Dynamic landscape of protein occupancy across the *Escherichia coli* chromosome

Peter L. Freddolino[1,2,*], Thomas J. Goss[1], Haley M. Amemiya[3], and Saeed Tavazoie[4,5,6,*]

1: Department of Biological Chemistry, University of Michigan Medical School, Ann Arbor, MI, USA
2: Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
3: Cellular and Molecular Biology Program, University of Michigan Medical School, Ann Arbor, MI, USA
4: Department of Biological Sciences, Columbia University, New York, NY, USA
5: Department of Systems Biology, Columbia University, New York, NY, USA
6: Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA

*: To whom correspondence should be addressed. Email: petefred@umich.edu, st2744@columbia.edu

Abstract

Free living bacteria adapt to environmental change by reprogramming gene expression through precise interactions of hundreds of DNA-binding proteins. A predictive understanding of bacterial physiology requires us to globally monitor all such protein-DNA interactions across a range of environmental and genetic perturbations. Here, we show that such global observations are possible using a modification of *in vivo* protein occupancy display technology (IPOD-HR) applied to *E. coli.* We observe that the *E. coli* protein-DNA interactome organizes into two distinct prototypic features: (**1**) highly dynamic condition-dependent transcription factor occupancy by dedicated transcriptional regulators, and (**2**) condition-invariant kilobase scale occupancy by nucleoid factors, forming silencing domains analogous to eukaryotic heterochromatin. We show that occupancy dynamics across a range of conditions can rapidly reveal the global transcriptional regulatory organization of a bacterium. Beyond discovery of previously hidden regulatory logic, we show that these observations can be utilized to computationally determine sequence-specificity models for the majority of active transcription factors. Our study demonstrates that global observations of protein occupancy combined with statistical inference can rapidly and systematically reveal the transcriptional regulatory and structural features of a bacterial genome. This capacity is particularly crucial for non-model bacteria which are not amenable to routine genetic manipulation.

Introduction

Transcriptional regulation plays a central role in establishing adaptive gene expression states. In bacteria, the dominant regulators are transcription factors (TFs) [1,2] and sigma factors, which direct the activity of RNA polymerase holoenzyme to a specific subset of promoters [3,4]. The phenotypic state of the bacterial cell is determined in large part by its transcriptional regulatory state which, in turn, is dictated by the binding pattern of TFs and sigma factors across the chromosome, likely in interplay with structural factors such as the local supercoiling state [5].

At present, however, our knowledge of the complete wiring of bacterial transcriptional regulatory networks remains insufficient to fully predict or design regulatory responses to arbitrary environmental conditions. *Escherichia coli* provides an illustrative case study: due to its status as a pre-eminent model organism and human pathogen, the *E. coli* transcriptional regulatory network has been an intense subject of investigation for several decades. As a result, researchers have obtained an increasingly comprehensive and detailed map of the binding specificities and physiological roles of transcriptional regulators in this organism [6]. However, roughly one quarter of the ~250 transcription factors in *E. coli* have no available binding or regulatory data [7], and many more are virtual unknowns in terms of the signals that might alter their regulatory activity. Likely as a result of this knowledge gap, Larsen and colleagues recently found that despite our broad knowledge of the potential regulatory targets of *E. coli* transcription factors, our ability to predict regulatory behavior on the basis of expression levels of transcription factors is no better than it would be for random networks. The authors attribute this partly to the fact that even when a TF is expressed, in many cases it will not bind its targets in the absence of additional signals [8]. Furthermore, *E. coli* represents a best-case scenario in terms of our knowledge state for a bacterial transcriptional regulatory network, and for most species current databases lag far behind.

Expanding our capability to predict, and ultimately design, bacterial regulatory responses will be critical for controlling bacterial pathogenesis and engineering synthetic microbes in biotechnology applications. Achieving such a complete predictive understanding, however, requires substantial additional information both on the binding sites of as-yet uncharacterized TFs, and the actual physical occupancy of sites for both known and uncharacterized factors across conditions. Widely used methods such as ChIP-seq pose difficulties on both fronts: they demand a combinatorial explosion of experiments to study many transcription factors across a variety of conditions, and require either an antibody against each TF of interest or genetic manipulation sufficient to add an epitope tag to each target TF.

In order to significantly advance our understanding of transcriptional network dynamics and chromosomal structure, we sought to monitor, in parallel, the occupancy states of all DNA-binding proteins across a set of genetic and environmental perturbations. We argue that such comprehensive observations are critical for defining the global modes of transcriptional regulation and determining the regulatory logic that underlies adaptive reprogramming of gene expression, particularly given the importance of combinatorial logic by many transcription factors in dictating transcriptional output [9]. In order to achieve our goal, we decided to employ the

concept of *in vivo* protein occupancy display (IPOD) which we, in a previous proof-of-concept study, demonstrated to reveal global occupancy of protein binding sites across the *E. coli* chromosome [10]. However, we had to introduce critical modifications and enhancements in order to deconvolve distinct contributions from sequence-specific TFs and RNA-polymerase and define binding sites at high-resolution. We will refer to this second generation IPOD technology as IPOD-HR. IPOD-HR enables efficient coverage of a large range of physiological conditions in relatively few experiments (one experiment per condition, rather than the one experiment per TF per condition that would be required for ChIP-seq). As we demonstrate below, a single IPOD-HR experiment can reveal the occupancy dynamics of dozens of known and novel active TFs genome-wide, permitting rapid profiling of global transcriptional regulatory logic across different conditions. Furthermore, the comprehensive nature of IPOD-HR profiles enables efficient statistical inference of sequence specificity models (TFBS motifs) for active transcription factors, both recapitulating well-known regulatory logic and revealing the presence and condition-dependent activities of novel regulatory elements.

Here, we characterized the dynamics of the global protein-DNA interactome of *E. coli* across a range of three physiological conditions and three genetic perturbations. Our observations allowed us to infer, in parallel, the activities of most annotated transcription factors across conditions, and provided a catalogue of many additional likely regulatory sites and DNA sequence motifs for uncharacterized TFs. With the compact set of experiments, we reveal the dramatic regulatory dynamics of dozens of transcription factors that collectively shape the response of *E. coli* to changing environments. In sharp contrast, we find that at the kilobase scale, the genome is characterized by a set of relatively static structural domains, which consist of transcriptionally silent loci with dense protein occupancy that appear mostly constitutive across a range of physiological conditions. These regions, which we refer to as EPODs (extended protein occupancy domains) following the nomenclature of Vora *et al. [10]*, appear to act, at least partially, to suppress prophages and mobile genetic elements.

Because our approach does not rely on prior knowledge of TFs of interest or genetic manipulation of the target organism, but rather only on essential physico-chemical properties of protein-DNA complexes, we expect that it will be broadly applicable across bacterial species, even those which cannot be cultured or genetically manipulated. Our approach lays the technical and analytic foundation to rapidly characterize the regulatory and structural features of any bacterial chromosomes.

Results

Global high-resolution profiling of condition-dependent transcription factor occupancy across the *E. coli* chromosome

The IPOD-HR procedure is shown in schematic form in **Figure 1A**: cells are grown under a physiological condition of interest, fixed using formaldehyde, and then lysed. Heavy digestion of the chromosomal DNA provides minimized DNA fragments that may be in either a protein

bound or unbound state. The protein bound DNA fragments are subsequently isolated using a phenol-chloroform extraction. Under appropriate buffer conditions, the amphipathic protein-DNA complexes are depleted from the aqueous phase and partition to a robust disc at the aqueous-organic interface [10].

As we will demonstrate below, the measurements enabled by IPOD-HR can subsequently be used for a broad range of downstream analyses, such as simultaneous monitoring of the activities of characterized TFs, large-scale inference of binding motifs for previously uncharacterized DNA-binding proteins, and identification of key occupancy sites driving previously unrecognized gene regulatory logic. To accomplish these objectives, it is essential to separate out the occupancy signal of RNA polymerase from that of specific regulatory factors of interest. Otherwise, the strong occupancy signal caused by RNA polymerase could mask changes in protein occupancy that in fact provide regulatory information (e.g., if a repressor becomes unbound but RNA polymerase subsequently binds to the same location, the occupancy signal would be nearly unchanged). To deconvolve occupancy caused by sequence-specific TFs and that of RNA polymerase, we subtract the normalized RNA polymerase ChIP-seq signal from that of the normalized raw IPOD-HR signal (see Methods for details), generating a corrected IPOD-HR profile that is a more precise representation of the cell's dynamic regulatory state (**Fig. 1B**). IPOD-HR is conceptually similar to the original IPOD method [10] in terms of overall workflow, but contains critical optimizations and extensions designed to permit genome-wide identification of binding by TFs and organizing factors such as nucleoid-associated proteins in a condition-specific manner; these optimizations arise both in the sample preparation procedure itself (through enhanced washing and additional measures to track and separately factor out RNA polymerase occupancy), and an expanded analytical framework for use in inferring both existing and new regulatory logic based on whole-genome protein occupancy profiles.

We note in passing that, at first glance, IPOD may seem to share superficial similarities with FAIRE (formaldehyde-assisted isolation of regulatory elements, originally described in [11]). However, FAIRE experiments were designed to detect regions of nucleosome-depleted DNA in eukaryotic chromosomes. IPOD was independently developed to detect occupancy of individual factors in prokaryotic chromosomes [10], and IPOD-HR contains further optimizations and additional experimental and computational steps to improve performance in detecting both localized and large-scale protein occupancy in bacteria.

An illustrative example of the ability of IPOD-HR to identify regulatory protein occupancy, its dynamics across conditions, and the importance of factoring out the RNA polymerase signal, is shown in **Figure 1C**. We consider the IPOD-HR occupancy profiles for the promoter region upstream of the *purC* gene in wild type (WT) and Δ*purR* cells during growth in rich defined medium. Based on the characterized behavior of PurR (which binds DNA in response to exogenous purine supplementation [12,13]), under this growth condition, transcription of *purC* should be repressed by binding of PurR to its promoter. However, if one considers only the raw IPOD-HR occupancy profiles (top panel), binding to the PurR site in this region is apparent in both WT and Δ*purR* cells. The resolution to this seeming paradox becomes apparent through

inclusion of the correction for RNA polymerase occupancy (middle panel), which is substantially higher in Δ*purR* cells. As expected, the resulting corrected IPOD-HR occupancy profiles (bottom panel) reveal a protein occupancy peak directly on top of the annotated PurR binding site in this region in the WT cells, and no detectable occupancy in the Δ*purR* cells. This demonstrates the ability of IPOD-HR to reveal condition-dependent TF occupancy dynamics even in regions that may overlap with RNA polymerase binding. In the following sections, IPOD-HR refers to the RNA polymerase-corrected occupancy signal, unless otherwise noted.

Local and large-scale protein occupancy patterns across the *E. coli* chromosome

To benchmark our ability to quantitatively profile protein occupancy at high spatial resolution, we performed IPOD-HR on *E. coli* cells from mid-exponential growth in rich defined medium (**Fig. 2A**). Over the length of the chromosome we observed a large number of small peaks, presumably corresponding to protein binding events at individual regulatory sites. In addition, we observed many large-scale (> 1 Kb) regions of high occupancy which we refer to as extended protein occupancy domains (EPODs), following the nomenclature of (Vora et al., 2009). An example of condition-dependent changes in binding of local TFs is shown in **Fig. 2B-C**. Examination of a ~50 kb slice of the genome reveals dozens of small occupancy peaks, with a visually apparent enrichment in intergenic regions (**Fig. 2B**). Many such peaks, which presumably correspond to individual protein binding events, coincide with known TF binding sites (TFBSs). For example, the region upstream of *argA* (**Fig. 2C**) shows strong occupancy at known ArgR binding sites, and condition-appropriate occupancy dynamics including weakening of binding in arginine-poor conditions [14,15] and loss of occupancy upon deletion of the *argR* gene. At the same time, similar occupancy patterns can be observed at many sites lacking an annotated TFBS, as seen in **Fig. 2D**, where conditionally dynamic binding sites are apparent upstream of *lgt* and *rppH*. These peaks likely indicate the presence of previously unrecognized TFBSs, as we will discuss in more detail below. As expected, at a genome-wide scale IPOD-HR signals show both higher occupancy in intergenic regions relative to coding regions, and higher occupancy at annotated TFBSs relative to other regions of the chromosome (**Fig. 2E**), demonstrating a strong overlap of the observed protein occupancy with transcriptional regulatory sites. Indeed, applying peak calling to the IPOD-HR signal demonstrates an increasingly strong overlap with known TFBSs as the threshold for peak calling is increased (**Fig. S1**).

It is also apparent by inspection of the genome-wide occupancy shown in **Fig. 2A** that many extended regions of high protein occupancy coincide with regions of relatively low transcription. For example, in **Figure 2F**, we show a typical ~300 kb region with alternating segments of high protein occupancy that have relatively low transcription, with those of low protein occupancy and relatively high transcription (also apparent in the higher-resolution plot in **Fig. 2G**). Thus, in addition to revealing occupancy at the level of individual regulatory sites, IPOD-HR enables tracking of the behavior of large, densely protein occupied regions of the chromosome that appear to coincide with transcriptionally silent loci. We will explore both of these prototypic classes of occupancy, in more detail, below.

Transcription factor and sigma factor occupancy dynamics across genetic and environmental perturbations

Since IPOD-HR occupancy profiles show highly enriched overlaps with known TFBSs (**Fig. 2E**), we asked whether IPOD-HR profiles can be used to reveal the occupancy dynamics for known *E. coli* TFs across a set of conditions. Indeed, we find that IPOD-HR reveals consistent and condition-appropriate regulatory logic at the level of individual regulons, and patterns of regulatory behavior across regulons. As expected, strains with each of three single TF deletions (*argR*, *lexA*, *purR*) show global loss of occupancy at the ensemble of annotated sites for the corresponding TFs (**Fig. 3A**). Analysis of condition-dependent changes in the occupancy of binding sites for single transcription factors likewise recapitulates expected behavior; for example, ArgR [16], PurR [12,17], and TyrR [18] all show enhanced binding to DNA in the presence of amino acid and/or nucleobase ligands which are supplied directly in our rich media conditions, and the IPOD-HR occupancy signal shows global loss of occupancy for binding sites of all three of these TFs in nutrient depleted conditions (minimal media and stationary phase) when compared with exponential growth in rich media (**Fig. 3A**). In contrast, RutR shows increased overall occupancy in minimal media relative to rich media conditions, consistent with the known inhibition of RutR binding by thymine and uracil [19]. Sites for MetR, which is dependent upon homocysteine as a co-regulator [20], likewise show large increases in occupancy in minimal media and in stationary phase (**Fig. 3A**). As homocysteine is the final intermediate in methionine biosynthesis [21], its levels would naturally be expected to rise upon methionine starvation, and consistently, transcript levels of the canonical homocysteine-dependent MetR-activated target *metE [22]* rise six-fold in our stationary phase RNA-seq data and more than 100-fold in our minimal media RNA-seq data (data not shown).

By applying an unsupervised clustering approach (see Methods for details), we identified transcriptional regulatory modules that show consistent co-regulation across the conditions in our study. We found clustering of TFs with highly similar behavior (**Fig. 3B**) that coordinate, for example, amino acid metabolism (orange), the core translational apparatus (pink), and iron homeostasis (red). We also observe several cases where regulatory cascades are clustered together; for example, YjjQ and its transcriptional activator LeuO (blue), or the tightly intertwined acid response regulators GadE, GadW, and GadX (yellow). We thus find that IPOD-HR occupancy profiles can provide detailed, site-level, condition-specific information on regulatory protein occupancy across the entire chromosome. By comparing changes in protein occupancy with changes in transcript levels across conditions, we can relate changes in protein occupancy to their positive or negative regulatory consequences. This can be seen for two nutrient sensing transcriptional repressors with sites annotated in RegulonDB, ArgR and PurR, across changes in nutrient conditions (**Fig. 3C**). As expected, given the annotated repressive role of these sites, in both cases we observe a strong anti-correlation between changes in protein occupancy and target transcription.

Since each IPOD-HR global protein occupancy data set is performed alongside an RNA polymerase ChIP-seq experiment, we can easily track promoter occupancy alongside TFBS occupancy. The use of rifampicin permits transcriptional initiation, but prevents promoter clearance. Thus, these data sets are ideal for identifying regulation at the level of RNA

polymerase (e.g. via different sigma factors). The differential patterns of RNA polymerase occupancy show strong correlations with transcript levels for each Sigma factor's regulon across a range of conditions. As shown in **Figure 3D**, when comparing logarithmic vs. stationary phase conditions, the changes in transcript abundance and RNA polymerase promoter occupancy show a Spearman correlation of 0.83 (p=0.042); a similar comparison for changes in occupancy vs. expression for cells grown in minimal media yields equivalent results (data not shown).

Global occupancy dynamics reveals the action of new DNA binding proteins

Despite extensive annotation efforts, at present fewer than 1,100 of the 3,560 annotated transcriptional units present in the RegulonDB database have any annotated regulation by transcription factors assigned to them [7]. While several recent notable efforts have sought to expand the completeness of these regulatory annotations by studying the DNA binding preferences of purified TFs [6,23,24], or via computational inference of likely additional regulation [25] and regulatory modules [26], none of these methods provides either direct evidence for binding *in vivo*, or information on condition-dependent changes in occupancy. IPOD-HR, in contrast, can provide both. Furthermore, the protein occupancy signals thus obtained provide information on occupancy of both well-characterized and uncharacterized proteins. In fact, a large fraction of dynamic IPOD-HR peaks occur in promoters with no previous annotation for TF binding sites, as we will discuss in detail in the following section.

A representative example of an orphan occupancy peak is seen upstream of the gene *sdaC* (**Fig. 4**). In our RNA-seq data, *sdaC* transcript levels are nearly twenty-fold higher during exponential growth in rich media (317.3 transcripts per million (TPM)) compared with either exponential growth in minimal media (17.9 TPM) or stationary phase in rich media (16.7 TPM). Despite a lack of annotated TFBSs upstream of *sdaC*, IPOD-HR occupancy profiles (**Fig. 4A**) show a likely transcriptional activator binding site upstream of the *sdaC* core promoter, which shows strong occupancy in the WT M9/RDM/glu conditions but not the related conditions where *sdaC* expression is lower. To identify the transcription factor(s) responsible for that occupancy, we used a biotinylated bait DNA matching the sequence of the *sdaC* promoter region to isolate proteins bound to that region from *E. coli* cells grown in the WT M9/RDM/glu condition (**Fig. 4B**). Mass spectrometry on isolated bait-dependent bands revealed two poorly characterized transcription factors, UlaR and YieP, that showed highly enriched binding to the *sdaC* promoter (see **Supplementary Table S1**). While UlaR proved difficult to purify due to poor solubility, and was thus excluded from further analysis, we found that purified YieP does indeed show specific shifting of the *sdaC* promoter in an electrophoretic mobility shift assay (**Fig. 4C**). Consistently, recent RNA-seq data on a Δ*yieP* strain shows a significant drop in *sdaC* transcript levels (2.7-fold change; $q=7.6 * 10^{-18}$) relative to isogenic cells with a plasmid-born reintroduction of YieP during growth in LB media (C. Bianco and C. Vanderpool, personal communication).

