RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array

Douglas W. Selinger[1], Kevin J. Cheung[2], Rui Mei[3], Erik M. Johansson[3], Craig S. Richmond[5], Frederick R. Blattner[5], David J. Lockhart[3,4], and George M. Church[1]*

*[1]Department of Genetics, Harvard Medical School, 200 Longwood Avenue Boston, MA 02115. [2]Harvard College, Cambridge, MA 02138. [3]Affymetrix Inc., 3380 Central Expressway, Santa Clara, CA. [4]Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, CA 92121. [5]Laboratory of Genetics, University of Wisconsin, Madison, WI 53706. *Corresponding author (church@arep.med.harvard.edu).*

Received 3 May 2000; accepted 30 August 2000

We have developed a high-resolution "genome array" for the study of gene expression and regulation in *Escherichia coli*. This array contains on average one 25-mer oligonucleotide probe per 30 base pairs over the entire genome, with one every 6 bases for the intergenic regions and every 60 bases for the 4,290 open reading frames (ORFs). Twofold concentration differences can be detected at levels as low as 0.2 messenger RNA (mRNA) copies per cell, and differences can be seen over a dynamic range of three orders of magnitude. In rich medium we detected transcripts for 97% and 87% of the ORFs in stationary and log phases, respectively. We found that 1,529 transcripts were differentially expressed under these conditions. As expected, genes involved in translation were expressed at higher levels in log phase, whereas many genes known to be involved in the starvation response were expressed at higher levels in stationary phase. Many previously unrecognized growth phase-regulated genes were identified, such as a putative receptor (b0836) and a 30S ribosomal protein subunit (S22), both of which are highly upregulated in stationary phase. Transcription of between 3,000 and 4,000 predicted ORFs was observed from the antisense strand, indicating that most of the genome is transcribed at a detectable level. Examples are also presented for high-resolution array analysis of transcript start and stop sites and RNA secondary structure.

Keywords: *Escherichia coli*, stationary phase, gene expression, functional genomics, DNA chips, oligonucleotide arrays, microarrays

The ability to simultaneously measure RNA abundance for large numbers of genes has revolutionized biological research by allowing the analysis of global gene expression patterns. Oligonucleotide arrays have been used to examine differential gene expression in many organisms, including yeast, human, mouse, and bacteria[1–5]. Various analytical approaches have been developed and applied to these datasets to further characterize transcriptional regulation and the connectivity of genetic networks[6–10]. Global gene expression analyses in prokaryotes have lagged behind those in eukaryotes, in part because of the lack of polyadenylation of prokaryotic mRNA, which has thwarted separation or selective labeling of mRNA in the presence of the much more abundant transfer RNA (tRNA) and ribosomal RNA (rRNA)[1,11–13].

We describe here a "genome array," on which both coding and noncoding regions of the *Escherichia coli* genome are represented, and describe a genome-wide analysis of RNA at sub-transcript-level resolution. We developed a labeling protocol based on random priming of total RNA that is reproducible, quantitative over three orders of magnitude, and sufficiently sensitive to detect as few as 0.2 copies per cell. When used to compare gene expression in log versus stationary phase, this method yields results that both agree with the literature and identify novel sets of co-regulated genes. We also present evidence that sub-transcript-level resolution paired with complete genomic representation of *E. coli* on the array allows for analysis of operon structure, identification of small RNAs and antisense RNAs, and some aspects of RNA secondary structure.

Results and discussion

Array design. The array consists of a 544 × 544 grid of 24 × 24 μm regions that each contain ~10^7 copies of selected 25-mer oligonucleotides (295,936 total) of defined sequence. The oligonucleotides on the array are synthesized in situ on a derivatized glass surface using a combination of photolithography and combinatorial chemistry[2,14]. Probe oligonucleotides are arranged in pairs, or probe pairs, one of which is perfectly complementary to the target sequence (the perfect match, or PM oligonucleotide) and one with a single base mismatch at the central position (the mismatch, or MM oligonucleotide), which serves as a control for nonspecific hybridization. Oligonucleotides on the array are further organized into groups, or probe sets, which are complementary to a single putative transcript. Probe sets are present for 4,403 "b-numbers," which include all 4,290 predicted ORFs (ref. 15), as well as all rRNAs and tRNAs. Both strands of intergenic regions at least 40 base pairs in length are represented, whereas only the strand predicted to be transcribed is represented for the ORFs. Most probe sets have 15 probe pairs, although certain selected RNAs, such as *lpp* and *Bacillus subtilis* control transcripts have 60 or more.

Oligonucleotides are arranged in alternating rows of PM and MM features (Fig. 1). The top half of the array contains oligonucleotides targeting ORFs and miscellaneous untranslated RNAs, and the bottom half targets intergenic regions. The extreme bottom has probes for tRNAs and rRNAs. A biotinylated control oligonucleotide is added to the hybridization mixture and binds to the checkerboard border, corners, the AFFX-E COLI-1 logo, and 100 pairs of features

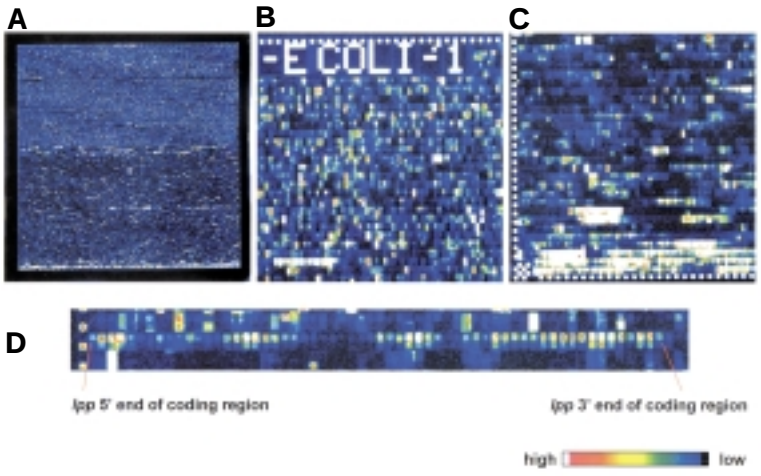

Figure 1. False-color images of scanned *E. coli* genome array hybridized with a sample derived from a stationary-phase culture growing in LB. (A) Whole array (top half: ORFs, bottom half, intergenic regions; very bottom, rRNAs and tRNAs). (B) Close-up of coding regions. The bright streak on the lower left is *rmf*. (C) Close-up of intergenic regions, rRNAs, tRNAs. (D) *lpp* coding region. Note: Apparent saturation (esp. in C) is due to display settings and not signal saturation.

in a regularly spaced grid across the array. These patterns are used for grid alignment and to correct for spatial variations in array brightness (see Experimental Protocol).

