Molecular Cell, Volume *24*

Supplemental Data

The Transition between Transcriptional Initiation and Elongation in *E. coli* Is Highly Variable and Often Rate Limiting

Nikos B. Reppas, Joseph T. Wade, George M. Church, and Kevin Struhl

SUPPLEMENTAL RESULTS

Annotation and validation of identified transcripts

Comparison of the location of transfrags to the gene map of *E. coli* MG1655 (Keseler et al., 2005) to a gene map rotated 1 Mb (essentially random data) confirms a highly significant concordance between the coordinates of transfrags and the boundaries of annotated genes (χ^2 p = 1.2e-316). We identified 210 transfrags that did not match annotated transcripts (see Supplementary Methods). We manually selected 58 of these that appeared to represent *bona fide* novel transcripts (9 intragenic, 25 antisense, and 24 intergenic; Table S2); the remainder consisted largely of fragments of very lowly expressed transcripts and were thus deemed artifacts. 16 of the 58 unannotated transcripts have either been previously reported as novel transcripts (Tjaden et al., 2002) and/or computationally predicted (Carter et al., 2001; Chen et al., 2002; Rivas et al., 2001; Saetrom et al., 2005).

Promoters in *E. coli* and other prokaryotes are enriched in relatively low melting temperature (T_m) regions that we refer to as T_m troughs (Kanhere and Bansal, 2005). We first determined the central positions of all T_m troughs in the genome whose depth is greater than 2°C. We then determined the minimum distance between each of the 58 novel transfrag 5’ ends and a T_m trough, comparing this distribution of distances to that generated using rotated (i.e. randomized) transfrag coordinates. The 5’ ends of novel transfrags are highly enriched in their proximity to T_m troughs relative to the rotated data (Mann-Whitney p = 2.6e-6), strongly suggesting that these novel transfrags are transcribed from genuine promoters.

Annotation and validation of identified σ^{70} promoters

We compared our σ^{70} ChIP peak coordinates to 484 known σ^{70} binding sites and the same number of negative control sites and determined for each σ^{70} peak whether a control coordinate was found within a series of absolute distances (see Supplementary Methods). The resulting data were used to obtain conservative estimates for the specificity, sensitivity, and false discovery rate (FDR) of our peak calls (Figure S1). Using a 180 bp threshold (see Supplementary Methods) for the distance allowed between observed and expected σ^{70} sites, the specificity, sensitivity, and FDR are 98.8%, 62.6%, and 1.9%, respectively. The 62.6% sensitivity is comparable to that observed in a ChIP-chip study of RNAP immobilized at promoters by treatment with rifampicin (Herring et al., 2005).

We observe a highly significant overlap between the location of the 1286 σ^{70} peaks identified here and the 1111 rifampicin-immobilized RNAP ChIP peaks (Herring et al., 2005): 58% of σ^{70} peaks are within ±300 bp of immobilized RNAP peaks, compared to 15% for a random model in which the coordinates of the former are rotated 1 Mb around the chromosome. We also performed wavelet analysis (Herring et al., 2005) to identify any periodic patterns in the binding of σ^{70} across the entire genome. We identified a statistically significant periodicity of ~700 kb^{-1} (Figure S2), similar to the one observed previously in studies of genome-wide transcription and immobilized RNAP (Allen et al., 2003; Herring et al., 2005; Jeong et al., 2004).

We conservatively annotated σ^{70} peaks with respect to known and predicted transcripts, and to annotated genes with no known or predicted σ^{70}-dependent promoters (see Supplementary Methods; Tables S5, S6). Given the density of σ^{70} peaks, the density of potential transcript starts across the *E. coli* genome, and the effective accuracy and resolution of the ChIP-chip data, σ^{70} peaks were often associated with more than one possible transcription unit; 555 peaks were of this kind (Class D peaks). Approximately the same number of peaks could be associated with a unique transcript start site, with most (405) corresponding to known or predicted σ^{70} promoters (Class C) and the remaining 165 peaks being associated with the 5' ends of genes that are not known or predicted to be transcribed by $E\sigma^{70}$ (Class B) (Bockhorst et al., 2003; Keseler et al., 2005). 51 genes associated with Class B σ^{70} peaks (e.g., *malY*, *cheA*, and *nuoF*) lie within known $E\sigma^{70}$ operonic transcripts (Tjaden et al., 2002; Keseler et al., 2005); the rest lie within solely computationally predicted operons (Bockhorst et al., 2003). Thus, many genes are transcribed both from a dedicated promoter immediately upstream, and as a downstream gene in an operon.

We also observe 161 σ^{70} peaks within the coding sequences of genes (Class A; e.g., *uhpT*, *xylG*, and *yagX*; Figure 2) and in the intergenic regions between convergently transcribed genes (e.g., between *yncC* and *yncD*). Although these peaks tend to be of lower height relative to those of the entire set of σ^{70} peaks (Table S5), as a whole they tend to be proximal to T_m troughs relative to rotated data (Mann-Whitney p = 3.4e-7), indicating that these ChIP peaks are likely highly enriched in real promoters. Significantly, 39% of them were also noted (±300 bp) in the rifampicin RNAP ChIP-chip study (Herring et al., 2005). Also, as would be expected, Class A peaks are observed at a much lower frequency than would be expected by chance: 12.5% versus 62.7%, respectively (Table S5). We speculate that most of these unusual binding sites represent the promoters of unknown/misannotated protein-coding genes or small noncoding RNAs that have not been hitherto described. Indeed, some of these σ^{70} peaks are associated with novel transfrags identified in our transcript analysis, e.g., at the 3' end of *uhpT* (Figure 2A).