YieP was recently (and independently) selected by Palsson and co-workers as a validation case to be used in their consideration of computational methods for identifying the binding sites of orphan TFs, and subjected to ChIP-exo analysis on cells grown in glucose minimal media using epitope-tagged YieP [25]. Indeed, their data demonstrate both strong direct YieP occupancy,

and a high confidence YieP motif match, at the precise position of the occupancy peak detected in our IPOD-HR data set (**Fig. 4D**). Based on the relative intensity at that position across conditions, combined with the expression data noted above, we infer that YieP binds to the *sdaC* promoter in nutrient-replete conditions and acts as a transcriptional activator (explaining the solitary strong peak in our "WT,rich" condition), whereas in other conditions, YieP binding is weakened (but not abolished) and additional factors likely bind downstream of the YieP site to repress *sdaC* transcription. We must emphasize that the discovery of YieP binding sites through IPOD-HR and subsequent mass spectrometry experiments (by us) occurred in parallel with the ChIP-exo experiments of Gao and colleagues, and indeed, represent highly complementary paths for identification of the binding sites for orphan transcription factors, with one centered on a candidate protein and the other on candidate sites.

The example presented here of regulation of *sdaC* by the uncharacterized transcription factor YieP highlights the broad potential for using IPOD-HR to rapidly identify and characterize previously cryptic regulatory connections. IPOD-HR thus complements the multitude of other approaches noted above (based on, e.g., promoter libraries or computational inference), and provides the unique benefit of directly assessing binding to DNA *in vivo*, at native loci, under physiological conditions of interest.

The utility of IPOD-HR in identifying the activity of previously un-characterized transcription factors motivates its extension to a genome-wide scale, providing an *in vivo* complement to high-throughput *in vitro* screening methods such as genomic SELEX [6]. By applying peak calling to our IPOD-HR data sets across the six conditions considered in the present study, we were able to identify thousands of likely TFBSs, many of which are not identifiable based on existing databases. A comprehensive listing of peak calls across conditions and thresholds is given in **Supplementary Data S1.** To compare the peak sets identified from IPOD-HR data with our existing state of knowledge, we divided the peak calls obtained from IPOD-HR into a set of annotated TFBSs from RegulonDB, and a set of binding sites predicted using all known PWMs available in the SwissRegulon database (see Methods for details). We find that across a range of thresholds, approximately half of the binding sites identified by IPOD-HR overlap with either known or predicted sites, whereas the other half represent novel binding sites which likely (as in the case of the YieP site described above) reflect the activity of poorly annotated or orphan TFs. Pooling the newly identified binding sites across conditions, our IPOD-HR data sets are able to provide a total of 19,068 putative TFBSs which are occupied *in vivo* under at least one condition (and track the dynamics of occupancy of those sites across conditions). This extensive map of chromosomal occupancy and its dynamics provides the community with a wealth of known and putative novel regulatory interactions that can be further explored and validated by follow-up experiments such as those shown in **Fig. 4**.

Global *de novo* discovery of sequence-specificity motifs for active transcription factors

While the peak calls obtained from IPOD-HR data show strong enrichments with known TFBSs (**Fig. S1**), roughly half of the called peaks do not match any known or predicted transcription

factor binding sites (as detailed in **Figure 5A**), and likely correspond either to unknown sites for well-characterized TFs or binding sites for previously uncharacterized TFs. Given that the majority of the newly inferred binding sites appear not to correspond to known or predicted sites for annotated TFs, we hypothesized that the regulons corresponding to those motifs would likely show enrichments for poorly annotated genes, as we expect here to reveal the regulatory logic driving typically under-studied pathways. We thus identified likely regulatory targets of each newly called peak, divided them between poorly annotated genes (those with UniProt annotation scores of 1 or 2 out of 5 [27]) and well-annotated genes, and then examined the proportion of poorly annotated targets for occupancy peaks matching RegulonDB binding sites compared with all other peaks. As shown in **Figure 5B**, peaks that do not correspond to RegulonDB-annotated binding sites are strongly enriched upstream of poorly annotated genes, whereas those matching annotated binding sites are enriched for well-annotated genes. Thus, examination of occupancy peaks derived from IPOD-HR enables identification of a large number of new putative regulatory sites, with a particular abundance of possible regulators of poorly-annotated genes.

Our large-scale identification of new TFBSs also raises the important possibility that new TF binding motifs might likewise be identifiable through *de novo* computational motif discovery in the set of all sequences within IPOD-HR peaks. Indeed, the application of the FIRE [28] motif discovery algorithm to peak locations obtained from IPOD-HR data reveals dozens of *de novo* discovered sequence motifs that are informative of strong occupancy sites, even after pruning of redundant motifs (**Fig. 5C**). Upon cross-referencing with a database of known *E. coli* TFBS motifs using TOMTOM [29], we find that approximately 25% of the discovered motifs can be matched with known motifs (87/97 of the annotated motifs in the *E. coli* SwissRegulon database are matched by at least one inferred motif from the set present prior to redundancy pruning, and 63/97 match at least one motif present in our inferred set after pruning), while at the same time nearly 200 novel motifs are called with similar confidence (**Fig. 5D**). A comprehensive list of redundancy-pruned motifs called across all conditions in our study is given in **Supplementary Data S2**. To provide estimates of the false discovery rate (FDR) arising from our motif inference, we performed an identical motif discovery procedure for each biological condition on 20 “decoy” data sets in which the underlying *E. coli* genomic sequence was rotated by a random distance relative to the peak calls, thus preserving the correlation structure of both the data and sequence with respect to themselves (light bars in **Figure 5C**). Our decoy data sets gave rise to no more than 11 motifs under any condition, and usually far fewer, giving rise to an average effective FDR (across shuffles and conditions) of 0.2% for the unpruned motifs or 1.0% for the pruned motifs. Using only the novel motifs (that is, motifs which did not have detectable similarity to any motifs in the SwissRegulon database) in a genome-wide search for potential binding sites using FIMO, we find that 32.1% of all IPOD occupancy peaks at a peak calling threshold of 4 can be explained by binding sites for the novel motifs, compared with 9.3% that can be explained by annotated binding sites from RegulonDB. Thus, the newly inferred motifs provide a substantially expanded ability to assign the observed profile of protein binding across the chromosome. Of the occupancy peaks not identifiable based on either our novel motifs or RegulonDB binding sites, 43.7% fall in EPODs, and thus are attributable to locally concentrated binding of the EPOD constituent proteins (likely H-NS and other nucleoid-associated proteins,

as discussed elsewhere). Taken together, the combination of newly called motifs, known binding sites from RegulonDB, and EPODs accounts for 64.3% of all occupancy peak locations at a calling threshold of 4.

In **Figure 5E** we show two representative examples of discovered motifs that show strong matches with annotated motifs, demonstrating that the motifs for well-characterized transcriptional regulators such as IHF and NanR can be inferred directly from IPOD-HR data. For comparison, in **Figure 5F**, we show two newly inferred motifs that do not match any known motifs in the *E. coli* SwissRegulon database. Intriguingly, the pattern of binding sites across the *E. coli* chromosome for both of these novel motifs illustrates a potential regulatory function, with the first motif associating with a substantial fraction of the genes involved in iron ion acquisition, and the second apparently involved in alanine metabolism (and in particular synthesis of the cell wall constituent D-alanine).

We further assessed the regulatory capacity of all newly called sequence motifs by comparing their genome-wide distribution of binding sites with annotated genes (coding regions) and promoters. We would expect that binding sites for functional transcriptional regulators would be enriched within promoters and depleted from coding regions, as was the case for overall IPOD-HR occupancy (**Fig. 2E**). Indeed, the overlap distributions of binding sites for our newly inferred motifs are uniformly enriched for annotated promoters and depleted for ORFs (**Fig. 5G**), demonstrating that motifs inferred directly from IPOD-HR occupancy data occur primarily in likely regulatory regions. Equivalent results were obtained even after excluding all of the newly inferred motifs with identifiable similarity to SwissRegulon motifs (as assessed using TOMTOM; data not shown).

Given that the majority of the newly inferred motifs appear not to correspond to annotated TFs, and our findings above regarding the enrichment of poorly annotated genes downstream of orphan binding peaks, we hypothesized that the regulons corresponding to our newly inferred motifs would likely show enrichments for poorly annotated genes, as we expect here to reveal the regulatory logic driving typically under-studied pathways. We thus calculated the fractions of the hypothetical regulons of each newly inferred motif that consist of poorly annotated genes (defined as noted above). As shown in **Figure 5H**, we found that the regulons of the newly inferred motifs were significantly enriched for poorly annotated genes when compared with both the annotated *E. coli* transcriptional regulatory network in RegulonDB ($p<2.2*10^{-16}$, Wilcoxon signed rank test), and the overall average rate of poorly annotated genes throughout the chromosome ($p<2.2*10^{-16}$, Wilcoxon signed rank test). Taken together, we see that IPOD-HR enables inference of a large number of sequence motifs, many of which likely correspond to functional, but currently under-studied, transcriptional regulators in *E. coli*, providing a substantial resource for ongoing investigation of this transcriptional regulatory network.

Extended protein occupancy domains define distinct and largely stable transcriptionally silent regions with unique sequence features

One of the most striking findings enabled by the original application of IPOD was the discovery of extended protein occupancy domains (EPODs): large regions of the *E. coli* chromosome that show unusually dense levels of protein occupancy over kilobase or longer scales [10]. EPODs are also clearly apparent in all our IPOD-HR data sets, and appear to correspond functionally to the transcriptionally silent tsEPODs of Vora *et al. [10]*. The profile of protein occupancy and EPODs, along with the accompanying impacts on transcript levels, for a representative region of the genome is shown in **Figure 6A**. Indeed, we found that many highly protein occupied regions measured using the original IPOD method (in particular, the highly expressed extended protein occupancy domains, or heEPODs) represent RNA polymerase occupancy, whereas the EPODs now revealed by IPOD-HR consist solely of large domains of occupancy by proteins other than RNA polymerase (which typically also exclude RNA polymerase). We discuss these differences and the details of the approach used in IPOD-HR to remove contributions from RNA polymerase in **Text S1.** The specific resolution of tsEPODs afforded by the IPOD-HR method, and the coverage of multiple genetic and nutrient perturbations in the present data sets, allow us to fully investigate the nature and condition-dependent occupancy of these chromosomal structures.

The identified EPODs show remarkable stability (**Fig. 6B**), with ~180 EPODs in each condition, and similar fractions of the genome contained in EPODs in each case (a comprehensive listing of EPODs identified across our conditions is given in **Supplementary Data S3**). Furthermore, the locations of individual EPODs are likewise well maintained, even across very different physiological conditions. For example, in **Figure 6C** we show IPOD-HR occupancy across the same region as shown in **Figure 6A**, comparing exponential growth in rich vs. minimal media, and stationary phase cells. In contrast with the condition-dependent occupancy of individual TFs, at the ~kilobase scale the occupancy traces are nearly superimposable, and show that most EPODs called under the various conditions overlap. Furthermore, out of the subset of EPOD calls that are missing from the 'WT,Rich' condition but present in the others, all but one are also present among calls made in the 'WT,Rich' condition using a relaxed threshold, suggesting that the small differences in EPOD locations that do appear between EPOD calls under different conditions are in fact due to thresholding effects. We observe the same trends genome-wide: 72-85% of genomic locations (at the base pair level) that are called as EPODs under any one condition are likewise EPOD calls under any other condition (**Fig. 6D**); furthermore, at least 89% (and typically much more) of the EPODs called in one condition are contained within the relaxed threshold calls under any other condition (*n.b.* the 'relaxed' threshold used here corresponds with the original EPOD definition from [10]). It is also worth noting, in this context, that 90% of the tsEPOD-occupied locations from [10] are contained within the new "WT,Rich" relaxed threshold EPOD set, in line with the observed concordance across experimental conditions in our new data sets.

Several defining characteristics of EPODs are readily apparent upon cross-referencing with other genome-wide datasets (**Fig. 6E**): they represent regions of high AT content, which are both associated with low levels of native transcripts and decreased transcriptional propensity (that is, expression of standardized integrated reporters [30]). Consistent with our original

findings [10], EPODs also show high occupancy of H-NS, HU, and LRP; low occupancy of Fis; and are associated with high efficiency of Tn5 integration (**Fig. 6E**). While the latter might seem surprising given that highly protein occupied regions on eukaryotic chromatin tend to exclude Tn5 (as is used to great effect in ATAC-seq [31]), we note that bacterial H-NS occupancy has previously been shown to facilitate Tn5 insertion [32]. Additional characteristics of EPODs, such as reduced densities of possible Dam methylation sites (consistent with the expected blocking of Dam methylase by bound proteins, previously shown in *in vivo* methylase protection experiments [33]) and a characteristic pattern of DNA structural parameters including decreased minor groove width, are shown in **Table S2**.

The remarkable condition-invariance of the locations of EPODs outlined above, even across such dramatic changes as transition from exponential to stationary phase, suggests that EPODs predominantly represent fixed structural features of the *E. coli* chromosome, rather than highly dynamic regulatory structures. We thus examined the classes of genes (assessed using gene ontology, or GO, terms) most strongly enriched or depleted in EPODs. As illustrated in **Figure 7A**, EPODs show strong enrichments for mobile elements (GO:0006313) and prophage genes (specifically lytic pathways; GO:0019835), and are depleted for core metabolic pathways such as ribosome components (GO:0030529). Indeed, EPODs are associated with the silencing of many prophages (e.g., **Fig. 7B**) and even smaller operons of unknown function (e.g., **Fig. 7C**).

Our findings regarding EPODs, particularly the high levels of H-NS binding in EPODs, and the known role of H-NS as a xenogeneic silencer [34,35], are highly consistent with prior information regarding the silencing role of H-NS. In order to determine the extent to which H-NS silenced regions and EPODs as defined here overlap, we compared the distribution of EPODs across the genome with H-NS ChIP-seq data from [36]. Using an unsupervised clustering method to divide genomic intervals into high, medium, and low levels of H-NS occupancy. We found that 66.7% of EPODs fall into the high H-NS category, compared with 3.9% of non-EPOD regions (**Fig. S3A**). Nevertheless, when considering the average transcript levels observed as arising from the same genomic intervals, the EPODs from the low H-NS and medium H-NS categories still showed significantly lower expression than non-EPOD regions with similar H-NS levels (**Fig. S3B**), and the small number of highly H-NS bound regions which are not part of EPODs are in fact more silent than highly H-NS bound EPODs. Taken together, we thus observe that while many EPODs represent chromosomal regions silenced by H-NS, roughly one-third of EPODs do not show the characteristics of highly H-NS occupied regions, but are nevertheless transcriptionally silenced by an extended stretch of high protein occupancy. The possibility of course exists that H-NS is repositioned in the conditions of our study, which differ from those of [36], to cover the remainder of the EPODs identified here; however, we have shown in practice that this is not the case, as the non H-NS EPODs identified here persist even in a *hns*/*stpA* strain (Amemiya and Freddolino, manuscript in preparation). The mechanism of silencing at these non-H-NS dependent EPODs will likely be a fruitful area for future investigation.

Discussion

The study of bacterial transcriptional regulatory networks has long benefitted from bottom-up approaches such as DNase footprinting, ChIP-chip, and ChIP-seq to map the behavior of individual factors and regulons. At the same time, however, the insight provided by such approaches has been inherently limited by the need to specify *a priori* the target of investigation, either in terms of the regulator, regulated gene, or both. However, as we hope to have demonstrated here, a global agnostic strategy (as exemplified by IPOD-HR) provides a unique top-down complement to existing methods by permitting rapid profiling of the protein occupancy landscape of a bacterial chromosome. We have demonstrated that IPOD-HR simultaneously enables resolution of individual changes in transcription factor binding at specific sites, inference of new regulatory motifs which likely correspond to functional but poorly characterized transcriptional regulators, and large-scale patterns of protein occupancy indicative of constitutively silenced genomic regions. IPOD-HR thus falls into the same family as methods such as DNase I hypersensitivity [37], MNase-seq [38], and ATAC-seq [31], but developed, tuned, and validated for the unique molecular and biophysical features of bacterial chromosomes.

We expect that all three key capabilities of IPOD-HR highlighted above will prove to be of substantial utility in investigating all cultivable bacterial transcriptional regulatory networks, and could potentially even be applied to environmental samples to study occupancy landscapes in uncultivable bacteria. The ability to directly track the occupancy of TFBSs for a large set of transcriptional regulators in parallel provides the missing link that has previously stymied efforts to predict the transcriptional output of *E. coli* across conditions, as consideration of only the expression levels of TFs to predict the behavior of their regulons has yielded mixed results [8,39]. Furthermore, the ability to identify likely regulatory sites even in the absence of prior knowledge, as shown both for isolated promoters (**Fig. 4**) and inference of entire regulons (**Fig. 5**), will substantially accelerate our ability to complete a wiring diagram for the *E. coli* transcriptional regulatory network and to rapidly approach the networks of other less well-characterized bacterial species. Here IPOD-HR provides a powerful high-throughput *in vivo* approach tracking occupancy at native sites, complementing methods based on screening with purified proteins [23], computational inference [25], or reporter assays [24].

Our study of diverse experimental conditions across different genetic and physiological states provides a comprehensive view of the protein-DNA interactome of *E. coli*. As we have shown, the majority of discovered occupancy events do not correspond to previously known or annotated sites of protein-DNA interactions. We have further shown that these global occupancy profiles can be used for wholesale discovery of sequence specificity for the set of TFs active under these conditions. These occupancy maps and the corresponding DNA motifs provide the community with a rich catalogue of likely regulatory events to study, targeting either particular genes or larger pathways. Indeed, our finding that the novel occupancy sites and DNA motifs are highly enriched upstream of genes that are under-studied promises to discover and expand the physiological and regulatory modules of *E. coli*, beyond those that have been targeted by decades of previous research.

550 Our study also provides significant additional evidence for the presence of large, transcriptionally silent, high occupancy chromosomal domains in *E. coli.* Many such EPODs clearly correspond to regions of H-NS binding, which has previously been shown to form several types of filaments that silence horizontally acquired DNA [5,35,40–42]. On the other hand, we also observe a substantial fraction of EPODs that do not correspond to H-NS binding, and yet
555 are still associated with transcriptionally silent regions of the chromosome. Numerous questions regarding the nature and role of those EPODs remain for future work, including: what is the protein composition of non H-NS EPODs? What rules dictate their formation on specific sites? We are also tempted to speculate that in some contexts non-H-NS EPODs may undergo condition-dependent changes in occupancy that drive transcriptional regulation, although no
560 such cases could be definitively identified in the conditions studied here. Such behavior has already been observed for H-NS filaments in various enterobacteria [43,44]. Ongoing application of IPOD-HR to a broader range of physiological conditions in *E. coli* should provide further insight into the overall landscape of large-scale protein occupancy across conditions, allowing tracking both of occupancy associated with H-NS (and the related protein StpA) and
565 other classes of EPODs in a single experiment.

