Published in final edited form as:

Methods. 2009 January ; 47(1): 63–72. doi:10.1016/j.ymeth.2008.10.017.

Technical Considerations in using DNA Microarrays to Define Regulons

Virgil A. Rhodius^{1,*} and Joseph T. Wade²

¹ Department of Microbiology and Immunology, University of California at San Francisco, San Francisco, California CA 94143, USA

² Wadsworth Center, NY State Dept. of Health, 120 New Scotland Ave., Albany, NY 12208, USA

Abstract

Transcription is the major regulatory target of gene expression in bacteria, and is controlled by many regulatory proteins and RNAs. Microarrays are a powerful tool to study the regulation of transcription on a genomic scale. Here we describe the use of transcription profiling and ChIP-chip to study transcriptional regulation in bacteria. Transcription profiling determines the outcome of regulatory events whereas ChIP-chip identifies the protein-DNA interactions that determine these events. Together they can provide detailed information on transcriptional regulatory systems.

Keywords

ChIP-chip; DNA microarray; DNA binding site; Gene expression; Transcription; Transcription profiling; Transcription factor; Sigma factor; Regulon

1 Introduction

Gene expression is a highly regulated process. In bacteria, the major target of regulation is transcription. Transcription is regulated by proteins and RNAs according to the growth conditions of the cell. Much work has focused on the regulation of transcription of individual genes. With the advent of microarray technology it is now possible to study transcriptional regulation on a genomic scale. Here we describe the application of two microarray-based techniques, transcription profiling and ChIP-chip, to the study of transcriptional regulation *in vivo*.

Transcription profiling uses DNA microarrays to measure differences in RNA levels between strains grown under different conditions and/or between strains with different genotypes. These differences are used to infer mechanisms of transcriptional regulation. ChIP-chip combines chromatin immunoprecipitation (ChIP) with DNA microarrays to determine the genome-wide association of a DNA-binding protein. The most common application of ChIP-chip is to determine the genome-wide association of a transcription factor; This distribution is used to infer its regulatory targets.

*Corresponding author. Fax: +1 415 514 4080. E-mail address: virgil.rhodius@ucsf.edu (V. Rhodius).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

2 Identifying Regulons

2.1 Transcription networks

A regulon is a group of genes or transcription units that are controlled by a single regulatory protein. This can be a transcription factor or sigma factor (the regulon of a sigma factor is also referred to as a “sigmulon”). The transcription network of an organism is comprised of many overlapping regulons. Such networks contain a few global regulators that have large regulons and regulate many target genes, and many regulators with small regulons that regulate one or a few genes. Many genes are co-regulated by several transcription factors, forming overlapping regulons. This also means that not all genes within a regulon are coherently regulated under different environmental conditions. A stimulon is the transcriptional response of an organism to an environmental perturbation such as heat shock or starvation. These responses can be complex, often involving multiple transcription factors and regulons, with cascades of gene expression both during and after the perturbation.

2.2 Identifying regulons using transcription profiling

To identify transcription factor regulons using transcription profiling, RNA is purified from two strains, reverse transcribed, and the cDNA is labeled and hybridized to a DNA microarray. Thus, RNA levels from one strain can be compared to RNA levels in the other strain. It is important to use growth conditions where the transcription factor is present and induced in its active DNA binding and regulatory form. Note that the two strains should have comparable growth rates and be compared at similar growth states to reduce any indirect effects on gene expression. To identify regulons using transcription profiling, the expression profile of strains carrying the active form of the regulator is compared with strains without the regulator or carrying a less active form. This is achieved by comparing either: (i) a wild type strain *vs.* a null mutant strain for the regulator, both grown under inducing conditions; or (ii) a null mutant or strain with little active regulator *vs.* a strain that overproduces or has a mutant form of the regulator that is constitutively active. “Steady-state” experiments are where the transcription factor is constitutively active in one of the strains either due to growth conditions, expression levels, or the use of constitutive mutants. These experiments are less than ideal, since constitutive expression of the target regulon may result in cascades of gene expression involving other transcription factors, thereby obscuring or even muting the expression of the original target genes. The optimal approach is to induce the activity the transcription factor at a particular time point, either by environmental perturbation and/or over-expression, and then monitor the changes in expression patterns using a time course. Genes that quickly change their expression levels are likely to be directly regulated, whereas those that change more slowly are more likely to be indirectly regulated. Note that some regulators at high cellular levels may bind DNA cooperatively or bind to weak physiologically irrelevant sites resulting in aberrant regulation of large numbers of genes.

Since some genes are regulated by multiple transcription factors, activation of one regulator may not be sufficient to alter expression levels of the target gene without appropriate inputs from the other regulator(s), which are likely to be active only under certain environmental conditions. Consequently, it can be difficult to identify the complete regulon of a transcription factor without employing a variety of environmental conditions to target co-regulated genes.

2.3 Identifying regulons using ChIP-chip

To identify transcription factor regulons using ChIP-chip, cells are crosslinked with formaldehyde, lysed, and sonicated to fragment DNA. The protein of interest is then immunoprecipitated, thus enriching for this protein and any DNA to which it is crosslinked. The enriched DNA is labeled and hybridized to a DNA microarray. Typically, enriched DNA levels are compared to a genomic DNA or a mock immunoprecipitation (IP) control. In most

cases the control sample is generated from genomic DNA purified from the sonicated, crosslinked cell extracts used for ChIP.

ChIP-chip does not require a comparison between wt and mutant cells. Therefore, ChIP-chip can usually be performed in a wt strain under growth conditions optimal for the function of the protein of interest. ChIP-chip requires an antibody that is specific to the protein of interest. Often such an antibody is not available. In these cases it is usually possible to epitope tag the protein at either the N- or C-terminus, and use an antibody specific to the tag. Epitope tagging can interfere with protein function, although this is generally not the case as most essential proteins can be tagged without affecting the growth rate of the cell (1,2). If possible, however, protein function should be confirmed by phenotype or functional assay. Epitope tagging is a relatively straightforward procedure in many bacterial species, thanks in part to recent advances in recombineering (3–5). A collection of >1,000 epitope-tagged strains is now available for *Escherichia coli* (2).

2.4 Microarray design

The early microarray platforms consisted of PCR (polymerase chain reaction) products of whole ORFs (open reading frames). This has transitioned into the use of single or multiple oligonucleotide probes for each ORF, to high density tiled arrays that have many overlapping probes that cover the entire genome on both strands, both in ORFs and intergenic regions. Recent advances in microarray synthesis mean that it is now possible to purchase custom-designed microarrays. This allows for customization of microarrays to any genome and any application. For ChIP-chip, greater probe densities improve the resolution when identifying the target DNA binding sites. For transcription profiling, most platforms are sufficient. Oligonucleotide probes are generally superior to whole ORF PCR products since they can be designed to reduce cross hybridization between homologous transcripts. In addition, multiple probes per ORF provide an averaged signal for each ORF; this is more reliable and less sensitive to any potential hybridization or secondary structure issues of individual probes.

High density tiled arrays, especially those that cover intergenic regions, can also provide information of the 5' and 3' ends of transcripts, and whether transcripts are contiguous between adjacent genes in the same orientation (6,7). This information aids computational location of promoters and can indicate whether adjacent genes are co-transcribed. High density tiled arrays are also essential for identifying small RNAs (50 to 500 nt in length), many of which are encoded in intergenic regions or antisense to protein-coding genes (8,9).