Half-life analysis of genes that bind "poised" RNAP

It is possible that sites of "poised" RNAP represent regions are in fact transcribed, but that express transcripts of particularly short half-life. However, based on the results of a genome-wide determination of *E. coli* transcript half-lives (Selinger et al., 2003), genes associated with non-transfrag-associated σ^{70} peaks do not have statistically significant shorter half-lives than those with transfrag-associated ones (Figure 3). Alternatively, non-transfrag-associated peaks could be somehow artifactual. However, they are in fact strongly enriched for being proximal to the 5' ends of annotated gene starts (χ^2 p = 1.2e-13), strongly suggesting that most of them are biologically meaningful. On average, non-transfrag-associated peaks are of lower σ^{70} ChIP signal, but there is an extensive region of overlap between the peak height distributions of the two classes (Figure 3), demonstrating that transcriptional activity is not completely dependent on the level of promoter-bound RNAP.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Probe BLAST analysis. We BLASTed all 382177 50-mer probe sequences against the MG1655 genome to generate for each probe a cross-hybridization score, defined as $-\log_{10}$(E-value) of its second-best BLAST hit. When plotting genome coordinate versus cDNA or ChIP signal, probes with multiple exact hits were treated as multiple points on the coordinate axis.

Data Preprocessing. cDNA and ChIP datasets exhibited good raw reproducibility (Spearman $r \geq 0.84$). Raw cDNA data were linearly scaled such that both datasets had the same total hybridization intensity; no other normalization was performed. For ChIP data, we computed the genomic-DNA-normalized $\log_2$ IP/control ratio for each probe, positivized the data by adding to each log ratio the negative of the lowest such ratio (due to a varying local background $\log_2$ IP/control ratio signal), and then linearly scaled with respect to total $\log_2$ ratio signal. We smoothed the ChIP vs. genome coordinate signal by two rounds of sliding window averaging over 300 bp.

Transfrag Calling. Only probes with BLAST cross-hybridization scores < 2.7 were used to delineate transcription fragments (transfrags). Watson- and Crick-stranded data were analyzed separately (there was apparent cDNA double-stranding activity present in the reverse transcriptase used). We first subtracted 1.5× the modal intensity value from each probe intensity as an initial background correction step, zeroing negative intensities, and extracted all probes of nonzero intensity across whose 50 nt span all probes on the opposite strand were of lower intensity – both criteria had to be satisfied in both replicates. This procedure generated a list of expressed probes (EPs). Having removed isolated, weakly expressed EPs, we merged the remaining ones within 200 nt of each other into type 1 fragments, merging up to the nearest EP on the opposite strand. For each type 1 fragment, we computed the strand-specificity factor (SSF) as the ratio of the mean (replicate-averaged) median expression (MME) value of the EPs contained within its bounds to that of the corresponding EPs on the opposite strand, updating the SSF and MME after each merging. Having removed short type 1 fragments of low SSF, we merged the remaining ones within

600 nt of each other into type 2 fragments if their MMEs were sufficiently similar, merging up to the nearest type 1 fragment on the opposite strand; SSFs and MMEs were updated as above. There were two merging exceptions: (1) if the ends of consecutive type 1 fragments are located in the same annotated gene, but separated by a single opposite strand EP, they were still merged, and (2) mergeable consecutive type 1 fragments were not merged if they are separated by an intergenic region containing ≥1 opposite strand EP. We culled type 2 fragments that were of low MME/low SSR/EP density, were predicted weakly expressed and predicted to lie entirely within a predicted antisense gene region, or were apparent "side-noise" artifacts (short, non-uniform cDNA profiles abutting the 5′ and 3′ ends of highly expressed transcripts). Finally, some manual adjustments were made to type 2 transfrag boundaries based on visual inspection of the cDNA data, generating a final set of 1815 transfrags.

We identified outlier transfrags by comparing transfrag position to that of known genes (Keseler et al., 2005). Outlier transfrags are defined as those that (1) overlap an annotated gene by <25% of the gene's length or (2) overlap an intergenic region by >60% of the intergenic region's length or (3) are antisense totally across or internally to a gene, or, if partially antisense, the antisense distance is ≥ 250 nt.

ChIP Peak Calling. For every probe midpoint coordinate (PMC), we located the leftmost and rightmost PMCs 150 bp away. If the corresponding triad of $\log_2$ ChIP ratios represented a turning point in the genomic ChIP profile, we labeled the central PMC as the location of a maximum/minimum. Maxima within 300 bp of each other were merged, as were minima; in these cases, an average turning point genome coordinate and log ratio were computed. Peak height was computed as the larger of (ChIP signal at peak - ChIP signal at left flanking trough) and (ChIP signal at peak - ChIP signal at right flanking trough); a peak height threshold of 0.8 $\log_2$ units was imposed at this stage. We accepted a preliminary peak in replicate #1 if there existed a single preliminary peak in replicate #2 called within a window of 300 bp centered at the former, computing an average genome coordinate and average peak height as the *x*,*y* coordinates of the final peak.

We split these peaks into bins defined by peak height intervals, i.e. 0.85-0.95, 0.95-1.05, 1.05-1.15, etc., and determined whether the peak genomic coordinates in each bin were enriched in being proximal to the starts of annotated genes (Keseler et al., 2005) relative to those coordinates rotated around the chromosomal gene map. Such enrichment is expected to be a signature of true σ^{70}/β binding. The Mann-Whitney test was used to evaluate the statistical significance of the actual vs. rotated peak-to-gene-start distance distributions. For all rotation distances selected, peaks whose $\log_2$ heights were 1.25-1.35 were significantly closer to gene starts than random data. A peak height threshold of 1.25 was therefore selected, yielding a final set of 1286 σ^{70} and 1032 β ChIP peaks. There is good correlation between the raw σ^{70} and β ChIP peak profiles, the four pairwise σ^{70}-β dataset pairs displaying Pearson *r* values of 0.65, 0.54, 0.65, and 0.65. 57% of σ^{70} ChIP peaks were within ±300 bp of a β ChIP peak ("copeaks"). A BLAST score was computed for each ChIP peak as the fraction of probes with cross-hybridization scores > 2.7 falling within a 600 bp window centered at the peak. A BLAST score ≥ 0.15 was considered as potentially problematic − 15% of σ^{70} ChIP were of this variety.