Our IPOD-HR strategy for mapping the global dynamics of the *E. coli* protein-DNA interactome relies only on simple physico-chemical principles for isolating protein-DNA complexes. As such, it is easily transferable to other bacterial species. The rich and comprehensive data sets
570 generated by such studies, and application of statistical inference during data processing as exemplified here, will provide particularly important regulatory roadmaps in organisms with less well studied transcriptional regulatory networks. In the future, more applications to a broader range of physiological conditions (in *E. coli*) and to other bacterial strains and species will provide important information on the role of large-scale nucleoprotein assemblies on gene
575 regulation, and pave the way for more comprehensive and predictive models of transcriptional regulatory logic, particularly for non-model bacterial species of clinical and industrial importance.

580

Figures

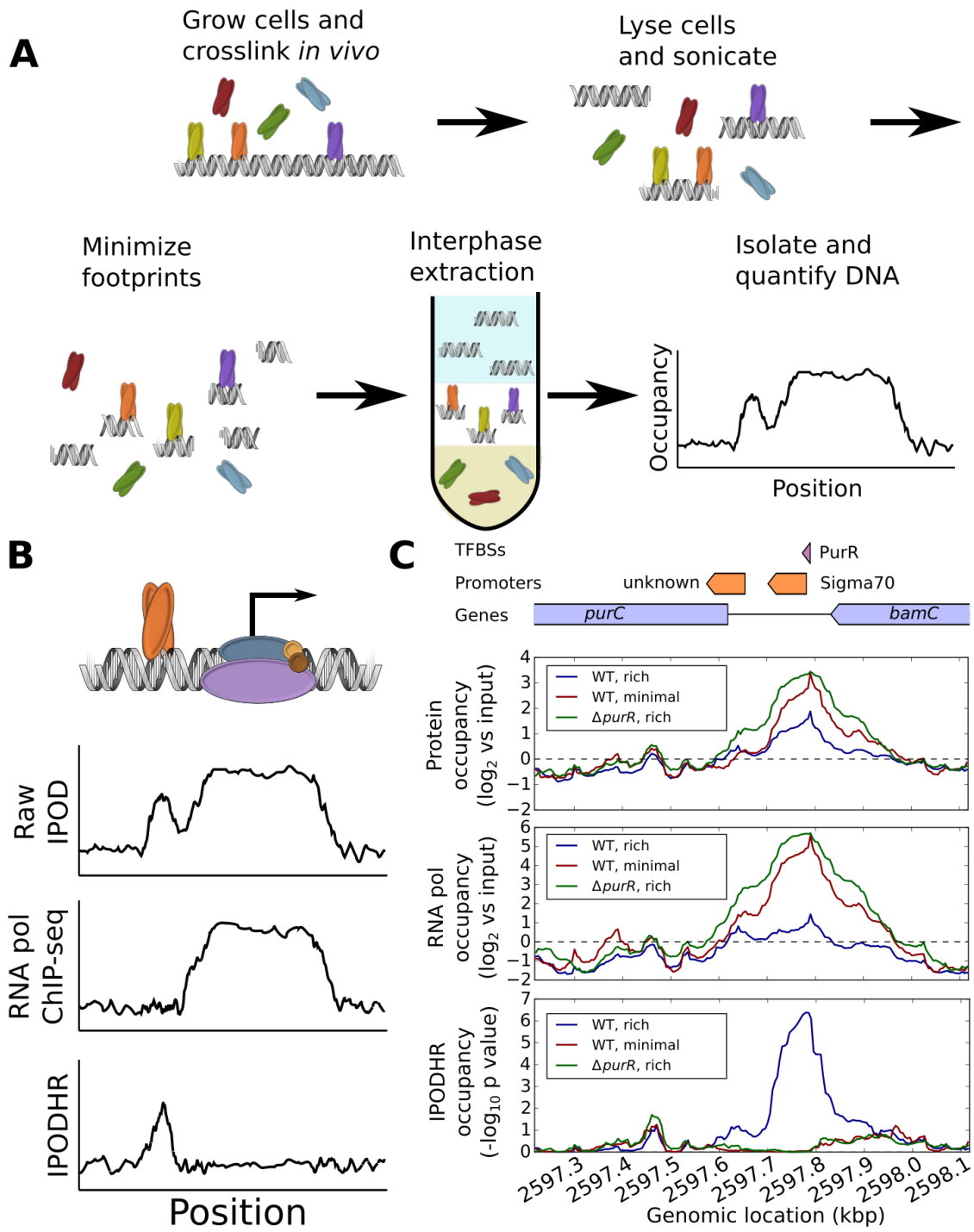

A
Grow cells and crosslink *in vivo*
Lyse cells and sonicate
Minimize footprints
Interphase extraction
Isolate and quantify DNA
Occupancy
Position
B
Raw IPOD
RNA pol ChIP-seq
IPODHR
Position
C
TFBSs
PurR
Promoters
unknown
Sigma70
Genes
purC
bamC
Protein occupancy (log2 vs input)
RNA pol occupancy (log2 vs input)
IPODHR occupancy (-log10 p value)
WT, rich
WT, minimal
ΔpurR, rich
2597.3
2597.4
2597.5
2597.6
2597.7
2597.8
2597.9
2598.0
2598.1
Genomic location (kbp)

Figure 1: Schematic of IPOD-HR technology and detection of context-dependent binding by transcription factor PurR. (A) Overall workflow for isolation of the IPOD-HR fraction and quantification of total protein occupancy. **(B)** The final IPOD-HR signal is obtained by subtracting a normalized RNA polymerase occupancy signal from the raw IPOD-HR protein occupancy, resulting in a polymerase-corrected signal. **(C)** Example of RNA-polymerase corrected IPOD-HR profile upstream of the *purC* gene, where subtraction of RNA polymerase occupancy from the raw IPOD-HR signal properly reveals a PurR binding site in rich media that is lost upon deletion of *purR* or transition to minimal media. In the schematic above the plots, blue regions show genes, orange regions show promoters, and purple regions show annotated transcription factor binding sites.

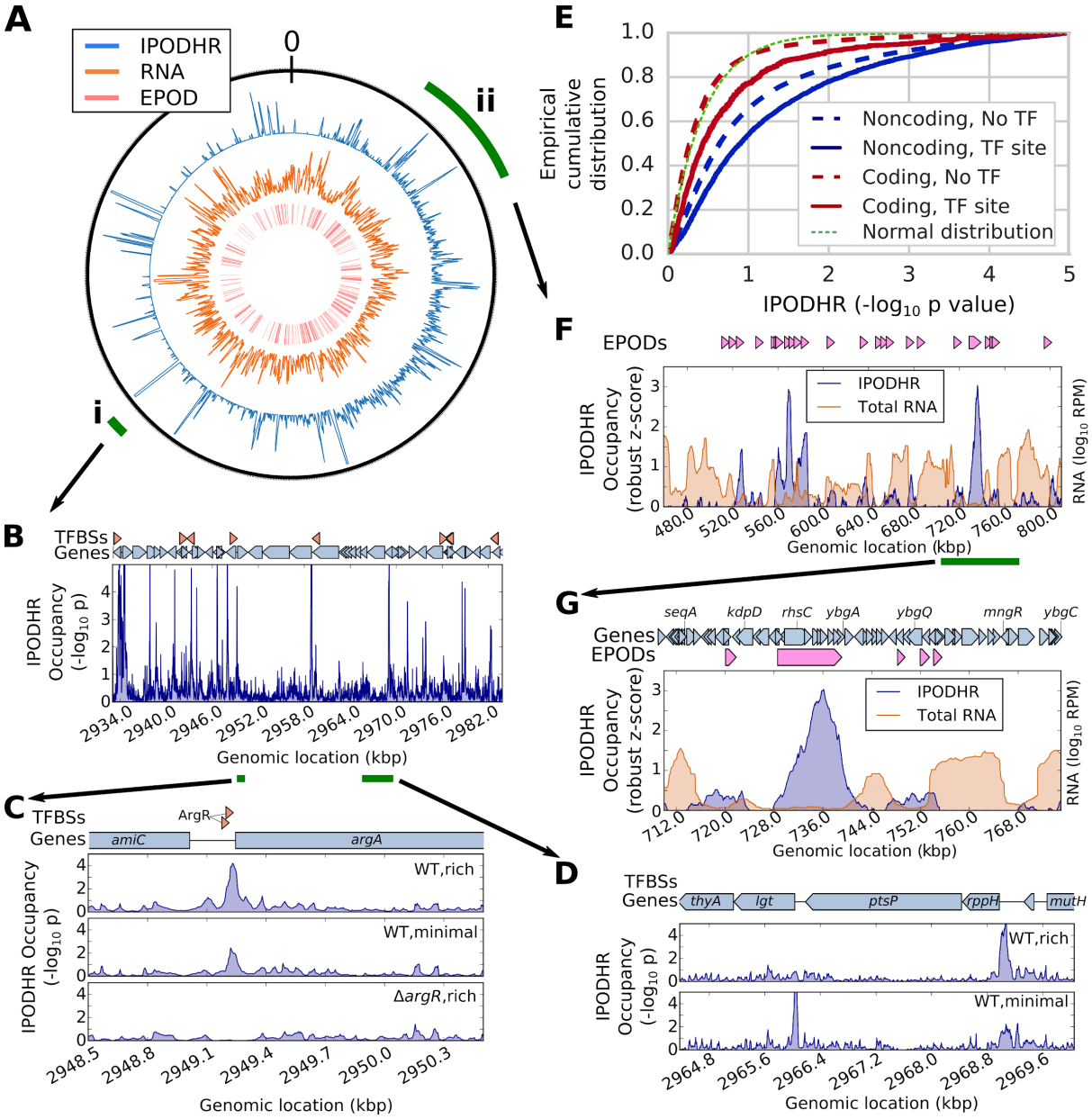


Figure 2: IPOD-HR profiles reveal rich high-resolution occupancy dynamics and large-scale structural features across the chromosome. (A) *Outer track:* IPOD-HR occupancy (robust Z-scores, 5 kb moving average); *middle track:* total RNA read density (5 kb moving average); *inner track*: locations of inferred EPODs. The outer green wedges mark the portion of the chromosome shown in subsequent panels. The origin of the coordinate system is oriented at the top of the plot. **(B)** IPOD-HR occupancy measured during growth in glucose rich defined medium, in the vicinity of wedge **i** from panel **A**. Green segments below the genomic coordinates indicate the regions highlighted in panels **C-D**. **(C)** Condition-dependent occupancy changes at the ArgR binding sites upstream of *argA*. **(D)** Identification of condition-specific occupancy of likely TFBSs upstream of *lgt* and *rppH*. **(E)** Cumulative histograms showing RNA polymerase ChIP-subtracted IPOD-HR occupancy in coding vs. noncoding regions, and at sites

that match known transcription factor binding sites from RegulonDB [7], compared with the
610 curve that would be expected from a standard normal distribution of scores. **(F)** Occupancy (blue) and total RNA abundance (orange) for a selected sector of the genome (wedge **ii** from panel **A**), showing the presence of several EPODs in regions corresponding to low RNA abundance; rolling medians over a 5 kb window are plotted, with RNA read densities shown in units of reads per million (RPM). **(G)** Magnification of the region highlighted by the green bar in
615 panel **F**, illustrating a silenced region in and around *rhsC*, alongside flanking areas of low IPOD-HR occupancy and high gene expression. A 5 kb rolling median is plotted.

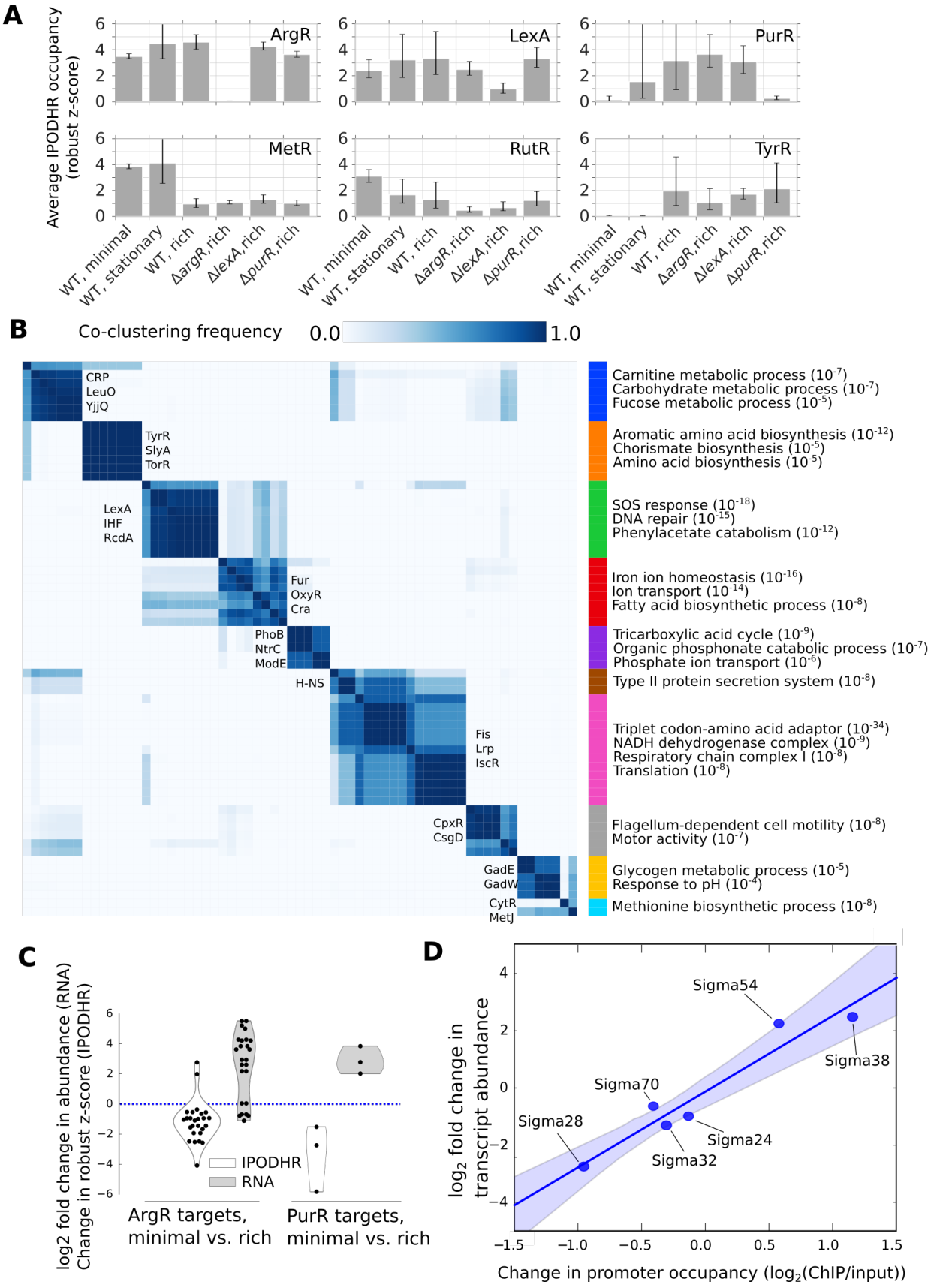

A
Average IPODHR occupancy (robust z-score)
ArgR
LexA
PurR
MetR
RutR
TyrR
WT, minimal
WT, stationary
WT, rich
ΔargR,rich
ΔlexA,rich
ΔpurR,rich
B
Co-clustering frequency
0.0
1.0
CRP
LeuO
YjjQ
TyrR
SlyA
TorR
LexA
IHF
RcdA
Fur
OxyR
Cra
PhoB
NtrC
ModE
H-NS
Fis
Lrp
IscR
CpxR
CsgD
GadE
GadW
CytR
MetJ
Carnitine metabolic process (10-7)
Carbohydrate metabolic process (10-7)
Fucose metabolic process (10-5)
Aromatic amino acid biosynthesis (10-12)
Chorismate biosynthesis (10-5)
Amino acid biosynthesis (10-5)
SOS response (10-18)
DNA repair (10-15)
Phenylacetate catabolism (10-12)
Iron ion homeostasis (10-16)
Ion transport (10-14)
Fatty acid biosynthetic process (10-8)
Tricarboxylic acid cycle (10-9)
Organic phosphonate catabolic process (10-7)
Phosphate ion transport (10-6)
Type II protein secretion system (10-8)
Triplet codon-amino acid adaptor (10-34)
NADH dehydrogenase complex (10-9)
Respiratory chain complex I (10-8)
Translation (10-8)
Flagellum-dependent cell motility (10-8)
Motor activity (10-7)
Glycogen metabolic process (10-5)
Response to pH (10-4)
Methionine biosynthetic process (10-8)
C
log2 fold change in abundance (RNA)
Change in robust z-score (IPODHR)
IPODHR
RNA
ArgR targets, minimal vs. rich
PurR targets, minimal vs. rich
D
log2 fold change in transcript abundance
Sigma54
Sigma38
Sigma70
Sigma28
Sigma24
Sigma32
Change in promoter occupancy (log2(ChIP/input))

Figure 3: IPOD-HR profiles reveal global binding activity of known transcription factors and sigma factors. (**A**) Average (geometric mean) occupancies for all annotated binding sites of the six indicated transcription factors under each indicated condition. Error bars indicate a 95% confidence interval based on parametric bootstrapping with pessimistic assumptions; see Methods for details. (**B**) Heat map showing the consensus clustering (co-occurrence frequencies) of the pattern of occupancy dynamics for the regulons of all considered TFs across the varied nutrient conditions in this study (see Methods for details). Consensus division into ten clusters via agglomerative clustering is shown at right; for each cluster, representative TFs (on matrix) and regulated gene ontology terms (right) are shown, with numbers in parentheses indicating the $\log_{10}$ p-value for enrichment of that GO term. A full listing of p values is given in **Table S4.** **(C)** Changes in occupancy and target gene transcript level for all annotated repressive binding sites of ArgR and PurR (for minimal media vs. rich media), in each case demonstrating the strong anti-correlation of binding and regulatory effects across the regulons. (**D**) Correlation of promoter-level occupancy changes (measured by RNA polymerase ChIP-seq) and changes in transcript abundance, shown for the WT stationary phase condition compared with exponential phase. Shaded area shows a bootstrap-based 95% confidence interval.