Microarrays are costly, but it is now possible to purchase custom-designed microarrays that contain multiple copies of the same set of probes, e.g. one microarray that is divided into 4, with each quadrant containing the same set of probes. Individual samples can be hybridized to each section of the microarray. Thus, a single microarray can be used for multiple experiments. The small size of bacterial genomes means that the probe density of these multiplex microarrays is sufficient to generate high quality data for both transcription profiling and ChIP-chip. Table 1 lists some common microarray platforms and suppliers.

3 Comparing transcription profiling arrays and ChIP-chip

3.1 Caveats of expression arrays and ChIP-chip

Transcription profiling and ChIP-chip can both be used to determine the genome-wide targets of a transcription factor. However, there are a number of important differences between the datasets generated by the two techniques:

1. If the transcription factor binds to the promoter of an operon and regulates transcription of all genes within the operon, ChIP-chip will identify only the site of

transcription factor binding. Transcription profiling, on the other hand, will identify changes in RNA levels for all genes within the operon (10–12).

2. Transcription profiling determines all changes in RNA levels regardless of whether those changes are directly due to the transcription factor of interest or are as a result of pleiotropic effects. Pleiotropic effects can be reduced by rapidly inducing transcription factor expression and identifying only those genes whose RNA levels change directly afterwards.
3. Transcription profiling cannot distinguish between changes in transcription and changes in RNA stability. ChIP-chip, on the other hand, provides no information on either transcription or RNA stability. Indeed, in some cases, transcription factor binding, as determined by ChIP-chip, may not be associated with changes in RNA levels, and hence it is impossible to assign a function to these sites without further investigation.
4. The magnitude of changes in RNA levels, as measured by transcription profiling, is somewhat dependent on the absolute RNA level, e.g. an increase in RNA level of 10 arbitrary units will result in a large fold-change if the initial absolute RNA level is 1 as opposed to 100. ChIP-chip, on the other hand, cannot determine changes in RNA levels although the level of binding that is measured is an absolute number.
5. For ChIP-chip, the target of regulation for a transcription factor must be inferred from the genomic location of binding. For some binding sites, e.g. those located between divergently transcribed genes, it is not possible to predict which gene (if any) is regulated by the transcription factor without further investigation.

3.2 Case study comparing different methods to identify the *E.coli* σ^{32} regulon

Often multiple approaches are required to identify transcription factor regulons. Figure 1 illustrates a comparison between transcription profiling, ChIP-chip and promoter mapping experiments designed to identify the regulon of the *E. coli* heat shock factor, σ^{32} (10,11). Transcription profiling identified 68 significantly induced transcription units (TUs) upon over-expressing σ^{32} ; this data was used as a guide to experimentally confirm existing known σ^{32} -dependent promoters, as well as identifying new promoters, to give a total of 50 confirmed promoters (10) (Figure 1). ChIP-chip against σ^{32} identified 87 significant peaks after exposing *E. coli* cells to heat shock. Note that not all of the ChIP-chip peaks have been experimentally tested for functional σ^{32} promoters. Both transcription profiling and ChIP-chip identified TUs or peaks that correlated with 37/50 confirmed promoters. Interestingly, 12 σ^{32} promoters were unique to the transcription profiling data and 43 ChIP-chip peaks did not match any of the confirmed promoters or TUs from expression profiling. It is still possible that some of these 43 peaks identify functional promoters: some of these discrepancies between the datasets may instead reflect differences in experimental design, for example: 1) different methods were employed to induce σ^{32} : some promoters may only be functional in certain environmental conditions due to the presence or absence of transcription factors; 2) the ChIP-chip study used a high density microarray platform enabling detection of non-annotated transcripts such as small RNAs, whereas the transcription profiling experiments used whole ORF PCR products as probes, thereby restricting analysis to known genes; 3) genes weakly regulated by σ^{32} but also have high basal expression rates from another promoter will not show large changes in mRNA levels upon σ^{32} induction, therefore may not be reliably detected using transcription profiling (in fact, some of these genes are weakly induced in the transcription profiling experiment, but are not statistically significant). 9 TUs from the transcription profiling with no confirmed promoters also contained ChIP-chip peaks upstream (Figure 1), suggesting that they may contain σ^{32} promoters that were missed during experimental validation. Interestingly, 3 of these TUs contain ChIP-chip peaks within the coding region, suggesting internal

promoters. One experimentally confirmed σ^{32} promoter was missed by both methods, and 12 TUs identified by transcription profiling contained no confirmed promoters or ChIP-chip peaks (Figure 1) suggesting that these genes were indirectly regulated after over-expressing σ^{32} .

The comparison of these studies illustrate that currently no single approach may completely and exclusively identify regulon members for a particular transcription factor. Combining the two approaches increases the number of target promoters identified and also allows for accurate identification of both the binding sites for σ^{32} and the genes whose transcription is induced by σ^{32} .

4 Transcription profiling methods

Transcription profiling using microarrays is well established. Hence, there are a variety of protocols available for RNA purification, cDNA synthesis, labeling, and sample hybridization. In addition, some of these protocols are specific for particular microarray and hybridization platforms. Rather than presenting all of these protocols, we highlight some of the main criteria that will affect choice of sample preparation, and present a common procedure developed and successfully used by many laboratories. For a more in depth discussion of the microarray procedure, see (13).

4.1 Sample Harvesting

For transcription profiling analysis, it is essential that the integrity of the RNA and its profile (i.e. relative amounts of each transcript) is maintained during sample harvesting and RNA purification. The process of harvesting bacterial cells (*e.g.* centrifuging a bacterial culture) will often stress the cells, hence altering their expression profile, and can also result in differential degradation of RNA. Consequently it is essential to “freeze” the RNA profile during sample harvesting. This is achieved by mixing the cells with a reagent designed to inactivate transcription and prevent RNA degradation; *e.g.* RNeasy Protect (QIAGEN), or a solution of 40% methanol, 62.5 mM HEPES, pH 6.5 at -45°C (14). We use a 5% acid phenol in ethanol stop solution as outlined below:

1. Transfer 10 ml of culture ($0.3\text{--}1 \times 10^9$ cells/ml) into a 15-ml conical tube containing 1.25 ml of ice-cold ethanol/phenol stop solution.
2. Harvest cells by centrifugation at 6,700 *g* for 2 minutes at 4°C . Remove media by aspiration.
3. Rapidly freeze cell pellet in liquid nitrogen. Store at -80°C until required.

4.2 Total RNA preparation. Estimated length of procedure: 2 hours

There are multiple RNA purification methods and kits available, choosing between them will depend primarily on the properties of your strain (growth conditions, ease of lysis, and level of RNases), but also on experimental design and quantity of available sample vs yield of RNA required for each array experiment. Consequently, it may be necessary to modify protocols to enhance lysis or to cope with high levels of endogenous RNases. It is also important to note that some kit protocols that have affinity purification steps may not give representative yields of small RNAs. Total RNA from bacterial cultures can be isolated using the hot phenol method outlined below, or by using several commercial reagents (*e.g.* TRIzol, Invitrogen), or kits (*e.g.* RNeasy, QIAGEN; RiboPure, Ambion). Also, there are protocols and kits available that: 1) enrich mRNA from total RNA preparations by using oligos that bind to 16S and 23S rRNA, enabling these RNAs to be subtracted from the RNA prep (*e.g.* MICROBExpress, Ambion); 2) purify bacterial RNA from complex host-bacterial samples using a similar oligo subtraction method to remove contaminating rRNA and poly-A mRNAs from select eukaryote hosts (*e.g.* MICROBEnrich, Ambion); and 3) amplify RNA from low yield samples using a linear cDNA

amplification step (15). We find that for *E. coli* cultures the hot phenol method outlined below yields the best quality RNA preps in terms of RNA integrity, purity and yield.

