Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 322 (2004) 347–354
BBRC
www.elsevier.com/locate/ybbrc

Genome-scale identification of conditionally essential genes in *E. coli* by DNA microarrays

Xin Tong[a], John W. Campbell[b], Gábor Balázsi[a], Krin A. Kay[a], Barry L. Wanner[c], Svetlana Y. Gerdes[b], Zoltán N. Oltvai[a,*]

[a] *Department of Pathology, Northwestern University, Chicago, IL 60611, USA*
[b] *Integrated Genomics Inc., Chicago, IL 60612, USA*
[c] *Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA*

Received 9 July 2004
Available online 6 August 2004

Abstract

Identifying the genes required for the growth or viability of an organism under a given condition is an important step toward understanding the roles these genes play in the physiology of the organism. Currently, the combination of global transposon mutagenesis with PCR-based mapping of transposon insertion sites is the most common method for determining conditional gene essentiality. In order to accelerate the detection of essential gene products, here we test the utility and reliability of a DNA microarray technology-based method for the identification of conditionally essential genes of the bacterium, *Escherichia coli*, grown in rich medium under aerobic or anaerobic growth conditions using two different DNA microarray platforms. Identification and experimental verification of five hypothetical *E. coli* genes essential for anaerobic growth directly demonstrated the utility of the method. However, the two different DNA microarray platforms yielded largely non-overlapping results after a two standard deviations cutoff and were subjected to high false positive background levels. Thus, further methodological improvements are needed prior to the use of DNA microarrays to reliably identify conditionally essential genes on genome-scale.

Keywords: Essential gene; Transposon mutagenesis; DNA microarray; *E. coli*; Aerobic and anaerobic conditions

With the rapidly increasing numbers of fully sequenced genomes, there is an increasing need for a comprehensive understanding of the roles of the thousands of genes that make each organism unique. However, the function of more than 35% of the genes in even the most closely studied model organisms, such as *Escherichia coli*, remains poorly understood, while their individual analysis represents a substantial challenge. An important step toward understanding the function of uncharacterized genes is to identify which gene plays an essential role in cell growth and survival, and under what conditions such genes are essential [1].

In recent years, a number of new experimental approaches have been developed to determine genome-wide gene essentiality. These include systematic knockouts in *Saccharomyces cerevisiae* [2,3] and *Caenorhabditis elegans* [4], RNA interference in *C. elegans* [5], and genetic footprinting in several microorganisms [1,6,7]. Standard genetic footprinting consists of three main steps that include random transposon mutagenesis of a large number of cells, selective outgrowth of the mutagenized population, and determination of transposon insertion sites in the genome using PCR and gel electrophoresis [8,9]. Although this is a rather precise, semi-automatic analysis, transposon insertions are detected individually using position-specific PCR primers followed by electrophoretic resolution of the amplification products on agarose gels. Thus, the readout phase of

* Corresponding author. Fax: +1 312 503 8240.
E-mail address: zno008@northwestern.edu (Z.N. Oltvai).

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.07.110

this process is time- and labor-intensive, which substantially limits the utility of the approach.

DNA microarrays have been used for genome-wide monitoring of gene expression [10], DNA copy-number changes [11], and the detection of DNA–protein interaction [12]. Recently, several attempts to accelerate the transposon insert detection phase of genetic footprinting by utilizing DNA microarrays for detecting transposon insertions have been reported [13–15]. The general strategy involves digesting transposon-mutagenized genomic DNA and adding a linker to the end of the DNA fragments to serve as a priming site for subsequent PCR amplification. DNA sequences within the transposon serve as the corresponding matching primer sites. Mutagenized genomic DNAs from different growth conditions are fluorescently labeled and compared by hybridization to DNA microarrays.

In this study, we report the results of a genetic footprinting approach to identify *E. coli* MG1655 genes that are selectively essential under aerobic vs. anaerobic growth conditions by using two different custom-built *E. coli* DNA microarrays. Our results demonstrate that although this approach provides high throughput readout of putative essential and non-essential genes, the assortment of essentiality is affected by the large number of false positive data points. However, the procedure did allow us to hypothesize that the genes for a number of hypothetical proteins were essential for anaerobic growth, some of which were confirmed experimentally. Thus, although the method lacks precision it can be useful for discovering conditionally essential genes.

Materials and methods

Bacterial strain, growth conditions, and transposon mutagenesis. *Escherichia coli* strain MG1655 (F^-, λ^-, *ilv*G, *rfb*50, *rph*1) [16] was used throughout this work. Generation of the transposon mutant library and outgrowth of the mutagenized population were performed, as described [17]. Briefly, a transposon library was constructed by incubating transposon DNA EZ::TN<KAN-2> with the hyperactive Tn*5* EZ::TN transposase [18] (Epicentre Technologies) to form a transposome complex, which was subsequently transformed to electrocompetent *E. coli* cells by electroporation. Cultures were immediately diluted with LB-based rich medium with supplements, incubated at 37 °C for 40 min, and then used to inoculate a BIOFLO 2000 fermentor (New Brunswick Scientific) containing the same medium supplemented with kanamycin (10 μg/ml). For aerobic growth, dissolved oxygen was held at 30–50% of saturation. For anaerobic growth, N_2 gas was continuously sparged into the medium throughout the fermentation. The medium and growth conditions were designed to minimize the number of genes required for cell survival. After ~20 population doublings cells were collected and genomic DNA was isolated.