Choice of a metric for RNA abundance. Signals from the 15 probe pairs in each probe set must be quantitated and combined into a measure of RNA concentration. The significant systematic differences in signal within a probe set for a given RNA led us to investigate RNA abundance metrics that could be used as alternatives to the previously reported "average difference" metric[2,3] (AD), which uses the mean of all PM-MM pairs after outliers are discarded. When probe pairs of the probe sets were ranked by intensity difference (PM – MM), and probe pairs of different ranks were used to represent the entire probe set, we found that the number of genes detected increased as brighter probe pairs were used. An exception was the brightest probe pair, which gave fewer detected transcripts because of the high variability of the maximal probe pair of the negative controls, which were used to establish the detection threshold. Transcripts were considered detected if the probe pair intensity difference of a given rank was at least 3 standard deviations above the mean of probe pairs of the same rank taken from control probe sets for which no transcript was present (see Experimental Protocol). Using the second maximal probe pair, 87% of the ORFs were detected in log phase, compared to 23% for the maximal and 70% for the third maximal. The use of the second maximal signal also led to the detection of more RNAs than measures of centraltendency such as the median intensity (20%) and AD (18%). We therefore chose to use the second maximal probe pair intensity, or "2max," as a metric for RNA abundance.

The three metrics investigated (2max, the median, and AD) had a sensitivity of <0.2 copies per cell, were approximately linear for relative changes less than 10-fold and nonlinear over a dynamic range of three orders of magnitude, and were about equally precise ($R \approx 0.94$)(Fig. 2). The lowest concentration of RNA for which a twofold concentration could be detected was a change from 0.2 to 0.4 copies per cell, which was called significant in 4/4 probe sets with an average measured fold change of 1.65 ± 0.35. We detected spiked RNAs from 100% (12/12) of probe sets at 0.2 copies/cell and 25%

Table 1. ORFs with significant changes in probe set intensity, previously known to be differentially regulated in stationary phase[a]

Gene	Absolute change[b]	Fold change[c]	Annotation
rmf	120,465	17	Ribosome modulation factor
glgS	118,425	160	Glycogen biosynthesis, *rpoS* dependent
hdeA	104,184	41	ORF, hypothetical protein
dps	91,763	55	Global regulator, starvation conditions
hdeB	34,968	5	ORF, hypothetical protein
osmY	21,914	9	Hyperosmotically inducible periplasmic protein
himA	19,920	23	Integration host factor (IHF), α-subunit; site-specific recombination
csgB	19,385	>30	Minor curlin subunit precursor, similar to CsgA
clpA	17,369	8	ATP-binding component of serine protease
wrbA	15,845	7	trp repressor-binding protein; affects association of trp repressor and operator
fic	14,900	26	Induced in stationary phase, recognized by *rpoS*, affects cell division
htrE	14,893	>24	Probable outer membrane porin protein involved in fimbrial assembly
cstA	13,475	11	Carbon starvation protein
sspA	13,076	4	Regulator of transcription; stringent starvation protein A
ftsA	11,171	>5	ATP-binding cell division protein, septation process, complexes with FtsZ, associated with junctions of inner and outer membranes
hyaE	10,406	>4	Processing of HyaA and HyaB proteins
dacC	10,064	8	D-alanyl-D-alanine carboxypeptidase; penicillin-binding protein 6
emrA	8,433	>4	Multidrug resistance secretion protein
otsB	8,276	2	Trehalose-6-phosphate phosphatase, biosynthetic
cfa	7,896	>4	Cyclopropane fatty acyl phospholipid synthase
iciA	7,506	>4	Replication initiation inhibitor, binds to 13-mers at oriC
rpoH	–26,713	0.4	RNA polymerase, sigma(32) factor; regulation of proteins induced at high temperatures
hns	–170,027	0.04	DNA-binding protein HLP-II (HU, BH2, HD, NS); pleiotropic regulator

[a]Altogether, 1,529 genes (including tRNAs and rRNAs) were significantly changed. Of these, 926 were increased in stationary phase and 603 were decreased. Annotations are from the University of Wisconsin Genome Project[15,19]. The complete dataset can be found at ExpressDB (refs 21,22).
[b]Genes are ranked by absolute change, given as Δ2max in arbitrary fluorescence units. Signal was normalized to total array intensity.
[c]Fold changes were adjusted based on calibration with spiked transcripts (Fig. 2B). For those transcripts that were called absent in one condition, the fold change was estimated (indicated by >) by substituting the mean of the negative controls + 3 standard deviations for the undetected transcript. Out of 69 transcripts that are known to be differentially expressed[15] and that are present on the array, 23 were called significantly changed. The remaining 46 were not significantly changed. Of the significant changes, 22 out of 23 agree with the direction of change reported in the literature. *rpoH*, the heat-shock sigma factor, is reported to increase in stationary phase, although RNA levels decreased about threefold in our experiment. This may be a result of translational control, which is known to play a role in the regulation of *rpoH* (ref. 16).

(2/8) at 0.02 copies/cell.

Stationary-phase versus log-phase expression analysis. We compared the expression profiles of cells grown in rich media (Luria–Bertani, LB) to either mid-log phase (OD_{600} = 0.6) in a fermentor or to late stationary phase in an overnight-shaken culture. As expected, log-phase cells showed increased RNA levels for genes involved in protein synthesis (rRNAs, tRNAs, and ribosomal proteins) and cell membrane synthesis (*lpp*), whereas stationary-phase cells showed increases in stress/starvation response genes such as *dps* and *rmf*. Of 69 genes known to be differentially regulated in stationary phase[16], 22 of these were called significantly changed in agreement with the literature (Table 1). One gene, *rpoH*, which is known to be regulated post-transcriptionally[17], was called significantly changed in the reverse direction from that reported. The remaining 46 were not significantly changed. Some discrepancies and apparent "missed" changes are expected because most of the changes reported in the literature were detected at the protein level (usually by activity of *lacZ* fusions), and the correlation between gene transcript levels and protein product activity is expected to be imperfect. A notable transcript that was not called changed is the gene for the stationary-phase sigma factor, *rpoS*. This is expected because the transcript is known to peak in early stationary phase and decrease thereafter, and therefore may not be significantly elevated by late stationary phase. *rpoS* is also known to be regulated at the level of translation and protein stability[18]. However, the mRNA levels of 16 genes known to be *rpoS* regulated are increased in stationary phase, indicating that *rpoS* activity has, in fact, increased.

