Selection analyses of insertional mutants using subgenic-resolution arrays

Vasudeo Badarinarayana[1], Preston W. Estep III[1], Jay Shendure[1], Jeremy Edwards[1,2], Saeed Tavazoie[3], Felix Lam[1], and George M. Church[1]*

We describe a method of genome-wide analysis of quantitative growth phenotypes using insertional mutagenesis and DNA microarrays. We applied the method to assess the fitness contributions of *Escherichia coli* gene domains under specific growth conditions. A transposon library was subjected to competitive growth selection in Luria–Bertani (LB) and in glucose minimal media. Transposon-containing genomic DNA fragments from the selected libraries were compared with the initial unselected transposon insertion library on DNA microarrays to identify insertions that affect fitness. Genes involved in the biosynthesis of nutrients not provided in the growth medium were found to be significantly enriched in the set of genes containing negatively selected insertions. The data also identify fitness contributions of several uncharacterized genes, including putative transcriptional regulators and enzymes. The applicability of this high-resolution array selection in other species is discussed.

Since 1995 we have witnessed exponential growth in genome sequencing to hundreds of projects (http://wit.integratedgenomics.com/GOLD). About 30–50% of the genes in each sequenced genome are classified as “uncharacterized” (e.g. GenProtEC; ref. 1), and nearly all genes are poorly characterized with respect to their quantitative contributions to the survival of the organism under various environmental conditions. An important step toward understanding the function of a gene would be to identify the conditions under which the gene is required for optimal growth. Many genes have multiple subgenic domains (e.g., small RNA and protein motifs) and multiple alleles for each domain. We would like a highly parallel method to quantitate the effects of such alleles on selection/fitness coefficients. Several methods have been developed to do this, including genetic footprinting using transposon insertions[2] and deletion mutant analysis using molecular bar-coding[3]. Genetic footprinting using transposon insertions has been applied to several microorganisms, including *Saccharomyces cerevisiae*[2], *Haemophilus*[4], *Salmonella*[5], *Mycoplasma*[6], and *E. coli*[7]. Transposon insertions permit simultaneous mutagenesis of an entire genome and allow for high-resolution mapping of the fitness contributions of all genomic loci. However, transposon insertions have generally been detected individually using gene-specific PCR followed by electrophoretic resolution on gels. This process is time- and labor-intensive and limits the throughput of the genetic footprinting approach.

DNA microarrays have been used for genome-wide monitoring of diverse processes such as messenger RNA (mRNA) expression[8], DNA–protein interaction[9,10], and changes in DNA copy number[11]. Here we apply DNA microarrays to the detection, quantitation, and analysis of transposon insertions on a genome-wide scale. We evaluate the method by studying the quantitative growth phenotypes of a set of well-characterized *E. coli* genes under well-characterized growth conditions. We demonstrate that the method produces highly reproducible results that correlate with the expected behavior of insertions in these genes. We have also identified condition-specific, quantitative growth phenotypes for several unknown genes, including putative transcriptional regulators and enzymes. By combining the sensitivity of competitive growth assays with the high throughput of transposon mutagenesis and microarray readout, our approach allows the rapid identification of condition-specific, quantitative growth phenotypes on a genome-wide scale.

Results and discussion

Strategy for genetic footprinting using microarrays. The insertion element used to generate the transposon insertion library is a derivative of a broad host range transposable element described in Alexeyev and Shokolenko[12]. Figure 1A displays the salient features of the plasmid (pJA1) containing the insertion element. The transposable element contains a kanamycin resistance cassette flanked by IS10 inverted repeat sequences. The insertion element is carried on a suicide vector containing the R6K mutant origin of replication, thereby preventing the plasmid from replicating in wild-type *E. coli*. The vector also contains the *tnp* gene, which encodes the Tn10 transposase. The *tnp* allele on the plasmid encodes a mutant transposase that has a 100-fold lower frequency of insertion at hot spots[13]. The kanamycin resistance cassette and the suicidal properties of the vector allow for selection of chromosomal insertions, while the mutant transposase enables more random coverage of the genome. We have cloned a T7 promoter facing outward from the transposon. This T7 promoter is used in the amplification and labeling protocol described below.

Using pJA1, we generated an *E. coli* transposon insertion library on LB medium that was estimated to contain clones representing ~10^5 independent insertions. This insertion library was subjected to a competitive growth selection in M9 minimal medium containing 2% glucose. The library was grown in a chemostat and maintained constantly in log phase, by dilution, for ~30 generations. Subsequently the cells were harvested and genomic DNA was isolated. We also isolated genomic DNA from the initial unselected insertion library and from the insertion library selected for 30 generations

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115. [2]Current address: Department of Chemical Engineering, University of Delaware, Newark, DE 19716. [3]Department of Molecular Biology, Princeton University, Princeton, NJ 08544. *Corresponding author (church@arep.med.harvard.edu).