Sensitivity-Specificity Analysis. To determine the specificity and sensitivity of ChIP peak calling, we performed ROC analysis of our predicted ChIP peak genome coordinates with respect to 484 nominally positive control coordinates and the same number of nominally negative control coordinates. Positive control coordinates were drawn from all known σ^{70} promoters (EcoCyc and RegulonDB; (Keseler et al., 2005; Salgado et al., 2006). In cases where such ≥2 sites were clustered, we calculated the length of the cluster as the distance between the extreme left and right sites; if this cluster distance was <300 bp, then site coordinates were averaged; if not then all sites within the cluster were ignored. (Large cluster distances might cause the corresponding σ^{70} ChIP signal to be too diffuse for defined peaking). Negative control coordinates were the midpoint coordinates of the 484 longest ORFs in *E. coli*, i.e., regions where there should be no RNAP binding sites. Note that it is highly probable that we are overestimating both the number of positive control binding sites – not all these promoters are expected to be active under the experimental conditions employed – and the number of negative control sites – at least a few long genes could harbor internal transcripts. We computed the distribution of absolute distances between the location of each σ^{70} ChIP peak and its closest positive/negative control coordinate for actual data and

1 Mb-rotated data. By varying the magnitude of this absolute distance and determining the resulting number of positive and negative coordinates falling within that distance from called ChIP peaks, we determined the specificity (# true negatives/(# true negatives + # false positives)), sensitivity (# true positives/(# true positives + # false negatives)), and false discovery rate (# false positives/(# true positives + # false positives)) of peak calling.

By inspecting the subset of 484 positive control σ^{70} sites that were both bound and could be unambiguously associated with gene starts (i.e., those starts > 1kb away from any other), the called ChIP peak lay between 350 bp upstream to 50 bp downstream of the gene start in 93% of cases (and upstream 94% of the time). In 95% of these cases, the distance between the known and predicted σ^{70} binding sites was ±180 bp, with a standard deviation of 58 bp.

σ^{70} ChIP Peak Annotation. σ^{70} ChIP peaks were first divided into σ^{70}-β copeaks − if the σ^{70} peak was within ±300 bp of a β ChIP peak − or orphan peaks. We constructed a comprehensive list of all known and predicted transcription start points (TSPs) for *E. coli* from databases and a variety of published studies. TSPs encompass annotated gene start codons (Keseler et al., 2005), known and computationally predicted σ^{70}-dependent transcription units (TUs) (EcoCyc, RegulonDB (Keseler et al., 2005; Salgado et al., 2006), (Bockhorst et al., 2003), transcripts inferred from high-density (although not tiled) microarray analysis (Tjaden et al., 2002), as well as known and computationally predicted small intergenic RNAs (Argaman et al., 2001; Carter et al., 2001; Chen et al., 2002; Kawano et al., 2005; Rivas et al., 2001; Saetrom et al., 2005; Vogel et al., 2003; Wassarman et al., 2001; Zhang et al., 2003). All feature coordinates were updated to the most recent version (U00096.2) of the *E. coli* MG1655 genome sequence.

The goal of the annotation process was to associate each ChIP peak with all TSPs within a given distance away. For gene start TSPs, this distance was 350 bp upstream to 50 bp downstream, and for all other TSPs (i.e., *bona fide* transcriptional start sites), it was ±160 bp (see above). Accordingly, 4 annotation classes could be defined. Class A: not associated with the 5' end of a known/predicted gene, e.g., intragenic, between convergently transcribed genes, or only associated with computationally predicted small intergenic RNAs; Class B: associated with the 5' end of an annotated gene that is not known or predicted to be transcribed by Eσ^{70}; Class C: associated with only 1 known/predicted σ^{70} promoter; Class D: associated with ≥2 known/predicted σ^{70} promoters.

Annotated σ^{70} ChIP peaks were first classified as being coincident with one of 1111 reliably called rifampicin ChIP RNAP peaks from (Herring et al., 2005) if the former was within ±300 bp of one of the latter.

Melting Temperature Analysis. We extracted every 15th Watson-strand 30-mer of the MG1655 genome and computed its melting temperature (T_m) using the program MELTING (Le Novere, 2001). We smoothed the chromosomal T_m profile by two rounds of sliding window averaging over 300 bp and computed troughs in an exactly analogous way was we determined peaks for the ChIP data. We reported the genomic coordinate and temperature depth of troughs >2 °C in depth.

ChIP-Transfrag Analysis. To compute traveling ratios (TRs), we selected transfrags with the following characteristics: (1) ≥1200 nt in length, (2) had no other transfrag 5′ ends within ±1.3 kb, (3) had a σ^{70}-β ChIP copeak within ±160 bp (relative to σ^{70} peak coordinate) from its 5′ end and no other σ^{70} or β peaks within ±1.3 kb, (4) had the beta peak downstream of the sigma peak relative to the strandedness of the transfrag (indicative of active transcription (Wade and Struhl, 2004). There were 59 such transfrags (Table S3). For both σ^{70} and β, we then computed the replicate-averaged absolute ChIP signal at the peak (p), at a point 800bp upstream – relative to the transfrag – of the peak (u_{800}), and at a point 800bp downstream (d_{800}); the TR is computed as $(u_{800}\text{-}u_{800})/(p\text{-}u_{800})$. 800 bp was chosen as this is sufficiently far from the promoter that the ChIP signal will be due entirely to elongating RNAP.