Table 2. ORFs with the largest significant increases in probe set intensity in stationary phase[a]

b-number	Gene[b]	Absolute change	Fold change	Annotation
b1005	*ycdF*	135,446	102	ORF, hypothetical protein
b0836	–	130,009	>1,000	Putative receptor
b0953	*rmf*	120,465	17	Ribosome modulation factor
b3049	*glgS*	118,425	160	Glycogen biosynthesis, *rpoS* dependent
b4045	*yjbJ*	117,238	9	ORF, hypothetical protein
b3510	*hdeA*	104,184	41	ORF, hypothetical protein
b0812	*dps*	91,763	55	Global regulator, starvation conditions
b1480	*rpsV*	74,063	48	30S Ribosomal subunit protein S22
b2665	*ygaU*	71,120	60	ORF, hypothetical protein
b3555	*yiaG*	67,426	12	ORF, hypothetical protein
b3239	*yhcO*	64,840	140	ORF, hypothetical protein
b1240	–	53,219	4	ORF, hypothetical protein
b1635	*gst*	51,788	81	glutathionine *S*-transferase
b1051	*msyB*	51,334	11	Acidic protein suppresses mutants lacking function of protein export
b0966	*yccV*	50,782	16	ORF, hypothetical protein
b1318	*ycjV*	48,950	75	Putative ATP-binding component of a transport system
b1154	*ycfK*	46,949	>180	ORF, hypothetical protein
b1566	*flxA*	45,987	13	ORF, hypothetical protein
b2212	*alkB*	43,206	6	DNA repair system specific for alkylated DNA
b1492	*xasA*	42,971	85	Acid sensitivity protein, putative transporter
b2266	*elaB*	42,249	>140	ORF, hypothetical protein
b1164	*ycgZ*	41,961	3	ORF, hypothetical protein
b3183	*yhbZ*	41,925	7	Putative GTP-binding factor
b1262	*trpC*	41,711	7	*N*-(5-phosphoribosyl) anthranilate isomerase and indole-3-glycerolphosphate synthetase
b1739	*osmE*	40,691	24	Activator of *ntrL* gene

[a]Same analysis as given in notes to Table 1.
[b]The genes *rmf, glgS, hdeA,* and *dps* are known to be differentially regulated in stationary phase[16]. The products of *yjbJ*, *dps*, and *hdeA* are the first, fifth, and sixth most abundant proteins, respectively, in stationary-phase *E. coli*[20].

Altogether, there were 1,529 RNAs (including tRNAs and rRNAs) in which the abundance significantly changed (see Experimental Protocol), representing about 35% of the putative 4,403 RNAs in the genome. Of the 926 that were increased in stationary phase and the 603 that were decreased, 77% were changed by more than twofold. It is unclear how many of these changes have biological significance and whether the size of the absolute change (copies per cell) or relative change is more important in the regulation of genetic networks, although it is likely to be gene- and condition-dependent. For genes with post-transcriptional regulation, changes in transcript level may have little effect on the final activity of the gene product. Still, the sheer number of

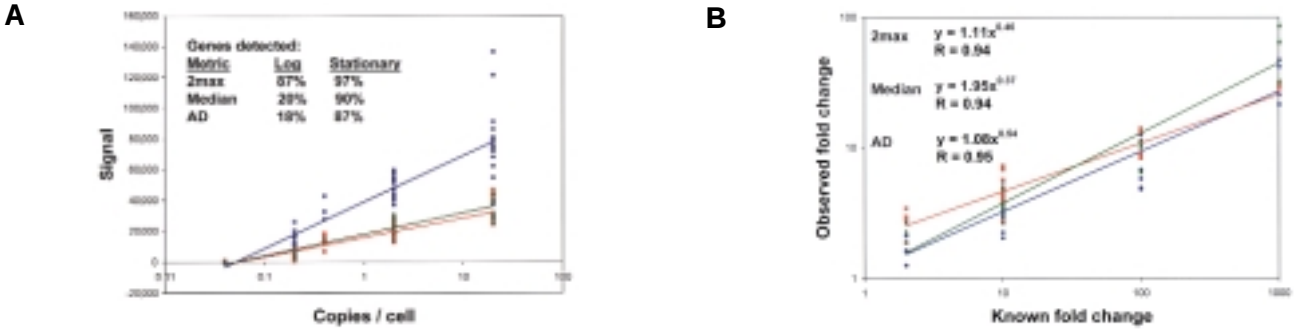

Figure 2. Comparison of 2max (blue circle), median (red square), and average difference (green triangle) abundance metrics using *Bacillus subtilis* control RNAs. (A) Abundance measurement vs. RNA concentration, with present calls. Genes are considered detected by the 2max and median metrics if they are at least 3 standard deviations above negative controls for which no RNA is present. Detection using the average difference metric is determined using an algorithm implemented in the GeneChip 3.2 software package. No false positives were detected for any of the metrics (see Experimental Protocol). (B) Plot of observed fold changes measured by various metrics vs. known fold changes. The relationship between observed and known fold change is nonlinear for all three metrics over a dynamic range of three orders of magnitude, and approximately linear for changes less than 10-fold.

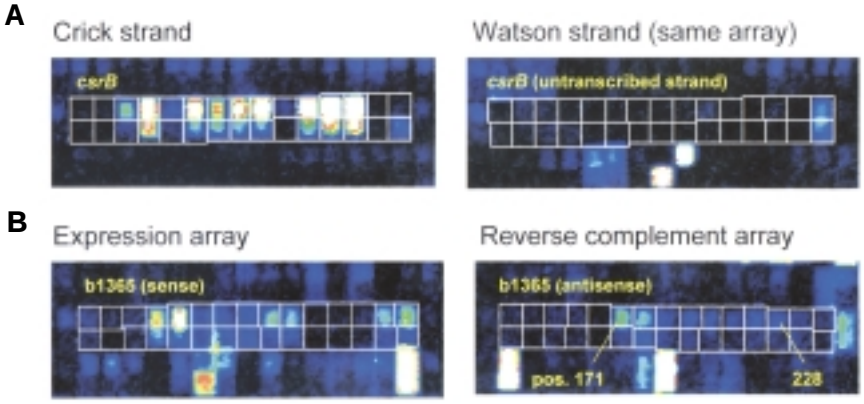

Figure 3. The *E. coli* array can detect strand-specific transcription and can be used to identify (A) small untranslated RNAs, such as *csrB*, and (B) detection of a previously unidentified antisense RNA in the Rac prophage. Transcription was detected on the strand opposite b1365 from positions 171 to 228. Position is given as the number of base pairs from the central nucleotide of the oligonucleotide probe to the translation start of b1365. The oligonucleotides are closely spaced, but three of them are nonoverlapping. The 15 probe pairs in these probe set are outlined by the white grid, with the PM features on the top row. Probe sets for the untranscribed strands show the background signal typical of undetected transcripts. The oligonucleotides in A and on the expression array of B are tiled from left to right in the 5' to 3' direction.

