nature biotechnology

High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons

Patrick T McGrath[1,2], Honglak Lee[3], Li Zhang[2,4], Antonio A Iniesta[2], Alison K Hottes[5], Meng How Tan[2], Nathan J Hillson[2], Ping Hu[6], Lucy Shapiro[2] & Harley H McAdams[2]

Using 62 probe-level datasets obtained with a custom-designed *Caulobacter crescentus* microarray chip, we identify transcriptional start sites of 769 genes, 53 of which are transcribed from multiple start sites. Transcriptional start sites are identified by analyzing probe signal cross-correlation matrices created from probe pairs tiled every 5 bp upstream of the genes. Signals from probes binding the same message are correlated. The contribution of each promoter for genes transcribed from multiple promoters is identified. Knowing the transcription start site enables targeted searching for regulatory-protein binding motifs in the promoter regions of genes with similar expression patterns. We identified 27 motifs, 17 of which share no similarity to the characterized motifs of other *C. crescentus* transcriptional regulators. Using these motifs, we predict coregulated genes. We verified novel promoter motifs that regulate stress-response genes, including those responding to uranium challenge, a stress-response sigma factor and a stress-response noncoding RNA.

The *C. crescentus* cell cycle and asymmetric cell division (**Fig. 1a**) are implemented by the ordered activation of multiple genetic modules[1–4]. *C. crescentus* cells control the timing of cellular events, in large part, by turning on the transcription of the relevant gene cohorts only when they are needed[5]. At least three master transcriptional regulators, whose concentrations and activities oscillate in time and space, govern many facets of *C. crescentus* cell cycle progression, including DNA replication, DNA methylation, polar organelle biogenesis and cell division[6–8]. Although these three proteins regulate the expression of ~200 genes[7–10], this control accounts for less than half of the total number of cell cycle–regulated genes, indicating that additional transcriptional regulators also control cell cycle progression. Besides the transcriptional changes that occur throughout the cell cycle, global transcriptional responses also occur in response to changing environmental conditions, particularly those invoking stress responses. Only a few of the regulators controlling these responses are known[11–13].

Most *C. crescentus* cell cycle regulators have been identified through genetic screens with subsequent high-throughput global identification of their regulons facilitated by microarray assays[7,8] and by chromatin immunoprecipitation assays[10]. Each of these methods involves indirect inference of the directly controlled downstream genes with substantial error rates owing to experimental noise.

An alternative method to identify regulons is through the identification of DNA motifs near the transcriptional start sites of genes. Genes coregulated by the same transcriptional factor will share the same motif. DNA motifs have been identified by clustering genes into groups with similar expression profiles, and then searching the region upstream of their translational start site for conserved DNA motifs[14]. Motifs are often short or degenerate and thus difficult to identify. However, when transcription start sites are known, the DNA region most likely to contain a regulatory binding site is circumscribed, and motif search effectiveness is greatly enhanced.

Bacterial transcription sites have been determined by primer extension experiments and by S1-nuclease protection assays. These techniques require separate experiments to identify the start site(s) for each individual gene; consequently, relatively few bacterial transcriptional start sites are known. With the availability of microarray technology, simultaneous measurement of the span of many mRNA messages is possible using tiled arrays. A hidden Markov model method was used to analyze tiled array data to estimate transcript boundaries in *Escherichia coli* to an accuracy of ± 30 bp[15].

We report here a method for identification of transcription start sites of genes by analysis of pair-wise correlation of probe-level signals from multiple microarray experiments under differing conditions using an Affymetrix array designed for this application. Because the potential accuracy of start-site estimates from array data are limited by the spacing of probes upstream of genes, we designed CauloHI1, a custom *C. crescentus* Affymetrix array that includes tiling at 5-bp spacing upstream of 2,367 predicted *C. crescentus* operons. Start-site identification using the CauloHI1 array requires profiles of gene transcripts under a variety of experimental conditions so that correlation between probe pairs within the span of the same messenger RNA

[1]Department of Physics, Stanford University, Varian Physics, 382 Via Pueblo Mall, Stanford, California 94305, USA. [2]Department of Developmental Biology, Stanford University, B300 Beckman Center, 279 Campus Drive, Stanford, California 94305, USA. [3]Department of Computer Science, Stanford University, Gates Building, Stanford, California 94305, USA. [4]Department of Applied Physics, Stanford University, 316 Via Pueblo Mall, Stanford, California 94305, USA. [5]Lewis-Sigler Institute for Integrative Genomics, Princeton University, L206 Carl C. Icahn Laboratory, Princeton, New Jersey 08544, USA. [6]Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, 1 Cyclotron Road Mail 70A3317, Berkeley, California 94720, USA. Correspondence should be addressed to H.H.M. (hmcadams@stanford.edu).

Received 14 June 2006; accepted 1 March 2007; published online 1 April 2007; doi:10.1038/nbt1294