Specific mutant strains of *E. coli* MG1655 with transposon inserts in individual gene were obtained from Frederick Blattner's lab (University of Wisconsin, Madison, WI; described at http://www.genome.wisc.edu). Non-stringent anaerobiosis (microaerophilia) was performed by growing cells on plates incubated in a GasPak jar under a reduced H_2 and CO_2 atmosphere. Strict anaerobiosis was obtained by growing cells in Hungate tubes (Bellco Glass) completely filled with LB agar containing 3.2 mM sodium sulfide, using Resazurin (0.2 mg/100 ml) as a redox indicator [19].

Construction of E. coli MG1655 microarrays. We used two different microarrays for the experiments. For the construction of the first set of microarrays, unique cDNA fragments corresponding to each predicted ORF of *E. coli* MG1655 were produced by PCR amplification from genomic DNA. To ensure that each primer pair would amplify the most unique 200–350 base pair region within each ORF, the PCR amplification product sequences were compared with all others in the genome by BLAST analysis [20]. Sequences producing the minimum *E* values, relative to all other ORFs in *E. coli*, were selected. Of the 4485 *E. coli* ORFs listed in the ERGO database [21], unique primers were found for 4442 ORFs (sequences available upon request). Most of the remaining ORFs represent genes that are shorter than 240 bp. Thus, each predicted ORF of *E. coli* MG1655 was represented by a fragment predicted to minimally cross-hybridize with other *E. coli* MG1655 sequences. These fragments varied in length from 200 to 350 bp, with a median length of ~300 bp.

After two rounds of PCR amplification the final products were purified on 384-well format ArrayIt PCR purification Kits (TeleChem International). PCR products were resuspended in 15 μl spotting buffer (3× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate), 1.5 M betaine) and printed in triplicate onto amino-alkylsilane coated slides (Sigma) with an OmniGrid arrayer (Gene Machines) equipped with Telechem SMP3 split pins. The DNA was cross-linked to the slides with UV light and baked in a vacuum oven at 60 °C for at least two hours. Residual salt and unbound DNA were removed by rinsing the slides with 0.5% sodium dodecyl sulfate and water. The slides were stored desiccated at RT until use.

The second set of microarrays containing full-length *E. coli* ORFs was printed as previously described [22] and was augmented to contain stable RNAs as well as various other controls, e.g., *E. coli* genomic DNA.

Generation of sample RNA. Five micrograms of mutagenized genomic DNA was partially digested with *Hin*P1 I or *Hpy*CH4 IV (New England Biolabs), 0.5 μl of 10 U/μl for 15 min at 37 °C. Fragments from both enzymes' digestion (larger than 500 bp) were purified from a 1.2% agarose gel with a QIAquick Gel Extraction Kit (Qiagen). The resulting size-selected, purified DNA fragments were ligated to a Y-shaped adaptor modified from Badarinarayana et al. [13] (CGG ACGCTACGTCCGTGTTGTCGGTCCTG and ACTACGCACCG GACGAGACGTAGCGTC) using T4 DNA ligase (Invitrogen) at 16 °C overnight.

After the Y-shaped adaptor was ligated to the transposon-mutagenized DNA fragments, the DNA was re-digested with *Pfl*M I (New England Biolabs), 1 μl of 8 U/μl for 1 h at 37 °C. A 1 μl aliquot of the digest was used as template in two separate PCRs containing 0.5 μM the adaptor primer (ACTACGCACCGGACGA) and 0.5 μM of one of the transposon-specific primers (primer 1: GTTCCGTGGCAAAG CAAAAGTTCAA or primer 2: CCGACATTATCGCGAGCCC ATTTAT), 0.2 mM dNTP mix, PCR buffer (Invitrogen), and 0.5 μl of 5 U/μl *Taq* polymerase (Invitrogen). Cycling conditions were 95 °C for 1 min; 25 cycles of 94 °C for 30 s, 55 °C for 30 s, and 72 °C for 1 min; and 72 °C for 5 min. Two microliters of the undiluted PCR was used as template for secondary PCR amplification using the same adaptor primer and one of the second, nested transposon-specific primers containing T7 promoter sequence (primer 3: GCGAAATTAATA CGACTCACTATAGGGTTCCGTGGCAAAGCAAAGTTCAA or primer 4: GCGAAATTAATACGACTCACTATAGGGCAAGACG TTTCCCGTTGAATATGGC). PCR conditions were 95 °C for 1 min; 5 cycles of 94 °C for 30 s, 72 °C for 1.5 min; 5 cycles of 94 °C for 30 s, 70 °C for 30 s, and 72 °C for 1 min; and 25 cycles of 94 °C for 30 s, 67 °C for 30 s, and 72 °C for 1 min; and 72 °C for 5 min. PCR products (200–500 bp) were purified from agarose gel and precipitated and resuspended in 15 μl TE.

One microgram of the secondary PCR product was used as template in an in vitro transcription reaction using T7 polymerase from the

MEGAscript kit (Ambion). After transcription, template DNA was removed by incubation with DNase I (Ambion), followed by phenol:chloroform extraction and isopropanol precipitation. The RNA from these reactions (about 100 μg) was stored frozen at −70 °C until use.