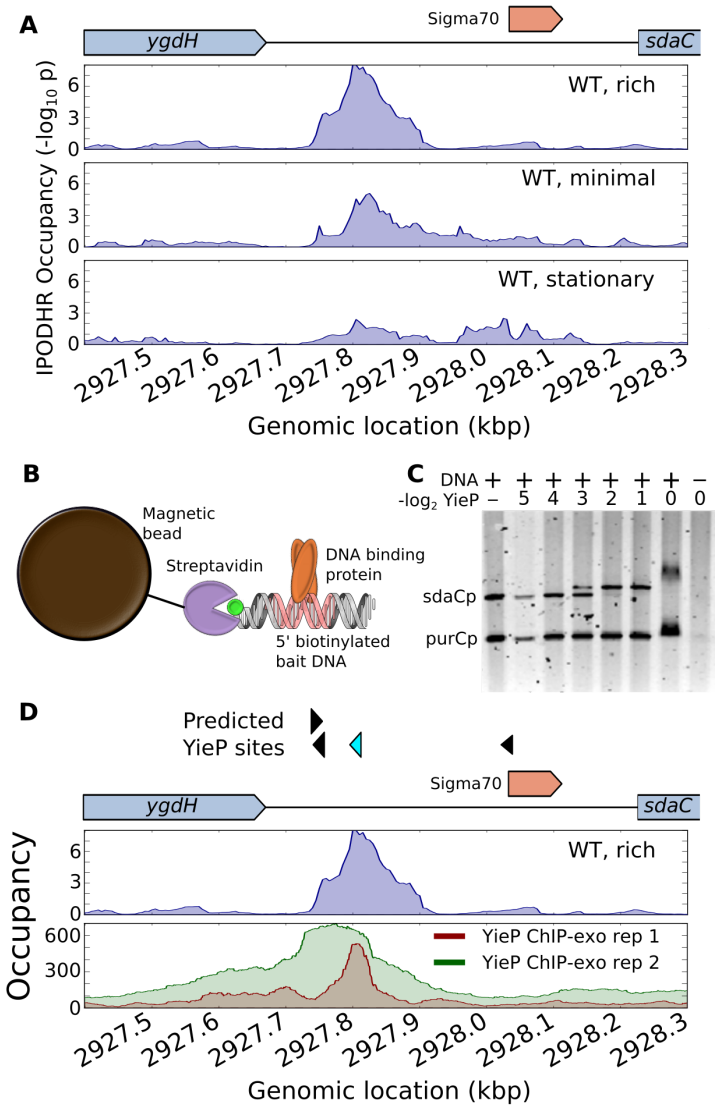


Figure 4: Experimental identification of the protein bound to a novel occupancy peak upstream of the *sdaC* promoter. (A) IPOD-HR profiles upstream of *sdaC* in rich (M9/RDM/glu) media, minimal (M9/glu) media, and in rich media in stationary phase. No annotated transcription factor binding sites are present in the displayed region. **(B)** Schematic of pulldown/mass spectrometry experiments used to identify factors binding the *sdaC* promoter. **(C)** Gel shift experiments showing specific interaction of YieP with the *sdaC* promoter. Increasing concentrations of purified His_6-YieP are incubated with a mixture of fluorescein-labeled promoter regions from *sdaC* and *purC* and then run on a gel, demonstrating specific shifting of the *sdaC* promoter region. YieP concentrations are given as the number of two-fold dilutions relative to full strength. **(D)** Comparison of IPOD-HR occupancy profile (as in panel **A**) with ChIP-exo data from [25], with the latter given as total read counts (parsed from GEO accessions GSM3022131 and GSM3022132). The top track of predicted YieP sites shows significant hits for the YieP motif identified based on that ChIP-exo data set. Out of 1,025 potential YieP sites in the genome, the location highlighted in cyan is tied for 10th highest score

(identified using FIMO; see Methods for details). Occupancy signal is given as $-\log_{10}(p)$ for the IPOD-HR track, or raw counts (averaged across strands) for the ChIP-exo tracks.

655

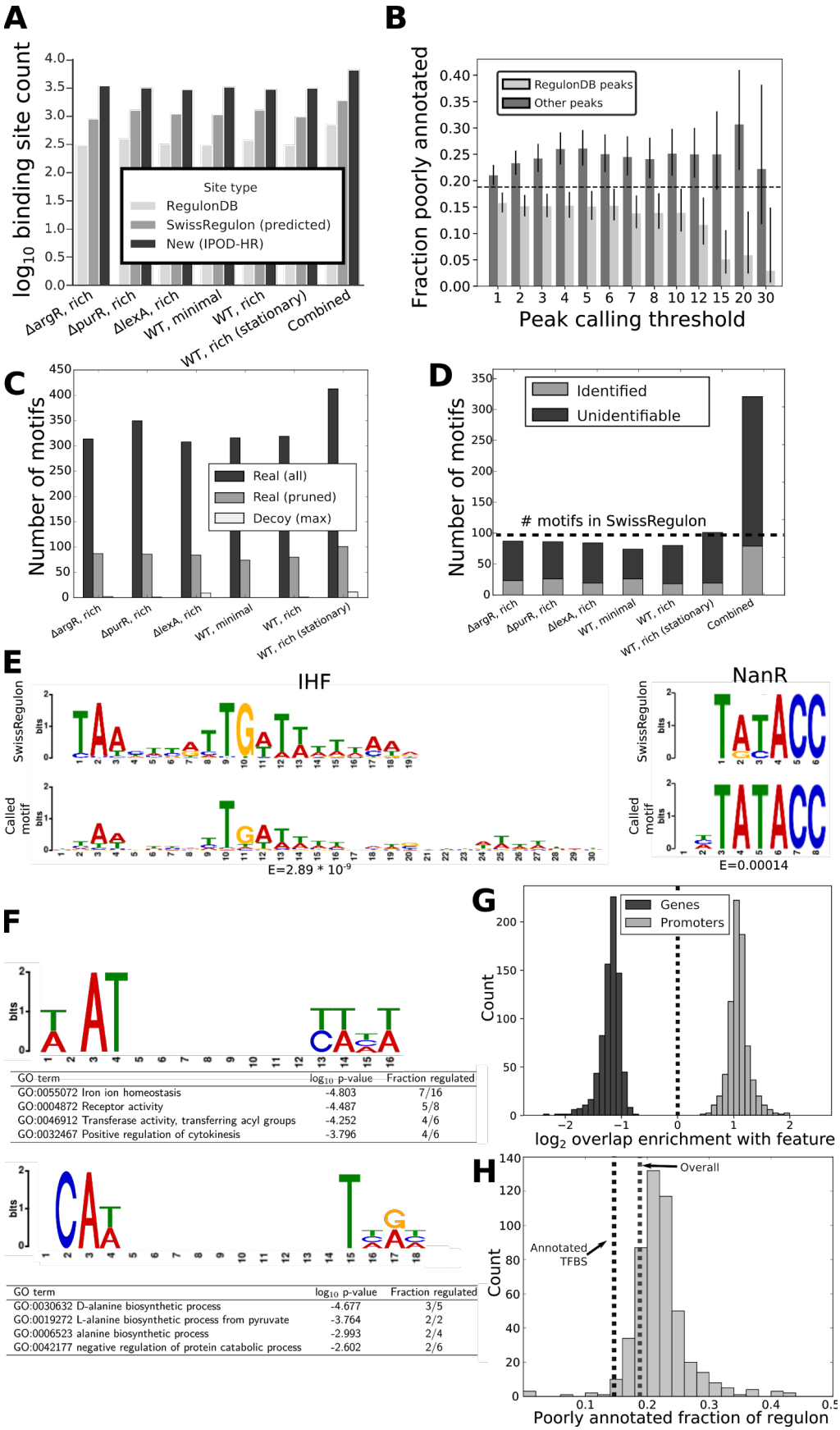

A
log10 binding site count
Site type
RegulonDB
SwissRegulon (predicted)
New (IPOD-HR)
ΔargR, rich
ΔpurR, rich
ΔlexA, rich
WT, minimal
WT, rich
WT, rich (stationary)
Combined
B
Fraction poorly annotated
RegulonDB peaks
Other peaks
Peak calling threshold
C
Number of motifs
Real (all)
Real (pruned)
Decoy (max)
D
Number of motifs
Identified
Unidentifiable
motifs in SwissRegulon
E
IHF
SwissRegulon
Called motif
E=2.89 * 10-9
NanR
E=0.00014
F
GO term | log10 p-value | Fraction regulated
GO:0055072 Iron ion homeostasis | -4.803 | 7/16
GO:0004872 Receptor activity | -4.487 | 5/8
GO:0046912 Transferase activity, transferring acyl groups | -4.252 | 4/6
GO:0032467 Positive regulation of cytokinesis | -3.796 | 4/6
GO:0030632 D-alanine biosynthetic process | -4.677 | 3/5
GO:0019272 L-alanine biosynthetic process from pyruvate | -3.764 | 2/2
GO:0006523 alanine biosynthetic process | -2.993 | 2/4
GO:0042177 negative regulation of protein catabolic process | -2.602 | 2/6
G
Genes
Promoters
Count
log2 overlap enrichment with feature
H
Overall
Annotated TFBS
Count
Poorly annotated fraction of regulon

Figure 5: Genome-wide *de novo* discovery of sequence specificity motifs for actively bound transcription factors. (**A**) At a peak calling threshold of 4 (c.f. **Figure S1**), we show the number of identified binding sites that overlap with annotated sites in RegulonDB ("RegulonDB"), motif-based predicted binding sites ("SwissRegulon"), or novel ("New"). The "Combined" category represents peak sets where the peaks at a given threshold identified across all conditions are merged, prior to comparisons with the RegulonDB and predicted databases. Qualitatively similar results are observed at all tested peak calling thresholds (data not shown). (**B**) All called IPOD-HR occupancy peaks across the conditions shown in panel **A** were combined, and then partitioned based on whether they overlap with a known or inferred binding site in RegulonDB (RegulonDB peaks) or not (Other peaks). Peaks were then considered to have regulatory potential if they fell within 50 bp of an annotated transcription start site, and the fraction of the genes potentially regulated by each peak category plotted across different peak calling threshold. Error bars show 95% credible intervals calculated assuming that the incidence of poorly annotated genes in the inferred regulon is a binomial random variable, using Bayesian inference with a Beta(1,1) prior. The dashed line shows the overall fraction of poorly annotated genes included in the analysis (that is, those belonging to transcripts regulated by at least one annotated transcription start site in RegulonDB). (**C**) Number of motifs discovered *de novo* using IPOD-HR occupancies under each condition in our study. "All" and "pruned" refer to all discovered motifs and those surviving cluster-based filtering by RSAT (see Methods for details), respectively. "Real" shows the motif counts discovered in real data, and "Decoy" shows the maximum discovered motif count across 20 independent circular permutations of the data under each condition. (**D**) Classification of non-redundant motifs across conditions as "Identified" (match to an existing motif from the SwissRegulon database, via TOMTOM, with E-value < 0.5) or "Unidentified" (no matches found with E<0.5). "Combined" refers to the full set of motifs discovered after pooling all motifs across all conditions and redundancy filtering; a horizontal dashed line shows the total number of known motifs present in SwissRegulon. (**E**) Example cases of "Identified" matches of IPOD-HR-inferred motifs with motifs from the SwissRegulon database, showing good correspondence with annotated IHF (left) and NanR motifs. E values arising from the TOMTOM search pairing newly discovered motifs with similar known motifs are shown beneath each inferred motif. y axes for motifs in this and the following panel show information content in bits. (**F**) Examples of two newly inferred motifs that do not have identifiable hits in the SwissRegulon database (as assessed using TOMTOM). In each case, the gene ontology (GO) terms showing most significant enrichments amidst the predicted regulon associated with that motif are shown (see Methods for details). (**G**) Overlap of predicted binding sites for IPOD-HR inferred motifs with either coding regions (genes) or promoters (both as annotated in RegulonDB); shown are the $\log_2$ fold enrichment or depletion of the overlap as compared with that expected by chance. (**H**) For the predicted regulon of each newly inferred motif, we show the fraction of regulon members that are poorly annotated (Uniprot annotation score of 1 or 2 out of 5); for comparison, dashed lines are shown for the values obtained when the same statistic is calculated for all annotated TF-gene interactions in RegulonDB ("Annotated TFBS"), and for the genome as a whole ("Overall").

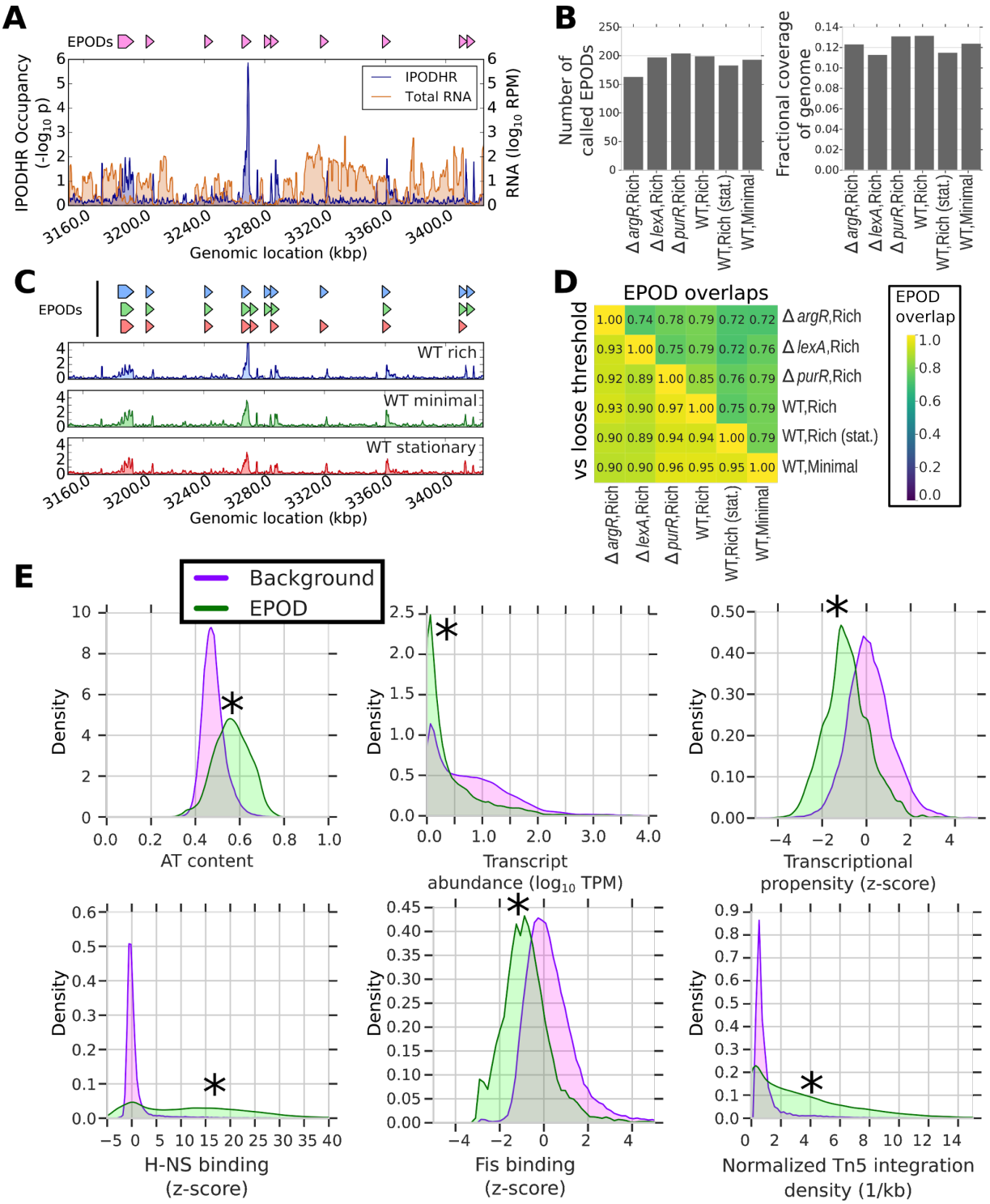


Figure 6: EPODs define stable genomic structures and are associated with many distinct features. (**A**) EPOD calls from a representative genomic region in the WT Rich media condition, along with protein occupancy and RNA levels smoothed with a 1 kb rolling median. (**B**) Number of called EPODs by condition (left) and fraction of the genome covered by EPODs (right). (**C**) IPOD-HR occupancies (shown over a 1 kb rolling median) and associated EPOD calls under three different conditions, in the same genomic region shown in panel **A**. EPOD calls are shown

above the occupancy, in the same order as the data tracks. (**D**) Upper triangle: Overlap of
710 EPOD calls (see text for definition) between each pair of the studied conditions. In the lower triangle, each entry shows the fraction of the EPOD calls (at a 5 bp resolution) from the sample defining that row that is contained in a relaxed set of EPOD calls (see text) of the sample defining that column. (**E**) Density plots showing normalized histograms (smoothed by a kernel density estimator) of the specified quantities for regions of the genome that are in EPODs
715 versus those that are not (Background), as assessed in the WT M9/RDM/glu (WT, rich) condition. '*' indicates FDR-corrected $p<0.005$ via a permutation test (against a null hypothesis of no difference in medians). Significance calling and additional comparisons are shown in **Table S2**.

720

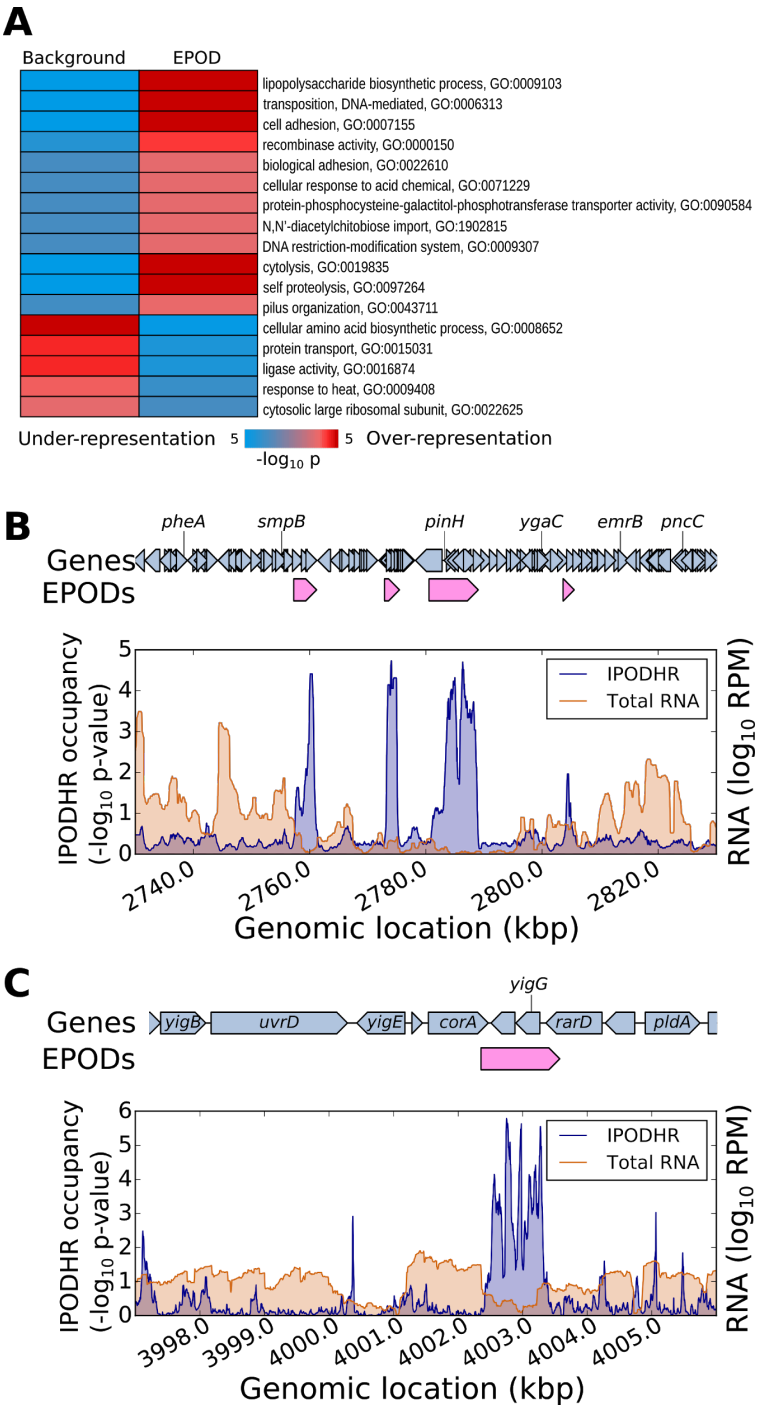


Figure 7: EPODs are statistically enriched for genes in specific functional categories. (**A**) The genome was split into EPOD and background regions as in **Figure 6**; we then applied iPAGE [45] to identify gene ontology terms showing significant mutual information with occupancy in EPODs. All shown GO terms were significant according to the built-in tests in iPAGE. **(B)** Multiple EPODs are associated with silencing of the CP4-57 prophage. Shown are the IPOD-HR occupancy and transcript levels in the vicinity of the prophage locus during growth in rich defined media with glucose, with EPOD locations indicated above the plots. **(C)**

730 Association of a small EPOD with two genes of unknown function, *yigE* and *yigF*; data tracks defined as in panel B.