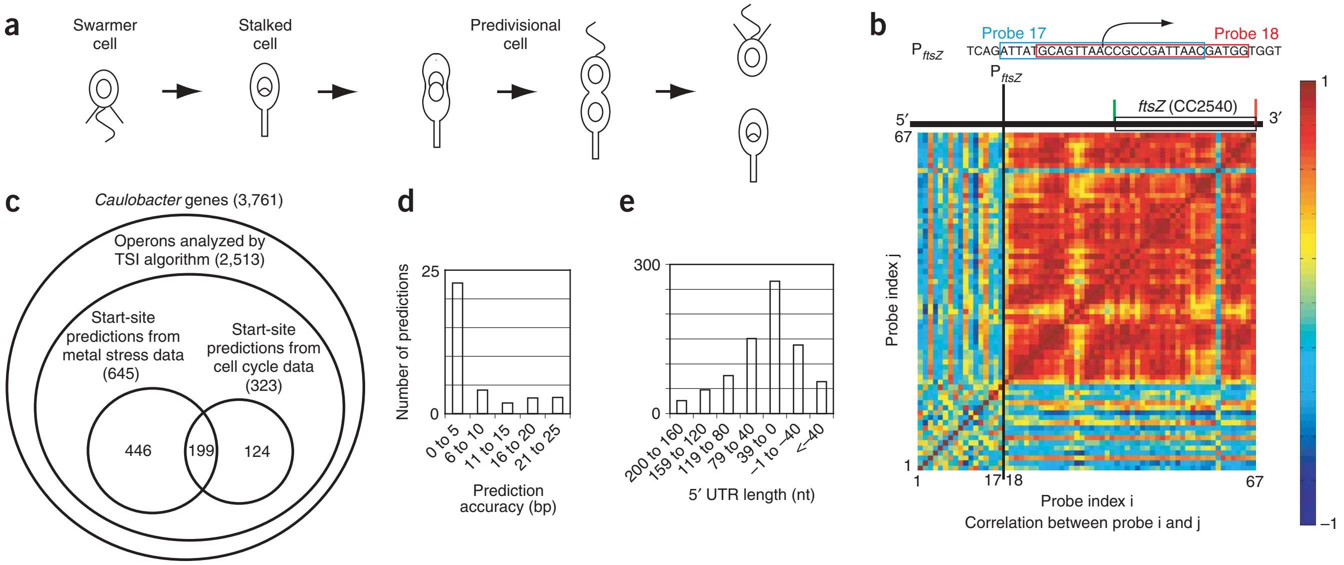

Figure 1 Transcriptional start-site identification. (**a**) *C. crescentus* executes a complex development program each cell cycle to produce an asymmetric cell division. The motile swarmer cell has a single polar flagellum and multiple pili, and chromosome replication is blocked. Swarmer cells differentiate into stalked cells by ejecting the flagellum, retracting the pili, growing a stalk at the previously flagellated pole and initiating chromosome replication. About halfway through chromosome replication, biogenesis of a new flagellum and pili at the pole opposite the stalk starts. After completion of chromosome replication and segregation, the cell divides asymmetrically into the swarmer and stalked progeny cells. (**b**) Correlation matrix of the *ftsZ* (CC2540) gene. The color of the square at (i,j) encodes the correlation between signals from probe i and probe j across 13 samples from cell-cycle synchronized *C. crescentus* cells. Probe indices are ordered from 5′ to 3′ with the most upstream probe labeled 1. The boundary between probe 17 and 18 (black line) is the location of a transcriptional start site. The biochemically identified transcriptional start site of *ftsZ* and the sequence of the 17th and 18th probe are shown at the top. (**c**) Start-site predictions from two experiment classes. Two *C. crescentus* datasets, one measuring the cell cycle transcriptional variation and another measuring stress response to heavy metal exposure, were analyzed. We made 323 start-site predictions from the cell cycle data and 645 predictions were made from the stress response data, yielding start-site predictions for 769 genes. (**d**) Comparison with biochemically determined start sites. Start sites for 34 of the 769 predictions have been previously determined biochemically. The histogram shows the difference between the prediction from the CauloHI1 probe analysis and the biochemically determined start sites; 65% are within ± 5 bp. (**e**) Histogram of the separation between the predicted (annotated) translational start sites and the 769 predicted transcriptional start sites.

can be distinguished. Here we used 62 different *C. crescentus* gene expression assays with the CauloHI1 chip from two types of experiments: (i) periodic samples from synchronized cell populations collected for this study, and (ii) expression profiles before and after heavy metal exposure[13]. Using these datasets, we identified transcriptional start sites for 769 *C. crescentus* genes. Motif searching upstream of the transcriptional start sites of coexpressed genes (determined by cluster analysis) revealed 27 conserved promoter motifs, 17 of which have not, to our knowledge, been described before. Several of these motifs were shown to mediate promoter activity. Genes that are coregulated by the proteins binding these motifs were identified by searching promoter regions of coexpressed genes for the motifs.

RESULTS

Identification of 769 transcriptional start sites

Probe-level signals, computed as the difference in 'perfect match' and 'mismatch' signals for each probe, from 62 microarray experiments were used to predict transcriptional start sites. For each perfect match and mismatch probe pair on the CauloHI1 chip, the perfect match-mismatch difference signals from the *m* experiments comprise an *m*-element vector. The cross-correlations between each of the *n* respective signal vectors from probe pairs found within the predicted gene and up to 200 bp upstream of the predicted translational start site, create an $n \times n$ matrix of probe correlation values for each gene. We expect much higher correlation between the signal vectors from probe pairs spanned by the expressed mRNA for the gene, compared to signals from probe pairs that are not within the same message. The boundary between probes with low cross-correlation and probes with high cross-correlation should distinctively identify the transcription start site with resolution determined by the interval of probe tiling. In practice, observation of a distinctive signature for a gene's transcription start site requires two conditions: (i) the gene must exhibit a range of different response levels over the experimental dataset, and (ii) the probe signals must be adequately above random experimental noise and spurious signals from probe cross-hybridization.

The color-coded display of the probe correlation values produces a striking visualization of the boundary of the correlated probes (**Fig. 1b**). The most upstream probe pair is the first index of the correlation matrix. High correlation is red, and the boundary (between probes 17 and 18) of the large red square (between probes 18 and 67) corresponds to the known *ftsZ* transcriptional start site[16]. This probe cross-correlation analysis method also frequently produces a distinctive pattern when the gene has multiple start sites so that individual start-site locations can be identified (see below).

A Matlab program, transcriptional start identification (TSI) program, was applied separately to the cell cycle and the metal-response datasets, resulting in start-site predictions for 323 genes from the cell cycle data and for 645 genes from the metal stress data (**Fig. 1c**; correlation data for each of these genes are available in T1 and T2 at http://www.stanford.edu/group/caulobacter/startsite/, (*startsite*)). For 199 genes, start sites were predicted from both datasets. Altogether, there are reliable start-site predictions for 769 *C. crescentus* genes. Start-site predictions for 65% of the 199 genes with predictions from both datasets agreed with each other within 5 bp. For 9, the predictions were