Synthesis of fluorescent cDNA from RNA. Twenty micrograms of RNA template and 10 μg of random hexamer primers (Invitogen) were mixed together in 14 μl RNase-free water. Primer annealing was accomplished by incubating for 10 min at 70 °C, followed by quenching at least 1 min on ice. cDNA was synthesized with SuperScript II reverse transcriptase (Invitrogen) in the presence of deoxynucleoside triphosphates (dATP, dGTP, dCTP, each at 0.5 mM, dTTP at 0.3 mM) and 0.2 mM amino-allyl dUTP (Sigma). In successive order 6 μl of 5× SuperScript II reaction buffer (Invitogen), 3 μl of 10× dNTP mix, 3 μl of 0.1 mM DTT, 1 μl RNaseOUT (40 U/μl, Invitrogen), and 3 μl of SuperScript II reverse transcriptase (200 U/μl) were added to the annealing mix. The reaction was incubated at 25 °C for 10 min, then at 42 °C for 2 h before terminated by heating at 70 °C for 10 min. The RNA template was hydrolyzed with 10 μl of 1 N NaOH and 10 μl of 0.5 M EDTA for 15 min at 65 °C. The reaction was neutralized by adding 25 μl of 1 M Tris–HCl (pH 7.4). Unincorporated amino-allyl dUTP and Tris were removed by filtration (Microcon YM-30, Millipore). cDNA was transferred into a dark tube and dried in a Speed Vac concentrator (Eppendorf). The cDNA pellet was resuspended in 9 μl of 0.1 M sodium bicarbonate buffer (pH 8.5–9), and added to a dry aliquot of Fluorolink Cy5 or Cy3 Monofunctional Dye (Amersham), then mixed and incubated for 1 h at room temperature in the dark. 4.5 μl of 4 M hydroxylamine (pH 8.5–9.0) was added for 15 min at room temperature to quench unreacted Cy5 or Cy3 dye derivatives. The labeled cDNAs were purified with a PCR purification kit (Qiagen), dried, and stored at −20 °C.

Microarray hybridization. Arrayed *E. coli* slides were prehybridized in 5× SSC, 0.1% SDS, and 0.1 mg/ml BSA for 1 h at 42 °C, rinsed with water, and dried. Labeled cDNA were resuspended in 10 μl water and mixed with 115 μl ArrayHyb Hybridization Buffer (Sigma), 2 μl blocking buffer (salmon sperm DNA (10 mg/ml)), denatured at 95–100 °C for 2 min, and hybridized to the slides. Hybridization was performed in a Gene TAC Hybridization Station (Genomic Solutions). The slides were rinsed with 1× SSC, 0.1% SDS, washed sequentially at room temperature for 5 min in 1× SSC, 0.1% SDS, and 0.1× SSC, 0.1% SDS, 1 min in 0.1× SSC, and finally rinsed with water. Slides were dried and scanned using the GenePix 4000B scanner (Axon Instruments), the data were analyzed with GenePix Pro 4.0 software (Axon Instruments).

The hybridizations of microarrays containing the full-length *E. coli* ORFs were performed, as described previously [22].

Data analysis. Each microarray experiment was performed twice, with dye swapping, and each ORF-specific probe was printed in triplicate on both microarray slides. The spot-specific log-ratio [SSLR] for each probe spot was defined as the base-10 logarithm of the fraction I_{635}/I_{532}, where I_λ represents the fluorescence intensity at wavelength λ (measured in nm). The numerator and the denominator of this fraction were defined as follows: $I_{635} = F_{635} - B_{635}$ and $I_{532} = F_{532} - B_{532}$, where F_λ and B_λ represent the median foreground and background fluorescence intensities, respectively, estimated at wavelength λ (nm). We eliminated the data corresponding to flagged spots, or to spots with less than 50% of the pixels 2 SD above both backgrounds. Next, we computationally improved the log-ratio of the microarray data to take into account the effects of the printing procedure, using MATLAB (MathWorks), as follows [23,24]. Three copies of the *E. coli* genome were printed on each slide by means of an 8-tip print head, resulting in 24 groups of probe spots. Therefore, we averaged all SSLR-s within every one of the 24 groups of probe spots on the slide and subtracted the result from all SSLR-s in that group. Next, we rescaled the resulting SSLR-s making sure that all SSLR-s within all 24 groups have the same standard deviation; we averaged all the improved SSLR-s corresponding to the same ORF-specific probe within one slide and used the resulting experiment-specific log-ratios (ESLR-s) in our further analysis. For one microarray experiment, the number of averaged SSLR-s was typically equal to 3, but varied from 1 to 6 in some cases (due to flagged entries and duplicate probe spots on the slide).

To take dye swapping into account before comparing the results of the repeated experiments, we reversed the sign of all ESLR-s in one of the experiments. At this point, we also applied a "replicate trim" as described [25] to eliminate ESLR-s which were significantly different in the replicate experiments. Shortly, we assessed the distribution of the quantity $\Delta\mathrm{ESLR} = \mathrm{ESLR}_{1,i} - \mathrm{ESLR}_{2,i}$ for all data points in common between the replicate experiments 1 and 2, and eliminated data for which ΔESLR was more than two standard deviations away from the mean. To assess the reproducibility of our results, we calculated the cross-correlation coefficient between the ESLR-s of all ORF-specific probes common in the two experiments, defined as:

$$\rho = \frac{\langle \mathrm{ESLR}_{1,i}\mathrm{ESLR}_{2,i}\rangle - \langle \mathrm{ESLR}_{1,i}\rangle\langle \mathrm{ESLR}_{2,j}\rangle}{\sigma_1\sigma_2},$$

where $\mathrm{ESLR}_{1,i}$ represents the ESLR of the gene with ORF i in experiment 1, the brackets represent averages over all ORFs, and σ_1 (σ_2) represents the standard deviation of ESLR-s in experiment 1 (2), defined as:

$$\sigma_1 = \sqrt{\frac{1}{N_c - 1}\sum_{i=1}^{N_c}\left(\mathrm{ESLR}_{1,i} - \langle \mathrm{ESLR}_{1,j}\rangle\right)^2},$$

where N_c represents the total number of gene entries in common between the two datasets.

