Supplementary Methods for

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

Patrick T. McGrath, Honglak Lee, Li Zhang, Antonio A. Iniesta, Alison K. Hottes, Meng How Tan, Nathan J. Hillson, Ping Hu, Lucy Shapiro, and Harley H. McAdams

Nature Biotechnology (2007)

Online sites

Supplementary figures cited below are available online at the Nature Biotechnology website.

Data tables cited below are available online at http://www.stanford.edu/group/caulobacter/startsite/ which is hereafter referenced as startsite.

Experimental data sets

Caulobacter cell cycle time series. Swarmer cells were isolated using standard procedures and allowed to proceed through the cell cycle (150 minutes in these conditions) in minimal media (M2G). Samples were taken every 12 minutes for 156 minutes. RNA was extracted from each sample, reverse transcribed into cDNA, fractionated, labeled, and then hybridized onto the Caulobacter array using the method of Lizewski et al. The 72 minute time point assay failed. After scanning and normalizing the intragene probe signals from the CauloHI1 arrays, we generated a time series of expression for each gene by averaging the difference between perfect match and mismatch of each probe pair within a gene after discarding outliers. Outliers were defined as probe pairs with an average expression less than zero or greater than three standard deviations from the other intragene probe pairs. Prior to this averaging, probe pairs within 25 bp of each other were averaged together to prevent over sampling of either end of the gene. The cell cycle data set from the chips for the 13 time points (excluding the 72 min point) is in Table T8 online at startsite.

Caulobacter environmental metal response data. Hu et al. measured Caulobacter transcriptional response to five different heavy metal salts: cadmium sulfate, potassium chromate, potassium dichromate, sodium selenite, and uranyl nitrate in 49 experiments using the CauloHI1 chip. Hu et al. also exposed Caulobacter to lead nitrate, however, the addition of lead nitrate had a minimal effect on the Caulobacter transcriptome so this data was excluded from their published report. In the current analysis, we have included the lead nitrate data. Hu et al. focused on changes in gene expression levels and thus used only the intra-gene probe data. Here, we use the probe cross-correlation data for upstream probes to estimate transcriptional start sites.
Data analysis

Each array was scaled by a normalization constant so that the average signal from all probe pairs (PM-MM) on the chip was one. The average cell cycle expression for each gene (excluding the 72 minute time point) was created from the probe pairs taken from the predicted coding region as follows. Due to the design of the chip, probe pairs are not distributed uniformly throughout the gene (the guaranteed 200bp tiling upstream of predicted operons created heavy 3’ tiling of some genes). Therefore, the PM-MM data from probe pairs within 25 bp of each other were averaged together into probe pair groups. Outliers (defined as probe pairs groups with average expression less than zero or greater than 3 standard deviations above the mean of all probe pair groups) were discarded, and probe pair groups were then averaged together into a single cell cycle profile for the entire gene.

Primer extension

Primer extension was performed as described by Ausubel, et al.⁴, with the Superscript II Reverse Transcriptase (Invitrogen). The oligonucleotide primer 5’-CAGTTCCGATCGCGAAATC-3’ is located 55 to 75 bp downstream of the rcdA predicted translational start site, and the annealing temperature used was 45ºC. RNA was extracted as previously described for NA1000 or rcdA₅ mutant cells (LS4185).