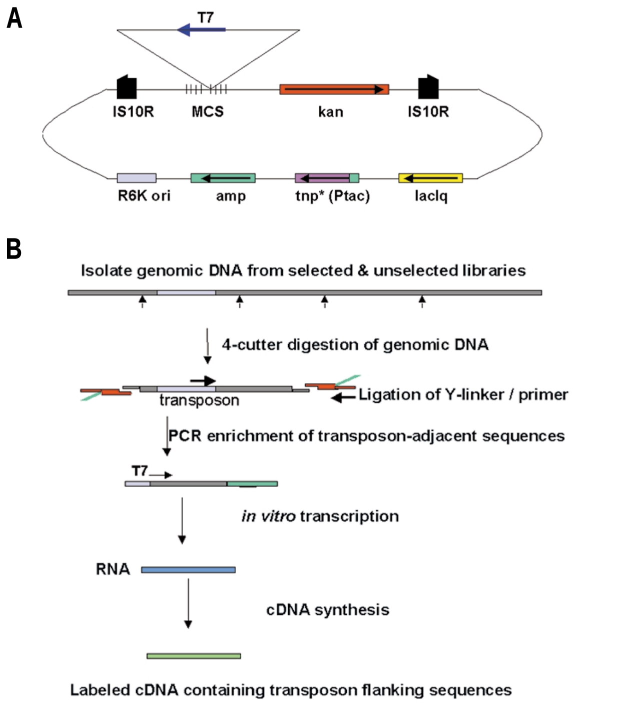
A
T7
IS10R
MCS
kan
IS10R
R6K ori
amp
tnp* (Ptac)
laclq
B
Isolate genomic DNA from selected & unselected libraries
4-cutter digestion of genomic DNA
transposon
Ligation of Y-linker / primer
PCR enrichment of transposon-adjacent sequences
T7
in vitro transcription
RNA
cDNA synthesis
Labeled cDNA containing transposon flanking sequences

Figure 1. Schematic diagram of experimental design. (A) Schematic diagram of the transposable element and the suicide vector. Based on the figure provided in Alexeyev and Shokolenko[12]. (B) Schematic diagram of the protocol for amplification and labeling of transposon-containing genomic DNA fragments.

on LB medium (see Experimental Protocol).

The genomic DNA samples were amplified and labeled according to the protocol shown in Figure 1B. Genomic DNA was digested with *Hin*p1, generating DNA fragments of an average size of 256 base pairs with 5′ or 3′ overhangs. These fragments were then ligated with a Y-shaped, partially double-stranded linker sequence[14]. The ligated fragments were amplified by PCR using a transposon-specific primer and a linker-specific primer. The Y-shaped linker enabled specific amplification of transposon-containing fragments. The PCR products were then used as templates for transcription by T7 RNA polymerase. The polymerase transcribes into the genomic DNA flanking the transposon insertion, thereby marking the location of each insertion. The RNA from the *in vitro* transcription reaction was reverse-transcribed in the presence of modified deoxynucleotide triphosphates (dNTPs) to generate fluorescently labeled complementary DNA (cDNA). The labeled cDNA was hybridized to microarrays to determine the presence of insertions within or near specific *E. coli* genes.

To assess the validity of our approach, we initially limited our microarray analysis to genes likely to be required for growth in M9 minimal medium and to other well-characterized genes known to encode components of metabolic pathways. We also included 70 uncharacterized genes primarily consisting of putative transcription factors[15] and genes conserved across several bacterial species[16]. The probes on the microarray were on average 350 base pairs in length and represent the most unique region of each open reading frame (ORF) (see Experimental Protocol). We used one probe for every 1.5 kilobases of each ORF. Therefore, longer genes are represented by multiple, nonoverlapping probes. This type of probe design not only reduces cross-hybridization problems but also potentially permits the detection of differential fitness effects of transposon insertions in different segments of a given ORF. In total, our *E. coli* microarray consists of 860 probes representing 680 genes.

To test the reproducibility of our approach, we replicated the entire experiment. Figure 2 shows a log plot of the intensities of the two insertion libraries selected for 30 generations in M9 minimal medium. The plot indicates that the competitive growth selection and the amplification/labeling of transposon-containing fragments is highly reproducible, with a correlation coefficient of 0.98.