1. Resuspend the cell pellet (from 10 ml of culture) in 800 μ l lysozyme solution. Transfer lysate to 2-ml microfuge tube containing 80 μ l of 10% SDS, mix by inversion and incubate at 64 °C for 2 min.
2. Add 88 μ l 1 M sodium acetate solution (pH 5.2) and mix by inversion.
3. To the lysate add an equal volume (~1 ml) of H₂O-saturated phenol (pH <7.0). Mix by inverting 10 times. Incubate in a 64 °C water bath for 6 minutes, continuing to mix the tube contents by inverting every 40–60 seconds.
4. Place tube on ice to chill for 2 minutes. Afterwards, centrifuge at 16,000 *g* for 10 min at 4 °C.
5. Remove the upper aqueous phase into a fresh tube, taking care not to disturb the interface (this is a common point of RNase contamination of preps). Also, perform this step quickly since the aqueous layer can rapidly become cloudy after centrifugation, making it difficult to separate the layers. If this happens, re-centrifuge the sample.
6. Add to the solution an equal volume (~ 1 ml) of 1:1 mix of H₂O-saturated phenol:chloroform. Invert the tube 6–10 times to mix and centrifuge at 16,000 *g* for 2 min.
7. Carefully remove the upper aqueous phase to a fresh microfuge tube. Repeat the H₂O-saturated phenol:chloroform extractions until the interface is clear (usually $\geq$ 2–3-times). Some strains may require extensive phenol:chloroform extractions to completely remove contaminating RNases.
8. Divide the final extracted solution equally between two 1.5-ml microfuge tubes. Precipitate by adding 0.1 volume 3 M sodium acetate (pH 5.5) and 2.5 volumes 100% cold ethanol. Incubate at –80 °C for 30 min.
9. Recover the RNA by centrifugation at 16,000 *g* for 30 min at 4 °C.
10. Wash the RNA pellet with 1 ml 80% cold ethanol. Centrifuge at 16,000 *g* for 5 min at 4 °C. Carefully remove the ethanol solution by aspiration and dry the RNA pellet in a speed vacuum.
11. Redissolve and pool each pair of pellets in a final volume of 87 μ l and place in a fresh 1.5 ml microfuge tube.

4.3 DNase treatment. Estimated length of procedure: 1 hour

1. To each RNA preparation (87 μ l) add 10 μ l 10 $\times$ TURBO DNase Buffer and 3 μ l TURBO DNase. Incubate reaction at 37 °C for 30 min.
2. Add an additional 3 μ l TURBO DNase and incubate a further 30 min at 37°C.
3. Add 10 μ l DNase Inactivation Reagent and incubate at room temp, mixing 4 times.
4. Centrifuge at 10,000 *g* for 1.5 min and transfer the supernatant containing the RNA to a fresh tube. Store at –20 °C until required.

4.4 Assessing RNA quality and yield

1. Determine the concentration of the RNA by measuring the absorbance of a 1:100 dilution in H₂O at 260 nm (concentration, *c* (μ g/ μ l), in a 1 ml quartz cuvette with a 1 cm path length: $c = A_{260} \times f \times 0.04 \mu\text{g}/\mu\text{l}$, where *f* is the dilution factor). Typical yields

are 70–300 µg RNA from 10 ml of culture, depending on strain of *E. coli*, growth conditions and culture density upon harvesting.

2. Check purity by measuring the absorbance ratio (A_{260}/A_{280}) of a 1:100 dilution in 10 mM Tris-HCl buffer, pH 7.5 (note that the absorbance ratio is sensitive to pH: since RNA is acidic the ratio must be measured in a low salt neutral buffer). Good RNA preps free from protein contamination give values between 1.8 and 2.1.
3. If required, the integrity of the RNA can be analyzed on a denaturing formaldehyde 1% agarose gel (16). Upon visualizing the gel, the 23S and 16S ribosomal RNA should be easily observed. For good RNA, the 23S species should be twice as intense as the 16S with little or no smearing between or below these bands.

4.5 cDNA synthesis. Estimated length of procedure: 3 hours

cDNA is synthesized using random octamer primers and a dNTP mix containing amino-allyl dUTP (aa-dUTP).

1. In 0.2 ml PCR tube, mix 20 µg total RNA with 12 µg random octamer primer to give a final volume of 15.5 µl. Incubate at 70 °C for 10 min and then chill on ice for 10 min.
2. cDNA synthesis reaction: Prepare a cocktail on ice containing for each reaction: 6 µl 5× First-Strand Synthesis Buffer; 1.2 µl 25× aa-dUTP/dNTP mix; 3 µl 0.1 M DTT, 0.75 µl SuperScript III RT; 3.55 µl H₂O. Add 14.5 µl cocktail to annealed RNA sample and incubate at 50 °C for 2 hours.
3. RNA is removed from the completed reverse transcription reaction by hydrolysis. To the sample add 10 µl 0.5 M EDTA and 10 µl 1 N NaOH and incubate for 15 min at 65 °C.
4. Neutralize the reaction by adding 50 µl 1M HEPES, pH 7 and mix well.

4.6 Cy3/Cy5 coupling. Estimated length of procedure: 2.5 hours

Free amines and unincorporated amino-allyl dUTP must be removed from the sample to enable successful coupling of the cDNA with the Cy3/Cy5 dyes. Consequently, each sample is cleaned using the QIAGEN MinElute Reaction Cleanup Kit. Note that the Cy dyes are shipped as a desiccant in sealed packs. They are extremely sensitive to light and moisture, therefore each pack is only opened and resuspended in DMSO immediately prior to use. If required, the efficiency of Cy labeling of the completed reaction can be determined spectroscopically (13).

1. Remove the reverse transcription reactions from the PCR tube to a 1.5 ml microfuge tube.
2. Add 300 µl Buffer ERC to each sample.
3. Load samples on to MinElute Columns and centrifuge at $\geq 10,000$ g for 1 min.
4. Discard the flow-through and add 750 µl Buffer PE to each column and centrifuge at $\geq 10,000$ g for 1 min.
5. Discard the flow-through and centrifuge again at $\geq 10,000$ g for 1 min.
6. Add 10 µl H₂O to the center of each column matrix, incubate for 1 min, and then elute using a fresh collection tube by centrifugation at $\geq 10,000$ g for 1 min.
7. To each cDNA sample, add 1 µl 1M sodium bicarbonate, pH 9.0 (note the bicarbonate becomes carbon dioxide with time, therefore use solution < 3 months old).
8. Resuspend each fresh tube of Cy3 or Cy5 in 17 µl DMSO.

9. Add 2 μ l of either Cy3 or Cy5 solution to each cDNA sample. Incubate for 2 hours at room temperature in the dark.
10. Unincorporated Cy dyes are removed using the QIAGEN MinElute Reaction Cleanup Kit following the procedure described in steps 2–6, this time eluting the sample in 13 μ l H₂O or 10 mM Tris-Cl pH 8.5.