changes detected suggests there are many transcriptionally regulated genes important for adaptation to stationary phase, or stresses in general, which have previously gone unrecognized. It is interesting to note that of the 25 RNAs most increased in stationary phase (ranked by absolute change), 14 are genes of unknown function (Table 2). This includes a gene (b0836), annotated as a putative receptor[19], which is measured to increase in stationary phase by more than 1,000-fold, and 30S ribosomal protein subunit S22, which increases 48-fold. Also found in the top 10 most increased in stationary phase are *yjbJ*, *hdeA*, and *dps*, for which the protein products were reported to be the first, sixth, and fifth most abundant in stationary phase, respectively[20]. Of the 10 genes of "known" function, only 3 were already known to be increased in stationary phase. The complete results of this analysis are in an expression database[21,22].

New applications of a genome array: identification of small and antisense RNAs. Inclusion of probes for predicted intergenic regions allows genome-wide scanning for previously unidentified RNAs (Fig. 3). *csrB*, a small (360 bases) untranslated RNA that is known to be abundant in stationary phase[23] but was not present in our annotation database, was easily detected by probes targeting the region between loci b2793 and b2792.

Genome arrays made by in situ synthesis of oligonucleotides also present an opportunity for the identification of antisense RNAs. By simply inverting the synthesis, a complementary array can be synthesized containing probes that will bind to antisense RNAs[24]. Hybridization of a stationary-phase sample to such a reverse-complement chip resulted in the detection of antisense transcription of between 3,000 and 4,000 predicted ORFs, suggesting that there is a low level of transcription throughout the *E. coli* genome. The physiological significance of this transcription is unclear. An example of a detected antisense RNA is b1365 (Fig. 3B), a predicted ORF located in the Rac prophage. This transcript may be from an overlapping gene encoded on the opposite strand, a common occurrence in phage and viruses. Alternatively, it could result from readthrough transcription of an upstream IS5 insertion. Consistent with this is the detection of IS5 transcription as well as antisense transcripts for the intervening ORFs, b1366–b1369.

It is important to note that transcription at a given locus may be part of a long 5′ or 3′ untranslated region (UTR), a spacer within an operon, an untranslated RNA, an ORF, or the result of an incorrectly predicted ORF start or stop site. The ability to establish transcript start and stops would aid in the interpretation of these RNAs, and is discussed in the next section.

Sub-transcript resolution. The large number of oligonucleotides (295,936) on the array allowed transcripts to be probed at high resolution. Intergenic regions were probed, on average, every 6 bases, whereas ORFs and known RNAs were probed on average every 60 bases. This makes it possible to obtain reasonably high-resolution information on transcript starts and stops and operon structure.

Analysis of oligonucleotide probes for selected transcripts revealed a large amount of intensity variation across the probes within a probe set, but also a striking consistency to the patterns (Fig. 4). A highly reproducible pattern was seen for all probe sets inspected. The intensity variation is likely due to sequence-dependent differences in hybridization affinity and accessibility and to the effects of secondary structure on hybridization. The similarity of the pattern obtained using RNA samples labeled by random primers and genomic DNA labeled directly with terminal transferase, suggests that the pattern is not a result of variations in priming or labeling efficiency. The signal pattern correlates well with regions of experimentally confirmed RNA secondary structure, such as the *ompA* 5′ stemloop[25] (data not shown), but poorly with G/C content or hypothetical hairpin formation of the probe oligonucleotides[26,27]. It is currently being investigated whether the signal is correlated with other predicted local RNA secondary structures. It has been shown that secondary structure can strongly affect oligonucleotide hybridization[24,28]. Locations of known secondary structures in the *lpp* and *rpsO* 3′ UTRs are highlighted in Figure 4. It must be noted, however, that lack of signal may indicate early transcription termination. Signal from flanking regions and/or independent information about transcription starts and stops can be used to rule out this possibility.

Analysis of transcription in predicted intergenic regions allows 5′ and 3′ UTRs to be mapped. Transcriptional start and stops derived from array data for *lpp* and *rpsO* (Fig. 4) agree well with those determined with other methods. *lpp* is known to be transcribed from –33 to 284, ending in a hairpin[29,30], and *rpsO* starting from –100 and continuing through a 3′ stemloop structure into *pnp*, with which it is co-transcribed[31]. To map transcription endpoints with the array, the ability of each oligonucleotide to hybridize to its target was determined. Oligonucleotides were considered "reliable" if, when hybridized to genomic DNA, their intensity difference (PM – MM) was at least 3 standard deviations above noise. Oligonucleotides below this cutoff are referred to as "unreliable." Transcription was considered detectable at positions that had reliable oligonucleotides if the mean intensity difference at that position was greater than its standard deviation. Signal from *lpp* was detected starting between oligonucleotides centered at positions –30 and –37 and can be detected until the last reliable probe at position 250. The probes from 274 to 284 are unreliable and correspond to the location of a known hairpin. Transcription of *rpsO* is first detected at position –94 and begins no earlier than –117, the first reliable oligonucleotide for which no transcription is detected. *rpsO* transcription is detected, albeit irregularly, throughout the 3′ UTR, where it presumably continues into *pnp*. Probes for *pnp*, however, are located only at the 3′ end of the ORF, so this continuation was not directly observed.

rpsO and *pnp* are co-transcribed and contain a structured attenuator sequence between them that causes a high frequency of rho-independent termination before the *pnp* coding region. This