Functional category analysis. We compared DNA from the initial unselected transposon library and from the M9-selected library to identify transposon insertions that are deleterious for growth on M9 minimal medium. The labeled DNA fragments derived from the final selected library were hybridized to a microarray along with random-primed end-labeled genomic DNA from wild-type *E. coli* as a control. For comparison, DNA from the unselected transposon insertion library and control *E. coli* genomic DNA were hybridized to a separate microarray. Intensities from the two microarrays were normalized according to the signal from the *E. coli* genomic DNA control, and fold change was calculated.

In the initial unselected library we detected transposon insertions in 710 out of 860 probes represented on the microarray. Analysis of replicate experiments showed the maximum variance to be 1.8-fold. Calculations based on an exponential decay equation indicated that a 2% growth defect would result in an approximately twofold decrease in population after 30 generations of competitive growth. Following the 30-generation selection on minimal medium, 240 of the 710 genes were reduced twofold or more, suggesting that insertions in these loci conferred a growth defect of at least 2%. Insertions in 70 genes were negatively selected following the 30-generation selection on LB medium. The two selection experiments had 62 negatively selected genes in common.

We would expect that insertions in genes required to synthesize essential biomass components not provided in the medium would be deleterious for growth in that medium. For example, genes involved in the biosynthesis of amino acids should be critical for growth in glucose M9 minimal medium, which lacks amino acids. The 10 genes with the largest fold decrease after M9 selection are shown in Table 1. As expected, the most deleterious insertions are in genes involved in the biosynthesis of amino acids, nucleotides, lipopolysaccharides, and polyamines.

We systematically analyzed the selection pattern of the insertions with respect to the genes' functional classification. To avoid insertion mutants that cause a general slow-growth phenotype irrespective of the

Table 1. *Escherichia coli* genes exhibiting largest fold decrease in signal

Gene[a]	Fold decrease[b]	Functional subcategory	Functional category
rfaC	140	Lipopolysaccharide	Macromolecule synthesis, modification
speF	125	Polyamine biosynthesis	Central intermediary metabolism
metB	114	Methionine	Amino acid biosynthesis
cysK	106	Cysteine	Amino acid biosynthesis
cydD	69	ABC superfamily (membrane)	Transport/binding proteins
gltA	64	TCA cycle	Energy metabolism, carbon
iciA	64	DNA replication, repair	Macromolecule synthesis, modification
aroA	63	Chorismate	Amino acid biosynthesis
purN	58	Purine ribonucleotide biosynthesis	Nucleotide biosynthesis
xylB	58	Carbon compounds	Degradation of small molecules

[a] Indicates the gene containing the insertion.
[b] Indicates the fold change derived from the ratio of the competitively selected library to the initial library.
[c] The complete list of genes analyzed is included in the supplemental data (http://arep.med.harvard.edu/).

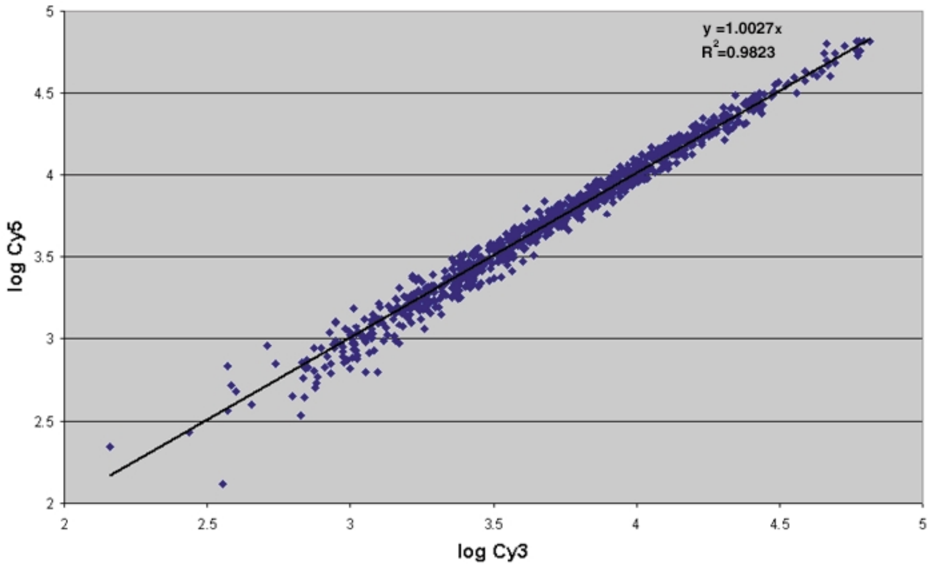

Figure 2. Log plot of the intensities from the two independent experiments of competitively selected libraries.