Methods

Strain construction

The base strain for all experiments used here is an MG1655 stock obtained from H. Goodarzi, which is isogenic with ATCC 700926 [46]. All specified gene knockouts were obtained by P1 transduction [47] of the FRT-flanked *kanR* marker from the corresponding knockout strain of the Keio collection [48], followed by Flp recombinase mediated excision of the marker using the pCP20 plasmid [49] to leave a small scar in place of the original open reading frame. Candidate isolates for each deletion were grown overnight at 42° C to drop the pCP20 plasmid, and then replica plated onto appropriate selective plates to ensure loss of both the plasmid and kanamycin resistance marker. Knockouts were confirmed by PCR fragment sizing and/or sequencing across the marker scar. Note that the Δ*lexA* strain that we refer to is in fact Δ*lexA*/Δ*sulA*, as loss of *lexA* is lethal in the presence of a functional *sulA* gene [50,51].

Media/culture conditions

For routine cloning applications and for recovery of cryogenically preserved cells, we used LB (Lennox) media (10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl), with bacteriological agar (15g/L) added as appropriate.

For physiological experiments, we made use of a variety of supplemented versions of M9 defined medium (6 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 1 g/L NH_4Cl, 0.5 g/L NaCl, 1 mM $MgSO_4$) [47]. Our M9 minimal media condition (M9/min) additionally includes 0.2% (w/v) glucose, 0.4 mM $CaCl_2$, 40 μM ferric citrate, and the micronutrient mixture typically incorporated in MOPS minimal media [52]. Our M9 rich defined medium condition (M9/rdm) instead incorporates into the M9 base 0.4% (w/v) glucose, MOPS micronutrients (as above), 4 μM $CaCl_2$, 40 μM ferric citrate, and 1x supplements ACGU and EZ as used in MOPS rich defined medium [52].

Cell growth and harvest for IPOD-HR

The cells of interest were grown overnight in the media of interest after inoculation from an LB plate. In the morning, the culture was back-diluted into fresh, prewarmed media to an OD600 of 0.003. The culture was then grown to the target OD600 (0.2, except in the case of stationary phase samples, which are described below), at which point a 200 μL aliquot was removed and preserved in 1 mL of DNA/RNA Shield (Zymo Research) following the manufacturer's instructions.

The remainder of the culture was treated with rifampicin to a final concentration of 150 μg/mL, and incubated for 10 minutes under the same culture conditions as the main growth. The culture was then rapidly mixed with concentrated formaldehyde/sodium phosphate (pH 7.4) buffer sufficient to yield a final concentration of 10 mM NaPO4 and 1% v/v formaldehyde. Crosslinking was allowed to proceed for 5 minutes at room temperature with vigorous shaking, followed by quenching with an excess of glycine (final concentration 0.333 M) for 5 minutes with shaking at

770 room temperature. The crosslinked cells were subsequently chilled on ice, and washed twice with ice cold phosphate buffered saline, 10 mL per wash. The fully washed pellets were carefully dried, any remaining media pipetted away, and then the pellets were snap-frozen in a dry ice-ethanol bath and stored at -80 C.

775 In the case of our stationary phase samples, cells were grown as described above in terms of back-dilution and growth to an OD600 of 0.2, and then grown for an additional three hours prior to RNA harvest, rifampicin treatment, and crosslinked as described above.

Cell lysis and DNA preparation

Frozen cell pellets were resuspended in 1x IPOD lysis buffer (10 mM Tris HCl, pH 8.0; 50 mM
780 NaCl) containing 1x protease inhibitors (Roche Complete Mini, EDTA free) and 52.5 kU/mL of ready-lyse (Epicentre); 600 µL per pellet (stationary phase cells were diluted 10x prior to lysis, and only 1/10 of the resulting material used, due to the much higher biomass of those pellets). We incubated the resuspended pellet for 15 minutes at 30 C, and then placed it on ice. We then sonicated the cells using a Branson digital sonicator at 25% power, using three 10 second
785 bursts with 10 second pauses between bursts. The cells were maintained in a wet ice bath throughout sonication.

We then performed a calibrated DNA digestion to sub-200 bp fragments, by adding to the sonicated lysates 60 µg RNase A (Thermo Fisher), 6 µL DNase I (Fisher product #89835), 5.4
790 µL 100 mM MnCl2, and 4.5 µL 100 mM CaCl2, and then incubating on ice. While the appropriate digestion time must be calibrated for each particular sample type and batch of DNase, 30 minutes of digestion proved appropriate for all samples here. Reactions were quenched after completion by the addition of 50 µL 500 mM EDTA (pH 8.0), typically yielding 50-200 bp fragments.
795

IPOD-HR Interface Extraction

Prior to interface extraction, samples were clarified by centrifugation for 10 minutes at 16,9000xg at 4° C. After clarification, a 50 microliter input sample was diluted 1:9 in elution buffer (50 mM Tris, pH 8.0; 10 mM EDTA; 1% SDS) and kept on ice until the reverse
800 crosslinking step. The remainder of the lysate was mixed with 1 volume of 100 mM Tris base and 2 volumes of 25:24:1 phenol:chloroform:isoamyl alcohol, vortexed, and then incubated for 10 minutes at room temperature. After incubation, the sample was spun at 21,130xg for two minutes at room temperature, allowing formation of a white disc at the aqueous-organic interface enriched for protein-DNA complexes [10,11].
805

The complete aqueous phases were removed and discarded, and the remaining disc washed again with 350 microliters TE (10 mM Tris, pH 8.0; 1 mM EDTA), 350 microliters 100 mM Tris base, and 700 microliters 24:1 chloroform:isoamyl alcohol. The resulting mixture was vortexed vigorously, and again centrifuged for 2 minutes at 21,130xg. All liquid was again removed, and
810 the wash was repeated using 700 microliters TE and 700 microliters 24:1 chloroform:isoamyl

alcohol. After vortexing, centrifugation, and removal of the final wash (exactly as above), any residual liquid was removed by wicking with a laboratory wipe (if any substantial pools of liquid were present). Finally, the interface was resuspended in 500 microliters of elution buffer (described above), vortexed vigorously, and kept on ice until reverse crosslinking (no more than a few hours).

We caution the reader that the separation of the interface layer from the liquid on either side of it is crucial to success with this method. We have found it most effective to tilt the microcentrifuge tube toward while pipetting out the organic layer from beneath, at which point the interface will adhere to the tube wall and allow easy removal of the aqueous layer. We have also found that the handling characteristics of the interface vary greatly with the plasticware in use. For the work described here, we have used 2 mL microcentrifuge tubes from USA Scientific for all interface handling, as the interfaces adhere nicely to the tube wall (other plasticware may yield variable results); at the same time, the use of low-retention pipette tips appears to reduce binding of the interface to the tip.

RNA polymerase chromatin immunoprecipitation

DNA for RNA polymerase ChIP-seq experiments was prepared as described above for IPOD-HR interface extraction up through the lysate clarification stage. Whenever possible, we used frozen pellets obtained from the same culture for matched IPOD-HR and ChIP-seq experiments, in which case the lysates were pooled and mixed immediately prior to removal of a single input sample. ChIP procedures here were modeled on those of [53].
The digested lysates were mixed 1:1 with 2x IP buffer (200 mM Tris, pH 8.0; 600 mM NaCl; 4% Triton X-100; 2x Roche Complete EDTA-free protease inhibitors), and then kept on ice for no more than a few hours prior to antibody addition. We added 10 microliters of purified anti-*E. Coli* RNA polymerase antibody (Neoclone WP023), and incubated overnight with rocking at 4 C. Near the end of the incubation period, we resuspended an aliquot of 50 microliters of protein G dynabeads (Invitrogen) and equilibrated the protein G beads with 1x IP buffer lacking protease inhibitors. The bead aliquot was added to the antibody-lysate mixture, and then incubated 2 hours with rocking at 4° C. The bead-antibody-target complexes were subsequently subjected to the following series of washes, with 1 mL used per wash. All washes were at room temperature, and involved manual resuspension of the beads in the new wash buffer followed by immediate re-separation.

- 1x Wash buffer A (100 mM Tris, pH 8.0; 250 mM LiCl; 2% Triton X-100; 1 mM EDTA)
- 1x Wash buffer B (100 mM Tris, pH 8.0; 500 mM NaCl; 1% Triton X-100; 0.1% sodium deoxycholate; 1 mM EDTA)
- 1x Wash buffer C (10 mM Tris, pH 8.0; 500 mM NaCl; 1% Triton X-100; 1 mM EDTA)
- 1x TE (10 mM Tris, pH 8.0; 1 mM EDTA)

The antigens were subsequently eluted by adding 500 microliters of elution buffer (composition described above) and incubating 30 minutes at 65° C, with vigorous vortexing every 5-10 minutes.

Crosslinking reversal and recovery of DNA

The DNA from the input, IPOD-HR, and ChIP fractions described above was recovered using identical procedures: samples diluted in elution buffer (see above) were incubated overnight (6--16 hours) at 65° C to reverse formaldehyde crosslinks. After allowing the samples to cool to room temperature, we then added 100 μg of RNase A (Thermo-Fisher), incubated 2 hours at 37° C, then added 200 μg of proteinase K (Fermentas) and incubated an additional 2 hours at 50° C. DNA was then recovered via standard phenol-chloroform extraction and ethanol precipitation, following protocols from [47]. We used Glycoblue (Ambion) as a co-precipitant, NaCl as a precipitating salt (due to the presence of SDS in our solution), and washed with ice-cold 95% ethanol to avoid loss of low molecular weight DNA.

Recovered DNA was quantified via fluorescent quantitation (using either the Invitrogen PicoGreen or Promega QuantIT system), and samples of sufficiently high concentration were also run on a 2% agarose gel for fragment size assessment. Typical total yields from the procedure above were on the order of 1 µg of DNA for the input samples, 100-200 ng for the IPOD-HR samples, and 1-10 ng for the ChIP samples.

Preparation of next-generation sequencing (NGS) libraries

Except as otherwise noted, all DNA samples were prepared for Illumina sequencing using the NEBNext Ultra DNA Library Prep Kit (NEB product #E7370), with either single index or dual index primers also obtained from NEB. We followed the manufacturer's instructions except for the following variations:

- Cleanups prior to adapter ligation were performed using a Zymo Clean&Concentrator 5 spin column kit or Zymo Oligo Clean&Concentrator spin column kit instead of Ampure beads, in order to avoid the loss of low molecular weight DNA
- We used Ampure and Axygen PCR cleanup beads interchangeably, having established in side-to-side comparisons that they were functionally identical for the steps in the NEB sequencing prep (data not shown). The final cleanup step was in some cases repeated to remove obvious populations of adapter dimers.

All libraries were sequenced on either an Illumina HiSeq or NextSeq instrument; detailed statistics on read lengths and counts are provided in Table S3. A small number of samples were prepared for sequencing using an Illumina Truseq Nano kit instead of the NEBnext kit noted above (those samples are identified in Table S3); we found that upon calculation of correlations between the coverages of a broad range of IPOD, input, and RNA polymerase ChIP-seq samples prepared using various sequencing preparation kits that the Truseq Nano samples were indistinguishable from NEBNext Ultra samples, whereas other sequencing preparation methods (notably including standard Illumina Truseq samples) lead to detectable non-biological differences in observed coverage.

RNA isolation and RNA-seq sample preparation

As noted above, samples for RNA isolation were preserved immediately prior to rifampicin addition by dilution in a 5x excess of DNA/RNA Shield (Zymo); the RNA samples were then stored at -80° C until purification. RNA was isolated using a Zymo QuickRNA microprep kit following the manufacturer's instructions, including the on-column DNase digestion. Purified RNA was quantified using RiboGreen (Invitrogen), and then ribosome-depleted using the Illumina RiboZero Gram-negative bacteria kit according to the manufacturer's instructions, with the input RNA amount and all reaction volumes cut in half. Final recovery of the ribo-depleted RNA was accomplished using the modified Zymo spin column protocol present in the RiboZero documentation. Ribo-depleted RNA was then prepared for sequencing using the NEBNext Ultra Directional RNA kit (NEB product E7420), and sequenced as described above for the DNA samples.

Analysis of NGS data

All NGS data was preprocessed using a common pipeline, after which DNA and RNA data sets were processed separately. The reference genome in all cases was the most recent version of the *E. coli* MG1655 genome (GenBank U00096.3), with gene, transcription factor binding site, and transcription start site annotations from RegulonDB [7]. Data processing was automated using in-house python and bash scripts, and parallelized where possible using GNU parallel [54] or the python multiprocessing library.

Read Quality Control and Preprocessing

All reads were subjected to adapter removal using cutadapt 1.8.1 [55] to cut the common sequence of Illumina Truseq adapters, and then trimmed to remove low-quality read ends with Trimmomatic 0.33 [56], using the trimming steps '`TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:10`'. Samples were subjected to additional manual quality checks using FastQC [57] and MultiQC [58] to identify any irregularities in terms of sequence content, quality, or duplication.

DNA Sequencing and Protein Occupancy Calling

Surviving DNA reads were aligned to the U00096.3 genome using bowtie2 version 2.1.0 with "very sensitive" end-to-end alignment presets, and dovetail alignments allowed. Only concordant paired-end reads were retained for subsequent quantitation. Separate occupancy tracks were calculated both for the aligned reads of each sample, using the parsing method of Kroner *et al. [59]*, and scaling the basepair-wise contribution of each read by the inverse of its length (thus, each read contributed the same total amount of occupancy signal to the traces). All sample-wise occupancy data were normalized by quantile-normalizing the original datasets (acting separately for the input, IPOD interface, and RNA polymerase ChIP-seq tracks). In order to correct for copy number variations, for each biological condition, we fitted a periodic smoothing spline with four evenly spaced knots to the input samples for that condition; all occupancy values were divided by the spline-smoothed abundances prior to further processing. After abundance normalization, all data tracks were rescaled to have matching means, and then

all replicates for each sample type/biological condition combination were averaged to generate a composite occupancy track (yielding, for example, one input data track for the WT M9/RDM/glu condition, one IPOD data track for the WT M9/RDM/glu condition, etc.).

The displayed IPOD and ChIP data tracks were then obtained as $\log_2$ ratios of the extracted (interphase or ChIP) to input samples for each condition; we refer to these tracks as the "IPOD" and "ChIP" signals below. Upon viewing the correlation between total protein occupancy and RNA polymerase occupancy, two protein-occupied subpopulations were apparent (**Figure S4**): a linear subpopulation where total protein and RNA polymerase are well-correlated, and a second subpopulation of positions where the total protein occupancy is much higher than expected based on the RNA polymerase occupancy. We interpret the former set of positions as protein occupancy due directly to RNA polymerase binding, and the latter as non-RNA polymerase occupancy (as schematized in Fig. 1B). To obtain the fully processed IPOD-HR signal for non-RNA polymerase occupancy, we applied a LOESS [60] fit to the scatter plot of IPOD vs ChIP signal for each sample type (fitting to a 500-fold downsampled data set for the sake of efficiency), and subtract from the IPOD signal at each position 1.1 times the value predicted from the LOESS model based on the corresponding ChIP occupancy signal (an arbitrary scaling factor that makes the calling of non-RNA polymerase occupancy more conservative); no subtraction was performed at positions where the observed RNA polymerase occupancy was negative. We refer to the resulting ChIP-subtracted IPOD signal as the IPOD-HR signal; for analysis and display we further standardize the signal by calculating robust z scores, where the robust z-score z_i at position i is defined as

$$z_i = (x_i - \text{median}(\mathbf{X})) / \text{mad}(\mathbf{X})$$

for a IPOD-HR data vector **X**, and mad() indicates the median absolute deviation. In many cases a more useful signal for visualization is a p-value for enrichment at each site; $\log_{10}$ p-values are calculated under the null hypothesis that the distribution of the robust z-scores is standard normal. To provide uncertainty estimates grounded in observed levels of biological variability across replicates, for each data point we also constructed an interval between the lowest and highest values that could have been obtained for our occupancy statistics using any combination of biological replicates (potentially different replicates for the IPOD, ChIP, and input samples to construct the largest possible range). To calculate the error bars shown in Figure 3A, we then used parametric bootstrapping to generate confidence intervals for the parameters of interest, assuming that the occupancy of each TFBS followed a normal distribution with a mean of the observed mean and standard deviation of one-quarter the range between the highest and lowest replicate-wise values (thus treating the range of the pessimistic replicate-wise possible values as an interval expected to contain ~95% of observed points); 95% confidence intervals for the average site-wise occupancies were then calculated from 1,000 bootstrap replicates.

Feature calling

To identify peaks in the robust z-scaled IPOD-HR data, we applied continuous wavelet-transform (CWT) peak calling [61] (as implemented in the scipy.signal package), with a range of widths from 25 bp to 125 bp (at 5 bp increments) used to generate the CWT matrix, and refer to peaks based on the minimum signal-to-noise ratio threshold at which they appear as peak calls. Each peak call was padded by 30 bp on each side to define the peak region used in subsequent analysis.

Extended protein occupancy domains (EPODs) were called using an approach similar to that in (Vora et al., 2009): we identified EPOD seed regions as any region at least 1,024 bp in length, over which a 512 bp rolling median exceeded the overall kth percentile of a 256 bp rolling median across the entire chromosome (in all cases acting on the robust z-scored IPODHR data); we used k=90 for the main EPOD calls made in the text, and k=75 for the relaxed version used in threshold analysis. In the case of overlapping seeds, only the one with the highest median was retained. Seed regions were then expanded in both directions as far as possible while maintaining the median over the entire EPOD call above the threshold noted above, and without crossing any location with a robust z-score ≤ 0.