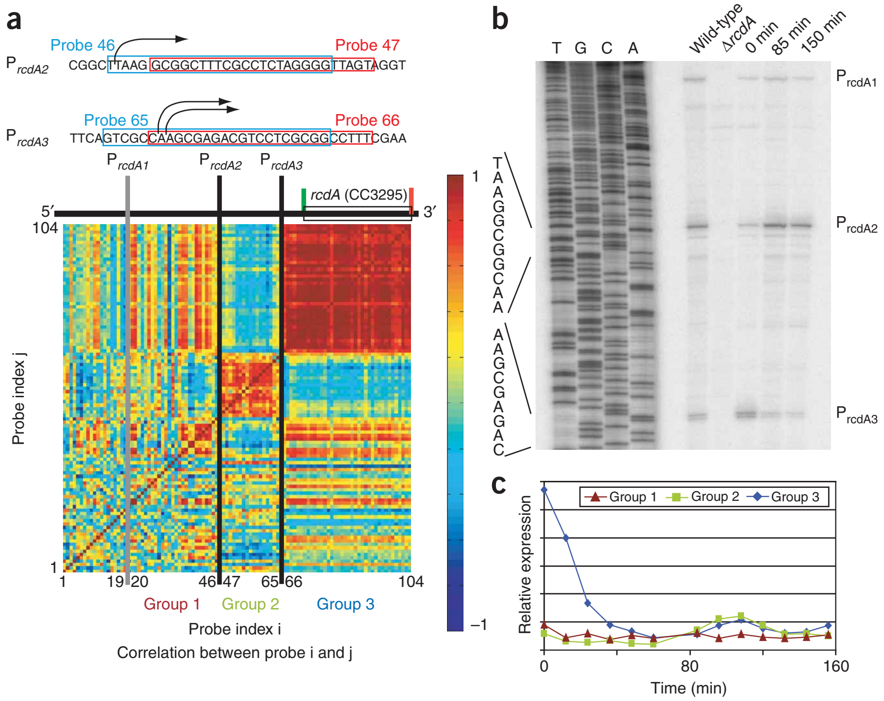

Figure 2 Genes with multiple start sites. (**a**) Genes transcribed from multiple transcriptional start sites have a distinctively patterned correlation diagram. The correlation matrix diagram is for *rcdA* (CC3295). The diagram shows cell cycle data for probe pairs interior to the gene to 500 bp upstream of the predicted translational start site. The boundaries (black lines between probes 46 and 47 and between probes 65 and 66) are the result of multiple transcriptional start sites. Group 3 probes (indices 67–104) are most highly cross-correlated. Group 2 probes (indices 47–66) are cross-correlated, but not correlated with group 3 probes. (**b**) Primer extension analysis shows that *rcdA* has three promoter start sites, two of which match the location of the boundaries (P_{rcdA2} and P_{rcdA3}). The nucleotide sequences of the regions surrounding P_{rcdA2} and P_{rcdA3} and related probes are above. The sequences of the two probes adjacent to the boundaries are highlighted. The approximate region of the probes that flank the start site of P_{rcdA1} is shown as a gray line on the correlation matrix between probes 17 and 18. (**b**) For primer extension analysis of *rcdA* we used purified RNA from five samples and a probe complementary to a region in the *rcdA* gene. Wild-type RNA yields three bands that are not present with RNA from an *rcdA* deletion strain, indicating that *rcdA* is transcribed from three different promoters. Cell cycle response of these three promoters was characterized by primer extension from RNA samples from synchronized wild-type cells at three time points: 0, 85 and 105 min in a 150-min cell cycle. P_{rcdA1} expression is relatively constant at the three time points, P_{rcdA2} expression peaks in predivisional cells and P_{rcdA3} expression peaks in swarmer cells. (**c**) Probe pairs from the three probe groups created by the three transcriptional start sites (**a**) were grouped and averaged. Plot shows the resulting average expression over the cell cycle. Average expression of the Group 1 probes, which measures expression only from P_{rcdA1}, is constant over the cell cycle, so that Group 1 probes yield a low correlation signal. The signal from Group 2 probes, which measures combined expression from P_{rcdA1} and P_{rcdA2}, peaks in predivisional cells. The signal from Group 3 probes, which measures expression from P_{rcdA1}, P_{rcdA2} and P_{rcdA3}, peaks in swarmer cells and predivisional cells. Analysis of the probe signals in different upstream regions can identify activity from each of the upstream promoters.

more than 40 bp apart, probably because different promoters, hence different start sites, were activated in the two different classes of experiments (see *startsite*, T3). For motif searching the two sets of start-site predictions were combined. Because multiple start sites for a single gene in *C. crescentus* are typically separated by over 40 bp (*startsite*, T4), predictions that agreed within 40 bp were averaged and predictions greater than 40 bp apart were used separately.

Published biochemically determined transcriptional start sites were found for 52 *C. crescentus* genes (*startsite*, T4). Thirty-five of these were among the 769 genes with predicted start sites in this study. **Figure 1d** shows a histogram of differences between our predictions and the published start sites: 22 (65%) agree within 5 bp, and 26 (76%) agree within 10 bp. Although we have used the biochemically determined start sites as a benchmark, in some cases, the chip-based predictions are known to be more accurate. For example, our prediction for CC0899 (*flaN*) is 23 bp upstream of the start site determined by nuclease S1 mapping[17]. However, promoter analysis of CC0899 led the authors to suggest that the transcriptional start site was actually 33 bp upstream of the nuclease S1–determined start sites and that the 5′ end was rapidly cleaved[17]. Comparison of all available biochemically determined start sites to those predicted in this study is on *startsite*, T4.

Figure 1e shows a histogram of the distance of the predicted transcriptional start sites from the predicted translational start site. The majority of predictions (72%) fall between –80 and 40 bp (the three highest bars). The 64 genes with predicted start sites greater than 40 bp downstream of the predicted translational start may indicate inaccurate start-codon predictions. Some of the 150 genes with large 5′ UTRs (>80 bp) could be post-transcriptionally regulated through sRNA regulation or as riboswitches regulated by metabolites.

Multiple start sites

In 45 of the 769 genes with start-site predictions, probe pairs immediately upstream of the predicted transcriptional start-site boundary are highly correlated with each other, but not with probe pairs within the gene. This produces adjacent squares (that is, adjacent probe sets) with high internal correlation. **Figure 2a** illustrates this for *rcdA*[18]. This pattern results from genes transcribed from multiple start sites. The level of correlation between probes in each square depends on the correlation between the expression profiles from the two promoters for the experimental conditions. To verify this, the *rcdA* transcriptional start sites were determined by primer extension. Different mRNA subspecies not present in an *rcdA* deletion strain were observed (**Fig. 2b**). The primer used for primer extension experiments lies within a portion of the *rcdA* coding region that is absent in the *rcdA* deletion strain. The two start sites found between probes 46/47 and 65/66 (**Fig. 2b**) are consistent with the correlation matrix predictions (**Fig. 2a**). The third start site identified by primer extension is ~400 bp upstream of *rcdA*. Two other genes (*clpX* and *dnaN*) independently known to have multiple +1 sites[19,20] have multiple squares in their correlation matrix images consistent with the biochemically determined start sites (**Supplementary Fig. 1** online). The CC2510 gene has two widely separated start sites with

Table 1 Function of identified promoter motifs[a]