growth conditions, we focused on genes that were negatively selected in only one of the two growth conditions (LB and minimal medium). We used the GenProtEC database[1], which contains the functional classification of all known genes and predicted ORFs in *E. coli*. Insertions that were reduced at least twofold following the competitive growth selection were called negatively selected. This analysis (Table 2) indicated that the functional categories with the largest proportion of negatively selected genes were related to the biosynthesis of biomass precursors (i.e., amino acids, fatty acids, and nucleotides). As mentioned above, the genes analyzed in this study were biased toward genes required for growth on minimal medium, and ~40% of the insertions detected in these genes were negatively selected. As a result of this bias, the statistical significance of any enrichment of negatively selected genes within a functional category will be underestimated. The results in Table 2 indicated that insertions in amino acid and nucleotide biosynthetic genes were enriched in the set of negatively selected insertions and that this enrichment is statistically significant. Of the 60 insertion-containing genes that are known to encode components of amino acid–biosynthetic pathways (KEGG metabolic database[17]), 42 were negatively selected in the competitive growth selection on minimal medium. In contrast, insertions in genes involved in the transport and binding of molecules and in genes of unknown function are underrepresented. In the genes negatively selected on LB, none of the functional categories are statistically enriched or underrepresented (see Supplementary Table 1 in the Web Extras page of *Nature Biotechnology* Online).

Of the 70 uncharacterized genes represented on the microarray, insertions were detected in 59 of them (http://arep.med.harvard.edu/VB_supp/). Insertions in several of these genes cause condition-specific growth defects. These genes are listed in Table 3A and B. Apart from the putative transcriptional regulators, genes required for growth on glucose minimal medium include a putative peptidase, aldolase, and deaminase (Table 3A). Identification of condition-specific growth defects for these uncharacterized genes should facilitate subsequent analysis of their function.

Comparison with expression data. We hypothesized that genes highly induced on minimal medium are required for optimal growth on minimal medium. To test whether our data supported this hypothesis, we compared the transcriptional profiles of wild-type *E. coli* grown on LB and on glucose minimal medium. The genes analyzed in our competitive growth experiment were then ranked by fold induction on minimal medium. The transposon insertion results for these genes are summarized in Table 4. The genes are organized into overlapping groups, ranging from the top 10 most highly induced to the total number of genes used in this comparison. As seen in Table 4, the genes displaying the greatest degree of fold induction are enriched for genes negatively selected in the footprinting assay. This enrichment is statistically significant for genes induced greater than fivefold on minimal medium. These results are consistent with the expectation that genes induced under specific growth conditions are likely to be required for growth under those conditions.

Comparison of footprinting results with FBA model predictions. As the majority of the genes analyzed in this study are included in the flux balance analysis (FBA) model developed for *E. coli*[18], we compared the FBA results with the results from the transposon insertion library. The flux balance model is based on the catalytic functions of enzymes in metabolic pathways. It uses a stoichiometric matrix to study the effects of gene deletions and various nutrient conditions on the metabolic capabilities of the cell. Several FBA predictions of the phenotypic behavior of *E. coli* mutants are consistent with previous experimental observations, and some have recently been experimentally confirmed[19]. We compared the growth defects caused by transposon insertions in 488 metabolic genes with the FBA growth

Table 2. Functional category analysis of *Escherichia coli* genes[a]

Functional category	M9 selection			LB selection		
	Negatively selected[b]	Percentage[c]	E < 0.05[d]	Negatively selected[b]	Percentage[c]	E<0.05[d]
Fatty acid biosynthesis	4	67	No	1	17	No
Nucleotide biosynthesis	15	58	Yes	6	23	No
Amino acid biosynthesis	42	56	Yes	5	7	No
Central intermediary metabolism	29	47	No	3	5	No
Cofactor biosynthesis	19	41	No	1	2	No
Energy metabolism	22	31	No	7	10	No
Degradation of small molecules	16	33	No	5	10	No
Transport and binding	30	28	No	13	12	No
Unknown function	6	10	Yes	10	17	No

[a]Classification of *E. coli* genes is taken from GenProtEC[1]. We have only looked at the categories represented on our microarray. Genes involved in ribonucleotide metabolism were included in the nucleotide biosynthesis category.
[b]Number of genes (within that category) containing insertions that were negatively selected on glucose M9 or LB medium.
[c]Percentage of insertion containing genes within each category that were negatively selected.
[d]This column indicates whether or not the distribution of negatively selected genes within a given category is statistically significant.