4.7 Sample hybridization

Hybridization protocols and volumes will vary depending on the microarray slide and hybridization chamber. Below we give a sample hybridization mix used for our arrays; volumes can be changed as necessary whilst maintaining the correct final concentration of SSC, HEPES and SDS. Note all solutions must be filtered (*e.g.* with a 0.2 μ m filter) to prevent small particles damaging the surface of the array.

1. Combine one Cy3 and one Cy5 labeled sample for each hybridization (~24 μ l) in a 0.5 ml or 0.2 ml microfuge tube.
2. To each hybridization reaction add 16 μ l H₂O (filtered), 7.5 μ l 20 $\times$ SSC (filtered), 1.25 μ l 1 M HEPES pH 7.0 (filtered) and 1.25 μ l 10% SDS (filtered). This gives a final mix of 3 $\times$ SSC, 25 mM HEPES, 0.25% SDS.
3. Incubate reaction at 99°C for 2 min and allow to cool at room temperature for 5 min. Apply to surface of microarray and follow the hybridization instructions for your microarray and hybridization chamber.

4.8 Slide washing and scanning

Prior to scanning, hybridized slides are washed to remove any sample non-specifically bound to the slide surface. Since slide washing protocols will vary according to the manufacturer we give a washing protocol commonly used for oligo and ORF PCR arrays printed onto polylysine coated slides. Note that all wash stock solutions should be filtered before using. After washing the Cy dyes are extremely unstable: Cy 5 is rapidly degraded by ozone in minutes (17,18). Slides should be dried and scanned in a low ozone chamber; alternatively, some companies supply wash solutions that stabilize the dyes (*e.g.* Agilent).

1. Prepare the following wash solutions in glass slide dishes: 2 glass slide dishes each containing 500 ml Wash Solution I (897 ml Milli-Q-water, 100 ml 20 $\times$ SSC, 3 ml 10% SDS). Place an empty slide rack in one of the dishes. If using oligo arrays, Wash Solution I should be preheated to 60 °C and poured into the slide dishes immediately prior to washing the slides. This is essential to remove non-specific hybridization on oligo arrays; 2 glass slide dishes each containing 500 ml Wash Solution II (950 ml Milli-Q-water, 50 ml 20 $\times$ SSC); 1 glass slide dish containing 500 ml Wash Solution III (495 ml Milli-Q-water, 5 ml 20 $\times$ SSC).
2. Carefully remove slide from the hybridization chamber; keeping the array level, submerge into the slide dish containing Wash Solution I with no slide rack.
3. Once submerged, using fine forceps carefully remove the cover slip or mixer-assembly following the manufacturer's instructions, taking care not to scratch the surface of the array.
4. After removing the cover, place the array on the rack in the second slide dish containing Wash Solution I.
5. Repeat steps 2–4 for any other remaining slides. When finished, plunge the rack up and down 10–20 times.

6. Immediately transfer the slide rack to Wash Solution II and plunge up and down for 60 seconds.
7. Drain the rack for 5 seconds, and then place in the second dish containing Wash Solution II and plunge up and down for 60 seconds.
8. Drain rack for 5 seconds, and then transfer to Wash Solution III and plunge up and down for 60 seconds.
9. Dry the arrays by centrifugation at 600 rpm for 2 min in a low ozone chamber.
10. Scan the arrays as soon as possible in a low ozone chamber to reduce degradation of Cy5.

There are several scanners and software available for processing slides and the generated image files, some are specific to certain slide platforms (e.g. Affymetrix). One of the most popular systems that handles several different slide platforms is the GenePix scanners and software from Molecular Devices, and also SpotReader from Niles Scientific. Users should scan their slides following the manufacturer's instructions.

5 ChIP-chip methods

ChIP-chip first requires the preparation of "chromatin": crosslinked, sonicated cell lysates. Chromatin can then be immunoprecipitated with an antibody raised against the protein of interest. It may be necessary to use an altered lysis procedure for some bacterial species that are more resistant to lysozyme.

5.1 Preparing crosslinked, sonicated cell lysates. Estimated length of procedure: 1 hour, not including growing cultures

1. Grow cells to a suitable density and crosslink with formaldehyde (1% final concentration for 20 minutes at room temperature; 10^{10} cells is typical). Add glycine (0.5 M final concentration) to quench excess formaldehyde. Centrifuge cells and wash twice with TBS. Cell pellets can be stored at -20°C .
2. Resuspend cell pellets in 1 ml FA lysis buffer containing 2 mg/ml lysozyme (higher concentrations of lysozyme may be required for some bacterial species), and incubate for 30 minutes at 37°C . Significant lysis should be visible following this incubation.
3. Chill crosslinked lysates on ice for 5 minutes and sonicate for 2×30 s with a Branson Sonifier 450 at 50% output using a microtip probe (or equivalent device), chilling on ice for ≥ 2 min between rounds of sonication. It is also possible to sonicate with a cup-horn sonicator. This requires a much longer time of sonication (≥ 20 min).
4. Centrifuge sonicated, crosslinked lysates to remove unlysed cells and cell debris. Keep supernatant and store at -20°C . It may be necessary to dilute the chromatin in order to reduce the likelihood of protein precipitation. The final volume should minimally be 2 ml for every 10^{10} cells used.

It is important to confirm that the DNA fragments in the chromatin are a suitable size. Typically, fragments will range from 100 to 1000 bp, with an average of ~ 400 bp. Increased sonication will reduce fragment size up to a limit. It is generally not possible to sonicate chromatin samples too much. For ChIP with eukaryotic cells, micrococcal nuclease is sometimes used to reduce fragment size beyond that achievable with sonication alone. This is generally not required for bacterial ChIP experiments. To determine the size range of fragments, a small fraction of chromatin is decrosslinked by boiling for 10 minutes, phenol extracted and ethanol precipitated, RNase treated, and analyzed by agarose gel electrophoresis. The crosslinked, sonicated lysate

is then immunoprecipitated (see below). A portion of this (typically 10–20 μ l) can be kept as a genomic DNA control sample. Alternatively, a mock IP sample can serve as a control.

5.2 Immunoprecipitation of crosslinked, sonicated cell lysates. Estimated length of procedure: 3 hours

Note: this part of the ChIP procedure is the same for all species. Many variations have been described in the literature.

1. Incubate crosslinked, sonicated lysate equivalent to $\sim 10^9$ cells in a total volume of 800 μ l FA lysis buffer with 25 μ l 50% slurry protein A sepharose beads in TBS, and an appropriate amount of antibody, for 90 minutes at room temperature with gentle agitation. The optimal antibody concentration varies according to the source of the antibody and must be determined empirically. Some antibodies give a more efficient IP if this step is performed overnight at 4 $^{\circ}$ C.
2. Pellet beads by centrifugation at low speed (1,500 $\times$ g). Aspirate the supernatant and resuspend the beads in 800 μ l FA lysis buffer. Incubate for 3 minutes at room temperature and repeat the wash. Wash 3 additional times with the following buffers: FA lysis buffer + 500 mM NaCl, ChIP wash buffer, and TE. After the final wash, collect the beads and resuspend in 100 μ l ChIP elution buffer. Mix by pipetting gently and incubate at 65 $^{\circ}$ C for 10 min. Centrifuge and collect the supernatant. Washes are best performed in Spin-X columns. Some antibodies give a more efficient IP if FA lysis buffer is used in all 5 wash steps.
3. Decrosslink the supernatant by boiling for 10 minutes or incubating at 65 $^{\circ}$ C for 6 h. Purify DNA using a QIAquick PCR purification kit or equivalent. Use one or multiple samples for amplification and labeling for microarrays (depending on the method used; see below). A typical yield of DNA is 10–30 ng.

