymes used in the protocol, determines which transposon insertion locations will contribute signal to the hybridization. Other considerations are a lab's familiarity with a particular platform and the availability of the needed infrastructure.
3. During competitive selections, the minimum population size should be kept large enough to avoid unwanted bottlenecking. Additionally, as determining the ideal length (generations) for an enrichment a priori is difficult, taking samples at multiple times is advisable.
4. If feasible, collect all samples at similar growth stages and conditions, such as stationary phase. Otherwise, cells from

the fastest growing cultures will have more DNA near the origin of replication, which will inflate the number of copies of insertions near the origin compared to the terminus of replication (20). If such growth rate differences are unavoidable, consult Vora et al. (21) for an example of a windowing approach that can be used to correct for the resulting chromosome position biases.

5. Samples are suspended in a slight excess of lysis buffer as the Triton X-100 causes bubbles that make it difficult to use the whole volume.

6. The inclusion of alkaline phosphatase, as suggested by Girgis et al. (10), prevents genomic DNA segments from ligating to each other instead of Y-linker.

7. Since transposon insertion sites too close to restriction enzyme cut sites do not yield identifiable DNA segments, two separate restriction digests are used. Ensure that the restriction enzymes do not cut between the T7 promoter and the end of the transposon. If the chosen restriction enzymes do not leave a 5′-CG overhang, then the Y-linker sequence will need to be adjusted.

8. Alkaline phosphatase cannot be heat-inactivated, and prolonged storage of the mixture at 4°C may result in DNA degradation. Either proceed immediately to the next step or store samples at −20°C.

9. DNA polymerase I repairs the nicks between the 5′-ends of the genomic DNA and the Y-linker, which exist because the genomic DNA was dephosphorylated. Unfortunately, the enzyme can be finicky, resulting in little or no PCR product in the next step. Omitting alkaline phosphatase from the restriction digest, similar to Badarinarayana et al. (9), obviates the need for DNA polymerase I repair and increases both the signal and the background.

10. The number of PCR cycles should be kept to a minimum to reduce nonlinear amplification biases.

11. Take standard precautions, such as using filter tips, to avoid introducing RNases into the sample.

12. Instead of comparing each sample to a common reference (genomic DNA), two samples can also be compared directly as was done in Goodarzi et al. (22). Typically, the use of a common reference facilitates the meta-analysis of data from a large number of competitions.

13. For simplicity, the analysis procedure discusses genes instead of probes. Repeating the analyses with the probes treated individually may provide insights into regions of genes, such as segments that code for protein domains, that affect fitness differentially. Additionally, many probes or probe sets represent intergenic regions, which can be treated similarly to genes.

14. A variety of commercial software performs all of the steps of Subheading 3.3.2. For example, the MAS5 algorithm (23–25) commonly used with Affymetrix arrays performs background corrections, combines all of the probes for each gene, and scales the final results so that sets of arrays will have similar scaling, which may reduce the need to perform between-array normalization (see Subheading 3.3.3).

15. The normalization procedure assumes that the abundance of most mutants in the population remains relatively constant throughout a selection. Some stringent selections that cause the abundance of all but the fittest mutants to decrease appreciably, however, can pose analysis problems if the level assigned to genes present in only negligible amounts shifts. A slight change in mean "signal" from absent genes coupled with the typically small standard deviation of the unselected library signal of essential genes (i.e., genes in which the cell cannot tolerate transposon insertions in the conditions used for library construction) can cause large *z*-scores to be associated with essential genes. The difficulty can be largely overcome by adding a small constant, which represents the global variability of the array, to all of the gene-specific standard deviations used in calculating *z*-scores (1). That is, each σ_i can be replaced with $\sigma_i + \sigma_{\text{global}}$, where σ_{global} is, for example, one-tenth to one half the average σ_i.

16. If only beneficial insertions are of interest, neglect the absolute value symbol and consider only positive *z*-scores.

17. The procedure described reduces the number of false positives at the possible expense of an increase in the number of false negatives. Averaging is avoided as the noise caused by spontaneous mutations that can cause some transposon insertions to hitchhike to prominence can result in extreme outliers that do not follow a Gaussian distribution. For similar reasons, all repetitions should be biologically independent and go through separate competitive enrichments.

Acknowledgments

We are grateful to Hany Girgis for developing and optimizing many of the techniques described here and to Hani Goodarzi for developing iPAGE. Work in the Tavazoie lab was supported by grants from NSF (CAREER), DARPA (BIOS), NIGMS (P50 GM071508), and the NIH Director's Pioneer Award (1DP10D 003787-01).