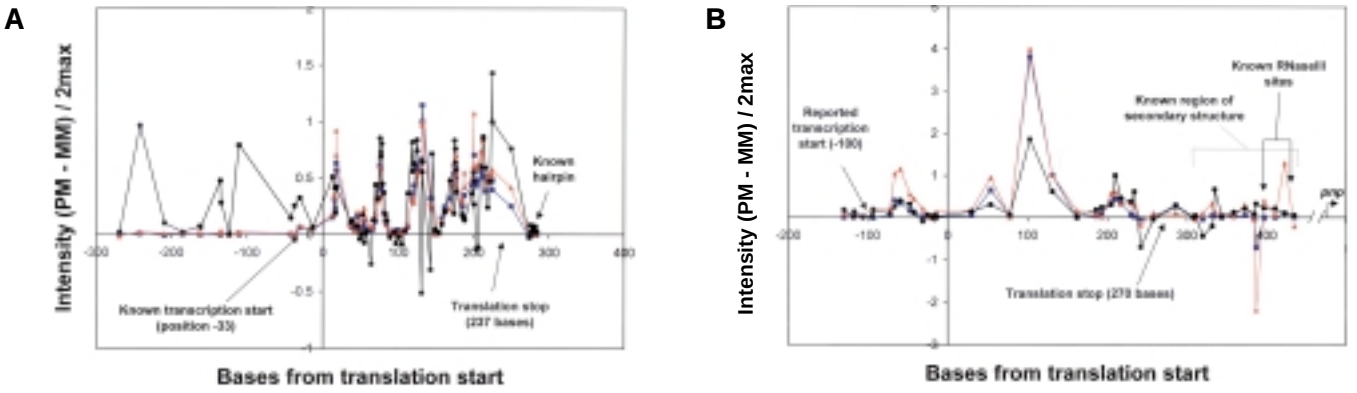

Figure 4. Determination of transcription starts of (A) *lpp* and (B) *rpsO*. Both genes exhibit reproducible hybridization patterns despite large log phase fold increases of 60- and 400-fold, respectively. 2max-normalized (PM - MM) fluorescence intensity of log phase (blue square), stationary phase (red triangle), and genomic DNA (black circle) arrays were plotted against distance from center of oligonucleotide to translation start site. Points for log and stationary phase are the means of duplicate experiments. Oligonucleotides that target both the open reading frame and the flanking intergenic regions allow this region to be probed at ~6 base pair average resolution for *lpp* and ~13 for *rpsO*. Transcription starts are detected between –30 and –37 for *lpp* (reported –33)[28,29] and between –94 and –117 for *rpsO* (reported –100) [30]. *lpp* is known to sometimes extend to position 284, ending in a hairpin structure. Oligonucleotides in this region showed no hybridization, suggesting early termination of transcription and/or sensitivity of the array to secondary structure. Variability in the hybridization pattern at the 3' end may reflect differential processing. The *rpsO* transcript has a 3' hairpin and can be co-transcribed with downstream *pnp*. The hairpin structure serves as a stabilizing element for both *rpsO* and *pnp* as well as a transcriptional attenuator[31,32]. Processing by RNaseIII may relieve secondary structure in this region and lead to the increased signal seen at the 3' end in stationary phase.

structured region also serves as a 3′ stabilizer for *rpsO* and a 5′ stabilizer for *pnp* and is targeted by RNaseE and RNaseIII, which lead to rapid degradation of both *rpsO* and *pnp* RNAs[32,33]. *rpsO* was seen to increase 400-fold in log phase, the largest relative fold increase in log phase, whereas *pnp* showed no change. Interestingly, the oligonucleotide hybridization pattern shows some differences between log and stationary phase toward the 3′ end of *rpsO* (Fig. 4B). This region is between two known RNaseIII sites and is increased in stationary phase relative to the other probe pairs in the probe set, perhaps indicating that RNaseIII processing at this site is increased in stationary phase, leading to a decrease in local RNA secondary structure and increased hybridization to the array.

Oligonucleotide arrays and cross-hybridization. Considerably more cross-hybridization is observed on *E. coli* arrays than on eukaryotic arrays, presumably because of the presence of large amounts of labeled rRNA and tRNA. Because PM features are tiled immediately above their MM counterparts, PM and MM features of equal intensity appear as rectangles in the image. These can be seen throughout the array images (Fig. 1B–D). If the MM feature were not used, a large number of cross-hybridizing PM oligonucleotides would be included in the analysis and increase the noise of the system. The combination of MM signal subtraction and removal of outliers has proved effective in quantifying RNA abundance changes with oligonucleotide arrays[2]. We considered using MM features to identify cross-hybridizing PM features, discarding them, and then using the raw PM intensities of the remaining features to derive abundance measures. Our preliminary analysis suggested that this approach yields results similar to those using PM – MM, so we did not pursue this line further.

The future of genome arrays. The noise present in a high-complexity hybridization reaction encourages use of increased statistical rigor to determine the significance of probe signal patterns. Corrections for systematic noise due to cross-hybridization, variability in probe efficiency, and spatial variability across the array surface can be used to increase the sensitivity and precision of the data. Because of the complexity of the factors influencing array signal, internal negative controls, such as probe sets that target absent RNAs, may be the best way to estimate the amount of signal that can be expected from all factors besides specific hybridization. Replicate array expression experiments, in combination with array hybridizations of genomic DNA, can be used to extract information from single oligonucleotides, allowing transcripts to be mapped at high resolution. The ability to interpret genome-wide transcription data at 10- to 100-base pair resolution has many potential applications for the study of gene regulation in both prokaryotes and eukaryotes, including identification of alternative promoters, and the ability to experimentally identify regions of transcription that are missed by ORF-predicting algorithms, a problem that is becoming more urgent as annotators deal with the difficult task of predicting genes in higher eukaryotic genomes[34].

There are a number of advantages of arrays that use short single-stranded probes over those that utilize longer double-stranded DNAs[35,36]. These advantages include higher resolution, better cross-hybridization controls, potential for paralog discrimination, splice variant identification, and strand-specific transcript detection. DNA arrays with probes covering entire genomes, rather than just ORFs, are a logical step in the evolution of arrays. Inclusion of intergenic regions allows arrays to be used as readouts for techniques that enrich for DNA sequences of interest, such as protein-bound sequences using whole-genome in vivo methylase protection[37] or chromatin immunoprecipitation (ChIP)[38,39]. If they are double-stranded they might be used as a direct in vitro assay of DNA–protein interactions[40]. Genome arrays should also be useful for genotyping both ORF and promoter sequences[41,42]. Integration of these data into an understanding of genetic networks and cell physiology will remain a central challenge in the post-genomic era.