Retention ratio profiles for σ^{70} and β ChIP peaks associated with the same set of 59 transfrags were computed as $\log_{10}((p\text{-}u_i)/(p\text{-}d_i))$, for all i = 50, 100, 150, …, 850, 900 bp. This ratio will be positive when the peak is skewed in the direction of transfrag transcription. (Logarithmic values cannot be determined for negative $(p\text{-}u_i)/(p\text{-}d_i)$ ratios). Retention ratios were estimated for σ^{70} assuming stochastic release with a half-life of 7 s and an elongation rate of 30 nt/sec by calculating: $\log_{10}(1/(1\text{-}(c^*(1\text{-}(1/(10^{R(\beta)}))))))$ where c is the fraction of σ^{70} estimated to associate with elongating RNAP at any given position and R(β) is the retention ratio for β at that position.

As a statistical control for the above analyses, we selected all σ^{70}-β ChIP copeaks that had no other σ^{70} or β peaks within ±1.3 kb and treated them as if they were associated with ≥1200 nt Watson-strand transfrags. From the 510 such copeaks thus selected, we extracted 1000 random subsets of 59 and computed for σ^{70} and β the TR and the peak skew ratio profiles as above. We could them query how many times 1000 random median values of the TR value distribution or of the peak skew value distribution were less than the corresponding median values of the actual dataset. If 10 or 990 random medians were lower than the actual, this would correspond to a *p*-value of 0.01 for the actual dataset being different from (*i.e.*, either greater or less than) a random model. For σ^{70}, neither the TR nor the peak skew distribution medians were significantly different to random at the $p = 0.1$ level; for β, both distribution medians were significantly greater than a random model at a $p < 0.001$ level.

If ≥1 transfrag start(s) were found within ± 300 bp of a σ^{70} ChIP peak, the corresponding transfrags were associated with the peak; if the σ^{70} ChIP peak was found >300 nt downstream into a transcript, it was deemed an internal peak. These two classes of peak are denoted transfrag-associated (TA) ChIP peaks; the remaining σ^{70} ChIP peaks are non-TA (NTA). 35% of transfrags could not be associated with a σ^{70} ChIP peak.

Wavelet Analysis. The 1286 σ^{70} peaks were first culled of any peaks where the fraction of probes with cross-hybridization scores > 2.7 falling within a 600 bp window centered at the peak was ≥ 15%, thereby yielding 1121 pairs of genome coordinates and associated ChIP peak heights. We discretized these peak data into 2.5 kb intervals. Autocorrelation analysis demonstrated that there was no correlation between consecutive data points, allowing us to use a random permutation of the data as a null model. Wavelet analysis was performed in MATLAB using a Morlet wavelet with wavenumber 6. We evaluated the number of times a 2-D coordinate in the resulting period vs. genome coordinate wavelet gave a higher or lower signal than the same coordinate in 1000 similar wavelet analyses of random permutations of the data; we took as significant only those 2-D coordinates which scored higher and lower across all 1000 times, yielding a p<0.0001 wavelet significance map. To further assess statistical significance, we repeated the above analysis with 20 random permutations of the data instead of the actual data, in each case generating a corresponding p<0.0001 wavelet significance map. In this way, we showed that the ~700 kb^{-1} periodicity observed in the actual wavelet was robust across the 20 data randomizations.

Data Visualization. All annotation, cDNA, and ChIP data was visualized in the program SignalMap (Nimblegen) and MATLAB (Math Works). Raw data can be obtained from http://arep.med.harvard.edu/~nreppas/RNAP.

SUPPLEMENTAL REFERENCES

Allen, T.E., Herrgard, M.J., Liu, M., Qiu, Y., Glasner, J.D., Blattner, F.R., and Palsson, B.O. (2003). Genome-scale analysis of the uses of the *Escherichia coli* genome: model-driven analysis of heterogeneous data sets. J Bacteriol *185*, 6392-6399.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., and S, A. (2001). Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr Biol *11*, 941-950.

Bockhorst, J., Qiu, Y., Glasner, J., Liu, M., Blattner, F., and Craven, M. (2003). Predicting bacterial transcription units using sequence and expression data. Bioinformatics *19*, 34-43.

Carter, R.J., Dubchak, I., and Holbrook, S.R. (2001). A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res *29*, 3928-3939.

Chen, S., Lesnik, E.A., Hall, T.A., Sampath, R., Griffey, R.H., Ecker, D.J., and Blyn, L.B. (2002). A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. Biosystems *65*, 157-177.

Herring, C.D., Raffaelle, M., Allen, T.E., Kanin, E.I., Landick, R., Ansari, A.Z., and Palsson, B.O. (2005). Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. J Bacteriol *187*, 6166-6174.

Jeong, K.S., Ahn, J.W., and Khodursky, A.B. (2004). Spatial patterns of transcriptional activity in the chromosome of *Escherichia coli*. Genome Biol *5*, R86.

Kanhere, A., and Bansal, M. (2005). A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinformatics *6*, 1.

Kawano, M., Reynolds, A.A., Miranda-Rios, J., and Storz, G. (2005). Detection of 5'- and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. Nucleic Acids Res *33*, 1040-1050.

Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., and Karp, P.D. (2005). EcoCyc: a comprehensive database resource for *Escherichia coli*. Nucleic Acids Res *33*, D334-337.