5.3 Validation of ChIP experiments prior to amplification of DNA for microarrays

It is advisable to confirm a successful ChIP experiment before amplification and labeling of DNA for microarrays. This can be done using quantitative PCR (qPCR) to determine the enrichment of known target regions for the protein of interest, as compared to control genomic regions that are expected to be unbound. Enrichment is calculated relative to a control sample that can be genomic DNA (prepared by decrosslinking cell extracts before the IP step) or a mock ChIP sample (e.g. no antibody included in IP). Enrichment above background of ≥ 5 -fold is likely to result in a successful ChIP-chip experiment. In some cases it may not be possible to validate the ChIP procedure if no target genomic regions are known.

5.4 Choice of amplification method for ChIP-chip

A variety of methods are available for amplification and labeling of ChIP samples for ChIP-chip. The four most used methods are (i) ligation-mediated PCR, (ii) random priming followed by PCR, (iii) T7 transcription followed by reverse transcription, and (iv) strand displacement amplification (SDA). Methods (i–iii) have been discussed in detail elsewhere (19), and each of these methods requires only a small amount of starting material (<10 ng in each case). Method (iv), SDA, uses the Klenow fragment of *E. coli* DNA polymerase to make multiple copies of the ChIP DNA fragments. Labeled nucleotide (or aa-dUTP which can later be coupled to Cy3 or Cy5) is incorporated during the amplification step. SDA results in ~ 10 -fold amplification, so relatively high levels of starting material are required. Typically, ≥ 4 ChIP samples (≥ 100 ng DNA) are required for a single amplification reaction. SDA is less error-prone than other amplification methods because the level of amplification is low and linear rather than exponential (20,21); it is also simpler to perform, but requires more starting material.

SDA can be performed using the BioPrime kit (Invitrogen) according to the manufacturer's instructions.

ChIP-chip compares an experimental sample to a control sample, either on a separate microarray (in the case of one-channel microarrays) or on the same microarray but labeled with a different dye. The control sample is usually either amplified genomic DNA, generated by decrosslinking a portion of the sonicated, crosslinked cell extract used for the IP, or a ChIP sample generated using a mock IP, e.g. with no antibody. Amplification and labeling of the control sample is the same as for the experimental sample. Hybridization of samples to microarrays and subsequent washing and scanning steps are identical to those for transcription profiling.

5.5 Interpretation of ChIP-chip data

ChIP generates fragments of ~400 bp on average. Hence, transcription factor binding to a <20 bp DNA site will result in ChIP signal up to 400 bp on either side of the binding site. The ChIP signal will decrease as the distance from the binding site increases, so a typical binding site will produce a peak ChIP signal with a width of ~800 bp (Figure 2). The center of this peak represents the location of the binding site. A common misconception is that the resolution of ChIP and ChIP-chip is limited by the DNA fragment size. Using a high density microarray, by identifying the center of the ChIP signal peaks it is possible to predict the position of transcription factor binding sites to a resolution of <50 bp. Comparison of the ChIP-chip signal peaks for σ^{32} with the known σ^{32} promoters reveals a median difference of 33 bp from the transcription start site (Figure 3) (10,11).

6 Analysis and validation of transcription profiling and ChIP-chip data

In order to reliably identify differentially expressed genes, it is essential to perform at least 3–4 independent biological replicates. It is insufficient to merely apply a fold cutoff to identify differentially expressed genes, since some genes may be inherently noisy due to low expression values. Most statistical programs for transcription profiling analysis employ the *t*-test to generate *p*-values from the replicated log intensity ratios to describe the likelihood of these values occurring by chance. However, whilst a *p*-value of 0.05 is highly significant for an individual gene/data point, the multiple testing scenarios of microarrays with, for example 4,000 data points, would result in 200 data points with a *p*-value ≤ 0.05 ; *i.e.* 200 false positives! Consequently it is necessary to control the false positive rate for microarray applications and to calculate an “adjusted” *p*-value to determine the significance level (see reviews (22–24)). The software Statistical Analysis of Microarrays (SAM; (25)) enables the user to determine the cutoff threshold for identifying significantly differentially expressed genes by controlling the false detection rate (FDR) and provides adjusted *p*-values (*q*-values) for each gene. This is extremely useful for the biologist since low or high stringency thresholds can be used, depending on the type of follow-up analysis to be performed on the dataset and the number of false positives the biologist is comfortable dealing with.

ChIP-chip data differs from transcription profiling data in that the regions of interest only have signal above the median value (one-tailed distribution). The unbound genomic regions that make up the majority of probes are all present at an equivalent level. Therefore, there are very few inherently noisy regions. ChIP-chip experiments are usually highly reproducible. Two independent biological replicates are often sufficient for high quality data, although three or more are desirable.

The simplest form of ChIP-chip data analysis is to rank order the data and select the genomic regions that score above a threshold value (26). ChIP-chip scores for non-target regions are expected to fit a roughly normal distribution so *p*-values can be estimated by determining the

difference between individual probe scores and the mean score. With multiple datasets it is possible to use multiple randomizations to estimate p -values or a FDR (11). A variation on this method is to assume that the distribution of values for non-target regions is symmetric. Since ChIP-chip data is one-tailed, the background distribution can be estimated using only the values below the mean (27). If a selection of genomic regions with a range of ChIP-chip values is tested by qPCR, a more accurate threshold value can be chosen. FDR and p -values can be estimated based on the known number of true and false positives (28).

Since ChIP-chip signal at target regions is a broad peak it is often useful to combine data from adjacent probes. This is especially true for high-density, tiled microarrays, where binding to a single target is likely to result in high ChIP-chip signal across multiple contiguous probes. In these cases, a powerful analysis method is to window the data such that values from several contiguous probes are averaged within a sliding window of fixed width that is moved across the genome, as illustrated in Figure 2 (9,29). This greatly reduces error from individual “rogue” probes. The ChIPOTle algorithm is an example of this approach (29). ChIPOTle is freely available as a Microsoft Excel Macro and requires very little technical expertise.

It is often important to validate some of the data from transcription profiling and ChIP-chip experiments. For transcription profiling, suitable validation methods include rtPCR (real time PCR) and Northern Blotting. For ChIP-chip, the best validation method is ChIP combined with rtPCR. *In vitro* binding assays, e.g. electromobility shift assays, can be used to validate ChIP-chip data, but it is important to note that transcription factors may often bind sites *in vivo* but not *in vitro*, for example, due to cooperative interactions with other proteins (30,31). Therefore, a negative result from an *in vitro* binding assay does not necessarily indicate that a predicted binding site is a false positive.

7 Computational methods of identifying transcription factor binding sites

A common use of transcription profiling and ChIP-chip data is to search for DNA sequence motifs that are enriched within the target regions. Several algorithms are available to search for conserved motifs from unaligned DNA sequences (Table 2). Both Motif Regressor (32) and W-AlignACE (33) are algorithms specifically designed for ChIP-chip and transcription profiling data. These approaches assume that sequences with high ChIP-chip scores or mRNA expression ratios contain strong and/or multiple transcription factor binding sites and are therefore likely to contain true motifs. Motif Regressor correlates statistically significant binding sites with ChIP-chip/gene expression scores, whilst W-AlignACE weights the input sequences according to their ChIP-chip/expression scores. In contrast, the other algorithms assume an equal probability of identifying sites from all of the input sequences. Below we highlight several strategies that improve the success of identifying specific sites.