The TSI algorithm

Cell cycle and metal stress data was analyzed separately. Matlab software to run the TSI algorithm is available online at startsite. Each chip was scaled by a normalization constant so that the average signal from all probe pairs (PM-MM) on the chip was one. Each gene with 5bp tiling at least 80bp upstream of the predicted translational start site was included for this analysis. For each gene meeting this criterion, we separated the set of probe pairs found either in the gene and or 200 bp upstream of the predicted start codon. Probe pairs found upstream of the predicted start codon were grouped together if they were within 5bp of each other to insure regular 5bp spacing throughout the upstream region (typically unnecessary, but sometimes required due to the nature of the tiling on the chip). Probe pairs found in the 3’ end of genes were grouped together every 50bp (the guaranteed tiling 200 bp upstream of each predicted operon often created heavily tiled 3’ ends). The exact grouping of probes can be found at www.stanford.edu/group/caulobacter/CauloHI1. We represent Affymetrix chip data of a single gene as an \( m \times n \) matrix \( X \) (where \( m \) denotes the number of probes, and \( n \) the number of experiments). Here \( X \) includes all the probes within the gene and its upstream region. Let \( X_i \) be the \( i \)-th row of \( X \) (response of probe \( i \)). Then correlation matrix is defined as:

\[
C_{ij} = \frac{\text{Cov}(X_i, X_j)}{\text{Stdev}(X_i) \cdot \text{Stdev}(X_j)}
\]

Then this correlation matrix can be represented in 2-D block contour diagram which can be used to represent the correlation between probes. Here the red block (Figure 2a in the
paper) represents a strongly correlated (consecutive) block of probes, and the boundary of the block will roughly correspond to the plus one site of this gene. To describe this idea more formally, we measure a score of a block (starting from probe \(k\) and ending with the last probe in the gene \(l\)) with the following formula:

\[
A_{k:l,\theta} = \sum_{k \leq i < j \leq l} (C_{ij} - \theta),
\]

where \(\theta\) is a cut-off constant obtained from the statistics of the probes known to be within the gene. More specifically \(\theta\) 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 (found by comparing a histogram of correlations between all probes (noise) throughout the entire chip data and a histogram of probes found inside of genes of comparable intercorrelation). \(\theta\) is empirically approximated by the equation

\[
\theta = .05 + .5 \times G,
\]

where \(G\) is the average of non-diagonal correlations for probes that are within the gene. Finally we pick the best \(k\) which maximizes the \(A_{k:l,\theta}\):

\[
k^* = \arg \max_k A_{k:l,\theta}
\]

We also evaluate the reliability of this prediction by calculating (i) the maximum score (the maximum value of \(A\) must be greater than 30), (ii) the boundary average score (the maximum value of \(A\) divided by the number of correlations within the boundary must be greater than .15), and (iii) the gene average score (\(G\) must be > .2). If these three criteria are satisfied the prediction is determined to be reliable. These cutoff scores were determined empirically by considering the accuracy of predictions as determined by previously biochemically determined start sites. Following the prediction of the boundary, the transcriptional start site was determined as the location of the 13\(^{th}\) nucleotide in the boundary probe minus 6 bp.

**Identification of promoter activity**

For genes predicted to have one transcriptional start site, the activity of the promoter was estimated by averaging the probe signal for probes downstream of the start site after first grouping and averaging expression of probe pairs within 25 bp of each other. Outliers were discarded (see above). For genes predicted to have two start sites, each promoter’s contribution to the overall gene expression was estimated. The activity of the upstream promoter was estimated by averaging the probe pair expression of probes between the two start sites after discarding outliers. The activity of the second promoter was estimated by averaging the probe pair expression in probes downstream of the second start site after grouping and averaging expression of probe pairs within 25 bp of each other, then the subtracting the estimated expression of the first promoter. If the averaged signal of the upstream promoter was greater than the averaged signal of the probe pairs
downstream of the second start site, the upstream promoter profile was limited to the expression profile observed downstream of the second start site.

### Clustering and motif identification

A normalized expression profile (see above for the normalization procedure) of the cell cycle transcriptional response was created for 450 promoters with (i) transcriptional start site predictions and (ii) minimal expression less than 0.35 times the maximum expression. A single normalized expression profile of the metal stress response was created for the 607 promoters with (i) transcriptional start site predictions, and (ii) minimal expression less than 0.5 times maximum expression. The 450 cell cycle profiles were hierarchically clustered into 14 groups with similar expression patterns (Supplementary Fig. 3 online, Table T7 online at `startsite`), while the 607 metal stress profiles were hierarchically clustered into 10 groups (Supplementary Fig. 4 online, Table T6 online at `startsite`). (Hierarchical clustering of gene expression profiles over multiple CauloHI1 experiments was done using Cluster 3.06.) Criteria for motif identification with MEME were: (i) zero or one occurrence per gene, (ii) minimum width of 8 bp, and (iii) maximum width of 35 bp. Motifs were not searched for on the reverse complement strand.