Table 3. Uncharacterized *Escherichia coli* genes required for growth on M9 glucose and on LB medium

A. Growth on M9 glucose

Gene[a]	M9 fold decrease[b]	LB fold increase[b]	Predicted function[c]
b2385	33	1	Putative peptidase
b2738	20	2	Putative fuculose phosphate aldolase
ybiH	6	1	Putative transcriptional repressor TetR family
ygaA	5	1	Putative transcriptional regulator EBP family
rtcR	4	4	Regulator of RNA 3′-terminal phosphate cyclase
yiaJ	4	1	Transcriptional repressor IclR family
yicP	2	1	Putative adenine deaminase

B. Growth on LB medium

Gene[a]	M9 fold increase[b]	LB fold decrease[b]	Predicted function[c]
feaR	3	174	Transcriptional activator of 2-phenylethylamine catabolism
ybeF	1	8	Putative transcriptional regulator LysR family
ydiB	1	8	Putative shikimate 5-dehydrogenase
ygjH	1	5	Putative tRNA synthetase
garK	1	5	Glycerate kinase I
yhiF	2	4	Putative transcriptional regulator LuxR family
yjdA	1	4	Putative vimentin
idnR	1	3	Transcriptional activator of L-idonate metabolism
yjiE	1	3	Putative transcriptional regulator LysR family
yjjQ	1	2	Putative transcriptional regulator LuxR family

[a] Assigned gene name.
[b] Selection behavior in M9 and LB, respectively.
[c] Predicted function of each gene, taken from GenProtEC[1].

rate predictions for knockouts of the same genes. Flux balance analysis predicts that 143 of the 488 genes are essential for growth on minimal medium; our data indicate that 80 of these 143 genes are deleterious for growth on glucose minimal medium (Table 5). Among the 46 gene deletions predicted to reduce growth rate, 24 contain insertions that are negatively selected. Similarly, for 180 out of 299 genes predicted to be nonessential for growth on minimal medium with glucose, our results indicate that insertions in these genes are unaffected or are positively selected in the competitive growth experiments (Table 5). Evaluation of this distribution by the χ^2 test indicates that it is statistically significant.

This result confirms that genes predicted to be essential are enriched for genes containing negatively selected insertions, and that negatively selected insertions are underrepresented in the genes predicted to be nonessential for growth on glucose minimal medium. Despite the statistically significant correlation between the FBA predictions and our genetic footprinting data, there are several outliers. There are several possible explanations for these interesting discordances. One possibility is that the diffusion of metabolites from neighboring cells during the competitive growth selection may enable a mutant cell to overcome the inability to synthesize the metabolite. Cross-feeding by metabolite or enzyme diffusion is known to occur in mixed cultures and even in petri plate colony phenotypes. Comparison of the same growth selection carried out at several different culture dilutions may allow identification of mutants affected by this phenomenon.

Another possibility is that redundant genes or pathways not incorporated into the FBA model can "substitute" for genes predicted to be essential. A third possibility is that a transposon insertion does not disrupt or only partially disrupts gene function and hence is not negatively selected as predicted by the FBA model. It is also likely that transposon insertions in a nonessential gene of an operon could cause polar effects by disrupting the expression of the downstream essential genes in the operon. The thrLABC operon is an example of such an operon: *thrA*, which encodes a predicted, nonessential enzyme, lies upstream of *thrB* and *thrC*, which encode enzymes essential for threonine and glycine biosynthesis. Hence, some of the insertions in *thrA* were probably negatively selected as a result of disruption of *thrB* and *thrC* (see Supplementary Table 2 in the Web Extras page of *Nature Biotechnology* Online). A table listing all the genes and their selection behavior as well as their position within operons is included in the supplemental data (http://arep.med.harvard.edu/VB_supp/). Incorporating a strong, outward-facing promoter and translation signals in the insertion element can minimize the polar effects of transposon insertions.

We have developed a high-throughput method that enables the genome-wide analysis of quantitative growth phenotypes using transposon mutagenesis and microarrays. We evaluated this method by analyzing the condition-specific growth phenotype of a large set of *E. coli* genes under well-characterized growth conditions. For most of the genes analyzed, our results are consistent with the expected phenotypes based on a priori knowledge of gene function. We have also identified condition-specific growth phenotypes for several uncharacterized genes (Table 3A,B); some of these genes are conserved across various bacterial species[16] and are predicted to have interesting enzymatic activities (Table 3A,B). These results should contribute to further characterization of gene function. For example, our data on putative transcriptional regulators may facilitate the design of experiments to identify the genes they regulate.