Finally, for every ORF-specific probe present in both datasets, we averaged the two ESLR-s to obtain the Average Log-Ratio (ALR). Using the ALR-s, we created a histogram for each experiment, as shown in Fig. 2.

The data from the full-length genomic microarrays were processed similarly, except that there was only one whole genomic probe printed on each slide.

Results

Analysis of genetic footprinting in E. coli by DNA microarrays

Recently, we completed a genome-wide essentiality study of *E. coli* MG1655 grown aerobically in rich LB-based medium, in which the analysis of transposon insertion sites in the surviving pooled mutant populations was determined using a PCR-based readout method [26]. At the same time, we also generated a transposon insertion library in *E. coli* cells that were subjected to growth selection in the same medium but under anaerobic condition. Both libraries contain about 2×10^5 independent insertion mutants. We have subsequently used the two transposon-mutagenized genomic DNA pools to identify conditionally essential genes that were specific for aerobic or anaerobic growth, respectively. The general experimental approach is illustrated in Fig. 1. Briefly, after transposon mutagenesis, genomic DNA was isolated, digested by restriction enzymes, and the DNA fragments were ligated to a specific Y-linker modified from that previously described by Badarinarayana et al. [13]. The DNA fragments flanking the

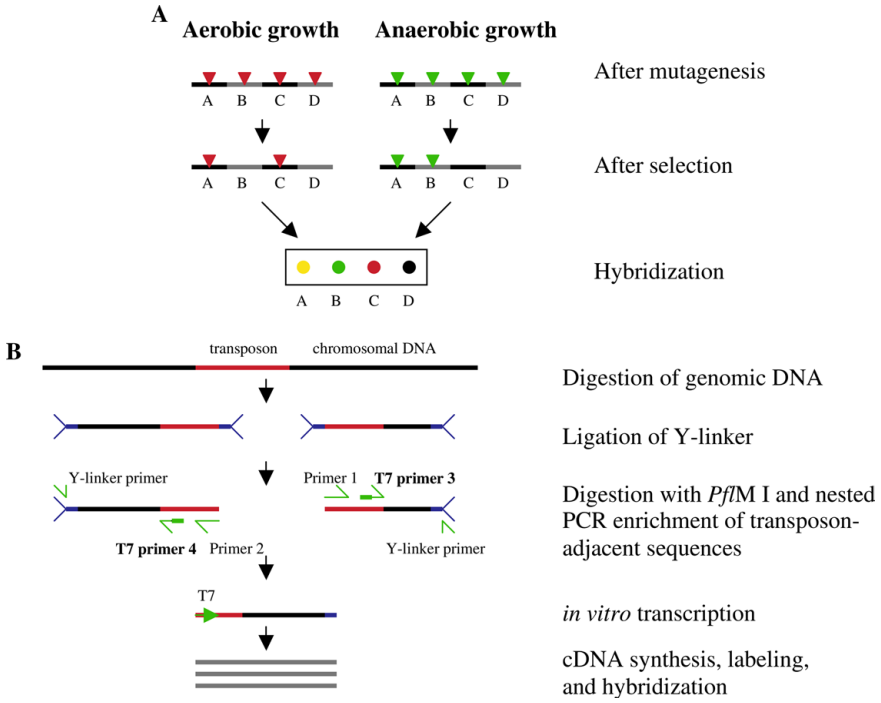

Fig. 1. Schematic diagram of the experimental protocol. (A) Immediately after the transposon mutagenesis, each mutagenized *E. coli* cell contains a single transposon insertion (triangle), and every gene of the *E. coli* genome (boxes with letters) at least one individual cell of the population is mutated. Pools of mutants are grown under aerobic or anaerobic conditions with antibiotic selection, ensuring the survival of only transposon-insert containing cells. In this example gene A is non-essential in either growth conditions (i.e., the transposon insert is tolerated); genes B and C are essential only under aerobic or anaerobic condition, respectively; and gene D is essential for both conditions. Target DNA that is complementary to chromosomal DNA flanking each transposon insertion is generated from the two pools, labeled with different fluorophores (Cy5 or Cy3), and hybridized to microarray slide. Probe A that represents gene A on the slide is hybridized by target DNA from both pools; probe B and probe C are hybridized by only targets from anaerobic and aerobic conditions, respectively; there is no hybridization on probe D because no target DNA is generated from either pools. (B) The protocol for amplification and labeling of chromosomal DNA flanked transposon insertions. See Materials and methods for details (Figures modified from [14]).

transposon insertions were enriched by PCR amplification with primers specific for the inserted transposon sequence and the Y-shaped adaptor. Subsequently, these target DNAs were labeled and hybridized to the two separate types of DNA microarrays. The reproducibility of the approach was ensured by repeating the whole experiment with the Cy3 and Cy5 labels exchanged. Figs. 2A and B depict the log ratios measured in one experiment vs. the other with correlation coefficient of 0.950 for partial cDNA microarray and 0.989 for full-length cDNA microarray, respectively, indicating that the procedure is highly reproducible.