<table>
<tr><th>Motif</th><th>Homology to identified promoters</th><th>Gene(s)</th><th>References</th></tr>
<tr><td>cc_2</td><td>TTCAGGC.CCGTTCAGGCGGG
||||||| |||||||||||||
TTCAGGC.CCGTTCAGGCGGG</td><td>sRNA stress response promoter element CC1840_CC1841</td><td></td></tr>
<tr><td>cc_3</td><td>   T
TTGAC.CG.ATCAA.GC
||||| ||||| ||
TTGAC.C.GATCAA.GC</td><td>FixK binding site</td><td>12</td></tr>
<tr><td>cc_5</td><td>   GC   G
TGGC.CCGGCCTTGCA
|||| |||.|||||||
TGGC CCGTCCTTGCC</td><td><i>flaN</i> σ⁵⁴ promoter</td><td>27,29</td></tr>
<tr><td>cc_7</td><td>   (n12)
GGAACC.............G..CGTT
|||||| | ||||
GGAACC.............G..CGTT</td><td>CC2883 <i>sigU</i> promoter element</td><td>28</td></tr>
<tr><td>cc_10</td><td>   (n6)
TTAA......GTTAACCAT
|||| |||||||||
TTAA......GTTAACCAT</td><td><i>fliQ</i> CtrA-dependent promoter</td><td>27</td></tr>
<tr><td>m_2</td><td>   (n16)   G
TTGAC................CCTA.A
||||| |||| |
TTGAC................GCTA.A</td><td>Consensus σ⁷⁰ promoter</td><td>30</td></tr>
<tr><td>m_5</td><td>CCC..CATTAC...A..TTAA...G.C
||| |||||| | |||| | |
CCC..CATTAC...A..TTAA...G.C</td><td>CC1891 promoter element</td><td></td></tr>
<tr><td>m_6</td><td>   (n7)   n10
C.CTTG.......CC..........CTA..T
| |||. || ||| |
C.CTTG.......CC..........CTA..T</td><td><i>rpoH</i> P₂ heat shock promoter</td><td>25</td></tr>
</table>

[a] In all cases, the upper sequence is the motif identified by MEME[21] or BioProspector[22], and the lower sequence is the conserved motif present in the promoter region of the indicated gene.

distinctive activation patterns in the cell-cycle and metal-stress experiments (**Supplementary Fig. 2** online and *startsite*, T5).

The cell cycle pattern of the three *rcdA* promoters was determined by primer extension at 0, 85 and 105 min. Transcription from the nearest promoter (A3) to the *rcdA* coding sequence peaked in swarmer cells; transcription from the next promoter (A2) peaked in predivisional cells; and the transcription from the furthest promoter (A1) was the same at all three time points (**Fig. 2b**), consistent with the probe-level cell cycle results (**Fig. 2c**). These *rcdA* results demonstrate identification of genes with multiple promoters from the probe-correlation analysis, when the transcription pattern of the different promoters differs over the experimental conditions. For these cases, activities of each promoter can be determined by grouping the probes based upon the start-site boundaries. For genes we identified as having one or two transcriptional start sites, the activity of each individual promoter was estimated (**Supplementary Methods** online and *startsite*, T6 and T7).

DNA binding motifs

After clustering promoters into coexpressed groups (**Supplementary Figs. 3** and **4** online) based upon their cell-cycle or metal-stress expression data, the MEME[21] motif finder was used to search two regions for motifs: (i) from 80 bp upstream to 15 bp downstream of the predicted transcriptional start sites, or (ii) from 130 bp upstream to 50 bp upstream of the predicted transcriptional start sites, whereas the BioProspector[22] motif finder was used to search the regions from 80 bp upstream to 15 bp downstream. For motifs with E-value <0.01, the MAST program[23] was used to identify each instance of the MEME-identified motifs within the cluster where the motif was found. MEME-identified motifs with E-value <1 are on *startsite*, T6 and T7.

Fourteen motifs with E-value <0.01 were found in 10 of the 14 cell-cycle clusters (**Supplementary Table 1** online and **Fig. 3**). These motifs are either not found or found at a lower E-value in a motif search of the entire set of the 450 genes, confirming the advantage of clustering the genes into coexpressed groups before motif searching. We confirmed that 3 motifs, cc_2, cc_3 and cc_7, mediate promoter activity (**Supplementary Figs. 5–7** online). The promoter regions of a novel class of small RNAs (sRNAs) contain a highly conserved cc_2 motif required for stress response–activation of the promoter (**Supplementary Fig. 5**). The cc_3 motif is the promoter site for the *C. crescentus* FixK global transcriptional regulator of respiration, present in multiple promoters in the FixK regulon[12] (**Supplementary Fig. 6**). The promoter region of the CC2883 gene encoding the ECF sigma factor SigU contains the cc_7 motif. Changes of conserved bases in the *sigU* cc_7 promoter motif abolish promoter activity (**Supplementary Fig. 7**). Sequences and functions of selected motifs are in **Table 1**. Analysis of another newly identified motif provided valuable insight into the control of temporally regulated genes. Gene cluster cc_g, activated in the early predivisional cell, includes genes with a novel highly conserved 14-base pair motif, cc_9, upstream of a CtrA binding site[9], motif cc_10 (**Fig. 3**). The genes in the cc_g cluster are involved in polar organelle biogenesis. We predict that cc_9 contributes to fine tuning of the time of expression of this gene cohort. Promoter regions (190) containing at least one of the 14 motifs and their locations are on *startsite*, T7.

Using MEME, thirteen motifs with an E-value <0.01 were found in 6 of the 10 metal stress clusters (**Supplementary Table 1** and **Fig. 4**). Ten of the 13 motifs are novel; three are similar to previously identified motifs. One of the novel motifs, m_5, was shown to control the promoter activity of a uranium-inducible promoter (**Supplementary Fig. 8**). Motif cc_5 is the RNA polymerase sigma 54 binding motif (**Supplementary Fig. 9**). Promoter regions (248) that contain one of these motifs and their locations are on *startsite*, T6.

We searched the same DNA regions within the different clusters with BioProspector. Differences in the motif search algorithms and implementations lead to somewhat different results. In some cases, BioProspector identifies shorter motifs that are included within the MEME motifs; in other cases, the two programs identify different motifs. The motifs found with BioProspector and their positions are available at on *startsite*, T9–T12.

To determine the extent that focused searching at the transcriptional start sites aided in identifying the motifs, we repeated the motif search without using the start-site information. **Figure 5** shows MEME results from focused search near the predicted start sites in coexpressed genes, compared with results from searching the region 200 bp upstream of the translational start sites in coexpressed genes. The unfocused search identified 11 motifs; focused search identified 27. In four cases (cc_3, cc_10, m_3 and m_4), the unfocused search identified more binding sites, suggesting that some binding sites lay outside the target region of the focused search.