The advantage of our method is that the mutagenesis, selection, and quantitation of individual mutants are done in parallel. Information about biological function can be inferred by quantitating the fitness contributions of insertion mutants under various selective growth conditions. The extent of information obtained will depend on the specificity of the selection conditions. Moreover, large data sets of selection experiments can be subjected to computational analysis such as clustering. Genes that display similar growth phenotype across various selections are likely to be involved in the same function or path-

Table 4. Growth selection behavior of *Escherichia coli* genes induced on glucose M9 minimal medium[a]

Genes induced on M9	Extent of fold induction	Negatively selected[b]	Percentage negatively selected[c]	*P* < 0.05[d]
Top 10	>68	7	70	Yes
Top 25	>35	17	68	Yes
Top 50	>10	34	68	Yes
Top 100	>5	54	54	Yes
Top 200	>2.6	91	46	No
Top 400	> 0.7	184	46	No
All	NA	225	46	NA

[a] The genes were ranked by fold induction on glucose M9 minimal medium and are listed in overlapping sets based on their ranking.
[b] Number of genes within each group that contain negatively selected insertions.
[c] Percentage of genes that were negatively selected in the footprinting experiment.
[d] This column indicates whether or not the enrichment of negatively selected insertions within each group is statistically significant. A *P* value < 0.05 indicates with 95% statistical significance the enrichment of negatively selected genes. *P* values were calculated using the hypergeometric distribution as described in Tavazoie *et al*[29]. NA, not applicable.

Table 5. Comparison of genetic footprinting data with FBA model predictions

Predictions from model	Number of genes within prediction class	Negatively selected[a]	Not negatively selected[b]
Essential	143	80	63
Reduced growth rate	46	24	22
Nonessential	299	119	180

[a]The number of genes within each class that contain negatively selected insertions. [b]The number of insertion containing genes within each class that were not negatively selected. The numbers in the last two columns were used to compute the χ^2 number and compute the P value. P value from $\chi^2 = 0.0039$.

way. The method can be easily adapted to any organism amenable to insertional mutagenesis and competitive growth selection, including large animal populations like *Caenorhabditis elegans*[20]. Such studies in diploid organisms could reveal growth defects caused by heterozygous genes. Previous studies using the bar-coding approach have shown that loss-of-function alleles of a heterozygous gene do confer a selective disadvantage in diploids[21]. Insertion-based mutagenesis can also be carried out in mammalian stem cells[22,23] and cancer cells.

Experimental protocol

The insertion element used to create the transposon insertion library was derived from pBSL181 (ref. 12). A T7 promoter created by annealing two oligos T7-Tn10 (5′-CTAGCACCTAACCGCTAGCACGTATACGACTCACTATAGGG AGGCGGATTCCTGAACGGTAGCATCTTGACGACGC-3′) and T7-Tn10-AS (5′-GTCGTCAAGATGCTACCGTTCAGGAATCCGCCTCCCTA TAGTGAGTCGTATTACGTGGCTAGCGTTAGGTG-3′) was cloned into the *Bam*HI site within the insertion element. The kanamycin resistance cassette within the insertion element also encodes –35 and –10 transcriptional signals.

The "suicide vector" bearing the insertion element was transformed into competent BL21 *E. coli* B (F^- dcm ompT hsdS gal). Transformed cells were incubated in LB containing 10 µg/ml kanamycin and 20 µM of isopropyl-β-D-thiogalactoside (IPTG). The library was amplified in LB containing kanamycin and frozen in 15% glycerol in aliquots of 2×10^9 cells. On the basis of colony-counting assays, 10^5 independent insertions were obtained.

An aliquot of 2×10^9 cells of the transposon insertion library was inoculated into 1 L of M9 minimal medium containing 10 µg/ml of kanamycin and 2% glucose. The recipe for M9 minimal medium is described by Sambrook *et al*[24]. The cells were cultured in a 2 L chemostat (New Brunswick Scientific, Edison, NJ) at 37°C with agitation at 300 r.p.m. and aeration at 10 L/h. At $OD_{600} = 0.7$, the input and output pumps were turned on. The flow rate for adding fresh medium was set at 900 ml/h. This dilution rate is slightly faster than the doubling time of the culture (50 min). The chemostat was maintained at a constant volume of 1 L by pumping out excess volume, as required.