7.1 Input data

The length of input sequences is the single most critical factor affecting prediction accuracy (34). The ability to detect correct motifs significantly decreases with sequence length, and also with the number of input sequences if some of these do not contain motifs. Therefore it is best to use short sequences that, based on the experimental data, are highly enriched for the correct motifs.

ChIP-chip data provides approximate genome coordinates for the location of the transcription factor, making it easier to search for conserved sites. The accuracy of site location depends predominantly on the resolution of the array. With high-density microarrays it is usually possible to accurately predict the position of a transcription factor binding site from ChIP-chip data alone (i.e. no sequence information), e.g. the median distance between σ^{32} binding site locations predicted from ChIP-chip data and the actual binding site is <50 bp (10,11). In

searching for sites, it is best to start with a small sequence window (e.g. 30 nt on either side of the predicted genome coordinate) and then repeat the searches gradually increasing the sequence window size.

Transcription profiling data identifies target-regulated genes, so the location of binding sites in the upstream regulatory regions has to be inferred. Rather than searching sequence windows upstream of every significantly regulated gene, it is better to organize the genes into possible transcription units (operons) and search upstream of every transcription unit, thereby restricting the search to likely promoter regions. Possible transcription units can be easily estimated by combining significantly regulated genes that are adjacent (separated by less than 50 – 100 nt), are in the same orientation, and have similar expression profiles (due to transcription attenuation sometimes genes towards the end of a transcription unit are more weakly expressed). Note that if a pair of adjacent genes are divergently transcribed and hence share the same intergenic region, the intergenic region should only be included once. The transcription start site of most promoters is within 100 nt upstream of the first transcribed gene (35), therefore it is usually sufficient to search a sequence window from 1 to 300 nt upstream of the first gene to identify most binding sites. Since these search windows are much larger than that of ChIP-chip data, it is still possible to miss or erroneously identify sites. Accuracy can be improved by incorporating transcription start site information, enabling a much smaller search window to be used (7). For high density tiled microarrays the approximate location of transcript 5' ends can be estimated from the individual probe intensities or ratios (6,7,9). Alternatively, for low density microarrays, some researchers map the 5' ends of transcripts using 5' RACE (10,36). Since transcription activators are located upstream of promoters; a window of 1 to 100 nt upstream of the transcription start point will identify most sites. Repressors, on the other hand, have a much broader distribution and can be upstream and downstream of the transcription start site. A window 100 nt upstream and downstream of the transcription start may be required. Sigma factors can be more precisely located using a window of 1 to 40 nt upstream of the transcription start, depending upon the precision of the start site data.

7.2 Motif width and orientation

Most of the DNA motif finding algorithms have options that control the motif search: *e.g.* motif size, palindromic sites, number of motifs per strand, search only one input strand or also the reverse complement. In addition, most programs provide statistics that describe the likelihood of output motifs occurring by random chance given the base composition of the input DNA sequences. The search results can be dramatically improved if use of the algorithm options is combined with some knowledge of the properties of the transcription factor binding site, and the output motifs assessed for their statistical significance. Most bacterial transcription factors are symmetric homodimers with a helix-turn-helix DNA binding domain, and hence bind to symmetric (palindromic) sites that cover two turns of the DNA helix. Hence, for unknown sites it is often best to start by searching one strand of each input DNA sequence for palindromic motifs between 16–24 nt wide. If there are no highly significant motifs, the site may be asymmetric. For ChIP-chip data, such sites will be present in both orientations, therefore search both strands. For transcription profiling data, these sites may be present in one or both orientations with respect to the promoter: therefore search just one strand and then try both strands. The binding sites of sigma factors and some transcription factors consist of two highly conserved motifs separated by a variable spacer. These are two-block motifs and ones with variable spacers are not captured well by most algorithms. However, BioProspector (37) is designed specifically for such sites and has input options to specify the width of each block and the variable spacing range.

7.3 Motif frequency

The motif finding algorithms request an estimate of the expected number of sites per input sequence. For ChIP-chip data there should be at least one site per sequence. For transcription profiling data, it is probable that some of the genes are indirectly regulated; hence choose one or less site per sequence. For some transcription factors, multiple sites per sequence may be appropriate if they are known to bind to multiple sites per promoter.

7.4 Background sequence composition

Motifs are identified based on the probability of their observed aligned nucleotide frequencies relative to a background model of nucleotide frequencies. The background model is derived either from the input DNA sequences, or by user input. Since the composition of DNA sequences can vary, the input sequences may not be representative of the genomic or intergenic regions. Consequently, the user can input frequencies, or upload a file that describes the background distribution of sequences for their genome. The appropriate choice of background model can help avoid spurious results, such as identification of polyA tracts that are common in intergenic regions.

7.5 Multiple search runs

The prediction of regulatory sites still remains a complex challenge. Consequently, the best approach is by cross-validation: identifying common motifs from several different algorithms. Likewise, for weak motifs, it is best to identify the top ranked hits from several algorithms than just the highest scoring motif from only one algorithm. Some algorithms such as AlignACE (38), Gibbs Sampler (39) and BioProspector (37) involve a stochastic process such that the output results may differ slightly from run to run. This is more apparent for noisy data and in these cases multiple runs are recommended to ensure robust results.

8 Analysis of target gene function

For most transcription profiling and ChIP-chip experiments it is important to determine whether functionally related groups of genes are regulated by the protein/condition being studied. This can be done by comparing the list of genes that are most highly bound or have the highest change in RNA level, with the list of all genes, and grouping the genes by functional category. Any functional category that is significantly over-represented or under-represented in the experimental list indicates a possible relationship between that function and the protein/condition being studied. This is known as Gene Ontology (GO) term analysis. GO term analysis can be performed manually or using freely available programs (e.g. DAVID, FuncAssociate) (40,41). Unfortunately, very few such programs currently exist for bacterial species.

9 Summary and Perspectives

Transcription profiling and ChIP-chip provide two different, but complementary methods for determining regulons. To date, transcription profiling and ChIP-chip have rarely been combined to study bacterial transcription factors. Equivalent work in yeast has provided important insights into transcriptional regulation (42,43). Bacterial genomes are relatively small and microarrays can now be synthesized at very high densities. Therefore, very high resolution transcription profiling and ChIP-chip experiments are now possible at an affordable price. Looking to the future, recently developed high-throughput sequencing methods can be used as an alternative to microarray-based approaches (44), although the cost of these methods is still very high.

Acknowledgements

We thank Emily Gogol for refining the RNA purification protocol. Protocols for transcription profiling using DNA microarrays were in part derived from Micorarray course notes developed by Adam Carroll and Paige Nittler at UCSF Centre for Advanced Technology and have been refined by many labs at UCSF. We thank Carol Gross for financial support and acknowledge funding from the National Institutes of Health (NIH), Grant GM57755.