References

1. Girgis H. S., Hottes A. K., and Tavazoie S. (2009) Genetic architecture of intrinsic antibiotic susceptibility. *PLoS One* 4, e5629.
2. Goodarzi H., Bennett B. D., Amini S., Reaves M. L., Hottes A. K., Rabinowitz J. D., and Tavazoie, S. (2010) Regulatory and metabolic rewiring during laboratory evolution of ethanol tolerance in *E. coli. Mol. Syst. Biol.* 6, 378.
3. Wang H. H., Isaacs F. J., Carr P. A., Sun Z. Z., Xu G., Forest C. R., and Church G. M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature* 460, 894–898.
4. Akerley B. J., Rubin E. J., Novick V. L., Amaya K., Judson N., and Mekalanos J. J. (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae. Proc. Natl. Acad. Sci. U. S. A.* 99, 966–971.
5. Dziva F., van Diemen P. M., Stevens M. P., Smith A. J., and Wallis T. S. (2004) Identification of *Escherichia coli* O157 : H7 genes influencing colonization of the bovine gastrointestinal tract using signature-tagged mutagenesis. *Microbiology* 150, 3631–3645.
6. Gonzalez M. D., Lichtensteiger C. A., and Vimr E. R. (2001) Adaptation of signature-tagged mutagenesis to *Escherichia coli* K1 and the infant-rat model of invasive disease. *FEMS Microbiol. Lett.* 198, 125–128.
7. Jacobs M. A., Alwood A., Thaipisuttikul I., Spencer D., Haugen E., Ernst S., Will O., Kaul R., Raymond C., Levy R., Chun-Rong L., Guenthner D., Bovee D., Olson M. V., and Manoil C. (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa. Proc. Natl. Acad. Sci. U. S. A.* 100, 14339–14344.
8. Salama N. R., Shepherd B., and Falkow S. (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori. J. Bacteriol.* 186, 7926–7935.
9. Badarinarayana V., Estep P. W., 3rd, Shendure J., Edwards J., Tavazoie S., Lam F., and Church G. M. (2001) Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol.* 19, 1060–1065.
10. Girgis H. S., Liu Y., Ryu W. S., and Tavazoie S. (2007) A comprehensive genetic characterization of bacterial motility. *PLoS Genet.* 3, 1644–1660.
11. Winterberg K. M., and Reznikoff W. S. (2007) Screening transposon mutant libraries using full-genome oligonucleotide microarrays. *Methods Enzymol.* 421, 110–125.
12. Baldwin D. N., and Salama N. R. (2007) Using genomic microarrays to study insertional/transposon mutant libraries. *Methods Enzymol.* 421, 90–110.
13. Do J. H., and Choi D. K. (2006) Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol. Cells.* 22, 254–261.
14. Speed T., and Zhao H. (2009) Microarrays. *Stat. Methods Med. Res.* 18, 531–532.
15. Steinhoff C., and Vingron M. (2006) Normalization and quantification of differential expression in gene expression microarrays. *Brief Bioinform.* 7, 166–177.
16. Tusher V. G., Tibshirani R., and Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 5116–5121.
17. Goodarzi H., Elemento O., and Tavazoie S. (2009) Revealing global regulatory perturbations across human cancers. *Mol. Cell.* 36, 900–911.
18. Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., and Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.
19. Amini S., Goodarzi H., and Tavazoie S. (2009) Genetic dissection of an exogenously induced biofilm in laboratory and clinical isolates of *E. coli. PLoS Pathog.* 5, e1000432.
20. Cooper S., and Helmstetter C. E. (1968) Chromosome replication and the division cycle of *Escherichia coli* B/r. *J. Mol. Biol.* 31, 519–540.
21. Vora T., Hottes A. K., and Tavazoie S. (2009) Protein occupancy landscape of a bacterial genome. *Mol. Cell.* 35, 247–253.
22. Goodarzi H., Hottes A. K., and Tavazoie S. (2009) Global discovery of adaptive mutations. *Nat. Methods* 6, 581–583.
23. Hubbell E., Liu W. M., and Mei R. (2002) Robust estimators for expression analysis. *Bioinformatics* 18, 1585–1592.
24. Liu W. M., Mei R., Di X., Ryder T. B., Hubbell E., Dee S., Webster T. A., Harrington C. A., Ho M. H., Baid J., and Smeekens S. P. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 18, 1593–1599.
25. Affymetrix. (2002) Statistical Algorithms Description Document http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf Accessed June 22, 2010