Experimental protocol

Cell culture. *E. coli* MG1655 was grown to mid-log phase in LB in a fermentor at 37°C with constant aeration of 11 L/min and agitation of 300 r.p.m. Stationary-phase cultures were grown at 37°C overnight in culture flasks containing LB aerated by shaking at 225 r.p.m. Samples were taken in duplicate for the log-phase culture and sampled once from the stationary-phase culture. Each log-phase duplicate was labeled once, and the single stationary-phase RNA was labeled twice independently.

RNA Preparation. RNA was prepared by extraction with acid phenol:chloroform extraction. Briefly, samples of culture were transferred directly into acid phenol:chloroform, 5:1 (Ambion, Austin, TX) at 65°C to ensure rapid lysis and inactivation of RNases. Two additional acid phenol:chloroform extraction were performed, followed by ethanol precipitation, treatment with 1.25 U of DNase I (GIBCO BRL, Gaithersburg, MD) per milliliter of culture, 20 µg proteinase K (Boehringer Mannheim, Mannheim, Germany) per milliliter of culture, and a final ethanol precipita-

tion. The pellet was then washed with 70% ethanol, resuspended in water treated with diethyl pyrocarbonate, quantified by absorbance at 260 nm, and visualized on a denaturing polyacrylamide gel. We subsequently found that contaminating salts and sugars from the media were inhibiting the reverse-transcription reaction used to make labeled complementary DNA (cDNA). The yield was dramatically improved (see below) by removing salts and sugars after the first precipitation by three passes through Centricon PL-20 concentrator columns (Centricon, Beverly, MA), which have a cutoff at about 30 bases, and diluting the concentrate with DEPC water.

cDNA synthesis, biotinylation. The protocol currently supported by Affymetrix for prokaryotic expression analysis was not available at the time of this study, and limited direct comparison has been made with the protocol described here. In our labeling protocol 1.5 mg (see note below) of total RNA was fragmented in a high-magnesium buffer (40 mM Tris acetate, pH 8.1, 100 mM potassium acetate, 30 mM magnesium acetate) at 94°C for 30 min in the presence of random octamers (6.7 mM) and four control RNAs generated by in vitro transcription (*B. subtilis dapB*, *thrB*, *lysA*, and *pheB*).

Note: 50 μg of column-purified total RNA (RNA preparation section) yielded >10 μg of cDNA, enough for an array hybridization. Taking into account a 67% loss from the Centricon columns, 150 μg of RNA from a phenol:chloroform prep are enough for an array experiment. This hybridization sample can be recovered and reused at least three times without significant loss of signal[3]. The use of Centricon columns caused no noticeable changes in the nature of the resulting array data.

After fragmentation the sample was put immediately on ice. The reaction was then diluted twofold into the following reverse-transcription reaction: 1× Superscript II buffer, dNTPs (1.3 mM), dithiothreitol (10 mM), 3,000 units of Superscript II Reverse Transcriptase (GIBCO BRL), which was incubated at 42°C for 3 h. RNA was then degraded by treatment with 135 units of RNase One (Promega, Madison, WI). RNase One was then heat-inactivated, and unincorporated nucleotides and random octamers were removed by Centrisep Spin Columns (Princeton Separations, Adelphia, NJ). This reaction typically yields ~30 μg first-strand cDNA. A 10 μg aliquot was then biotinylated with 30 units of Terminal Deoxynucleotidyl Transferase (TdT; GIBCO BRL) and 50 μM Biotin-N6-ddATP (Dupont NEN, Boston, MA) in 1× One-Phor-All buffer (Pharmacia, Piscataway, NJ) and incubated at 37°C for 2 h. Genomic DNA was fragmented with DNaseI (Promega), 1.1 U per microgram of DNA in 1× One-Phor-All buffer to an average size of 100 bp and then biotinylated with TdT as above. A 10 μg aliquot of biotinylated cDNA or genomic DNA was then hybridized to an *E. coli* array (Affymetrix, Santa Clara, CA) at 45°C for 40 h, washed, and stained with streptavidin–phycoerythrin (Molecular Probes, Eugene, OR). Arrays used for expression analysis are denoted "antisense" by Affymetrix because they contain probes that will bind to the reverse complement of the transcript, for example cDNA, whereas "sense" arrays (Affymetrix part no. 900284) will bind to the transcripts themselves. Antisense arrays are not yet commercially available. It should be noted, however, that the commercially available sense chips can be used to analyze both strands: Affymetrix's RNA labeling protocol can be used for expression analysis, and our cDNA labeling protocol for reverse complement analysis. In this article, we refer to antisense arrays as "expression arrays" and sense as "reverse complement arrays." Most arrays were scanned after a single staining, but one stationary-phase array and the reverse complement array were signal-amplified with a biotinylated anti-streptavidin antibody, followed by a second streptavidin–phycoerythrin staining, according to standard Affymetrix protocols. This amplification increased the signal/noise ratio about two- to threefold, but did not result in a significant increase in the number of transcripts detected. The array was then scanned by a HP-Affymetrix array scanner.

Data processing and normalization. Background was determined using GeneChip 3.2, which divides the array into 16 sectors and takes the average of the lowest 2% of features of each sector. After background subtraction, MM features were subtracted from PM features, and the resulting difference was multiplied by a scaling factor derived from GeneChip software. For the analysis of spiked control RNAs the scaling factor was derived from setting the 16S ribosomal mean average differences to 50,000. For the log- versus stationary-phase analysis, intensities were scaled so that the mean average difference for all probe sets was 5,000 units. All array analyses after the derivation of background and scaling factors were done with a set of Perl scripts that we have dubbed "Genome Array Processing Software" or "GAPS". GAPS takes ".CEL" files, generated by GeneChip, as input. GAPS and the .CEL files used in this study can be found at ExpressDB (ref. 22).

The array contains a regularly spaced 10 × 10 grid of control feature pairs that all hybridize to the same control oligonucleotide, and should thus be of equal intensity. However, we found that fluorescence intensity of these features typically varied about two- to threefold across the surface of the array, possibly because of local differences in washing/staining efficiencies. To correct for this spatial variation, the control grid was used to estimate local deviations in fluorescence intensity. First, each pair of controls were averaged. Then experimental features were multiplied by a correction factor that is derived from control features representing the relative brightness of the region. Control features closer to the probe pair contributed more to the final correction factor than distant ones. This correction factor was determined by the following equation:

$$\text{Correction factor} = \frac{\bar{c}}{\sum_{i=1}^{4}\left(\frac{\frac{1}{d_i}}{\sum_{j=1}^{4}\frac{1}{d_j}}\right)c_i}$$

where $d_{i \text{ or } j}$ is the Euclidean distance from the PM feature to the four closest control features, c_i is the intensity of control feature i, and $\bar{c}$ is the mean of all control features on the array.