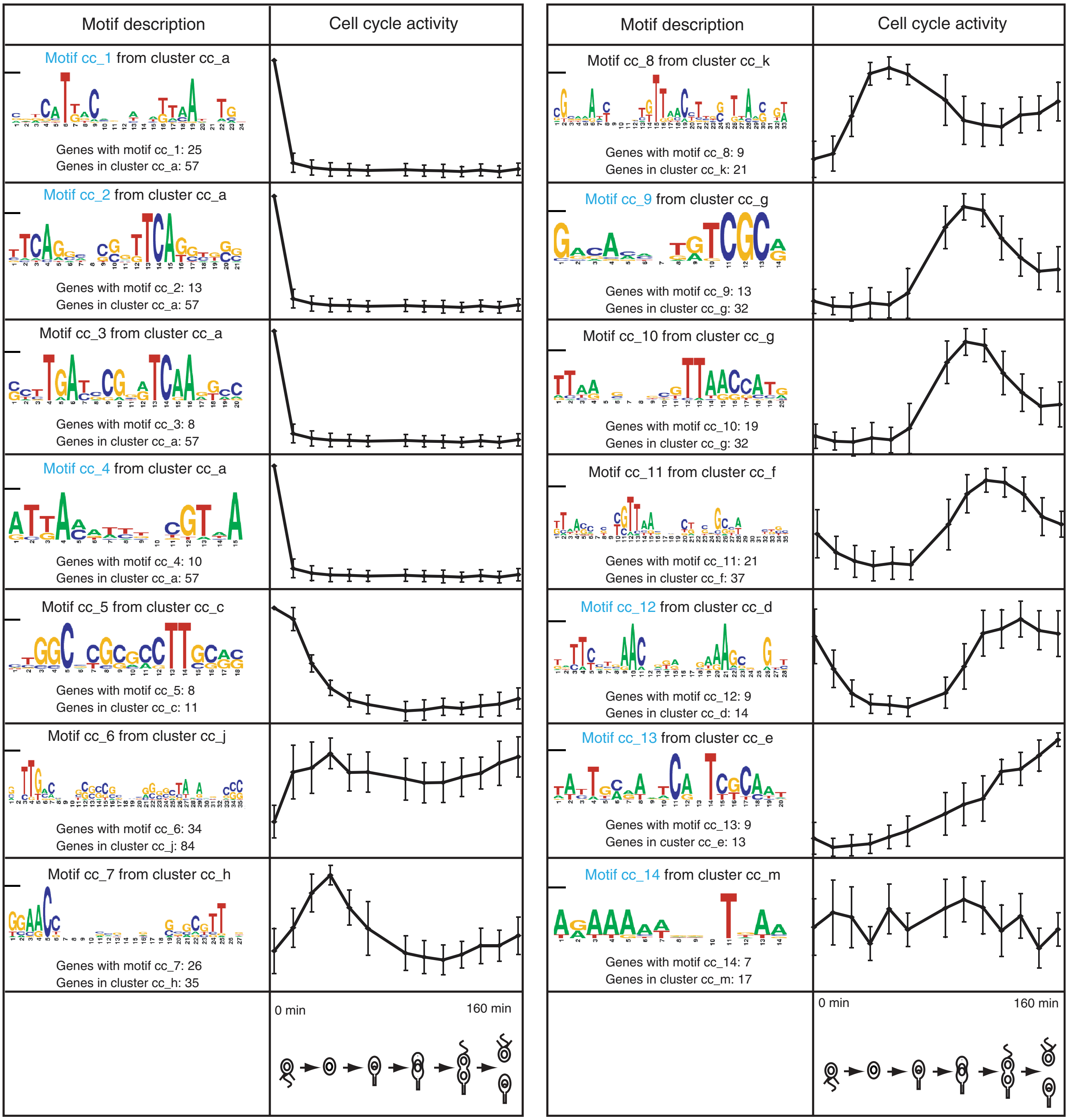

Figure 3 Motifs identified from cell cycle dataset. The normalized transcription levels for the 450 genes with transcriptional start sites identified using the cell cycle dataset were clustered hierarchically into 14 clusters of coexpressed genes. We found 14 motifs upstream of these genes using the MEME motif finder. The motif name, name of the cluster the motif was found in, sequence logo representation of the motif, number of occurrences and number of genes in the cluster containing the motif are in the motif description column. If the motif is novel, its name is blue. The average cell cycle profile and s.d. of the genes in the cluster containing the motif are in the cell cycle activity column.

DISCUSSION

We used the CauloHI1 chip, an Affymetrix chip we designed with 5 bp tiling upstream of the translational start sites of the first gene in 2,367 predicted operons, to identify 769 *C. crescentus* transcriptional start sites. Using targeted searching upstream of these start sites for regulatory-protein binding motifs in the promoter regions of genes with similar expression patterns, we then identified 27 motifs, 17 of which are novel.

In this report, we exploited probe-level data of the tiled CauloHI1 chip for transcriptional start-site identification. The incremental cost of designing Affymetrix (or similar) microarray chips with well-positioned intergenic tiling is small and then the start site, motif and regulon information is a by-product obtained by analysis of the microarray datasets from studies done for other purposes.

The 769 *C. crescentus* transcriptional start sites we have identified are a 14-fold increase over the 52 start sites previously identified by