Le Novere, N. (2001). MELTING, computing the melting temperature of nucleic acid duplexes. Bioinformatics *17*, 1226-1227.

Rivas, E., Klein, R.J., Jones, T.A., and Eddy, S.R. (2001). Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr Biol *11*, 1369-1373.

Saetrom, P., Sneve, R., Kristiansen, K.I., Snove, O.J., Grunfeld, T., Rognes, T., and Seeberg, E. (2005). Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming. Nucleic Acids Res *33*, 3263-3270.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J*., et al.* (2006). RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res *34*, D394-397.

Selinger, D.W., Saxena, R.M., Cheung, K.J., Church, G.M., and Rosenow, C. (2003). Global RNA half-life analysis in Escherichia coli reveals positional patterns of transcript degradation. Genome Res *13*, 216-223.

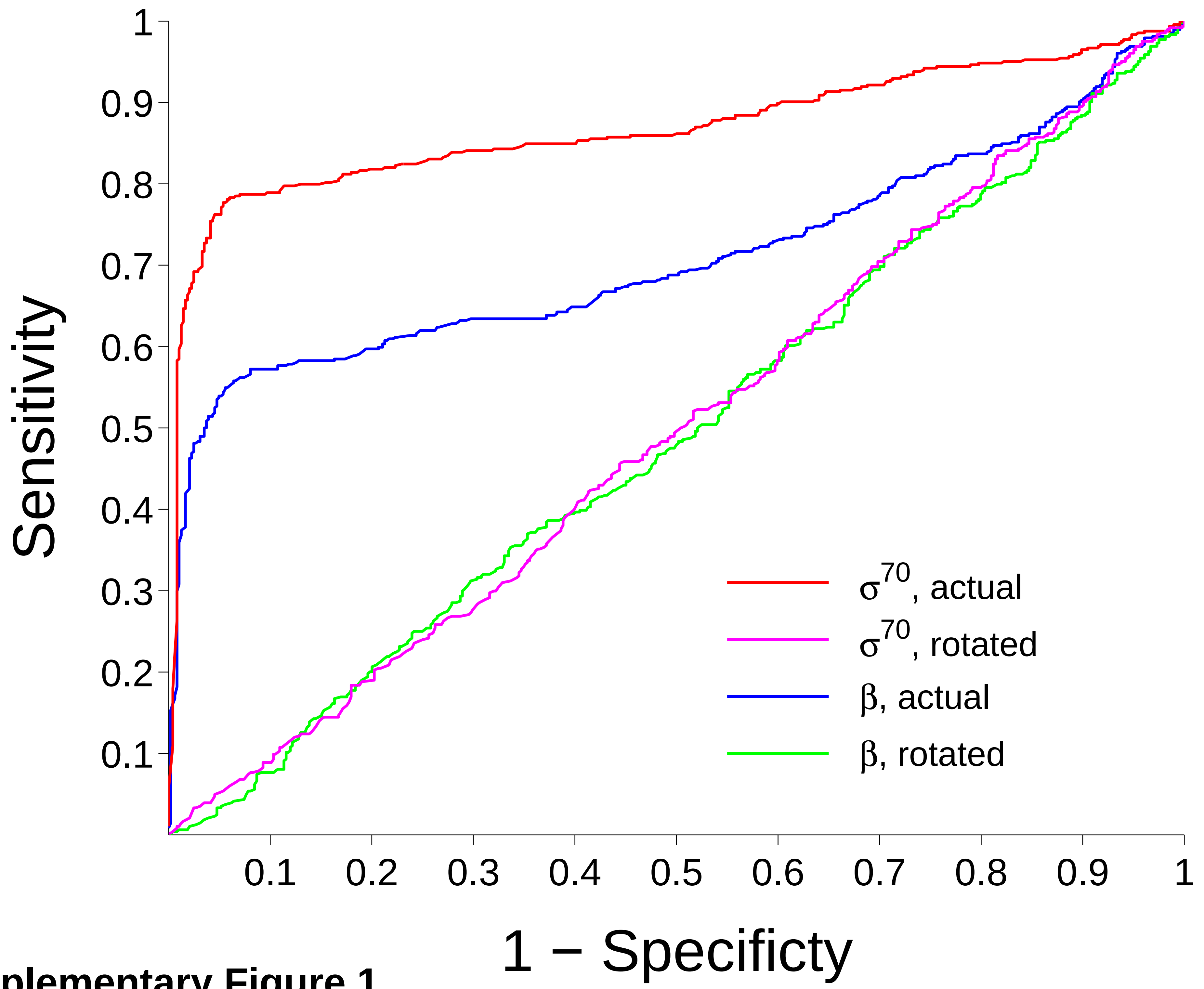
Tjaden, B., Saxena, R.M., Stolyar, S., Haynor, D.R., Kolker, E., and Rosenow, C. (2002). Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. Nucleic Acids Res *30*, 3732-3738.

Vogel, J., Bartels, V., Tang, T.H., Churakov, G., Slagter-Jager, J.G., Huttenhofer, A., and Wagner, E.G. (2003). RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. Nucleic Acids Res *31*, 6435-6443.

Wade, J.T., and Struhl, K. (2004). Association of RNA polymerase with transcribed regions in Escherichia coli. Proc Natl Acad Sci USA *101*, 17777-17782.

Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev *15*, 1637-1651.

Zhang, A., Wassarman, K.M., Rosenow, C., Tjaden, B.C., Storz, G., and Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. Mol Microbiol *50*, 1111-1124.