References

1. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. *Nature* 2003;425:686–91. [PubMed: 14562095]
2. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N, Davey M, Parkinson J, Greenblatt J, Emili A. *Nature* 2005;433:531–37. [PubMed: 15690043]
3. Uzzau S, Figueroa-Bossi N, Rubino S, Bossi L. *Proc Natl Acad Sci USA* 2001;98:15264–69. [PubMed: 11742086]
4. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL. *Proc Natl Acad Sci USA* 2000;97:5978–83. [PubMed: 10811905]
5. Cho BK, Knight EM, Palsson BO. *Biotechniques* 2006;40:67–72. [PubMed: 16454042]
6. Tjaden B, Saxena RM, Stolyar S, Haynor DR, Kolker E, Rosenow C. *Nucleic Acids Res* 2002;30:3732–38. [PubMed: 12202758]
7. McGrath PT, Lee H, Zhang L, Iniesta AA, Hottes AK, Tan MH, Hillson NJ, Hu P, Shapiro L, McAdams HH. *Nat Biotechnol* 2007;25:584–92. [PubMed: 17401361]
8. Altuvia S. *Curr Opin Microbiol* 2007;10:257–61. [PubMed: 17553733]
9. Reppas NB, Wade JT, Church G, Struhl K. *Mol Cell* 2006;24:747–57. [PubMed: 17157257]
10. Nonaka G, Blankschien M, Herman C, Gross CA, Rhodius VA. *Genes Dev* 2006;20:1776–89. [PubMed: 16818608]
11. Wade JT, Roa DC, Grainger DC, Hurd D, Busby SJW, Struhl K, Nudler E. *Nat Struct Mol Biol* 2006;806–14. [PubMed: 16892065]
12. Zhao K, Liu M, Burgess RR. *J Biol Chem* 2005;280:17758–68. [PubMed: 15757896]
13. Botwell, D.; Sambrook, J. *DNA Microarrays: A Molecular Cloning Manual*. Cold Spring Harbor Press; New York: 2003.
14. Pieterse B, Jellema RH, van der Werf MJ. *J Microbiol Methods* 2006;64:207–16. [PubMed: 15982764]
15. Gao H, Yang ZK, Gentry TJ, Wu L, Schadt CW, Zhou J. *Appl Environ Microbiol* 2007;73:563–71. [PubMed: 17098911]
16. Ausubel, FM.; Brent, R.; Kingston, RE.; Moore, DD.; Seidman, JGASJ.; Struhl, K. *Current Protocols in Molecular Biology*. John Wiley & Sons, Inc; 1998.
17. Fare TL, Coffey EM, Dai H, He YD, Kessler DA, Kilian KA, Koch JE, LeProust E, Marton MJ, Meyer MR, Stoughton RB, Tokiwa GY, Wang Y. *Anal Chem* 2003;75:4672–5. [PubMed: 14632079]
18. Branham WS, Melvin CD, Han T, Desai VG, Moland CL, Scully AT, Fuscoe JC. *BMC Biotechnol* 2007;7:8. [PubMed: 17295919]
19. Buck MJ, Lieb JD. *Genomics* 2004;83:349–60. [PubMed: 14986705]
20. Grainger DC, Hurd D, Harrison M, Holdstock J, Busby SJ. *Proc Natl Acad Sci USA* 2005;102:17693–98. [PubMed: 16301522]
21. Grainger DC, Overton TW, Reppas N, Wade JT, Tamai E, Hobman JL, Constantinidou C, Struhl K, Church G, Busby SJ. *J Bacteriol* 2004;186:6938–43. [PubMed: 15466047]
22. Pan KH, Lih CJ, Cohen SN. *Proc Natl Acad Sci U S A* 2002;99:2118–23. [PubMed: 11854507]
23. Armstrong NJ, van de Wiel MA. *Cell Oncol* 2004;26:279–90. [PubMed: 15623938]
24. Dudoit S, Shaffer JP, Boldrick JC. *Statistical Science* 2003;18:71–103.
25. Tusher VG, Tibshirani R, Chu G. *Proc Natl Acad Sci U S A* 2001;98:5116–21. [PubMed: 11309499]
26. Wade JT, Hall DB, Struhl K. *Nature* 2004;432:1054–58. [PubMed: 15616568]

27. Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B. *Proc Natl Acad Sci USA* 2003;100:8164–69. [PubMed: 12808131]
28. Yang A, Zhu Z, Kapranov P, McKeon F, Church GM, Gingeras TR, Struhl K. *Mol Cell* 2006;17:593–602. [PubMed: 17188034]
29. Buck MJ, Nobel AB, Lieb JD. *Genome Biol* 2005;6:R97. [PubMed: 16277752]
30. Belyaeva TA, Wade JT, Webster CL, Howard VJ, Thomas MS, Hyde EI, Busby SJ. *Mol Microbiol* 2000;36:211–22. [PubMed: 10760178]
31. Wade JT, Reppas NB, Church GM, Struhl K. *Genes & Dev* 2005;2619–30. [PubMed: 16264194]
32. Conlon EM, Liu XS, Lieb JD, Liu JS. *Proc Natl Acad Sci U S A* 2003;100:3339–44. [PubMed: 12626739]
33. Chen X, Guo L, Fan Z, Jiang T. *Bioinformatics* 2008;24:1121–8. [PubMed: 18325926]
34. Hu J, Li B, Kihara D. *Nucleic Acids Res* 2005;33:4899–913. [PubMed: 16284194]
35. Burden S, Lin YX, Zhang R. *Bioinformatics* 2005;21:601–7. [PubMed: 15454410]
36. Rhodius VA, Suh WC, Nonaka G, West J, Gross CA. *PLoS Biol* 2006;4:e2. [PubMed: 16336047]
37. Liu X, Brutlag DL, Liu JS. *Pac Symp Biocomput* 2001:127–38. [PubMed: 11262934]
38. Roth FP, Hughes JD, Estep PW, Church GM. *Nat Biotechnol* 1998;16:939–45. [PubMed: 9788350]
39. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. *Science* 1993;262:208–14. [PubMed: 8211139]
40. Dennis GJ, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. *Genome Biol* 2003;4:P3. [PubMed: 12734009]
41. Berriz GF, King OD, Bryant B, Sander C, Roth FP. *Bioinformatics* 2003;19:2502–04. [PubMed: 14668247]
42. Hu Z, Killion PJ, Iyer VR. *Nat Genet* 2007;39:683–87. [PubMed: 17417638]
43. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. *Nature* 2004;431:99–104. [PubMed: 15343339]
44. Schuster SC. *Nat Methods* 2008;5:16–18. [PubMed: 18165802]
45. Hertz GZ, Stormo GD. *Bioinformatics* 1999;15:563–77. [PubMed: 10487864]
46. Bailey TL, Williams N, Misleh C, Li WW. *Nucleic Acids Res* 2006;34:W369–73. [PubMed: 16845028]

Appendix A - Solutions, supplies and equipment

A1.1 Solutions for microarray transcription profiling sample preparation

1. All solutions should be RNase free, either by treating with 0.1% DEPC, or by purchasing as RNase free from suppliers such as Ambion.
2. Ethanol/phenol stop solution. H₂O-saturated phenol (pH<7.0) in ethanol (5% v/v).
3. Lysozyme solution. 500 µg/ml lysozyme in 10 mM Tris (pH 8.0), 1 mM EDTA (pH 8.0). Prepare fresh just before use.
4. Nuclease-free Water, Ambion #AM9938.
5. Sodium Acetate, pH 5.5, 3M (100 ml), Ambion #AM9740.
6. 25× aa-dUTP/dNTP mix: 12.5 mM dATP/dGTP/dCTP, 5 mM dTTP, 7.5 mM amino-allyl dUTP (Ambion, #8439). The 3:2 ratio of aa-dUTP:dTTP is optimized for *E. coli* based on the (G+C) content of the template cDNA. If labeling efficiency is low for your strain, it will be necessary to experiment with different ratios.
7. SSC 20× (1L), Ambion #AM9763.
8. SDS, 10% solution (100 ml), Ambion #AM9822.