The BioProspector search was done with parameters: (i) motif width of 8 bp with no gaps and (ii) the motif may occur on only some input sequences. In several clusters (cc_a, cc_d, cc_e, cc_f, cc_g, cc_k, cc_m), there was a fixed distance between the positions of the first two 8 bp long motifs. Then we also searched motifs of 16 bp long among those clusters. Both strands were searched with BioProspector.

#### Verification of motif cc_2

An independent study was carried out using the CauloHI1 Affymetrix whole-genome microarray that is tiled every five base pairs across all intergenic regions. Design details of the chip are online at www.stanford.edu/group/caulobacter/CauloHI1. Probe cross-correlation matrices were computed for all probe pairs in each intergenic region. Groups of probes with signals well above noise are strong candidates to be caused by expressed RNAs. In this manner, 28 sRNAs (small non-coding regulatory RNAs) were identified within *Caulobacter* intergenic regions. Among these was an 81 nucleotide RNA expressed from the region between genes CC1840 and CC1841 (Supplementary Fig. 5 online). An RNA product from this intergenic region was confirmed by Northern blot analysis. The +1 transcription start site was determined using 5’ RACE analysis. Using this start site, we identified a 21 nucleotide element 19 bases upstream of the start site that perfectly matched motif cc_2. This gene was not among those used to define the motif, and thus it is an independent identification of the sequence. The temporal expression profile of this sRNA is similar to the profile of the class of genes used to identify this motif, and it belongs to a group of stress response sRNAs. Another stress response sRNA identified in the Affymetrix analysis of small RNAs, CncR, also contained this highly conserved promoter motif. In this case, replacement of the promoter motif with a xylose-inducible promoter resulted in constitutive expression of
the CncR sRNA, verifying that the conserved promoter element is necessary for stress induced transcription.

**Verification of motif cc_3**

We have identified cc_3 as the binding motif for the FixK transcriptional regulator (Supplementary Fig. 6 online). FixK is a global regulator of respiration in *Caulobacter*\(^7\). FixK binding sites are present in the promoter regions of the *fixLJ* operon, *fixT*, and the operon for terminal oxidases Cytbb\(_3\), Cytbd, Cytbo\(_3\), and Cytaa\(_3\), as well as *ftrB* and *flmH*. The cc_3 motif is also present in the FixK-binding sites of rhizobia and the FNR-binding sites of *E. coli*\(^8\). We have shown that a deletion FixK alters the expression of all *Caulobacter* genes bearing cc_3 motifs in their promoter region.

**Verification of motif cc_5**

The sigma54 binding motifs is present in promoter regions of multiple flagellar and chemotaxis operons in *Caulobacter crescentus*. For example, Supplementary Figure 9 online shows the cc_5 motif in the promoter region of the *flaN* gene\(^9\).

**Verification of motif cc_7**

The CC2883 gene encodes ECF sigma factor SigU. The *sigU* gene is autoregulated by its product. The promoter region of *sigU* is homologous to motif cc_7 shown in Figure 3 of the paper. To verify that the conserved sequence is required for *sigU* transcription, we created two mutant versions of the *sigU* promoter (Supplementary Fig. 7B online) and constructed genetic fusions of the wild type and each of the two mutant promoters to *lacZ*. We then assayed β-galactosidase activity of each promoter construct (Supplementary Fig. 7A online). As shown, specific base changes in the highly conserved regions eliminated promoter activity, thus verifying that the cc_7 motif is a required promoter element. Twenty six *Caulobacter* genes have the cc_7 motif upstream of their +1 transcription start sites.