Supplementary Figure 1

Figure S1. Receiver Operating Curve

Sensitivity vs. 1-Specificity (or Receiver Operating Curve) of ChIP peak coordinate calls. We first computed the absolute distance between the location of each σ^{70}/β ChIP peak and its closest positive control coordinate and, separately, its closest negative control coordinate. For all absolute distances between 0 and 10000 bp, we then determined for each how many positive and how many negative controls were called in total. Thus the tradeoff between specificity and sensitivity was evaluated for σ^{70} (red) and β (blue) ChIP peak calls. We recomputed the corresponding plots for ChIP peak coordinates rotated by 1 Mb around the chromosome; as expected, both σ^{70} (magenta) and the β (lime) rotated ChIP data yield no discrimination between positive and negative control coordinates.

Supplementary Figure 2

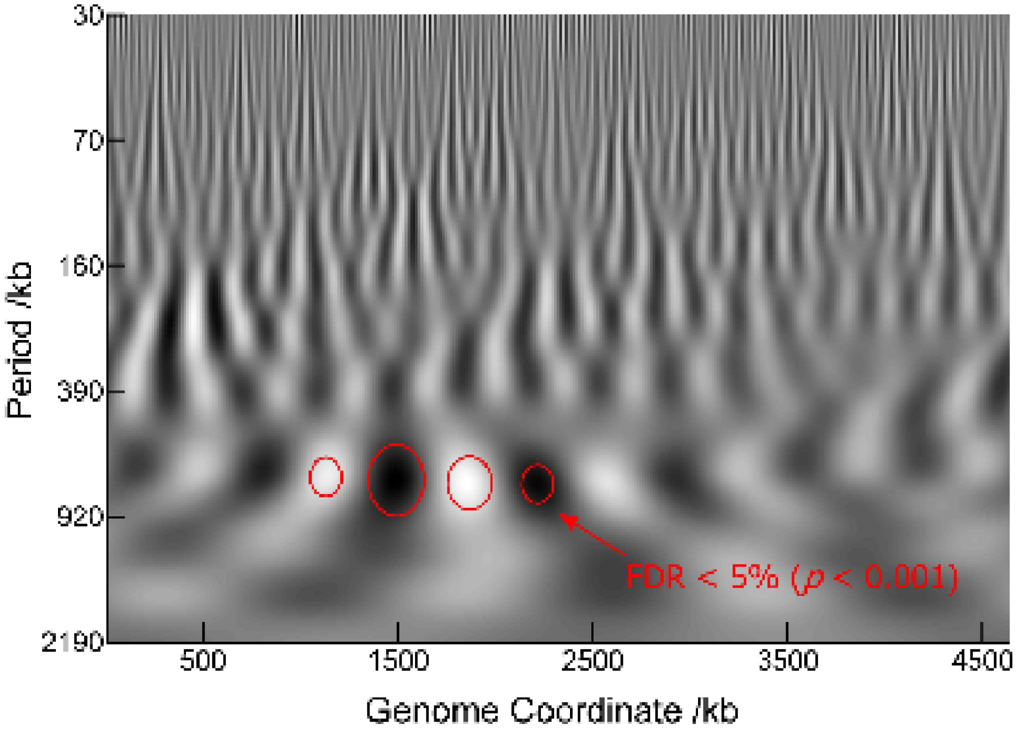


Figure S2. Periodicity of σ^{70} binding across the genome.

Wavelet analysis of discretized σ^{70} ChIP peak data. Only ChIP peaks in regions with minimal cross-hybridization potential were used in this analysis (see Methods). Discretized regions with no peak were assigned a ChIP peak height of zero. A Morlet wavelet with wave number 6 was employed to compute the wavelet plot shown. For each wavelet period vs. coordinate pixel, we calculated the number of times the wavelet signal was higher in the actual dataset than that for 1000 randomly permuted ChIP peak datasets. In black are those pixels where the actual signal was lower than random and in white where the actual signal was higher than random for all 1000 randomizations; grey shades represent intermediate significance. Significance plots were recomputed using 20 randomly permuted datasets as the actual data; this demonstrated that the ~700 kb^{-1} period centered at approximately 1.6 Mb at a wavelet significance level of $p < 0.001$ had an associated false discovery rate of < 5%.