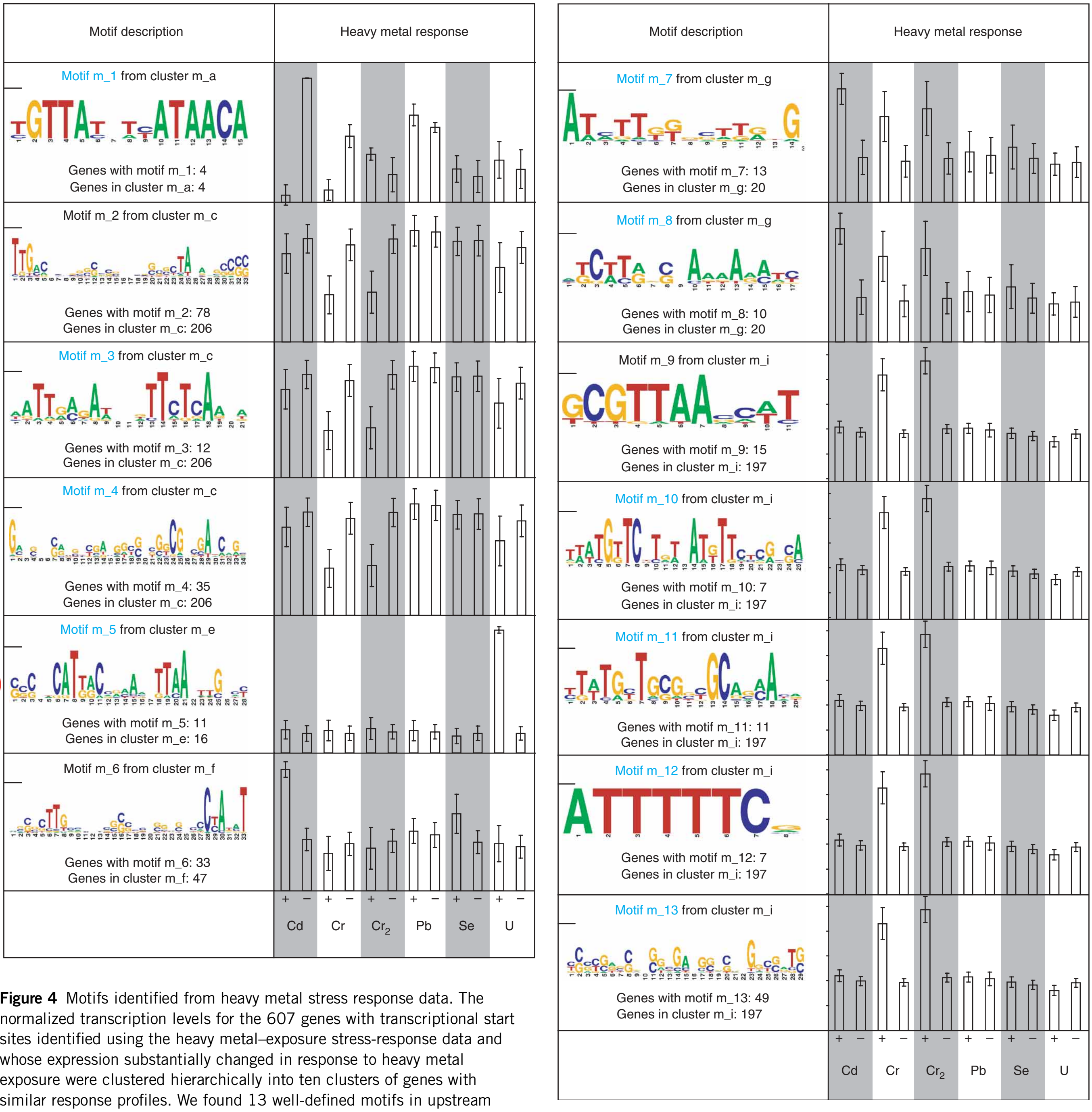

Figure 4 Motifs identified from heavy metal stress response data. The normalized transcription levels for the 607 genes with transcriptional start sites identified using the heavy metal–exposure stress-response data and whose expression substantially changed in response to heavy metal exposure were clustered hierarchically into ten clusters of genes with similar response profiles. We found 13 well-defined motifs in upstream regions of these genes using the MEME motif finder. The motif name, name of the cluster the motif was found in, sequence logo representation of the motif, number of occurrences and number of genes in the cluster containing the motif are in the motif description column. The height of each letter in the sequence logo indicates the bit information for that nucleotide. A small black line indicates the maximum bit score (2 bits). If the motif is novel, its name is blue. The average heavy metal stress-response profile and s.d. of the genes in the cluster containing the motif are in the heavy metal response column.

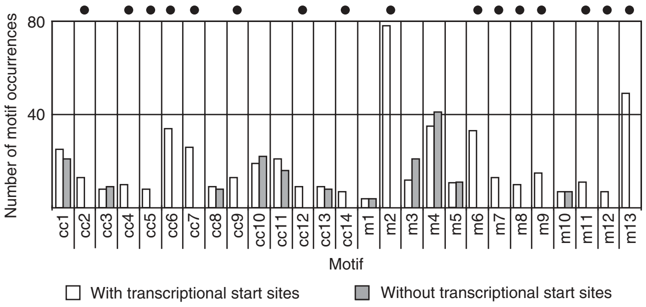

Figure 5 Comparison of motif searching with and without transcriptional start sites. We searched for motifs in the region 200 bp upstream of the translational start sites for genes in each of the clusters containing one of the motifs identified upstream of predicted start sites (**Figs. 3** and **4**) as would typically be done in the absence of identified transcriptional start sites. The bar graphs show the number of motifs found with this search method versus the more focused search (see text) possible when the transcription start site is known. Black dots over the graph indicate motifs that required the focused motif searching to be found.

primer extensions or S1-nuclease protection assays. Biochemical assays to determine start sites are performed on a gene-by-gene basis and are difficult to scale up to a genome-wide level, because they require radioactive assays, large sequencing gels and specific troubleshooting for each individual gene. However, biochemical assays can identify transcriptional start sites to single nucleotide resolution. In contrast, our method simultaneously identifies many start sites to within a few base pairs for the set of genes with substantial variation in expression across the set of available microarray experiments. These are the genes with greatest potential for physiological relevance.

In *E. coli*, RNA polymerase is immobilized on promoter locations by the addition of rifampicin. After immobilization, ChIP/Chip assays on a subunit of the RNA polymerase will then determine the location of the stalled polymerase and thus the location of promoters[24]. This method identified 1,139 *E. coli* promoters; our method is more accurate, however, as DNA pieces are typically sheared to a length of 500 bp for ChIP/Chip experiments[24].

The DNA binding motifs we identified nearly double the number of previously known binding motifs in *C. crescentus*. Almost half the motifs identified using this focused search method have been verified previously. We verified the functions of several of the motifs. Many of the remaining motifs contain either two reverse complementary sequences separated by a small gap (motif cc_4, cc_9, cc_13, m_1, m_3, m_11), or two highly similar sequences separated by a small gap (cc_2, cc_13, m_7, m_10) as would be expected for sequences that bind protein dimers. In some cases, known motifs were found in genes responding to a newly identified stimulus that controls the expression of the transcriptional regulator that binds to the motif. For example, the m_6 motif is similar to the motif recognized by RpoH[25], a sigma factor activated by heat shock and found here in gene promoters activated in response to cadmium.

Many of the promoter regions of the genes with identified start sites do not contain one of the 27 motifs identified here. For MEME-identified motifs, this was true for 260 of 450 genes from the cell cycle data, and 359 of 607 genes for the metal response data. Possible explanations include (i) the genes are members of small regulons and thus difficult or impossible to find by the motif search algorithms, (ii) the regulator protein binding site is far from the transcriptional start site, (iii) the regulator protein recognizes a short DNA motif or a feature of the DNA unrelated to its sequence (such as its flexibility) or (iv) the gene clustering algorithm grouped genes regulated by several different ligands together leading to ambiguous motif signatures.