Transcription factor co-clustering analysis

For the transcription factor occupancy and co-clustering data, we performed consensus clustering (inspired by [62]). For each biological condition, we assigned each transcription factor a score given by the geometric mean of the site-level occupancies (IPOD-HR $-\log_{10}$ p values) for annotated binding sites of that TF in that condition (using a minimum value of 0.01 for each site-level value); the condition-wise average occupancies for each TF were then divided by the highest average occupancy for that TF across all conditions, yielding an occupancy score on the interval (0,1] for each TF-condition combination. The occupancy profiles of TFs across conditions were clustered 100 times using K-means clustering at each number of clusters between 8 and 12 (inclusive); the 'co-clustering frequency' κ is defined as the fraction of those 100 trials in which a given pair of TFs were assigned to the same cluster. We then used the quantity (1-κ) as a distance measure in a final hierarchical clustering, assigning the TFs to 10 clusters, to provide the cluster identities shown in Figure 3B.

RNA Sequencing and Differential Expression Calling

RNA-seq data sets were subjected to the same initial preprocessing and quality control steps as outlined above for the DNA samples, and then gene-level expression was quantified using kallisto v0.43 [63] on a version of the MG1655 (GenBank NC_000913) genome with all ribosomal RNAs removed. Gene-level transcript per million (TPM) values from kallisto were

used for all downstream analysis unless otherwise noted. To generate high-resolution occupancy plots, reads were instead aligned with bowtie2 as described above for DNA reads, and read occupancies quantified using the genomecov command of bedtools2 [64].

TFBS Comparison

Binding sites identified from IPOD-HR peaks (as described above in the "Feature calling" paragraph) were cross-referenced with known and predicted TFBSs using bedtools v2.17.0 [64]. "Known" sites comprise all binding sites contained in the RegulonDB release 9.4 BindingSetSet.txt file [7]; "predicted" sites are identified by scanning the MG1655 (GenBank NC_000913) genome with FIMO [65] using all *E. coli* position weight matrices from SwissRegulon [66] (as distributed by the MEME project); each PWM was applied separately, and all sites with a q-value less than 0.2 were retained. Default settings were used for FIMO, except that the background was a second-order Markov model based on the NC_000913 genome, and the number of maximum stored scores was set to 10,000,000.

Motif Identification

Novel sequence motifs implied by IPOD-HR data were identified using an inference pipeline built off of FIRE [67]. Occupancy peaks and associated discrete threshold scores in the IPOD-HR traces were called using the CWT-based approach described above, using a score threshold of 4; each peak was assigned a discrete score corresponding to the average IPOD-HR occupancy score within that peak, rounding down. We then generated a background distribution of unbound sequences drawn from the portion of the genome not included in peaks, matching the length distribution of the peaks but with three times as many locations; all such background regions were assigned a score of 0 to distinguish them from the various thresholded peak regions.

Motifs were called using FIRE with two separate variations: FIRE_gapped (with parameters `--kungapped=6 --gap=0-10 --jn_t_gapped=4 --minr=0.5`), which searches for gapped motifs typical of prokaryotic transcription factor binding sites; and FIRE_maxdeg (with parameters `--jn_t=8 --minr=1.5 --maxdeg=1.8`), which searches for motifs while preserving information content above a specified threshold. We applied additional empirical filters to specifically enrich for peaks corresponding to binding sites: all peaks identified via FIRE were required to have the motif significantly depleted from the background population ($p<0.01$). To assess the false discovery rate of our methods, we also generated 20 decoy peak sets by shuffling the locations of the real peaks observed in each condition, along with corresponding randomized unbound sets for each, and then applied identical peak calling procedures to each decoy set.

To avoid repeated reporting of very similar motifs which might be identified by our pipelines, we applied the matrix-clustering module of RSAT [68] (using recommended thresholds `-lth cor 0.7 -lth w 5 -lth Ncor 0.4`) to obtain non-redundant motif sets for downstream analysis. We compared all called motifs with previously known motifs from the SwissRegulon

database using TOMTOM [29] with default parameters, requiring an E-value of 0.5 or lower for 'Identified' hits, and labeling other identified motif matches as 'Ambiguous' (counted together with TFs lacking any hits). For the identification of predicted regulons associated with each motif, we applied the FIMO program [65] to identify potential binding sites on the *E. coli* K12 genome, with a q-value threshold of 0.2. For the purposes of our analysis of the potential regulatory networks of novel motifs), we marked each transcriptional unit in *E. coli* as being regulated by a particular motif if and only if a predicted binding site for that motif overlapped the gene's core promoter.

In vitro pulldown of unidentified transcription factors

In order to identify the protein(s) binding to the *sdaC* promoter (as in **Fig. 4**), we first prepared biotinylated bait DNA by cloning a fragment of the *sdaC* promoter (running from positions 2927790 to 2927975 in the U00096.3 genome) into a pAZ3-based cloning vector [69], and then amplifying that region of the plasmid using a primer pair where one primer contained a 5' biotinylation. The resulting 486 bp fragment was treated with Exonuclease I (Affymetrix) according to the manufacturer's instructions to remove unreacted primer, and then purified using a Zymo Clean & Concentrate 25 kit.

The biotinylated bait DNA was then bound to equilibrated Dynabeads MyOne Streptavidin C1 beads (Invitrogen). Beads were equilibrated by washing three times with 1x B&W buffer (5 mM Tris Cl, pH 7.5; 0.5 mM EDTA; 1 M NaCl) and then resuspended in five volumes of 2x B&W buffer, using 42 µL of the original resuspended bead solution per reaction. The equilibrated beads were combined with 8 µg of biotinylated bait DNA plus an appropriate volume of water to yield a final 1x B&W solution, and incubated 15 minutes at room temperature with gentle rocking to allow for bait binding. The beads were then washed three times with 500 µL of 1x B&W buffer, twice in 500 µL of 1x BMg/THS buffer (5 mM HEPES, pH 7.5; 5 mM MgCl2; 50 mM KCl; 31 mM NaCl; 1x cOmplete EDTA-free protease inhibitors (Roche)), once with 500 µL of 1x BMg/THS/EP buffer (1x BMg/THS buffer supplemented with 20 mM EGTA (pH 8.0) and 10 µg/mL poly d(IC) (Sigma)). The beads were then resuspended in 200 µL of BMg/THS/EP buffer and gently mixed by hand for one minute to complete equilibration.

Cell extracts were prepared by growing to an OD_{600} of 0.2 in M9/RDM/glucose media (following the same procedures as those given for IPOD-HR experiments). Once reaching the target OD, the cells were chilled 10 minutes on ice, and then pelleted by spinning for 10 minutes at 5,500xg while at 4° C. Supernatant was removed, and the cells were flash-frozen in a dry ice/ethanol bath. Cells were then lysed by resuspending the frozen pellet resulting from 82 mL of culture in 160 µL of B-PER II bacterial protein extraction reagent (Thermo Scientific). We then added 3.8 mL of 1x BMg/THS buffer and 0.8 µL of ReadyLyse lysozyme solution (Lucigen), 40 µL of 10 mg/mL RNase A, 20 µL of $CaCl_2$, and 200 µL of micrococcal nuclease (NEB; 2,000,000 gel units/mL). The lysis/digestion reaction was allowed to proceed for 30 minutes at room temperature, and then clarified by centrifugation at 30 minutes at 16,100x g held at 4° C. The reaction was halted by the addition of 444 µL of 5 M NaCl, and then the entire volume applied to

a 3 kDa MWCO spin filter (Amicon Ultra; Millipore) and centrifuged at (3,200 x g held at 4° C) until ~400 µL of retentate remained. We then added 3.6 mL of BMg/THS (lacking NaCl and KCl, but containing 5 mM CaCl2) and filtered to 400 µL of retentate. Retained liquid was then recovered, and diluted to a final volume of 4.0 mL with addition of salt-free BMg/THS + 5 mM $CaCl_2$. The retained lysate was then incubated 30 minutes at room temperature to permit further activity of micrococcal nuclease on remaining DNA in the sample, and then quenched with 168 µL of 500 mM EGTA. The volume of the sample was reduced to approximately 1.6 mL by ultrafiltration as above, and further supplemented with 10 µg/mL of poly d(IC) and 1 mM dithiothreitol.

Probing of the lysates was then accomplished by combining the equilibrated bait-bead complexes (described above) with the lysates, and incubating 30 minutes with rocking at room temperature. The supernatant was then removed, and the beads washed twice with 200 µL of BMg/THS/20 mM EGTA/10 µg/mL poly d(IC), and once with 200 µL of BMg/THS/20 mM EGTA. Proteins were then eluted from the beads through progressive washes of elution buffer (25 mM Tris HCl, pH 7.5) with 100 mM NaCl, 200 mM NaCl, 400 mM NaCl, and 1 M NaCl, with 50 µL used for each elution.

We successively probed the lysates described here with probes containing promoter sequences from *lexA*, *purR*, and finally, *sdaC* (each containing identical plasmid-derived flanking sequences). A ~25 kDa band of interest appeared in the 400 mM and 1 M NaCl *sdaC* eluates but not eluates from a parallel experiment performed under identical conditions with a segment of the *thiC* promoter; these bands were excised from a silver-stained gel. The 400 mM gel slice was then subjected to proteomic analysis at the University of Michigan Proteomics & Peptide Synthesis core facility. The gel slice was processed using a ProGest robot (DigiLab) to wash with 25 mM ammonium bicarbonate followed by acetonitrile, reduce with 10 mM dithiothreitol at 60° C followed by alkylation with 50 mM iodoacetamide at room temperature, digested with trypsin (Promega) at 37° C for 4 hours, and then quenched with formic acid. The digest was then analyzed by nano LC-MS/MS with a Waters NanoAcquity HPLC system interfaced to a ThermoFisher Q Exactive. Peptides were loaded on a trapping column and eluted over a 75 µm analytical column at 350 nL/min; both columns were packed with Jupiter Proteo resin (Phenomenex). The injection volume was 30 µL. The mass spectrometer was operated in data-dependent mode, with the Orbitrap operating at 60,000 FWHM and 17,500 FWHM for MS and MS/MS respectively. The fifteen most abundant ions were selected for MS/MS. Data were searched using a local copy of Mascot, and Mascot DAT files were parsed into the Scaffold software for validation, filtering and to create a non-redundant list per sample. Data were filtered using 90% protein and 95% peptide probability thresholds (Prophet scores) and requiring at least two unique peptides per protein). The resulting mass spectrometry analysis is given in Table S1 after manual pruning by core staff of common contaminants (e.g. human keratin).

Data Availability

The raw and processed sequencing data used in this study have been deposited in the Gene Expression Omnibus under accession GSE142291.

Acknowledgments

This work was supported by NIH grants R00-GM097033 and R35-GM128637 (to PLF) and R01-AI077562 (to ST). We are grateful to Dr. Alison Hottes for technical assistance and suggestions on the original IPOD method.

Author Contributions

Conceptualization, S.T.; Methodology, P.L.F. and S.T.; Investigation, P.L.F., T.J.G., and H.A.; Data Analysis and Curation, P.L.F. and S.T.; Writing -- Original Draft, P.L.F. and S.T.; Writing -- Review & Editing, P.L.F., H.A., and S.T.; Funding Acquisition, P.L.F. and S.T.

Declaration of Interests

The authors have no competing interests.

Supplementary Text

Supplementary Text S1: Effects of RNA polymerase on large-scale protein occupancy

In the IPOD-HR method, we address the contribution of RNA polymerase occupancy to the overall protein occupancy signal through two steps: first, immediately prior to crosslinking, cells are treated with rifampicin for 10 minutes. We have calibrated the rifampicin treatment to allow sufficient time for in-progress transcripts to finish while minimizing perturbation of cellular physiology (building on data from [70], as well as direct calculation of the required time based on the elongation rate of RNA polymerase and lengths of *E. coli* transcripts). As rifampicin inhibits promoter clearance but not the transcriptional initiation nor completion of already-elongating transcripts [71], the result will thus be to cause an accumulation of RNA polymerase at active promoters, while clearing the majority of polymerase occupancy from gene bodies, and thus substantially simplifying the identification of RNA polymerase occupancy. In addition, all IPOD-HR experiments described here were performed in parallel with RNA polymerase ChIP-seq experiments under the same conditions, permitting calibrated subtraction of RNA polymerase occupancy from active promoters to reveal changes in regulatory protein occupancy (as schematized in **Figure 1B-C**). However, the fact that transcriptional inhibition is known to affect nucleoid condensation [72–75] prompted us to directly inspect the effects of rifampicin on genome-wide protein binding, with a particular emphasis on the effects on EPOD formation and transcription factor binding.

To illustrate the effects of rifampicin treatment on regions of extremely high or low RNA polymerase occupancy, we performed IPOD experiments under our baseline growth condition (WT cells undergoing exponential growth in M9/RDM/glu medium) following the same procedure as used for all other samples in the present study, but omitting the rifampicin treatment (*n.b.* the data sets of [10] were obtained under a slightly different condition, during exponential growth in LB medium). An overall analysis of the resulting EPOD calls (**Fig. S5**) demonstrates that the -RIF EPOD set arising from our methods show good correlations with both the heEPOD and tsEPOD calls from [10], whereas the +RIF EPOD set from our method (used throughout the rest of text) shows a much stronger correlation with the Vora tsEPOD set and a weaker correlation with heEPODs. It is also important to note that the fraction of locations from the Vora tsEPOD set that is contained within our relaxed +RIF EPOD calls (0.90) is in line with the equivalent overlaps between the EPOD sets observed across different conditions in our main data sets (lower triangle of **Fig. 6D**), further indicating that the EPODs identified via IPOD-HR closely resemble the original Vora tsEPODs.

To obtain a more detailed picture of the effects of rifampicin on protein occupancy in our assays, we show an example of a highly expressed EPOD (heEPOD) from the original IPOD data sets in **Figure S6**. In the original IPOD data set [10], strong protein occupancy was noted throughout the cluster of ribosomal protein operons running from *rplQ* to *rpsJ*. Here, we see from the rifampicin-omitted (-RIF) samples that the protein occupancy profile in this region is dominated by RNA polymerase occupancy, closely matching the bounds of the originally called heEPOD.

On the other hand, in the rifampicin-treated samples (+RIF), the vast majority of occupancy in this region is lost, with peaks only apparent at a few points within the region of interest (likely corresponding to highly active promoters). After subtraction of the scaled RNA polymerase occupancy to yield the IPOD-HR signal, the only prominent peak remaining in the region is in the *gspA-gspC* intergenic region, which has been demonstrated to be repressed by H-NS binding (although the precise binding location was previously unknown [76]).

The occupancy observed in the ribosomal protein operon cluster described above contrast strongly with that seen in the transcriptionally silent EPODs (tsEPODs) found in [10]. In the region of the tsEPOD shown in **Figure S7**, for example, we observe that in the -RIF samples, there is continuous protein occupancy but essentially no RNA polymerase occupancy throughout the large tsEPOD that was originally observed to span the *waaQGPSBOJYZU* operon, whereas both strong IPOD occupancy and strong RNA polymerase occupancy are apparent on a nearby heEPOD covering the *rpmBG* operon. Treatment with rifampicin does not substantially alter the high level of overall protein occupancy throughout the *waa* region, whereas it restricts occupancy near *rpmBG* to active promoters only. As a result, in the ChIP-subtract IPOD-HR signal, strong occupancy remains throughout the *waaQGPSBOJYZU* operon, resulting in an EPOD call nearly identical to the original tsEPOD in that region from [10]. One feature of the +RIF samples that requires consideration is the fact that several RNA polymerase occupancy peaks appear in regions such as tsEPODs where no comparable occupancy is apparent in -RIF samples. We attribute these additional peaks to the fact that during rifampicin treatments, concentrations of free RNA polymerase will rise substantially due to the immobilization of polymerase at transcription start sites; thus, occupancy at normally weak promoters through the chromosome will increase. We find, however, that the ChIP-subtraction step of our IPOD-HR data processing pipeline accurately removes RNA polymerase occupancy at both normally-active promoters and those showing RNA polymerase binding only in the presence of rifampicin treatment (as seen by the well-calibrated removal of RNA polymerase occupancy in the +RIF tracks of **Figures S6-S7**, for example). Overall, we find that even in the +RIF samples, RNA polymerase occupancy remains very well correlated with transcript levels (**Figure 3D**), can be cleanly subtracted to yield condition-appropriate changes in transcription factor occupancy (**Figure 1C**); furthermore, brief rifampicin treatment does not appear to substantively alter large-scale protein occupancy other than that directly attributable to RNA polymerase binding (**Figures S5-S7**). Given that rifampicin treatment permits cleaner subtraction of RNA polymerase occupancy, while not perturbing either the local or large-scale protein binding patterns that are the subject of our interest, we make use of it throughout the IPOD-HR data sets shown in the present work.

Supplementary Figures

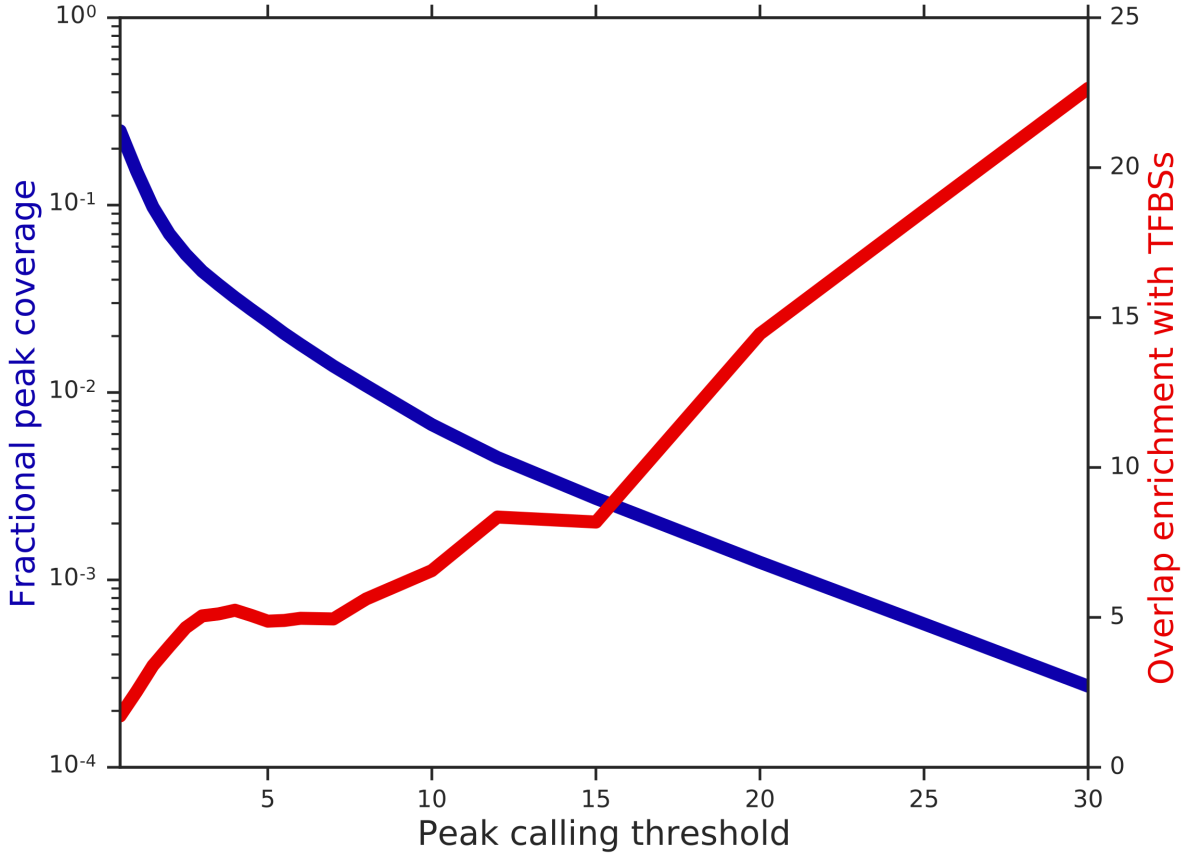


Figure S1: Effect of peak calling threshold on coverage and enrichment of known transcription factor binding sites (TFBSs). Data are shown for the wild type cells in the rich defined medium condition. Shown is the fraction of the entire genome contained in peak calls (left vertical axis, blue line) or the enrichment of TFBSs overlapping those peak calls relative to that expected by chance (right vertical axis, red line). Overlaps at all shown thresholds were statistically significant ($p<0.01$, permutation test in each case).