**Bacterial strains and growth conditions**

*C. crescentus* strains were grown in peptone-yeast extract (PYE) complex media or M2 minimal salts + 0.2% glucose (M2G) minimal media\(^10\) at 28°C for all strains. The concentration of the antibiotic chloramphenicol used was 2 μg/ml. Plasmids were introduced into CB15N (Evinger and Agabian, 1977) by electroporation.

**Construction of plasmids**

100bp upstream of the ATG translational start site of CC2883 was cloned upstream of the *lacZ* coding sequence in pPR9TT (Santos et al., 2001) using the XhoI and HindIII cut sites. For each promoter region, four oligos were phosphorylated by T4 polynucleotide kinase, mixed together, heated to 95°C, and then annealed. The oligos used are tabulated below. The wild type promoter region of CC2883, containing the motif GGAACCTTCGCCGCTGCGGGCCGTTGT, was generated using oligos σ\(_U\)-1, σ\(_U\)-2, σ\(_U\)-3, and σ\(_U\)-4. The front-mutated promoter region of CC2883, containing the motif
GGGTTCCTCGCCGCTGCAGCCGCTGCGTGT, was generated using oligos σU-2, σU-3, σU-5, and σU-6. The back-mutated promoter region of CC2883, containing the motif GGAACCTTCGCCGCTGCAGCTGACGT, was generated using oligos σU-1, σU-3, σU-7, and σU-8.

β-galactosidase assays

The β-galactosidase activity of strains containing the promoter-lacZ plasmids was assayed in log-phase cultures in PYE media using o-nitrophenyl-β-D-galactoside (ONPG) \(^{11}\). As a control, the activity of a strain containing the original pPR9TT vector was also assayed. The reported β-galactosidase activities represent the average of at least three independent assays.
Verification of motif $m_5$

The CC1891 gene encodes a protein that is upregulated upon exposure to uranium\textsuperscript{12}. The promoter region of CC1891 contains a sequence that is similar to the $m_5$ motif shown in Figure 4 of the paper. To verify that this motif is a functional promoter element, site-specific base substitutions were created in the highly conserved regions of the CC1891 promoter. The wild type and mutant versions of the CC1891 promoter were fused to a \textit{lacZ} reporter gene, and the $\beta$-galactosidase activity of each promoter construct was assayed (Supplementary Fig. 8A online). As shown, the mutant promoter was inactive and could not be activated in the presence of uranium, verifying that motif $m_5$ is a required promoter element. The $m_5$ motif is present in the promoter region of 11 \textit{Caulobacter} genes.

\textbf{Reagents}

T4 DNA ligase and endonucleases were purchased from Fermentas (Hanover, MD) and New England Biolabs (Ipswich, MA). DNA oligos were purchased from the Stanford Protein and Nucleic Acid Biotechnology Facility (Stanford, CA). One Shot Top10 chemically competent \textit{E. coli} and 0.1 cm electroporation cuvettes were purchased from Invitrogen (Carlsbad, CA). DNA sequencing was performed by Sequetech (Mountain View, CA). KOD Hot Start DNA polymerase was purchased from Novagen (Madison,
WI). DNA miniprep and gel extraction kits were purchased from Qiagen (Valencia, CA). Depleted uranyl nitrate was purchased from Sigma-Aldrich, and a stock solution was prepared as described previously.

**P<sub>cc1891</sub>-LacZ reporter strains and assays**

The promoter for gene cc1891, P<sub>cc1891</sub>, contains a perfect match to the consensus predicted metal response motif m<sub>5</sub> (Supplementary Fig. 8B online), extending from 63 to 37 bp upstream of the predicted transcriptional +1 site. To test the hypothesis that the predicted metal response motif m<sub>5</sub> is responsive to uranium stress (in accordance with our microarray experiments), and that mutating the m<sub>5</sub> motif would result in attenuating the uranium stress response, we set out to construct wildtype and mutant P<sub>cc1891</sub>-LacZ reporter strains and assay them in the absence or presence of uranium stress.