Assessment of microarray data for conditional gene essentiality

When using the partial cDNA microarray for readout, after omitting flagged and low-intensity spots, we detected transposon insertions in 3504 of the 4442 ORF-specific probes present on the microarray. Statistical analyses of the log ratio of all detected genes in the aerobic versus anaerobic conditions revealed a standard deviation of 0.58 from the mean, indicating that there is 95.5% confidence that any log ratio is significant if the value is greater than 1.16 (two standard deviations from the mean) (Fig. 2C). Based on this criterion, the majority of genes proved dispensable under both conditions, only 73 genes (2.1%) being asserted as indispensable under anaerobic-, and 118 genes (3.4%) as essential under aerobic condition (The list of all 3504 genes and their log ratios are available in Table S2 of the Supplementary Material). Note, if a gene is essential for both growth conditions, the signal of the corresponding probe on the microarray is absent after hybridization. Thus, these genes (319 total) are not included in the data above. Due to the failure of PCR amplification from both DNA pools or the failure of hybridization, these may include genes falsely labeled as essential. However, we found many genes in this group whose function is known to be essential for growth, such as those encoding ribosomal proteins (e.g., *rplD*, *rplL*, *rplS*, *rplT*, *rpsB*, *rpsK*, *rpsM*, and *rpsQ*) or proteins involved in DNA replication (e.g., *dnaB*, *dnaN*, *dnaQ*, and *dnaT*). A list of these 319 genes is provided in Table S3 of the Supplementary Material.

A potential drawback of partial cDNA microarray analysis arises from the characteristics of this type of microarrays. Specifically, the gene probes deposited on

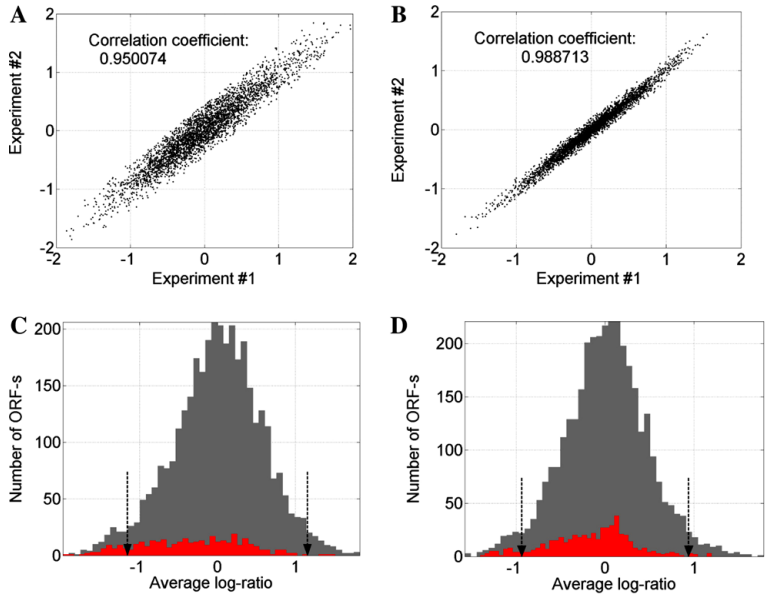

Fig. 2. Distribution of transposon insertions detected by DNA microarrys. (A,B) Plot of log ratio from two independent experiments with dye swapping using partial cDNA- (A) and full-length cDNA microarray (B). See Materials and methods for the calculation of the cross-correlation coefficient between the two experiments. (C,D) Histogram plot of the log ratio of aerobic probe to anaerobic probe for each ORF on the partial (C) and full-length (D) cDNA microarrays. Gray histogram: total microarray data; red histogram: microarray data for the genes asserted as essential under aerobic condition by PCR-based genetic footprinting [26]. The data represent the average of two replicated experiments. Significant log ratio change (log ratio = 1.16 for (C) and 0.94 for (D)) is indicated with dashed arrows. Genes negatively selected only under anaerobic condition are >1.16 (0.94), and genes negatively selected only under aerobic condition are <−1.16 (−0.94). Genes negatively selected under both conditions are not included in the analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

the microarray slides correspond to a unique ~300 bp segment of almost all predicted ORFs in *E. coli* MG1655. Although this design minimizes cross-hybridization among the various ORF sequences, it may result in false positive results when used to identify essential genes due to the fact that if the transposon insertion sites are not directly adjacent to the sequences printed on the array, there may be no hybridization with the target sequence on the microarray, even when the PCR amplification of the insert specific genes is successful.

To formally analyze the potential contribution of this disadvantage, we also used full-length cDNA microarray slides for the readout, and after data analysis were able to assort 3496 gene-specific spots on the slides for essentiality study (The names and corresponding log ratios are available in Table S4 of the Supplementary Material). Using the two standard deviations cutoff (with a cutoff value of 0.94), most genes were asserted as non-essential, only 87 genes (2.4%) appearing indispensable under anaerobic-, and 108 genes (3.0%) as essential under aerobic growth conditions (Fig. 2D) (The lists of these essential genes are in Tables S5 and S6 of the Supplementary Material).

To further interpret conditional gene essentiality, we analyzed the data in a functional context, and aerobically and anaerobically essential genes were compared by using the GenProtEC database [27], which contains functional classification of all known genes and predicted ORFs in *E. coli* (as of April, 2004). (The details of the functional categories are included in the Supplementary Material). Although more than one-third of essential genes are uncharacterized under both conditions, many genes directly involved in aerobic or anaerobic respiration are asserted as essential under the corresponding growth condition.

High ratio of false positives in both types of DNA microarray data

To determine the analytical power and overlap of the obtained data, we first compared the results of the two types of microarrays. We found that the log ratios measured in both microarrays have a correlation coefficient of 0.65, demonstrating moderate overlap between the two data sets. There were 73 genes and 118 genes asserted as essential under anaerobic- and aerobic-conditions by partial cDNA microarray, while the full-length cDNA microarray identified 87 genes and 108 genes were asserted as essential under anaerobic- and aerobic-conditions, respectively. Of these essential genes, only 19 anaerobic-, and 30 aerobic genes were commonly identified as essential by both types of microarray data (Fig. 3).