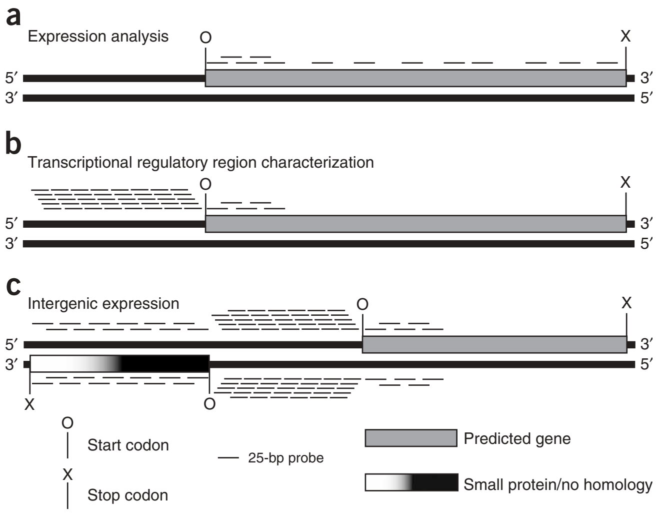

Figure 6 CauloHI1 Affymetrix microarray design. The arrangement of about 500,000 25-bp probes on the CauloHI1 microarray has three main purposes: (i) monitoring gene expression, (ii) characterizing transcription regulatory regions and (iii) identifying unannotated transcripts. (**a**) For purpose (i), five probes were tiled every 15 bp in the first 85 bp after the predicted start site, and eight probes were chosen in the coding region found after the first 85 bp. These eight probes are spaced throughout the region and chosen based upon thermodynamic properties using standard Affymetrix design methods. (**b**) For purpose (ii), probes were generally selected every 5 bp in the 200 bp upstream of the first gene in each predicted operon, and every 15 bp for the 85 bp downstream of the predicted start codon. These probes support identification of transcriptional start sites or identification of the DNA binding sites of transcriptional regulators. The five probes selected 85 bp downstream of the predicted start codon were chosen as a hedge against inaccurate start codon annotations. (**c**) For purpose (iii), intergenic regions were tiled every 5 bp on both strands, small putative genes with no known homology were tiled every 15 bp on both strands (as some of these small putative genes are likely to be noncoding intergenic DNA) and the first 85 bp of each gene were tiled every 15 bp on both strands.

The method described here for identifying transcription start sites, conserved promoter motifs and regulons using tiled microarrays enables direct identification of transcriptional regulatory networks. Although start sites can be found for any gene with variable activity across the experimental dataset used for the search, the motif and regulon identification procedure is most successful for genes that are members of large regulons.

METHODS

The cell cycle experiments using the CauloHI1 chip, the metal-response dataset, strains, growth conditions, primer extensions and data analysis are in **Supplementary Methods** online.

***C. crescentus* Affymetrix chip design.** The CauloHI1 microarray, constructed using the Affymetrix Genechip technology (http://www.affymetrix.com/technology/index.affx), consists of a grid of ∼501,075 18 μm × 18 μm features, with each feature (or probe) comprised of many 25-mer identical oligonucleotides. Probes are present in pairs. The perfect-match probe sequence exactly complements the target sequence, and, normally, the mismatch probe sequence matches the perfect-match sequence with the exception of the central (13th) nucleotide which is complementary to the central (13th) nucleotide of the perfect-match probe (http://www.affymetrix.com/support/technical/technotes/25mer_technote.pdf). However, in cases where the perfect-match probe matches at least 22 of 25 bp elsewhere on the *C. crescentus* genome, the mismatch probe sequence is chosen so that the most central nucleotide is changed to more perfectly match the similar 25-bp region. For example, if the perfect-match probe is AAAAAAAAAAAAAAAAAAAAAAAAA, and if a region of the *C. crescentus* genome contains a 25-nt sequence of ATAAAAAAAATAAAAAAAAAAAAAA, the mismatch probe chosen would be AAAAAAAAAATAAAAAAAAAAAAAA. The mismatch probe is used to estimate the amount of nonspecific hybridization to the perfect-match probe. Standard Affymetrix control features are included in the array design to support grid alignment and assessment of labeling and hybridization efficiencies (http://www.affymetrix.com/support/technical/other/custom_design_manual.pdf). Nucleotide sequences of the remaining features were chosen for three main purposes: (i) monitoring the expression levels of *C. crescentus* genes as well as a set of foreign genes that are frequently introduced genetically into the genome for experiments (such as antibiotic resistance genes); (ii) characterization of the *cis* regulatory regions of the *C. crescentus* genome (either by identification of transcriptional start sites or through ChIP/Chip experiments); and (iii) detection of unannotated transcripts that encode either sRNAs or mRNAs.

Details of the design of CauloHI1 are at http://www.stanford.edu/group/caulobacter/CauloHI1/. Here, we only describe features used for each of the preceding three purposes. For purpose (i), thirteen optimally chosen probe pairs are in each annotated *C. crescentus* gene for measurement of the gene's expression level (**Fig. 6a**). For purpose (ii), probe pairs are tiled every 5 bp upstream of 2,367 predicted operons (operons were predicted using a probabilistic learning approach[26]) and every 15 bp in the first 85 bp of each predicted gene. These probes are for use in prediction of transcriptional start sites or determination of transcription-factor binding sites through ChIP/Chip experiments (**Fig. 6b**). For 2,293 of the 2,367 predicted operons, probe pairs were tiled 200 bp upstream of the first gene of the operon. For the remaining 75 predicted operons, probe pairs were tiled 150 bp upstream of the first gene of the operon. Finally, for purpose (iii), probe pairs found every 5 bp on both strands in intergenic regions and every 15 bp on both strands of the first 85 bp of each predicted ORF, as well every 15 bp on both strands of small (that is, <700 bp) genes lacking homology to other genes (**Fig. 6c**).

Identification of transcriptional start sites. Transcriptional start sites were predicted by analysis of the correlation patterns from 2,513 *C. crescentus* genes. These 2,513 genes include the first gene of the 2,367 predicted operons and an additional 146 genes with 5-bp tiling at least 80 bp upstream of the start codon. (The 2,513 genes can be found at http://www.stanford.edu/group/caulobacter/CauloHI1/.) The start-site prediction involved two steps: (i) predict the transcriptional start site using the probe cross-correlations by scoring each possible transcript boundary and choosing the boundary with the highest score and (ii) assess the reliability of the prediction. The score for each possible boundary is found by summing each of the *m* nondiagonal $\text{probe}_{i,j}$ ($i \neq j$) correlations for the probes that fall downstream of the boundary (to the end of the ORF) minus an offset term. The offset term is specific to each gene, and is chosen as the correlation value that is equally likely to be produced by two probes inside the mRNA as by two probes in different mRNAs (**Supplementary Methods** online).

After identifying the boundary with maximum score, three values are calculated to determine if the boundary prediction is reliable (step 2 listed above): (i) the maximum score, (ii) the boundary average score, defined as the maximum score divided by *m*, the number of probe correlations used to calculate the maximum score, and (iii) the gene average score (**Supplementary Methods**). Each of these values must be greater than empirically determined cutoff values for the prediction to be deemed reliable. After determining the boundary probe, the final start-site prediction was determined by subtracting 6 bp from the position of the 13th nucleotide of the boundary probe. The 6-bp offset was found by training on 35 previously determined start sites (**Supplementary Methods**).