**Cloning of the transcriptional fusion LacZ reporter vector pNJH185**

The vector pPR9TT<sup>13</sup> is often used to make promoter activity reporter constructs by making a translational fusion of the protein of interest to LacZ. Reporter constructs based upon pPR9TT are dependent upon the ribosomal binding site (RBS) of the protein of interest. In the case of a weak RBS, a pPR9TT-based reporter may provide weakly detectable LacZ activity even if there is moderate transcription from the promoter of interest. We converted the pPR9TT translational fusion vector into a transcriptional fusion vector with a strong RBS, by placing the RBS through the start ATG codon of pRKLac290 in frame with the LacZ protein sequence of pPR9TT, yielding the transcriptional fusion LacZ reporter vector pNJH185. To do this, we digested pPR9TT with HindIII and PstI (protocol as directed by the manufacturer), and then purified the cut vector backbone by electrophoresis through 1.2% agarose, followed by gel extraction. Oligos NJH214 and NJH215 (see below), 100 pmol/μL each, were mixed 1:1 in a 50 μL total volume, heated at 94 ºC for 2 min and then annealed at room temperature. This annealed mixture of NJH214/215 was diluted 1:300, and then mixed 1:1 with HindIII/PstI digested pPR9TT vector backbone for ligation overnight at 16 ºC with T4 DNA ligase (protocol as directed by the manufacturer). This ligation mixture was then transformed into OneShot Top10 chemically competent E. coli cells (protocol as directed by the manufacturer). The sequence of resulting plasmid, pNJH185, was confirmed by primer extension sequencing using oligo NJH155 (see below).

**Cloning of wildtype and mutant cc1891 promoter activity LacZ reporters**

The genomic region containing the cc1891 promoter was amplified from Caulobacter CB15N genomic DNA with KOD Hot Start DNA polymerase and oligos NJH220 and NJH221 (see below). The 50 μL PCR reaction mixture, containing 5% DMSO, was made per the manufacturer's protocol. The PCR reaction was initiated by 1 min 45 sec melting at 94 ºC, followed by 32 cycles of 15 sec melting at 94 ºC, 30 sec annealing at 62 ºC, and 1 min 15 sec extension at 68 ºC. This PCR product was purified by electrophoresis through 1.2% agarose, and then gel extracted. The purified PCR product was then
reamplified as described immediately above except using oligos NJH216 and NJH217 (see below). This second round PCR product was then purified as above, then digested with BgIII and HindIII (protocol as directed by manufacturer), ligated (as described above) with similarly digested pNJH185 vector backbone, and then transformed (as described above) into OneShot Top10 chemically competent *E. coli* cells. The sequence of the resulting plasmid, pNJH187, was confirmed by primer extension sequencing using oligo NJH155.

The promoter for *cc1891*, P\textsubscript{*cc1891*}, contains a perfect match to the consensus predicted metal response motif m\_5 (Supplementary Fig. 8B online). To mutate the m\_5 motif, we made single base pair substitutions at 4 of the most conserved positions within the consensus motif (see Fig. 4 of the paper and Supplementary Fig. 8B online), all of which are upstream of the -35 region of P\textsubscript{*cc1891*}. We made these mutations in the m\_5 motif of P\textsubscript{*cc1891*} using the splicing by overlap extension (SOE) method\textsuperscript{14}. pNJH187 template was amplified with KOD Hot Start DNA polymerase using oligos NJH216 and NJH225 (5’ SOE PCR reaction, see below), or NJH224 and NJH228 (3’ SOE PCR reaction, see below), as described above. After purifying these first round SOE PCR products (as described above), they were mixed 1:1 as template for the second round SOE PCR reaction, and ten amplified using oligos NJH216 and NJH228, as described above, except the extension time was 1 min 30 sec. This second round SOE PCR product was then purified as above, then digested with BgIII and HindIII (as described above), ligated (as described above) with similarly digested pNJH185 vector backbone, and then transformed (as described above) into OneShot Top10 chemically competent *E. coli* cells. The sequence of the resulting plasmid, pNJH188, was confirmed by primer extension sequencing using oligo NJH155.