A1.2 Supplies for microarray transcription profiling sample preparation

1. All microfuge tubes and pipette tips should be RNase and DNase free.
2. TURBO DNA-free Kit, Ambion #AM1907. Contains TURBO DNase, 10× Buffer, and DNase Inactivation Reagent.
3. Random octamer oligonucleotide, any oligo company.
4. SuperScript III Reverse Transcriptase, Invitrogen #18080-093. Includes 5× First-Strand Synthesis Buffer.
5. MinElute Reaction Cleanup Kit (50), QIAGEN #28204.
6. CyDye Post-Labeling Reactive Dye Packs, GE Healthcare #RPN5661.

A1.3 Solutions for ChIP-chip

1. TBS; Tris-buffered Saline: 20 mM Tris, pH 7.4, 150 mM NaCl.
2. FA lysis buffer: 50 mM Hepes-KOH, pH 7, 150/500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS.
3. ChIP wash buffer: 10 mM Tris-HCl, pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% Nonidet-P40, 0.5% sodium deoxycholate.
4. ChIP elution buffer: 50 mM Tris-HCl, pH 7.5, 10 mM EDTA, 1% SDS.

A1.4 Supplies for ChIP-chip

1. Protein A sepharose, CL-4B, GE Healthcare Biosciences #17-0780-04.
2. Spin-X columns, Corning #8161.
3. BioPrime kit, Invitrogen #18095-012.
4. Cy3-dCTP, GE Healthcare #53021, and Cy5-dCTP, GE Healthcare #55021.
5. QIAquick PCR purification kit, QIAGEN #28104.

Figure 1. Venn diagram comparing transcription profiling, ChIP-chip and promoter mapping analysis of *E. coli* σ^{32} regulon members

Venn diagram comparing promoters bound by the *E. coli* heat shock factor, σ^{32} identified either by transcription profiling or by ChIP-chip. Transcription profiling compared *E. coli* cells 10 min after over expressing σ^{32} with a wild type control (10), identifying 68 significantly induced transcription units (TUs) (light grey circle). The 50 confirmed promoters includes known promoters, as well as promoters experimentally identified upstream of significant TUs from the transcription profiling data (10) (dark grey circle). Additional σ^{32} promoters are present in the literature, but they were not experimentally confirmed in (10), so are excluded in this diagram. ChIP-chip against σ^{32} from *E. coli* cells exposed to heat shock (transitioned from 30 °C to 50 °C for 10 min) identified 87 significant peaks (11) (clear circle). * Note 2 TUs each contain two σ^{32} promoters: each pair of promoters was represented by a single ChIP-chip peak.

Figure 2. Example ChIP-chip peak

(A) An example ChIP-chip peak for *E. coli* σ^{70} upstream of *rpsJ*. The height of the bars represents the level of signal from the ChIP-chip experiment at that genomic position. The center of the ChIP-chip peak is 2 bp from the known transcription start site of *rpsJ*. The average spacing of probes on this microarray was ~12 bp and all probes were 50 nt long. The position of the *rpsJ* gene is indicated by the large grey arrow. The position of the transcription start site for *rpsJ* is indicated by the black arrow. The short horizontal line has a width 100 bp. (B) The same ChIP-chip data was smoothed using a sliding window encompassing 9 contiguous probes. Each plotted value represents the mean of the 9 contiguous probes that are centered at the genomic position indicated. Data is plotted on the same scale as that in (A).

Figure 3. Histogram of position of σ^{32} ChIP-chip peaks relative to the transcription start site of known σ^{32} promoters

Comparison of σ^{32} ChIP-chip peak positions identified by Wade *et al.* (11) and known σ^{32} promoter transcription start sites from Nonaka *et al* (10). Negative and positive values indicate ChIP-chip peaks located upstream and downstream of known transcription start sites, respectively. Most ChIP-chip peaks map slightly upstream of the transcription start, corresponding to the sigma factor binding site. The average spacing of the 60-mer probes on the ChIP-chip microarray was 223 base pairs.

Table 1

Different Microarray Platforms

Probe ^a	Attachment ^b	Surface	c# Features/array ^d	Hybridization ^e	Supplier	Notes ^g
25 mers	Photolithography	Glass	1.3 × 10 ⁶ – 11 probes/gene PM and MM	One color	Affymetrix (NimbleExpress) www.affymetrix.com	<i>E. coli</i> /custom Requires specialist equipment.
60 mers	Photodeposition	Glass	1 × 385K, 4 × 72K	One or two color	NimbleGen www.nimblegen.com	All sequenced bacteria/custom – reusable ×3
35–40 mers	Electrochemical detritylation	Glass	12K	Two color	CombiMatrix http://combinatrix.com	Custom – reusable ×3
< 60 mers	Ink jet <i>in situ</i> synthesis	Glass	1 × 244K, 2 × 105K, 4 × 44K, 8 × 15K	Two color	Oxford Gene Technology www.ogt.co.uk	Select bacteria/custom
60 mers	Ink jet <i>in situ</i> synthesis	Glass	1 × 244K, 2 × 105K, 4 × 44K, 8 × 15K	Two color	Agilent www.agilent.com	No bacteria/custom
50 mers	Spotted	Glass	1 probe/gene	Two color	Ocimum www.ocimumbio.com	Select bacteria/custom/oligo sets
80 mers	Self spot	Glass	1 probe/gene	Two color	Operon www.operon.com	Select bacteria/custom/oligo sets
85 mers	Spotted Self spot	Nylon	9K 1 probe/gene	1 sample	Sigma Genosys www.sigma-genosys.com	Select bacteria/custom/oligo sets
ORFmers	PCR and self spot	Glass	1 probe/gene	Two color	Sigma Genosys www.sigma-genosys.com	Select bacteria/custom

^a Type of probe and probe length; mer, oligonucleotide; ORFmers, doublestranded PCR products of whole open reading frame.

^b Method of probe synthesis and attachment to array surface

^c Type of array surface: glass slide or nylon membrane

^d Number of features per array and number of probes per gene: K=1000 (e.g. 72K = 72,000); PM, perfect match probes; MM, mismatch probes

^e Sample hybridization: one color, one sample is hybridized to the array (often labeled Cy 3); two color, competitive hybridization of two samples labeled Cy3 and Cy 5; 1 sample, 1 sample radioactively labeled.

^g Notes on whether arrays are predesigned for specific organisms or can be custom designed. Some arrays can be reused upto ×3 times.

Table 2

Conserved motif finding algorithms

Algorithm	URL
AlignACE (38)	http://atlas.med.harvard.edu/cgi-bin/alignace.pl
BioProspector (37)	http://ai.stanford.edu/~xslu/BioProspector/
Consensus (45)	http://bifrost.wustl.edu/consensus/
Gibbs Motif Sampler (39)	http://bayesweb.wadsworth.org/gibbs/gibbs.html
MEME (46)	http://meme.sdsc.edu/meme/intro.html
Motif Regressor (32)	http://ai.stanford.edu/~xslu/MDscan/
W-AlignACE (33)	http://www1.spms.ntu.edu.sg/~chenxin/W-AlignACE/