In a previous gene essentiality study by genetic footprinting and PCR-based readout, of the 4442 ORFs in *E. coli* we assessed the essentiality of 3746 genes under aerobic growth condition [26]. Within that set 620 genes were identified as essential and 3126 genes as dispens-

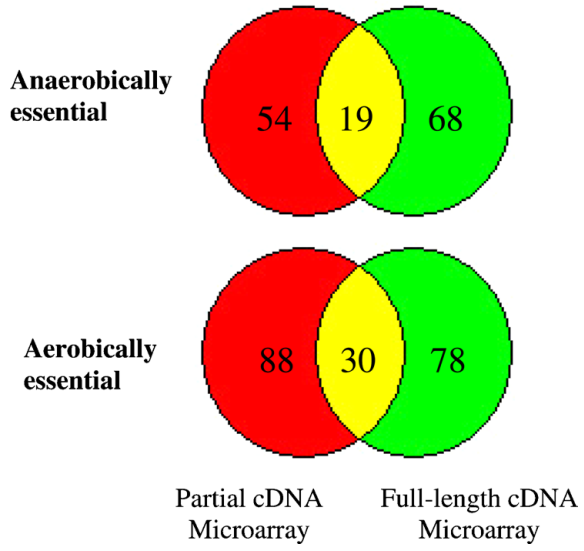

Fig. 3. Gene essentiality compared between microarray data. The Venn diagrams of the results obtained from the two types of DNA microarrays are shown.

able for cell survival (The detailed essentiality results are available at the authors' website at: www.oltvailab.northwestern.edu/pubs/JBact2003). Although the PCR-based analytical phase of that process is extremely time- and labor-intensive, transposon insertion sites can be defined very precisely, allowing this method to serve as a 'gold-standard' against which all other analytical techniques can be compared. Of the 620 essential genes, in the present study we asserted only 46 genes by the partial cDNA microarray (7.4%) and 38 genes by the full-length cDNA microarray (6.1%) as aerobically essential (Figs. 2C and D). Also, when we considered the genes that were asserted as essential by both PCR- and microarray-based readout approaches, the microarray approach recognized only ~39% (partial cDNA microarray) and ~35% (full-length cDNA microarray) of the true positives.

On the other hand, 73 (partial cDNA microarray) and 87 (full-length cDNA microarray) genes were asserted by DNA microaray approach as anaerobically essential. Of the 73 anaerobically essential genes identified by the partial cDNA microarray, 26 mutant *E. coli* MG1655 strains, in which that gene is individually disrupted, are available (http://www.genome.wisc.edu). All the mutant strains grew well under aerobic conditions, and their growth also appeared unimpaired under microaerophilic growth conditions. Only 7 of the 26 mutant *E. coli* strains did not grow under strictly anaerobic conditions (Table 1). While the positive control *nrdG* mutant strain, whose missing gene product has been shown to be essential for strict anaerobic growth [19], did not grow, the negative control wild type *E. coli* MG1655 grew normally. Of the 26 tested mutants, the gene products of 8 were also asserted by full-length ORFs DNA microarray as anaerobically essential, but only 3 of these did not grow under strictly anaerobic growth condition (Table 1). Thus, the ratio of true positives detected by both DNA microarrays was approximately 30% for identifying anaerobic essential genes.

Table 1
Growth of mutant strains with deleted gene products predicted to be essential by DNA microarrays

Gene name	Blattner number	Growth under strict anaerobic condition
agaB[a]	b3138	Yes
alr[a]	b4053	Yes
atpD[a]	b3732	No
fadL	b2344	Yes
fimB[a]	b4312	Yes
fimI[a]	b4315	No
gcvA	b2808	Yes
moaA	b0781	Yes
moaC[a]	b0783	Yes
menB	b2262	Yes
mrcA	b3396	Yes
mutL	b4170	Yes
nikC[a]	b3478	Yes
oppA	b1243	Yes
pgi	b4025	Yes
rnk	b0610	Yes
selA	b3591	Yes
soxR	b4063	Yes
tolC	b3035	Yes
uup	b0949	Yes
yacH	b0117	No
yafU	b0218	No
yciI[a]	b1251	No
yeiH	b2158	Yes
yhiM	b3491	No
yieO	b3754	No
nrdG[b]	b4237	No
[MG1655][c]		Yes

[a] These genes also predicted to be essential by full-length microarrays.
[b] *nrdG* mutant strain: positive control for anaerobic growth restriction [19].
[c] Wild type MG1655: negative control for anaerobic growth restriction.

Discussion

Although *E. coli* has been the focus of intense biochemical and genetic scrutiny, global genomic essentiality data have become available only recently for this organism [26]. In this study, we coupled the genetic footprinting approach to a procedure allowing a rapid, DNA microarray-based readout for genome-wide detection of conditionally essential genes. We tested the utility of this strategy by examining the differences in essential genes between aerobic and anaerobic *E. coli* cultures using two different DNA microarrays. Both microarray platforms identified a number of *E. coli* genes that appeared to be significantly negatively

selected under anaerobic or aerobic growth conditions, although the data from the two platforms were significantly non-overlapping, especially when the cutoff was applied (Fig. 3). Among these genes, many have been reported to be involved in anaerobic respiration (e.g., *fdhD*, *fdhF*, and *menB*) or aerobic respiration (e.g., *ubiA*, *ubiB*, *ubiF*, and *cydA*). Moreover, we identified some uncharacterized genes as essential for anaerobic growth (*yacH*, *yafU*, *yciI*, *yhiM*, and *yieO*) and confirmed the predicted phenotype by testing their growth under strictly anaerobic condition.