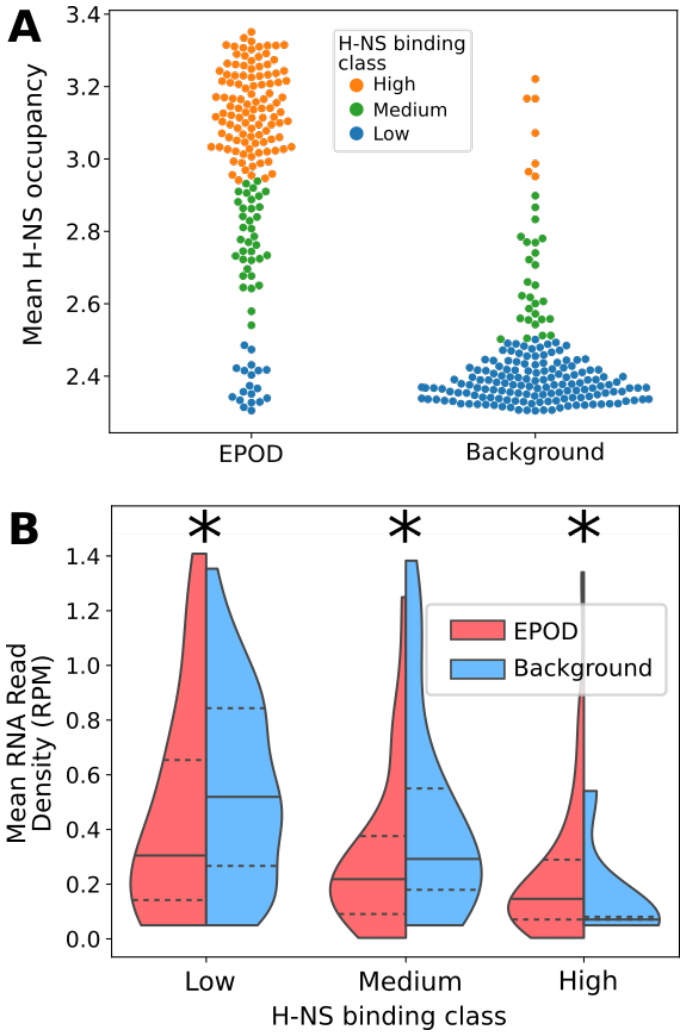


Figure S3: Interplay of H-NS occupancy, EPOD locations, and transcription. (A) Mean levels of H-NS binding (data from [36]) for all EPODs called in the WT rich media condition; each point shows either an EPOD or a single contiguous non-EPOD region. Each point is colored by its classification into high, medium, or low H-NS binding using a Gaussian mixture model with three groups, after removal of outliers using the local outlier factor [77] as implemented in the python scikit-learn module [78], using 25 neighbors and default settings for other parameters. **(B)** Distributions of mean RNA read density stratified by the H-NS binding categories shown in panel **A**, with each case divided by EPOD status. The median of each group is shown by a dashed line, and the 25th and 75th quartiles by dotted lines. '*' indicates a significant difference between the medians of the EPOD vs. background groups ($p<0.05$, Mann-Whitney U test).

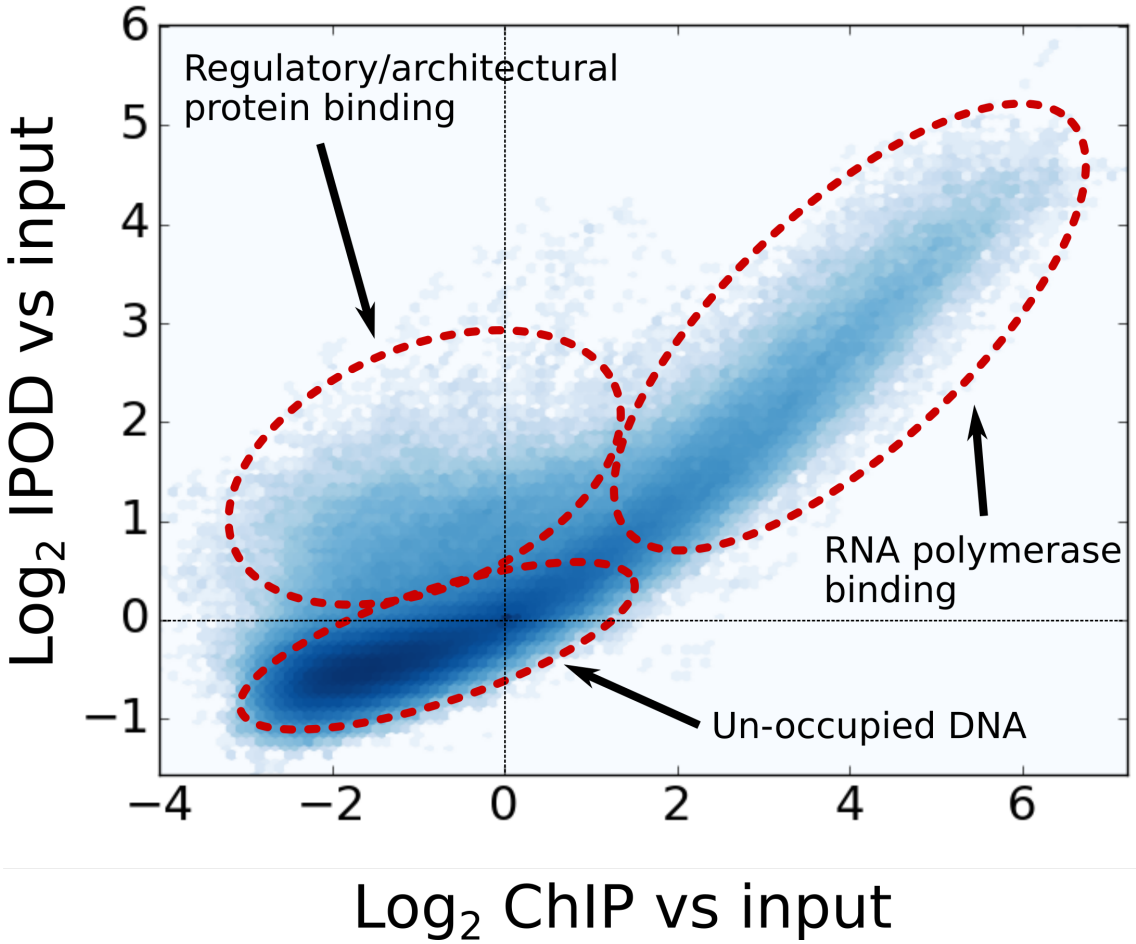


Figure S4: Identification of RNA polymerase vs. non-RNA polymerase protein occupancy. Shown is a density plot of the $\log_2$(IPOD/Input) signal vs. $\log_2$(RNA polymerase ChIP/Input) signal, demonstrating the presence of three subpopulations of genomic positions: unbound positions (without enrichment using either protein occupancy profiling method), RNA polymerase occupancy (part of a highly correlated region of high IPOD occupancy and high RNA polymerase occupancy), and occupancy with other proteins (which shows high IPOD occupancy but low RNA polymerase occupancy). Note that there is no corresponding population of high RNA polymerase occupancy but low IPOD occupancy; rather, the RNA polymerase-bound regions are a subset of the regions detected by IPOD.

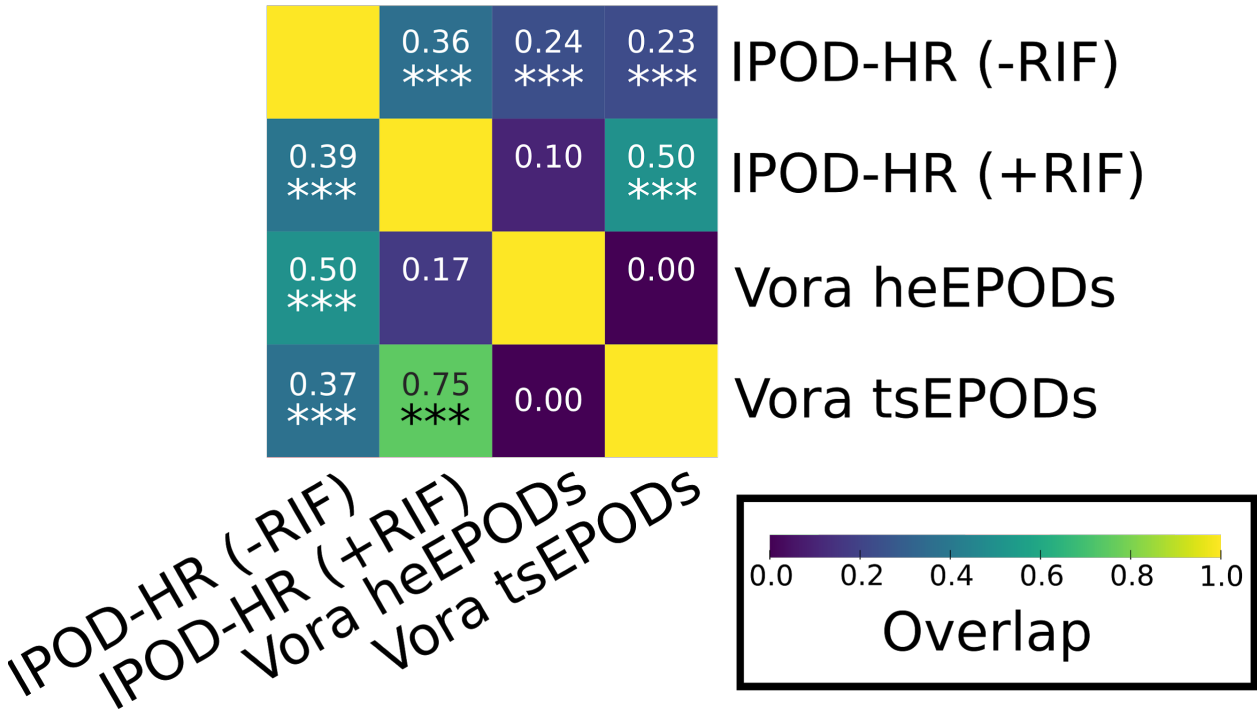
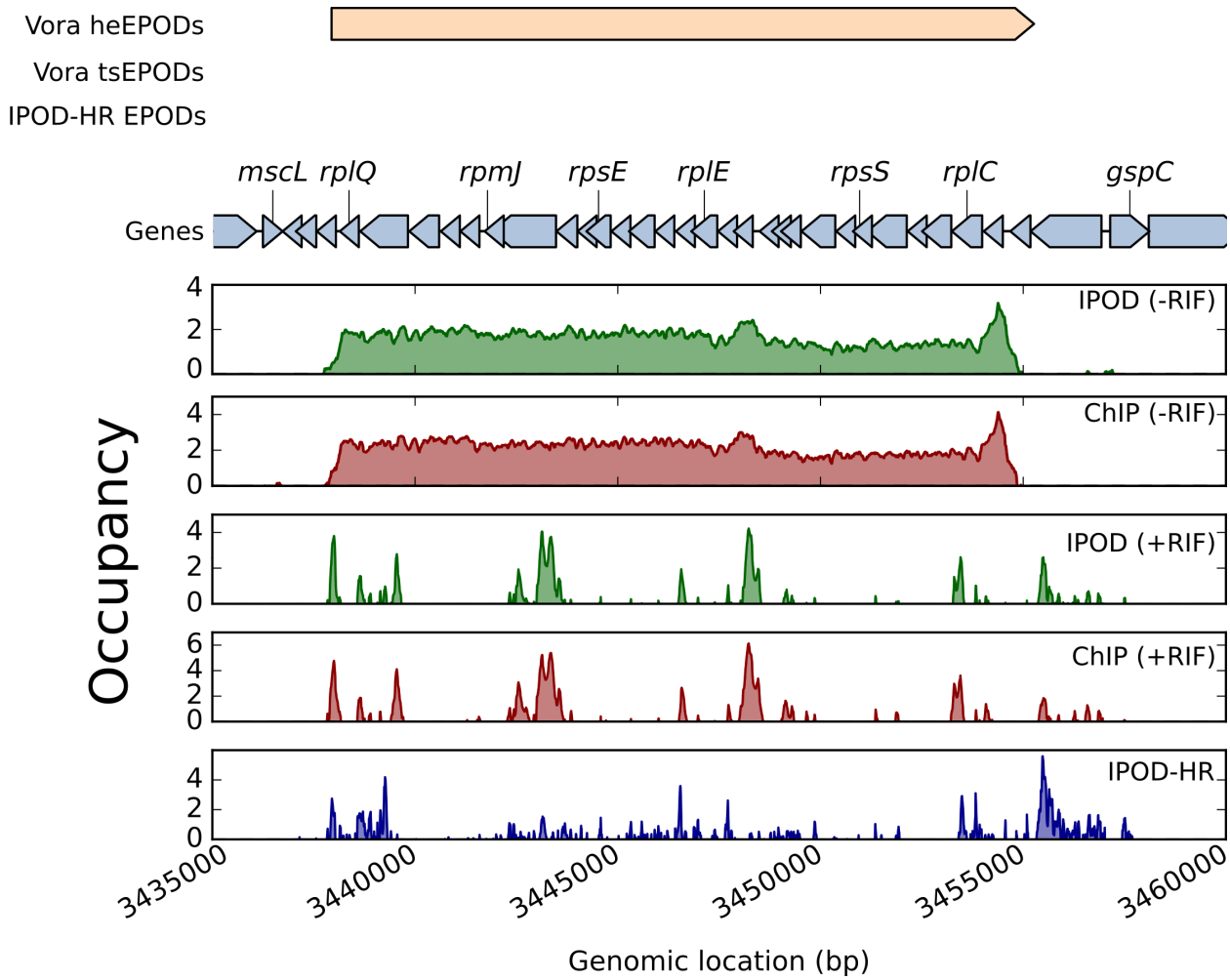


Figure S5: Overlaps of EPOD sets resulting from different calling methods. Shown in the heatmap are the fraction of EPODs from the EPOD set defined by the row label that overlap the EPOD set defined by the column label. Asterisks reflect p-values arising from a Monte Carlo permutation test (1000 random circular permutations; * p<0.05, ** p<0.01, *** p<0.001). p-values for the overlaps between the +RIF IPOD-HR EPOD set and the Vora heEPODs were >0.8 for both directions of comparisons; a full list of values is given in **Table S4**.



1280 **Figure S6: Effects of rifampicin treatment on protein occupancy of a highly transcribed region.** Shown are occupancy signals for interphase-extracted, RNA polymerase ChIP, and ChIP-subtracted IPOD occupancy (IPOD-HR) samples in the vicinity of a large cluster of ribosomal protein genes (running from *rplQ* to *rpsJ*). Signals are log2 extracted:input ratios (for IPOD and ChIP samples), or ChIP-subtracted robust z scores (IPOD-HR).

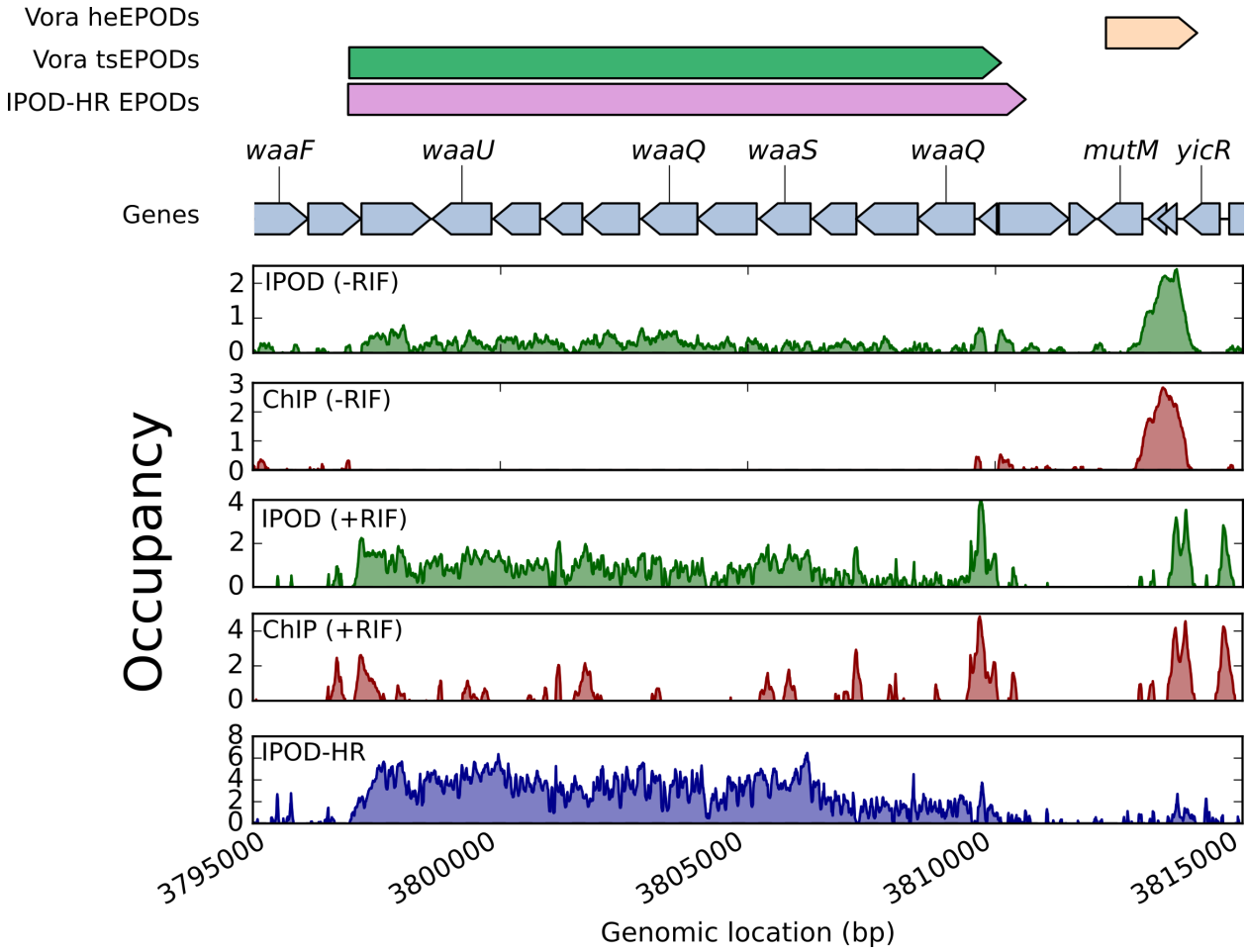


Figure S7: Effects of rifampicin treatment on protein occupancy of a transcriptionally silent region. Shown are occupancy signals for interphase-extracted, RNA polymerase ChIP, and ChIP-subtracted IPOD occupancy (IPOD-HR) samples in the vicinity of the *waaQGPSBOJYZU* operon, which was identified as a strong tsEPOD in [10]. Signals are log2 extracted:input ratios (for IPOD and ChIP samples), or ChIP-subtracted robust z scores (IPOD-HR).