RNA abundance metrics: average difference and 2max. Five control RNAs from *B. subtilis* that have four probe sets each on the array were analyzed at concentrations ranging from ~20 to ~0.0002 copies/cell, and no RNA, which served as a negative control. Each control probe set contained 15 probe pairs. Control RNAs were spiked into total cellular RNA before labeling. A total of 100 independent pairwise comparisons were made. Copies/cell was estimated by assuming cells have ~60 fg of total RNA (ref. 43). Copies per cell can be recalculated for different total RNA contents, which normally ranges from 20 to 200 fg/cell. For example, 1 copy per cell in a cell with 60 fg of total RNA is equivalent to 2 copies per cell in a cell with 120. The average transcript size of our spiked RNAs was 4.6 kb. Probe pairs were averaged over duplicates and then ranked by their mean intensity difference (PM – MM). The total intensity normalized values reported in the tables and the online data file are ~90% of the ribosomal normalized values of Figure 2A. The relationship between fluorescent signal and copies per cell is given by the equations of the regression lines of Figure 2A:

$$\text{2max Signal} = 13{,}000 \times \ln(\text{copies/cell}) + 39{,}000,\ R^2 = 0.76$$
$$\text{Median signal} = 5{,}500 \times \ln(\text{copies/cell}) + 16{,}000,\ R^2 = 0.80$$
$$\text{Average difference signal} = 6{,}000 \times \ln(\text{copies/cell}) + 18{,}000,\ R^2 = 0.86$$

Conversions from fluorescence intensity and copies per cell should be used with extreme caution. In addition to cell size issues noted above, there is a significant amount of error introduced by the large variability of probe signal, such that probes whose target RNA is present at equal concentration will have variable raw fluorescence intensity (see Fig. 2A). Experiments are in progress to use a hybridization of genomic DNA (where all genes are equimolar) to calibrate this conversion and allow more accurate measurement of absolute RNA levels. For the purposes of this study, we focus on the change in fluorescence of *identical* probe sets (thus bypassing inherent variability between *different* probe sets) and report "absolute change" and "fold change" (Tables 1, 2) rather than absolute RNA levels.

We found that by using the intensity difference of the second maximal probe pair to represent a probe set we maximized the number of detected genes. We therefore chose the second maximal probe pair intensity difference "2max" as a measure of RNA abundance. Using Excel, an exponential trend line was fit to a plot of observed versus expected fold change, and the equation was used to calibrate estimates of fold change in our stationary versus log expression comparison (Fig. 2B). The calibration equation is as follows: calibrated fold change = 1.2 × (measured fold change)$^{1.9}$. Pairwise comparisons of the 2max of the same probe sets on duplicate arrays yielded an average linear correlation coefficient of 0.85 ± 0.04.

Transcript detection. To determine which transcripts were detected, we used a set of four distinct *B. subtilis* probe sets for which the target RNA was not used in our spiking experiments. After normalization to total intensity, we determined the average 2max of these probe sets on the arrays used in the stationary versus log comparison. Transcripts were considered detected if their 2max was at least 3 standard deviations above the mean of the four probe sets for the absent *B. subtilis* RNA. We detected 97% and 87% of transcripts in stationary and log phase, respectively. We were unable to detect 1.7% of the transcripts in either condition. Because the negative controls were used to determine the detection threshold, they could not be used to estimate false positives. The false positive rates for the 2max and median metrics, therefore, were estimated by

using probe sets for which the RNAs were spiked at 0.004 copies/cell or less, well below the sensitivity of the assay. These metrics both yielded a false positive rate of 0% (0/20) by this method. For the AD metric, detection is decided by Affymetrix's calling algorithm that works independently of internal negative controls. We therefore used the negative controls to estimate the false positive rate, which was also 0% (0/15). The parameters used in Affymetrix's software package, GeneChip 3.2, were the following: SDT multiplier = 4, ratio threshold = 1.5, ratio limit = 10, horizontal zones = 4, vertical zones = 4, percentage background cells = 2, positive/negative min = 3, positive/negative max = 4, positive ratio minimum = 0.33, positive ratio maximum = 0.43, average log ratio minimum = 0.9, average log ratio maximum = 1.3.

It is important to note that 2max does not detect the maximal number of transcripts in every experiment. The maximum number of transcripts (4,033) on the reverse complement array was detected using the fourth brightest probe pair, or "4max." Averaging the fourth through eighth ranks "4–8max," which represents the peak of detection, gave 3,470 detected transcripts (78% of predicted RNAs). In this case 20 *B. subtilis* probe sets were used as negative controls, with a detection cutoff of 3 standard deviations above the mean. Widespread detection of transcription in *E. coli* with a reverse complement array has been confirmed in our lab on an independent RNA sample using the current Affymetrix labeling protocol in which 4,344 transcripts were detected (99% of predicted RNAs) using 4–8max (Daniel Janse, personal communication). The agreement is particularly striking considering the many differences between our original experiment and the confirmation experiment, which were, respectively: biotinylated total cDNA versus mRNA-enriched biotinylated RNA, antisense versus sense chip, and stationary-phase versus log-phase RNA samples. Both protocols include a DNaseI digestion to remove genomic DNA, and no genomic DNA contamination was detected by ethidium bromide staining.

Significance of changes. To determine which changes in 2max were significant, we devised a calling algorithm that uses both a *t*-test and a consensus measure. If either of the following criteria are fulfilled for transcripts that were detected in at least one condition, the transcript is called significantly changed: (1) mean 2max from duplicates is determined to be significantly different in the two conditions by a two-tailed Student's *t*-test with >95% confidence, or (2) after discarding the brightest and dimmest probe pairs, at least 11/13 of the remaining probe pairs are all changed in the same direction, by any amount. For transcripts with >15 probe pairs, the 15 brightest were identified and processed in the same way as the other probe sets. In the rare cases in which these two criteria conflicted, the decision based on the second maximal probe pair was used. It is important to note that the magnitude of the fold or absolute changes are not considered in deciding their significance, although 77% of the significant changes were greater than twofold. The number of differentially expressed transcripts is likely to increase with increasing experimental precision. The limits of biological significance may be set by subtle population genetic effects. Out of 100 independent pairwise comparisons of differentially spiked control RNAs, there were 52 in which at least one member of the pair was detected. The algorithm correctly assigned significant changes to all 52 of these probe sets, all of which had fold changes of at least twofold. Probe sets for control RNAs spiked at equal concentrations showed no significant changes (0/16).