In spite of these limited successes, in its present format DNA microarray readout seems not appropriate for the accurate assessment of conditional gene essentiality on a genome-wide scale. Compared to the results of standard genetic footprinting for genes asserted under aerobic growth condition [26], and the experimental results of anaerobic growth (Table 1), the DNA microarray platform has a high number of false positives. The error rate can arise from several possible reasons. First, the insertion of transposons into the genomic DNA is a random event, so the transposon insertion sites in the genomic DNA templates are variable under the two growth conditions being compared. Moreover, in some cases genes can tolerate transposon inserts within certain restricted loci without a detrimental effect on the corresponding gene product, especially when the transposon insertion occurs at or close to the gene's 3′ end. All of these variables contribute to the difference of two DNA pools after growth selection. Second, non-specific PCR amplification artifacts may also play an important role for false positives. For example, genes *thyA*, *ynaJ*, and *ybeY* were asserted as anaerobically essential, but we considered them likely to be false positives due to non-specific PCR amplification, because there is no transposon insertion in these genes under aerobic growth conditions based on previous genetic footprinting result [26]. Third, high-stringency PCR amplification is needed in order to avoid the generation of non-specific PCR products produced from the Y-linker-specific primer alone. However, high-stringency conditions may lead to an unavoidable loss of at least some of the specific PCR products amplified by the Y-linker-specific and transposon-specific primers (especially, when the PCR templates are from two different DNA pools, which may produce even more uneven amplification products). For example, the products of genes *dmsC* and *hyfD* (dimethyl sulfoxide reductase chain C and hydrogenase 4, respectively) are involved in anaerobic respiration [28,29], but in the microarray analysis they were asserted as essential under aerobic growth conditions. This discrepancy may be the result of failed PCR amplifications from the aerobic DNA pool, since our previous study found two transposon insertion sites in each gene under aerobic growth conditions [26]. Fourth, using a restriction enzyme that cleaves the *E. coli* genome with a relatively high frequency may separate the target site on the microarray from the transposon insertion site within a given gene, resulting in a lack of hybridization when using the partial cDNA microarray. However, the high rate of false positives using the full-length cDNA microarray suggests that this is a relatively minor problem. Lastly, in the previous reports the T7 promoter has been incorporated into the transposon [13,14]. The promoter sequence within the transposon could provide an extra degree of specificity in the complex PCR that amplifies transposon–chromosomal junctions, and reduce the non-specific background amplification.

To address these issues we modified several parameters of the original protocol [13,14]. First, we used partial digests instead of complete digests and used the restriction enzyme, *Hpy*CH4 IV, to cut genomic DNA. This restriction enzyme cleaves the *E. coli* genome less frequently (~3000 cuts/genome) than *Hin*P1 I (~8000 cuts/genome), but has the same overhang allowing the same Y-linker to be utilized. Second, to increase the specificity of PCR amplification, we applied a semi-nested PCR procedure, i.e., in the first round of amplification, an external transposon-specific primer and Y-linker primer were used, followed by amplification with an internal transposon-specific primer and Y-linker primer in the second round. In order to fully utilize the entire length of partially digested DNA fragments, we also used two pairs of transposon-specific primers to amplify regions both 3′ and 5′ ends of the transposon insertion site. Third, to reduce the non-specific PCR products produced from the Y-linker primer only, we cut the Y-linker tagged genomic DNA with a second restriction enzyme (*Pfl*M I) prior to PCR amplification. This restriction enzyme specifically recognizes the transposon DNA upstream from the transposon-specific primer binding-site and cleaves the *E. coli* genomic DNA relatively infrequently (~350 restriction sites/*E. coli* genome).

In summary, the present study indicates that although the DNA microarray-based approach for detecting essential genes following global transposon mutagenesis is conceptually straightforward, the technical challenges remain substantial. Despite these challenges the method has been proven to be useful for identifying an essential role for five hypothetical genes in *E. coli* grown under anaerobic conditions.

Note added in proof

An alternative experimental approach to detect transposon insertion sites by DNA microarray analysis in *E. coli* has been recently communicated: K. Winterberg, W.S. Reznikoff: Using HAIL, Hybridization Analysis of Insertion Libraries, to Map Chromosomal Locations of Tn5 Transposon Insertions in *Escherichia coli* K-12,

Abstract presented at the American Society of Microbiology general meeting, Washington D.C., 2003.

Acknowledgments

We thank Arkady B. Khodursky and Jaeyong Ahn for the help with the full-length DNA microarrays. Research at Northwestern University was supported by grants from the National Institutes of Health (NIGMS) and the Department of Energy Genomes to Life Program.

Appendix. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bbrc.2004.07.110.

References

[1] C.A. Hutchison, S.N. Peterson, S.R. Gill, R.T. Cline, O. White, C.M. Fraser, H.O. Smith, J.C. Venter, Global transposon mutagenesis and a minimal Mycoplasma genome, Science 286 (1999) 2165–2169.

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

[3] E.A. Winzeler, D.D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J.D. Boeke, H. Bussey, et al., Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis, Science 285 (1999) 901–906.