The selection on LB was done by serial dilution of 2×10^9 cells into 1 L of LB containing 10 µg/ml of kanamycin after every seven doublings (~3.5 h). The selection was continued up to 30 doublings, based on OD_{600} . Following the selection, samples 2×10^9 cells were harvested and used to isolate genomic DNA.

Isolation and purification of *E. coli* genomic DNA was carried out as described by Ausubel *et al*[25]. A 10 µg sample of genomic DNA isolated from the insertion library was digested with *Hin*P1 (New England Biolabs, Beverly, MA), 3 µl of 20 U/µl for 3 h at 37°C. Use of alternate restriction enzymes with similar cut site frequencies will cause minor differences in the transposons detected. The digested DNA was precipitated and resuspended in 25 µl of TE. A ligation was carried out using 2 µg of digested genomic DNA and 4 µl of 4 pmol/µl of the Y-linker (described in Tavazoie and Church[14]). A 0.75 µl aliquot of the ligation was used as template for a PCR reaction containing 1 µl of 20 µM newT7 primer (5′-GCACCTAACCGCTAGCACGTAATACGACTC-3′), 1 µl of 20 µM YCG linker primer 13, 2.5 µl of 10× PCR buffer (Sigma, St. Louis, MO), 2.5 µl of 2 mM dNTP mix, and 0.5 µl of 5 U/µl *Taq* polymerase. The PCR mix was initially heated to 95°C for 1 min and then cycled 25 times between 94°C (20 s) and 72.5°C (75 s).

1 µg of PCR product was used as template in an *in vitro* transcription reaction using the MEGASCRIPT kit from Ambion (Austin, TX). The RNA yields ranged from 40 to 80 µg.

Four micrograms of RNA and 10 µg of Random primers (Gibco BRL, Grand Island, NY) were used to generate labeled cDNA as described in Marton *et al*[26]. Detailed protocols are also listed at http://www.microarrays.org/protocols.html. The reference control sample of Cy3-labeled genomic DNA was prepared by mixing Cy3-end-labeled random primers and *E. coli* genomic DNA sheared by sonication to an average size of 500 bp along with Klenow (exo-) DNA polymerase (Gibco BRL). The reaction was incubated at 37°C for 2 h. The excess random primer was removed by passing the sample through a spin column (Chromaspin-30; Gibco BRL).

Array design. We have designed *E. coli* microarrays to specifically target the unique region of each *E. coli* ORF. We identified the 350 bp segment of each ORF with the lowest expectation value when BLAST[27] was used to align the 350 bp segment against the rest of the genome. If the lowest expectation value was identified for multiple 350 bp segments of the same genes, we selected the 5′-most 350 region. We selected one probe per 1,500 bp of ORF sequence; therefore, a 1,700 bp gene will receive two nonoverlapping 350 bp probes on our microarray. After all the probes were selected, primer 3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi/) was used to design primers to amplify the corresponding region of the genome. Because primer 3 was used to design the primers, each probe has a length that is slightly different from 350 bp. The synthesis of the probes was carried out as described in Livesey *et al*[28]. Production of spotted microarrays and slide post processing was carried out as described in Livesey *et al*[28]. The DNA at each spot was checked by staining the slide with Sybr Green 1. The spots where 50% of the pixels were 2 standard deviations above the background were considered suitable for use as probes. By this criterion we had 860 probes available for analysis.

The hybridization and washing protocols are described in detail at http://www.microarrays.org/protocols.html. In the experiment to test the reproducibility of the method, the DNA samples from the two selected libraries were labeled with Cy3 and Cy5, respectively, and hybridized to the same microarray. In all subsequent experiments, the cDNAs derived from transposon insertion libraries were labeled with Cy5. Samples from each insertion library (initial and selected) were hybridized to separate microarrays. Before hybridization, each of these Cy5-labeled samples was mixed with Cy3-labeled reference control derived from 2 µg of genomic DNA. The slides were scanned using the Scanarray 5000, and the data were analyzed using the GenePix 3.0 software (Axon Instruments, Foster City, CA).

Data analysis. The spots were called present only if 50% of the pixels were 2 standard deviations above the background. By this criterion, transposon insertions were detected by 710 out of the 860 probes. For comparison between the initial and selected library, the Cy5 intensity of each spot on the two slides was normalized to the Cy3 (genomic DNA control) intensities at the respective spots. The normalized Cy5 intensities were used to calculate the ratios (selected library:initial library) and derive the fold change.