Acknowledgments

We thank J. Edwards for improvements to the labeling protocol, D. Janse for sharing unpublished data, A. Derti and A. Petti for bioinformatics contributions, F. Lam for help with the fermentor, M. Mittmann for array design, P. Juels for impeccable computer tech support, W. Rindone and J. Aach for expression database support, B. Cohen, R. Mitra, M. Bulyk, P. Estep, M. Steffen, and the rest of the Church lab for the many helpful discussions and encouragement that made this work possible. We also thank the reviewers for significant improvements to the manuscript. This work was supported by grants from Aventis Pharma, Lipper Foundation, DOE, and NSF.

1. de Saizieu, A. et al. Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nat. Biotechnol.* **16**, 45–48 (1998).
2. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680 (1996).
3. Wodicka, L. et al. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat. Biotechnol.* **15**, 1359–1367 (1997).
4. Lee, C.K. et al. Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**, 1390–1393 (1999).
5. Zhu, H. et al. Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **95**, 14470–14475 (1998).
6. Wen, X. et al. Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA* **95**, 334–339 (1998).
7. Roth, F.P. et al. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945 (1998).
8. Tavazoie, S. et al. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
9. Eisen, M.B. et al. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
10. Tamayo, P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912 (1999).
11. Richmond, C.S. et al. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Res.* **27**, 3821–3835 (1999).
12. Tao, H. et al. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.* **181**, 6425–6440 (1999).
13. Chuang, S.E., Daniels, D.L. & Blattner, F.R. Global regulation of gene expression in *Escherichia coli*. *J. Bacteriol.* **175**, 2026–2036 (1993).
14. Pease, A.C. et al. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* **91**, 5022–5026 (1994).
15. Blattner, F.R. et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474 (1997).
16. Hengge-Aronis, R. Regulation of gene expression during entry into stationary phase. In Escherichia coli *and* Salmonella*: cellular and molecular biology*. (eds Neidhardt, F.C. et al.) 1497–1512 (ASM Press, Washington D.C.; 1996).
17. Yuzawa, H. et al. Heat induction of sigma 32 synthesis mediated by mRNA secondary structure: a primary step of the heat shock response in *Escherichia coli*. *Nucleic Acids Res.* **21**, 5449–5455 (1993).
18. Lange, R. & Hengge-Aronis, R. The cellular concentration of the sigma S subunit of RNA polymerase in *Escherichia coli* is controlled at the levels of transcription, translation, and protein stability. *Genes Dev.* **8**, 1600–1612 (1994).
19. *E. coli* genome project. http://www.genome.wisc.edu
20. Link, A.J., Robison, K. & Church, G.M. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18**, 1259–1313 (1997).
21. Aach, J., Rindone, W. & Church, G.M. Systematic management and analysis of yeast gene expression data. *Genome Res.* **10**, 431–445 (2000).
22. ExpressDB. http://arep.med.harvard.edu/cgi-bin/ExpressDBecoli/EXDStart
23. Liu, M.Y. et al. The RNA molecule CsrB binds to the global regulatory protein CsrA and antagonizes its activity in *Escherichia coli*. *J. Biol. Chem.* **272**, 17502–17510 (1997).
24. Southern, E.M., Milner, N. & Mir, K.U. Discovering antisense reagents by hybridization of RNA to oligonucleotide arrays. *Ciba Found. Symp.* **209**, 38–44 (1997).
25. Chen, L.H. et al. Structure and function of a bacterial mRNA stabilizer: analysis of the 5′ untranslated region of *ompA* mRNA. *J. Bacteriol.* **173**, 4578–4586 (1991).
26. SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* **95**, 1460–1465 (1998).
27. mfold. http://mfold2.wustl.edu/~mfold/dna/form1.cgi
28. Mir, K.U. & Southern, E.M. Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat. Biotechnol.* **17**, 788–792 (1999).
29. Taljanidisz, J., Karnik, P. & Sarkar, N. Messenger ribonucleic acid for the lipoprotein of the *Escherichia coli* outer membrane is polyadenylated. *J. Mol. Biol.* **193**, 507–515 (1987).
30. Cao, G.J. & N. Sarkar. Poly(A) RNA in *Escherichia coli*: nucleotide sequence at the junction of the *lpp* transcript and the polyadenylate moiety. *Proc. Natl. Acad. Sci. USA* **89**, 7546–7550 (1992).
31. Portier, C. & Regnier, P. Expression of the *rpsO* and *pnp* genes: structural analysis of a DNA fragment carrying their control regions. *Nucleic Acids Res.* **12**, 6091–6102 (1984).
32. Portier, C. et al. The first step in the functional inactivation of the *Escherichia coli* polynucleotide phosphorylase messenger is a ribonuclease III processing at the 5′ end. *EMBO J.* **6**, 2165–2170 (1987).
33. Regnier, P. & Hajnsdorf, E. Decay of mRNA encoding ribosomal protein S15 of *Escherichia coli* is initiated by an RNase E-dependent endonucleolytic cleavage that removes the 3′ stabilizing stem and loop structure. *J. Mol. Biol.* **217**, 283–292 (1991).
34. Pennisi, E. Are sequencers ready to 'annotate' the human genome? *Science* **287**, 2183 (2000).
35. DeRisi, J. et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**, 457–460 (1996).
36. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
37. 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).
38. Dedon, P.C. et al. A simplified formaldehyde fixation and immunoprecipitation technique for studying protein–DNA interactions. *Anal. Biochem.* **197**, 83–90 (1991).
39. Orlando, V. & Paro, R. Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell* **75**, 1187–1198 (1993).
40. Bulyk, M.L. et al. Quantifying DNA–protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.* **17**, 573–577 (1999).
41. Winzeler, E.A. et al. Whole genome genetic-typing in yeast using high-density oligonucleotide arrays. *Parasitology* **118**, S73–S80 (1999).
42. Gingeras, T.R. et al. Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. *Genome Res.* **8**, 435–448 (1998).
43. Neidhardt, F.C., Ingraham, J.L. & Schaechter, M. *Physiology of the bacterial cell: a molecular approach*. (Sinauer Associates, Sunderland, MA; 1990).