Supplementary Figure 3

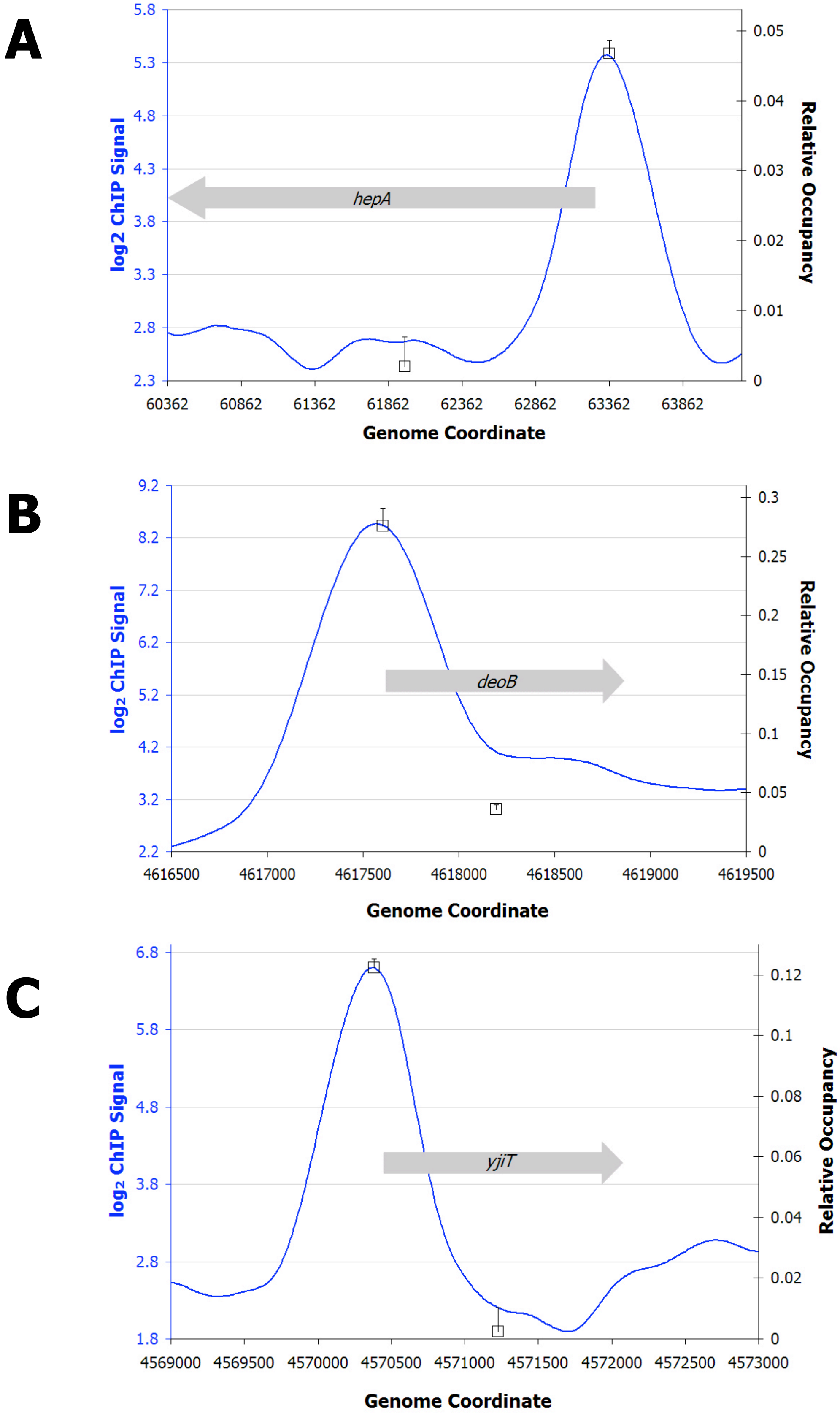


Figure S3. Representative smoothed ChIP-chip profile of β and ChIP/quantitative PCR validation.

(A-C) Averaged, smoothed $\log_2$ ChIP-chip profile (blue line) of β at the (A) *hepA*, (B) *deoB*, and (C) *yjiT* genes plotted against genome position. Genes are indicated by gray arrows. Association of β was also determined using ChIP and quantitative PCR at the promoter and distal coding sequences of each gene. These association values are shown as empty squares positioned at the center of the corresponding PCR product. ChIP association values are normalized to binding in the *bglB* coding sequence, background subtracted, and plotted relative to binding to a region within the rDNA locus. Error bars represent one standard deviation from the mean.

Supplementary Figure 4

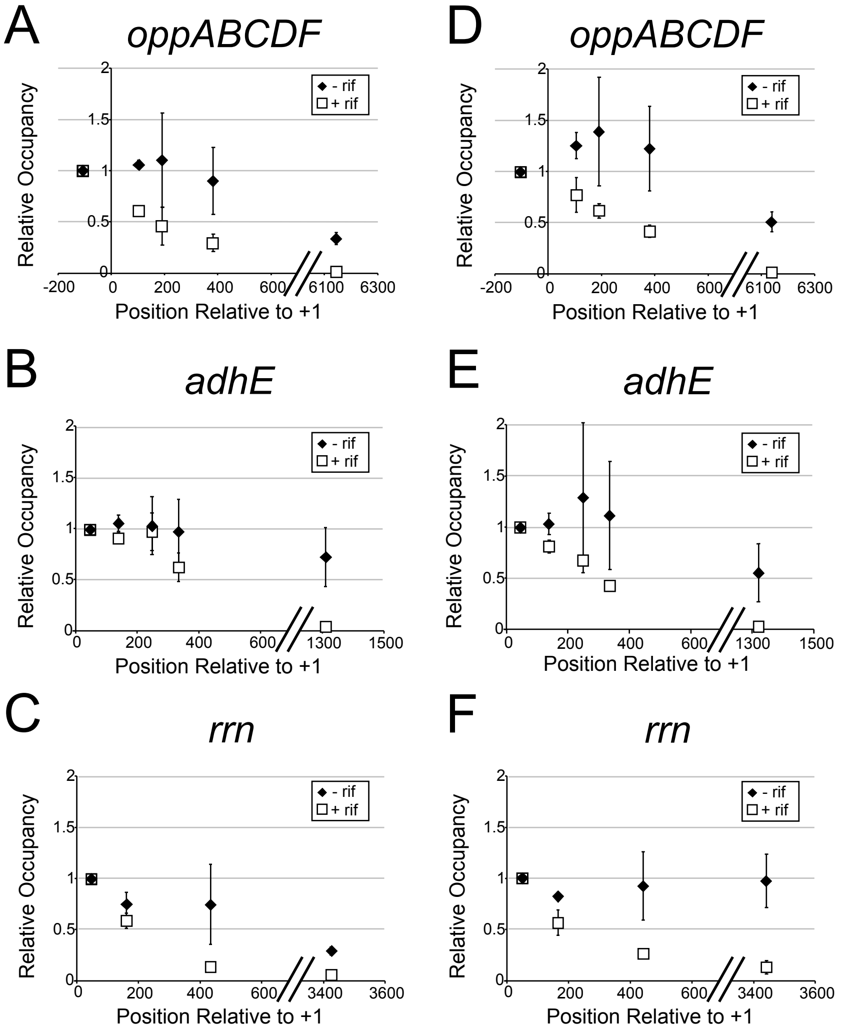


Figure S4. High-resolution mapping of β at three transcribed regions.