A Matlab program called the Transcriptional Start Identifier (TSI) program (**Supplementary Methods**) applied to the probe cross-correlation matrices for the cell cycle and the heavy metal response datasets implemented these operations.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS
We thank Esteban Toro for comments on the manuscript and Michael Laub and Michael Mittman for assistance in designing the CauloHI1 chip. This work was supported by National Institute of Health (NIH) grants GM32506 (L.S.) and 5R24GM73011-2 (H.H.M., L.S.), Department of Energy Office of Science grant DE-FG02-01ER63219 (H.H.M., L.S.), NIH training grant T32 HG00044 (P.T.M.) and Damon Runyon Cancer Research Foundation Fellowship DRG-1880-05 (N.J.H.). P.H. was funded by the DOE Genomes to Life program under the auspices of the University of California, Lawrence Berkeley National Laboratory, Contract no. DE-AC02-05CH11231. H.L. was supported by a Stanford Graduate Fellowship.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. McAdams, H.H. & Shapiro, L. A bacterial cell-cycle regulatory network operating in time and space. *Science* **301**, 1874–1877 (2003).
2. Ausmees, N. & Jacobs-Wagner, C. Spatial and temporal control of differentiation and cell cycle progression in *Caulobacter crescentus. Annu. Rev. Microbiol.* **57**, 225–247 (2003).
3. Skerker, J.M. & Laub, M.T. Cell-cycle progression and the generation of asymmetry in *Caulobacter crescentus. Nat. Rev. Microbiol.* **2**, 325–337 (2004).
4. Viollier, P.H. & Shapiro, L. Spatial complexity of mechanisms controlling a bacterial cell cycle. *Curr. Opin. Microbiol.* **7**, 572–578 (2004).
5. Laub, M.T., McAdams, H.H., Feldblyum, T., Fraser, C.M. & Shapiro, L. Global analysis of the genetic network controlling a bacterial cell cycle. *Science* **290**, 2144–2148 (2000).
6. Collier, J., Murray, S.R. & Shapiro, L. DnaA couples DNA replication and the expression of two cell cycle master regulators. *EMBO J.* **25**, 346–356 (2006).
7. Hottes, A.K., Shapiro, L. & McAdams, H.H. DnaA coordinates replication initiation and cell cycle transcription in *Caulobacter crescentus. Mol. Microbiol.* **58**, 1340–1353 (2005).
8. Holtzendorff, J. *et al.* Oscillating global regulators control the genetic circuit driving a bacterial cell cycle. *Science* **304**, 983–987 (2004).
9. Quon, K.C., Yang, B., Domian, I.J., Shapiro, L. & Marczynski, G.T. Negative control of bacterial DNA replication by a cell cycle regulatory protein that binds at the chromosome origin. *Proc. Natl. Acad. Sci. USA* **95**, 120–125 (1998).
10. Laub, M.T., Chen, S.L., Shapiro, L. & McAdams, H.H. Genes directly controlled by CtrA, a master regulator of the *Caulobacter* cell cycle. *Proc. Natl. Acad. Sci. USA* **99**, 4632–4637 (2002).
11. Gomes, S.L., Gober, J.W. & Shapiro, L. Expression of the *Caulobacter* heat shock gene *dnaK* is developmentally controlled during growth at normal temperatures. *J. Bacteriol.* **172**, 3051–3059 (1990).
12. Crosson, S., McGrath, P.T., Stephens, C., McAdams, H.H. & Shapiro, L. Conserved modular design of an oxygen sensory/signaling network with species-specific output. *Proc. Natl. Acad. Sci. USA* **102**, 8018–8023 (2005).
13. Hu, P., Brodie, E.L., Suzuki, Y., McAdams, H.H. & Andersen, G.L. Whole Genome Transcriptional Analysis of Heavy Metal Stresses in *Caulobacter crescentus. J. Bacteriol.* **187**, 8437–8449 (2005).
14. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, G.M. Systematic determination of genetic network architecture. *Nat. Genet.* **22**, 281–285 (1999).
15. Tjaden, B., Haynor, D.R., Stolyar, S., Rosenow, C. & Kolker, E. Identifying operons and untranslated regions of transcripts using *Escherichia coli* RNA expression analysis. *Bioinformatics* **18** Suppl 1, S337–S344 (2002).
16. Kelly, A.J., Sackett, M.J., Din, N., Quardokus, E. & Brun, Y.V. Cell cycle-dependent transcriptional and proteolytic regulation of FtsZ in *Caulobacter. Genes Dev.* **12**, 880–893 (1998).
17. Mullin, D.A. & Newton, A. Ntr-like promoters and upstream regulatory sequence *ftr* are required for transcription of a developmentally regulated *Caulobacter crescentus* flagellar gene. *J. Bacteriol.* **171**, 3218–3227 (1989).
18. McGrath, P.T., Iniesta, A.A., Ryan, K.R., Shapiro, L. & McAdams, H.H. A dynamically localized protease complex and a polar specificity factor control a cell cycle master regulator. *Cell* **124**, 535–547 (2006).
19. Roberts, R.C. & Shapiro, L. Transcription of genes encoding DNA replication proteins is coincident with cell cycle control of DNA replication in *Caulobacter crescentus. J. Bacteriol.* **179**, 2319–2330 (1997).
20. Osteras, M., Stotz, A., Schmid Nuoffer, S. & Jenal, U. Identification and transcriptional control of the genes encoding the *Caulobacter crescentus* ClpXP protease. *J. Bacteriol.* **181**, 3039–3050 (1999).

21. Bailey, T.L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
22. Liu, X., Brutlag, D.L. & Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*, 127–138 (2001).
23. Bailey, T.L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).
24. Herring, C.D. *et al.* Immobilization of *Escherichia coli* RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. *J. Bacteriol.* **187**, 6166–6174 (2005).
25. Reisenauer, A., Mohr, C.D. & Shapiro, L. Regulation of a heat shock sigma32 homolog in *Caulobacter crescentus*. *J. Bacteriol.* **178**, 1919–1927 (1996).
26. Craven, M., Page, D., Shavlik, J., Bockhorst, J. & Glasner, J. A probabilistic learning approach to whole-genome operon prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 116–127 (2000).
27. Gober, J.W. & Shapiro, L. A developmentally regulated *Caulobacter* flagellar promoter is activated by 3′ enhancer and IHF binding elements. *Mol. Biol. Cell* **3**, 913–926 (1992).
28. Helmann, J.D. The extracytoplasmic function (ECF) sigma factors. *Adv. Microb. Physiol.* **46**, 47–110 (2002).
29. Brun, Y.V. & Shapiro, L. A temporally controlled sigma-factor is required for polar morphogenesis and normal cell division in *Caulobacter*. *Genes Dev.* **6**, 2395–2408 (1992).
30. Malakooti, J., Wang, S.P. & Ely, B. A consensus promoter sequence for *Caulobacter crescentus* genes involved in biosynthetic and housekeeping functions. *J. Bacteriol.* **177**, 4372–4376 (1995).