Supplementary Tables and Data Files

1295 **Table S1:** Mass spectrometry identified peptide counts showing abundances of proteins pulled down by biotinylated bait DNA from the *sdaC* promoter region, after pruning of likely contaminants (see Methods for details).

Identified protein	Accession Number	Molecular Weight	Total Spectrum Count
UlaR	tr|C3SFV2|C3SFV2_ECOLX	28 kDa	16
YieP	tr|E2QHU3|E2QHU3_ECOLX	26 kDa	9
RpsC	tr|C3SQX2|C3SQX2_ECOLX	26 kDa	7
RpoC	tr|C3SIA2|C3SIA2_ECOLX	155 kDa	11
RpoB	tr|E2QJ13|E2QJ13_ECOLX	151 kDa	7
RuvA	tr|C3T5R2|C3T5R2_ECOLX	22 kDa	10
FliA	tr|C3SDE6|C3SDE6_ECOLX	28 kDa	6
RpoA	tr|C3SR67|C3SR67_ECOLX	37 kDa	7
RpsD	tr|C3SR62|C3SR62_ECOLX	23 kDa	3
RdgC	tr|E2QGD8|E2QGD8_ECOLX	34 kDa	3
UvrA	tr|C3SHF7|C3SHF7_ECOLX	104 kDa	5
CysB	tr|C3TC57|C3TC57_ECOLX	36 kDa	5
FabR	tr|E2QIZ8|E2QIZ8_ECOLX	24 kDa	4
GroL	tr|Q548M1|Q548M1_ECOLX	57 kDa	3
IhfB	tr|Q14F22|Q14F22_ECOLX	11 kDa	4
RplC	tr|C3SQU2|C3SQU2_ECOLX	22 kDa	2
AccB	tr|C3SRL7|C3SRL7_ECOLX	17 kDa	4
AmiA	tr|E2QPR8|E2QPR8_ECOLX	31 kDa	3
IhfA	tr|Q14F23|Q14F23_ECOLX	11 kDa	2
Ppx	tr|C3T027|C3T027_ECOLX	58 kDa	2
RplB	tr|C3SQV7|C3SQV7_ECOLX	30 kDa	2
RpoE	tr|Q0P6M2|Q0P6M2_ECOLX	22 kDa	2
FabZ	tr|C3TPH7|C3TPH7_ECOLX	17 kDa	3
PolA	tr|E2QI51|E2QI51_ECOLX	103 kDa	2
TufB	tr|E2QFJ4|E2QFJ4_ECOLX (+1)	43 kDa	2

Table S2: Summary of EPOD characteristics across experimental conditions. The "Median difference" column refers to the difference in median robust Z-scores between EPODs and all other sites in the genome, with positive values indicating higher levels within EPODs. P-values for a significant difference are obtained using a resampling test, with 1000 random circular permutations of the EPOD locations on the genome (thus preserving the correlation structure of genomic features); q-values are obtained by correction of the p-values using the Benjamini-Hochberg method [79]. All sequence features were subjected to a 500 bp rolling mean prior to overlap calculation.

Feature	Median difference	p value	q value	Reference
Transcriptional propensity	-1.0347	0.001	0.001	[30]
Supercoiling density	-0.1259	0.766	0.766	[80]
Normalized Tn5 integration density	4.5285	0.001	0.001	[30]
Fis binding	-1.0107	0.001	0.001	[36]
H-NS binding	11.8366	0.001	0.001	[36]
HU binding	-0.1193	0.042	0.048	[81]
LRP binding	0.3464	0.268	0.286	[59]
SeqA binding	-0.2629	0.003	0.004	[82]
Dam sites	-0.6745	0.001	0.001	Sequence analysis
AT content	1.8548	0.001	0.001	Sequence analysis
RNA	-0.6668	0.001	0.001	Present study
MGW	-0.0843	0.001	0.001	Calculated using DNAshapeR [83]
HelT	0.0707	0.001	0.001	
ProT	-0.1692	0.001	0.001	
Roll	-0.0910	0.001	0.001	

Table S3: List of IPOD-HR experiments performed over the course of the study, including concise names used to refer to each experiment in the remainder of the text. (included as a separate file). Note that "Aligned reads" refers specifically to the number of concordant, uniquely aligned read pairs arising from a given sample, or the number of pseudoaligned reads in the case of an RNA sample.

Table S4: P-values arising from the statistical tests shown in **Figure 3B** (calculated using iPAGE software), and in **Figure S3** (arising from permutation tests in which the genomic coordinates of one EPOD set were rotated in unison relative to the other, to preserve the internal correlation structures of each group).

Supplementary Data S1: GFF file containing the locations of all peak calls obtained from our IPOD-HR occupancy profiles through all conditions in the present study. The value in the 'score' column corresponds to the peak calling threshold in use. Note that the 'strand' and 'frame' fields convey no useful information.

Supplementary Data S2: MEME-formatted file containing the complete set of redundancy-pruned motifs discovered across all conditions in the present study.

Supplementary Data S3: GFF file containing the locations of all EPODs called on our IPOD-HR occupancy profiles (see Methods for details). Note that the 'score', 'strand', and 'frame' fields convey no useful information.

References

1. Browning DF, Butala M, Busby SJW. Bacterial Transcription Factors: Regulation by Pick "N" Mix. J Mol Biol. 2019. doi:10.1016/j.jmb.2019.04.011

2. Lee DJ, Minchin SD, Busby SJW. Activating transcription in bacteria. Annu Rev Microbiol. 2012;66: 125–152.

3. Gruber TM, Gross CA. Multiple sigma subunits and the partitioning of bacterial transcription space. Annu Rev Microbiol. 2003;57: 441–466.

4. Murakami KS, Darst SA. Bacterial RNA polymerases: the wholo story. Curr Opin Struct Biol. 2003;13: 31–39.

5. Shen BA, Landick R. Transcription of Bacterial Chromatin. J Mol Biol. 2019. doi:10.1016/j.jmb.2019.05.041

6. Ishihama A, Shimada T, Yamazaki Y. Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors. Nucleic Acids Res. 2016;44: 2058–2074.

7. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016;44: D133–43.

8. Larsen SJ, Röttger R, Schmidt HHHW, Baumbach J. E. coli gene regulatory networks are inconsistent with gene expression data. Nucleic Acids Res. 2019;47: 85–92.

9. Mejía-Almonte C, Busby SJW, Wade JT, van Helden J, Arkin AP, Stormo GD, et al. Redefining fundamental concepts of transcription initiation in bacteria. Nat Rev Genet. 2020. doi:10.1038/s41576-020-0254-8

10. Vora T, Hottes AK, Tavazoie S. Protein occupancy landscape of a bacterial genome. Mol Cell. 2009;35: 247–253.

11. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 2007;17: 877–885.

12. Meng LM, Nygaard P. Identification of hypoxanthine and guanine as the co-repressors for the purine regulon genes of Escherichia coli. Mol Microbiol. 1990;4: 2187–2192.

13. Cho B-K, Federowicz SA, Embree M, Park Y-S, Kim D, Palsson BØ. The PurR regulon in Escherichia coli K-12 MG1655. Nucleic Acids Res. 2011;39: 6456–6464.

14. Tian G, Lim D, Carey J, Maas WK. Binding of the arginine repressor of Escherichia coli K12 to its operator sites. J Mol Biol. 1992;226: 387–397.

15. Charlier D, Roovers M, Van Vliet F, Boyen A, Cunin R, Nakamura Y, et al. Arginine regulon of Escherichia coli K-12. A study of repressor-operator interactions and of in vitro binding

affinities versus in vivo repression. J Mol Biol. 1992;226: 367–386.

16. Van Duyne GD, Ghosh G, Maas WK, Sigler PB. Structure of the oligomerization and L-arginine binding domain of the arginine repressor of Escherichia coli. J Mol Biol. 1996;256: 377–391.

17. Rolfes RJ, Zalkin H. Purification of the Escherichia coli purine regulon repressor and identification of corepressors. J Bacteriol. 1990;172: 5637–5642.

18. Pittard AJ, Davidson BE. TyrR protein of Escherichia coli and its role as repressor and activator. Mol Microbiol. 1991;5: 1585–1592.

19. Shimada T, Hirao K, Kori A, Yamamoto K, Ishihama A. RutR is the uracil/thymine-sensing master regulator of a set of genes for synthesis and degradation of pyrimidines. Mol Microbiol. 2007;66: 744–757.

20. Urbanowski ML, Stauffer GV. Role of homocysteine in metR-mediated activation of the metE and metH genes in Salmonella typhimurium and Escherichia coli. J Bacteriol. 1989;171: 3277–3281.

21. Soda K. Microbial sulfur amino acids: an overview. Methods Enzymol. 1987;143: 453–459.

22. Cai XY, Redfield B, Maxon M, Weissbach H, Brot N. The effect of homocysteine on MetR regulation of metE, metR and metH expression in vitro. Biochem Biophys Res Commun. 1989;163: 79–83.

23. Shimada T, Ogasawara H, Ishihama A. Genomic SELEX Screening of Regulatory Targets of Escherichia coli Transcription Factors. Methods Mol Biol. 2018;1837: 49–69.

24. Belliveau NM, Barnes SL, Ireland WT, Jones DL, Sweredoski MJ, Moradian A, et al. Systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria. Proc Natl Acad Sci U S A. 2018;115: E4796–E4805.

25. Gao Y, Yurkovich JT, Seo SW, Kabimoldayev I, Dräger A, Chen K, et al. Systematic discovery of uncharacterized transcription factors in Escherichia coli K-12 MG1655. Nucleic Acids Res. 2018;46: 10682–10696.

26. Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, et al. The Escherichia coli transcriptome mostly consists of independently regulated modules. Nat Commun. 2019;10: 5536.

27. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 2006;34: D187–91.

28. Elemento O, Slonim N, Tavazoie S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. Molecular Cell. 2007. pp. 337–350. doi:10.1016/j.molcel.2007.09.027

29. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8: R24.

30. Scholz SA, Diao R, Wolfe MB, Fivenson EM, Lin XN, Freddolino PL. High-Resolution

Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. Cell Syst. 2019;8: 212–225.e9.

31. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109: 21.29.1–9.

32. Whitfield CR, Wardle SJ, Haniford DB. The global bacterial regulator H-NS promotes transpososome formation and transposition in the Tn5 system. Nucleic Acids Res. 2009;37: 309–321.

33. Tavazoie S, Church GM. Quantitative whole-genome analysis of DNA-protein interactions by in vivo methylase protection in E. coli. Nat Biotechnol. 1998;16: 566–571.

34. Navarre WW, McClelland M, Libby SJ, Fang FC. Silencing of xenogeneic DNA by H-NS-facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. Genes Dev. 2007;21: 1456–1471.

35. Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, et al. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. Science. 2006;313: 236–238.

36. Kahramanoglou C, Seshasayee ASN, Prieto AI, Ibberson D, Schmidt S, Zimmermann J, et al. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. Nucleic Acids Research. 2011. pp. 2073–2091. doi:10.1093/nar/gkq934

37. Ling G, Waxman DJ. DNase I Digestion of Isolated Nulcei for Genome-Wide Mapping of DNase Hypersensitivity Sites in Chromatin. Methods in Molecular Biology. 2013. pp. 21–33. doi:10.1007/978-1-62703-284-1_3

38. Mieczkowski J, Cook A, Bowman SK, Mueller B, Alver BH, Kundu S, et al. MNase titration reveals differences between nucleosome occupancy and chromatin accessibility. Nat Commun. 2016;7: 11485.

39. Fang X, Sastry A, Mih N, Kim D, Tan J, Yurkovich JT, et al. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. Proc Natl Acad Sci U S A. 2017;114: 10286–10291.

40. Singh SS, Singh N, Bonocora RP, Fitzgerald DM, Wade JT, Grainger DC. Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev. 2014;28: 214–219.

41. Kotlajich MV, Hron DR, Boudreau BA, Sun Z, Lyubchenko YL, Landick R. Bridged filaments of histone-like nucleoid structuring protein pause RNA polymerase and aid termination in bacteria. Elife. 2015;4. doi:10.7554/eLife.04970

42. Landick R, Wade JT, Grainger DC. H-NS and RNA polymerase: a love--hate relationship? Curr Opin Microbiol. 2015;24: 53–59.

43. Atlung T, Ingmer H. H-NS: a modulator of environmentally regulated gene expression. Mol Microbiol. 1997;24: 7–17.

44. Ono S, Goldberg MD, Olsson T, Esposito D, Hinton JCD, Ladbury JE. H-NS is a part of a thermally controlled mechanism for bacterial gene regulation. Biochem J. 2005;391: 203–213.

45. Goodarzi H, Elemento O, Tavazoie S. Revealing global regulatory perturbations across human cancers. Mol Cell. 2009;36: 900–911.

46. Freddolino PL, Amini S, Tavazoie S. Newly identified genetic variations in common Escherichia coli MG1655 stock cultures. J Bacteriol. 2012;194: 303–306.

47. Ausubel FM. Current Protocols in Molecular Biology. John Wiley & Sons; 1998.

48. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2: 2006.0008.

49. Cherepanov PP, Wackernagel W. Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. Gene. 1995;158: 9–14.

50. Huisman O, D'Ari R. An inducible DNA replication–cell division coupling mechanism in E. coli. Nature. 1981;290: 797–799.

51. Huisman O, D'Ari R, Gottesman S. Cell-division control in Escherichia coli: specific induction of the SOS function SfiA protein is sufficient to block septation. Proc Natl Acad Sci U S A. 1984;81: 4490–4494.

52. Neidhardt FC, Bloch PL, Smith DF. Culture medium for enterobacteria. J Bacteriol. 1974;119: 736–747.

53. Mooney RA, Davis SE, Peters JM, Rowland JL, Ansari AZ, Landick R. Regulator trafficking on bacterial transcription units in vivo. Mol Cell. 2009;33: 97–108.

54. Tange O, Others. Gnu parallel-the command-line power tool. The USENIX Magazine. 2011;36: 42–47.

55. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.

56. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

57. Andrews S, Others. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.

58. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048.

59. Kroner GM, Wolfe MB, Freddolino PL. Escherichia coli Lrp regulates one-third of the genome via direct, cooperative, and indirect routes. J Bacteriol. 2018. doi:10.1128/JB.00411-18

60. Cleveland WS, Devlin SJ. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. J Am Stat Assoc. 1988;83: 596–610.

61. Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics. 2006;22: 2059–

2065.

62. Slonim N, Atwal GS, Tkacik G, Bialek W. Information-based clustering. Proc Natl Acad Sci U S A. 2005;102: 18297–18302.

63. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017;14: 687–690.

64. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics. 2014;47: 11–12.

65. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017–1018.

66. Pachkov M, Balwierz PJ, Arnold P, Ozonov E, van Nimwegen E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. Nucleic Acids Res. 2013;41: D214–20.

67. Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol. 2005;6: R18.

68. Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M, van Helden J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. 2017;45: e119.

69. Kawano M, Aravind áL, Storz G. An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin. Mol Microbiol. 2007;64: 738–754.

70. Herring CD, Raffaelle M, Allen TE, Kanin EI, Landick R, Ansari AZ, et al. Immobilization of Escherichia coli RNA polymerase and location of binding sites by use of chromatin immunoprecipitation and microarrays. J Bacteriol. 2005;187: 6166–6174.

71. Campbell EA, Korzheva N, Mustaev A, Murakami K, Nair S, Goldfarb A, et al. Structural mechanism for rifampicin inhibition of bacterial rna polymerase. Cell. 2001;104: 901–912.

72. Cabrera JE, Cagliero C, Quan S, Squires CL, Jin DJ. Active transcription of rRNA operons condenses the nucleoid in Escherichia coli: examining the effect of transcription on nucleoid structure in the absence of transertion. J Bacteriol. 2009;191: 4180–4185.

73. Pettijohn DE, Hecht R. RNA molecules bound to the folded bacterial genome stabilize DNA folds and segregate domains of supercoiling. Cold Spring Harb Symp Quant Biol. 1974;38: 31–41.

74. Cabrera JE, Jin DJ. The distribution of RNA polymerase in Escherichia coli is dynamic and sensitive to environmental cues. Mol Microbiol. 2003;50: 1493–1505.

75. Dworsky P, Schaechter M. Effect of rifampin on the structure and membrane attachment of the nucleoid of Escherichia coli. J Bacteriol. 1973;116: 1364–1374.

76. Francetic O, Belin D, Badaut C, Pugsley AP. Expression of the endogenous type II secretion pathway in Escherichia coli leads to chitinase secretion. EMBO J. 2000;19: 6697–6703.

77. Breunig MM, Kriegel H-P, Ng RT, Sander J. LOF: identifying density-based local outliers. Proceedings of the 2000 ACM SIGMOD international conference on Management of data. New York, NY, USA: Association for Computing Machinery; 2000. pp. 93–104.

78. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12: 2825–2830.

79. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Series B Stat Methodol. 1995;57: 289–300.

80. Lal A, Dhar A, Trostel A, Kouzine F, Seshasayee ASN, Adhya S. Genome scale patterns of supercoiling in a bacterial chromosome. Nat Commun. 2016;7: 11055.

81. Prieto AI, Kahramanoglou C, Ali RM, Fraser GM, Seshasayee ASN, Luscombe NM. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. Nucleic Acids Res. 2012;40: 3524–3537.

82. Joshi MC, Magnan D, Montminy TP, Lies M, Stepankiw N, Bates D. Regulation of sister chromosome cohesion by the replication fork tracking protein SeqA. PLoS Genet. 2013;9: e1003673.

83. Chiu T-P, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2016;32: 1211–1213.