Functional category analysis. The GenProtEC database[1] was used to extract the functional annotations for all the *E. coli* genes. For this analysis, genes with multiple probes were counted as negatively selected, if at least one of the probes was reduced by twofold or more. The *P* values were derived using the hypergeometric distribution analysis as described in Tavazoie *et al*[29].

Expression profiling. BL21 cells from an overnight culture were inoculated into 100 ml of LB and 100 ml of glucose M9 minimal medium. The cells were harvested in mid-log phase at $OD_{600} = 0.6$. The RNA isolation, labeling, and hybridization were done as described in http://www.microarrays.org/protocols.html.

Detailed information about the genes in this study is available at http://arep.med.harvard.edu/VB_supp/ and upon request from the authors.

Note: Supplementary information can be found on the Nature Biotechnology *website in Web Extras (http://biotech.nature.com/web_extras).*

Acknowledgments
We would like to thank Martin Steffen and Xiaohua Huang for their tireless efforts to produce the E. coli *microarrays, Barak Cohen and Tzachi Pilpel for helpful discussion, and members of the Church lab for their support and encouragement. We would also like to thank the referees for their insightful comments. This work was supported by grants from the US Department of Energy and the National Science Foundation.*

Received 20 March 2001; accepted 17 August 2001

1. Riley, M. Genes and proteins of *Escherichia coli* K-12. *Nucleic Acids Res.* **26**, 54 (1998).
2. Smith, V., Chou, K.N., Lashkari, D., Botstein, D. & Brown, P.O. Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074 (1996).
3. Winzeler, E.A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
4. Akerley, B.J. *et al.* Systematic identification of essential genes by *in vitro* mariner mutagenesis. *Proc. Natl. Acad. Sci. USA* **95**, 8927–8932 (1998).
5. Wong, S.M. & Mekalanos, J.J. Genetic footprinting with mariner-based transposition in *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. USA* **97**, 10191–10196 (2000).
6. Hutchison, C.A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169 (1999).
7. Hare, R.S. *et al.* Genetic footprinting in bacteria. *J. Bacteriol.* **183**, 1694–1706 (2001).
8. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
9. Iyer, V.R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
10. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
11. Pollack, J.R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**, 41–46 (1999).
12. Alexeyev, M.F. & Shokolenko, I.N. Mini-Tn10 transposon derivatives for insertion mutagenesis and gene delivery into the chromosome of gram-negative bacteria. *Gene* **160**, 59–62 (1995).
13. Bender, J. & Kleckner, N. IS10 transposase mutations that specifically alter target site recognition. *EMBO J.* **11**, 741–750 (1992).
14. Tavazoie, S. & Church, G.M. Quantitative whole-genome analysis of DNA–protein interactions by *in vivo* methylase protection in *E. coli*. *Nat. Biotechnol.* **16**, 566–571 (1998).
15. Robison, K., McGuire, A.M. & Church, G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284**, 241–254 (1998).
16. Arigoni, F. *et al.* A genome-based approach for the identification of essential bacterial genes. *Nat. Biotechnol.* **16**, 851–856 (1998).
17. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
18. Edwards, J.S. & Palsson, B.O. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA* **97**, 5528–5533 (2000).
19. Edwards, J.S., Ibarra, R.U. & Palsson, B.O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**, 125–130 (2001).
20. Zwaal, R.R., Broeks, A., van Meurs, J., Groenen, J.T. & Plasterk, R.H. Target-selected gene inactivation in *Caenorhabditis elegans* by using a frozen transposon insertion mutant bank. *Proc. Natl. Acad. Sci. USA* **90**, 7431–7435 (1993).
21. Giaever, G. *et al.* Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.* **21**, 278–283 (1999).
22. Hicks, G.G. *et al.* Functional genomics in mice by tagged sequence mutagenesis. *Nat. Genet.* **16**, 338–344 (1997).
23. Zambrowicz, B.P. *et al.* Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature* **392**, 608–611 (1998).
24. Sambrook, J., Fritsch, E.F. & Maniatis, T. *Molecular cloning: a laboratory manual.* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY; 1989).
25. Ausubel, F.M. *et al. Current protocols in molecular biology.* (Wiley Interscience, New York, NY; 1994).
26. Marton, M.J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat. Med.* **4**, 1293–1301 (1998).
27. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
28. Livesey, F.J., Furukawa, T., Steffen, M.A., Church, G.M. & Cepko, C.L. Microarray analysis of the transcriptional network controlled by the photoreceptor homeobox gene *Crx*. *Curr. Biol.* **10**, 301–310 (2000).
29. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).