(A+D) Relative occupancy values for β at indicated positions throughout the *oppABCDF* operon before (black diamonds) and following (white squares) rifampicin treatment. Positions are indicated relative to the transcription start point. Values are normalized to *bglB* coding sequence and plotted relative to the value at the most upstream position.

(B+E) Relative occupancy values for β at indicated positions throughout the *adhE* gene before (black diamonds) and following (white squares) rifampicin treatment. Positions are indicated relative to the transcription start point. Values are normalized to *bglB* coding sequence and plotted relative to the value at the most upstream position.

(C+F) Relative occupancy values for β at indicated positions throughout the rDNA locus before (black diamonds) and following (white squares) rifampicin treatment. Positions are indicated relative to the transcription start point. Values are normalized to *bglB* coding sequence and plotted relative to the value at the most upstream position. (A-C) show data collected from cells grown in M9 minimal media at 30 °C. (D-F) show data collected from cells grown in LB media at 37 °C.

Error bars represent one standard deviation from the mean.

Supplementary Figure 5

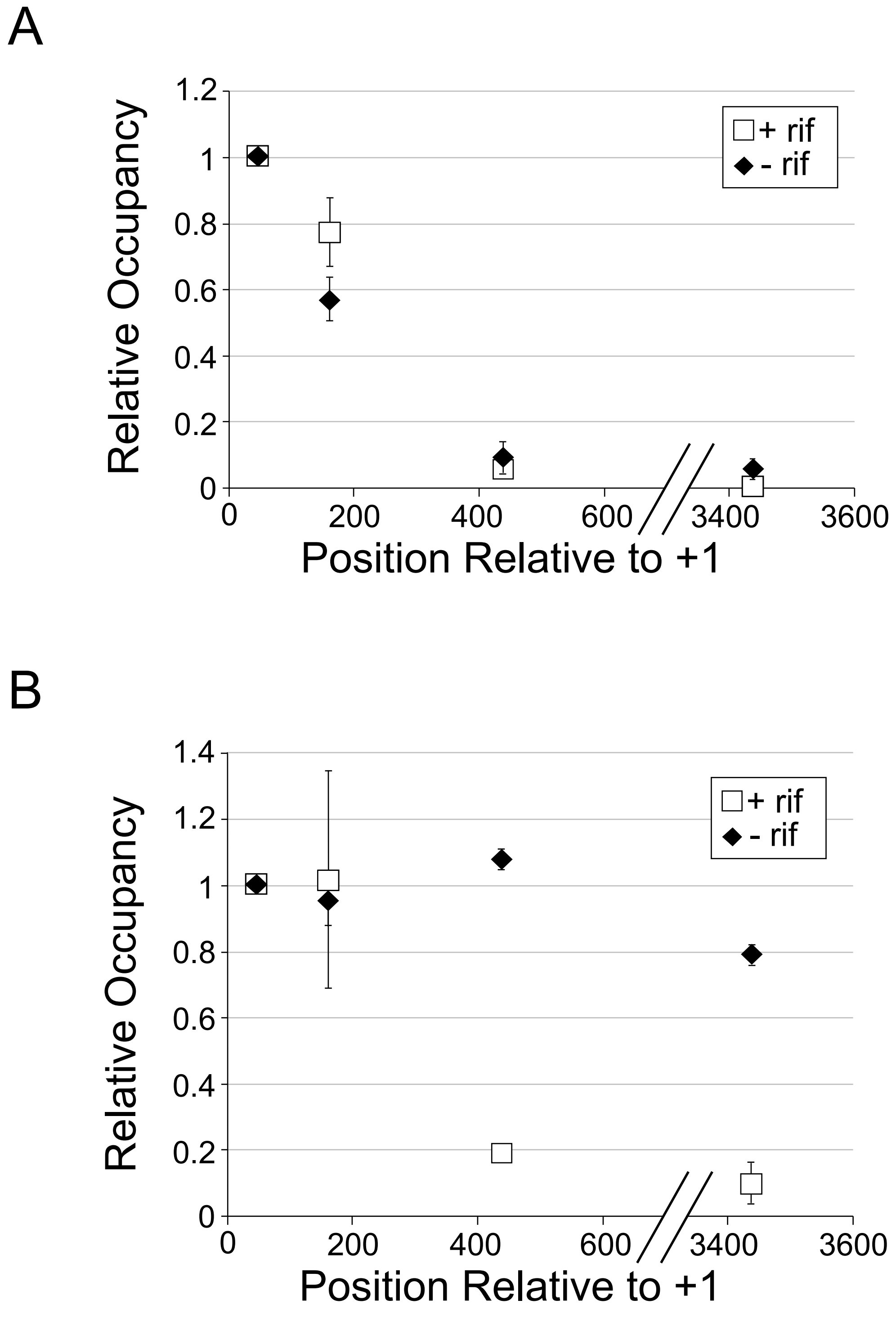


Figure S5. High-resolution mapping of σ^{70} and β at the rDNA locus.

(A) Relative occupancy values for σ^{70} at indicated positions throughout the rDNA locus before (black diamonds) and following (white squares) rifampicin treatment. Positions are indicated relative to the transcription start point. Values are normalized to *bglB* coding sequence and plotted relative to the value at the most upstream position.

(B) Relative occupancy values for β at indicated positions throughout the rDNA locus before (black diamonds) and following (white squares) rifampicin treatment. Positions are indicated relative to the transcription start point. Values are normalized to *bglB* coding sequence and plotted relative to the value at the most upstream position.

Error bars represent one standard deviation from the mean.