---

**Wildtype and mutant cc1891 promoter activity LacZ reporter Caulobacter strains**

*Caulobacter crescentus* CB15N Δacc1634 (strain LS4358) was transformed with plasmid pNJH187 or pNJH188 to yield the wildtype and mutant cc1891 promoter activity LacZ reporter strains NJH332 and NJH333, respectively. The in frame Δacc1634 deletion reduces the background β-galactosidase activity of *Caulobacter*. The LS4358 strain was transformed by electroporation as previously described\textsuperscript{10}.

**Wildtype and mutant cc1891 promoter activity LacZ reporter activity assays**

Cultures of strains NJH332 (wildtype) and NJH333 (mutant) were grown overnight at 28°C in M2G media\textsuperscript{10}. Stationary phase cultures were diluted to an OD\textsubscript{600nm} of 0.1 with fresh M2G, and then grown for an additional 2 hr at 28 °C to resume exponential growth before stressing the cells with either mock treatment or 200 μM uranyl nitrate. The stressed cultures were then grown for 2 hr before liquid culture β-galactosidase assays were conducted, as previously described by Miller\textsuperscript{11}.

**Constructs used in verification of motif m\_5**
DNA oligos$^a$.

NJH155  aagcaacggccggaggggtgg
NJH214  Pagtttacacaggaacagctatgaccatgtcctgca
NJH215  Pggacatgtcatagcttctcttgtaga
NJH216  ccAgcAgAgatctctgggacgtcaacgcctatgg
NJH217  tcAcaatAAGcTTGcatccccgcgaaggacgttagc
NJH220  ctcaggacgtagccgcacaaat
NJH221  tgctctcgacgcggacata
NJH224  taaatcccgtcGGtaccgcacttGGagtgtctctggtc
NJH225  ggccagagaccacctCCaattgctgtaCCgacgggattta
NJH228  t cacctgccataagaaactgttagttaccggtag

$^a$ Underlined portions indicate endonuclease recognition sites (NJH216: BglII; NJH217: HindIII) or sticky single-stranded ends of annealed oligo pairs that are complimentary to digested DNA (NJH214: 5' HindIII compatible end, 3' PstI compatible end). Bold portions indicate sites of mutation (within motif m_5: NJH224 and NJH225). The character "P" at the 5' end of the oligo sequence indicates 5' phosphorylation.

Plasmids.

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>pPR9TT</td>
<td>translational fusion LacZ reporter vector</td>
<td>13</td>
</tr>
<tr>
<td>pNJH185</td>
<td>pPR9TT with pRKLacl290 RBS/</td>
<td>This study</td>
</tr>
<tr>
<td></td>
<td>transcriptional fusion LacZ reporter vector</td>
<td></td>
</tr>
<tr>
<td>pNJH187</td>
<td>pNJH185-Pcc1891_LacZ</td>
<td>This study</td>
</tr>
<tr>
<td>pNJH188</td>
<td>pNJH185-Pcc1891mutant_LacZ</td>
<td>This study</td>
</tr>
</tbody>
</table>

Strains

<table>
<thead>
<tr>
<th>Strain</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>NJH332</td>
<td><em>Caulobacter</em> CB15N Δcc1634</td>
<td>pNJH187</td>
</tr>
<tr>
<td>NJH333</td>
<td><em>Caulobacter</em> CB15N Δcc1634</td>
<td>pNJH188</td>
</tr>
</tbody>
</table>

This study
References