[4] S.K. Kim, J. Lund, M. Kiraly, K. Duke, M. Jiang, J.M. Stuart, A. Eizinger, B.N. Wylie, G.S. Davidson, A gene expression map for *Caenorhabditis elegans*, Science 293 (2001) 2087–2092.

[5] R.S. Kamath, A.G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta, A. Kanapin, N. Le Bot, S. Moreno, M. Sohrmann, et al., Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi, Nature 421 (2003) 231–237.

[6] R.S. Hare, S.S. Walker, T.E. Dorman, J.R. Greene, L.M. Guzman, T.J. Kenney, M.C. Sulavik, K. Baradaran, C. Houseweart, H. Yu, et al., Genetic footprinting in bacteria, J. Bacteriol. 183 (2001) 1694–1706.

[7] B.J. Akerley, E.J. Rubin, V.L. Novick, K. Amaya, N. Judson, J.J. Mekalanos, A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*, Proc. Natl. Acad. Sci. USA 99 (2002) 966–971.

[8] V. Smith, D. Botstein, P.O. Brown, Genetic footprinting: a genomic strategy for determining a gene's function given its sequence, Proc. Natl. Acad. Sci. USA 92 (1995) 6479–6483.

[9] P. Ross-Macdonald, P.S. Coelho, T. Roemer, S. Agarwal, A. Kumar, R. Jansen, K.H. Cheung, A. Sheehan, D. Symoniatis, L. Umansky, et al., Large-scale analysis of the yeast genome by transposon tagging and gene disruption, Nature 402 (1999) 413–418.

[10] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[11] J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, P.O. Brown, Genome-wide analysis of DNA copy-number changes using cDNA microarrays, Nat. Genet. 23 (1999) 41–46.

[12] B. Ren, F. Robert, J.J. Wyrick, O. Aparicio, E.G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, et al., Genome-wide location and function of DNA binding proteins, Science 290 (2000) 2306–2309.

[13] V. Badarinarayana, P.W. Estep III, J. Shendure, J. Edwards, S. Tavazoie, F. Lam, G.M. Church, Selection analyses of insertional mutants using subgenic-resolution arrays, Nat. Biotechnol. 19 (2001) 1060–1065.

[14] C.M. Sassetti, D.H. Boyd, E.J. Rubin, Comprehensive identification of conditionally essential genes in mycobacteria, Proc. Natl. Acad. Sci. USA 98 (2001) 12712–12717.

[15] C.M. Sassetti, D.H. Boyd, E.J. Rubin, Genes required for mycobacterial growth defined by high density mutagenesis, Mol. Microbiol. 48 (2003) 77–84.

[16] K.F. Jensen, The *Escherichia coli* K-12 wild types W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels, J. Bacteriol. 175 (1993) 3401–3407.

[17] S.Y. Gerdes, M.D. Scholle, M. D'Souza, A. Bernal, M.V. Baev, M. Farrell, O.V. Kurnasov, M.D. Daugherty, F. Mseeh, B.M. Polanuyer, et al., From genetic footprinting to antimicrobial drug targets: examples in cofactor biosynthetic pathways, J. Bacteriol. 184 (2002) 4555–4572.

[18] I.Y. Goryshin, J. Jendrisak, L.M. Hoffman, R. Meis, W.S. Reznikoff, Insertional transposon mutagenesis by electroporation of released Tn5 transposition complexes, Nat. Biotechnol. 18 (2000) 97–100.

[19] X. Garriga, R. Eliasson, E. Torrents, A. Jordan, J. Barbe, I. Gibert, P. Reichard, nrdD and nrdG genes are essential for strict anaerobic growth of *Escherichia coli*, Biochem. Biophys. Res. Commun. 229 (1996) 189–192.

[20] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[21] R. Overbeek, N. Larsen, T. Walunas, M. D'Souza, G. Pusch, E. Selkov Jr., K. Liolios, V. Joukov, D. Kaznadzey, I. Anderson, et al., The ERGO genome analysis and discovery system, Nucleic Acids Res. 31 (2003) 164–171.

[22] A.B. Khodursky, B.J. Peter, D. Botstein, P.O. Brown, C. Yanofsky, DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*, Proc. Natl. Acad. Sci. USA 97 (2000) 12170–12175.

[23] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, T.P. Speed, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucleic Acids Res. 30 (2002) e15.

[24] G. Balazsi, K.A. Kay, A.L. Barabasi, Z.N. Oltvai, Spurious spatial periodicity of co-expression in microarray data due to printing design, Nucleic Acids Res. 31 (2003) 4425–4433.

[25] J. Quackenbush, Microarray data normalization and transformation, Nat. Genet. 32 (Suppl) (2002) 496–501.

[26] S.Y. Gerdes, M.D. Scholle, J.W. Campbell, G. Balazsi, E. Ravasz, M.D. Daugherty, A.L. Somera, N.C. Kyrpides, I. Anderson, M.S. Gelfand, et al., Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655, J. Bacteriol. 185 (2003) 5673–5684.

[27] M. Riley, Genes and proteins of *Escherichia coli* K-12, Nucleic Acids Res. 26 (1998) 54.

[28] S.C. Andrews, B.C. Berks, J. McClay, A. Ambler, M.A. Quail, P. Golby, J.R. Guest, A 12-cistron *Escherichia coli* operon (hyf) encoding a putative proton-translocating formate hydrogenlyase system, Microbiology 143 (1997) 3633–3647.

[29] D. Sambasivarao, J.H. Weiner, Dimethyl sulfoxide reductase of *Escherichia coli*: an investigation of function and assembly by use of in vivo complementation, J. Bacteriol. 173 (1991) 5935–5943.