Differential Roles of positive and negative supercoiling in compacting the *E. coli* genome

Ziqi Fu[1], Monica S. Guo[2#], Weiqiang Zhou[1#], Jie Xiao[3#]

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St., Baltimore, MD 21205
2. Department of Microbiology, University of Washington School of Medicine, 959 NE Pacific St., Seattle, WA 98195
3. Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, 725 N. Wolfe St., Baltimore, MD 21205, U.S.A.

[#]: Correspondences should be addressed to, msguo@uw.edu, wzhou14@jhu.edu, and xiao@jhmi.edu

Abstract

This study aims to explore whether and how positive and negative supercoiling contribute to the three-dimensional (3D) spatial organization of the bacterial genome. We used recently published GapR ChIP-seq and Topo-seq data, which marks positive and negative supercoiling sites, respectively, to study how GapR and TopoI binding sites correlate with the corresponding contact frequencies obtained from the Hi-C data of the *E. coli* chromosome. We found that chromosomal loci enriched with GapR or TopoI binding sites have overall higher Hi-C contact frequencies than sites that do not contain the corresponding binding sites, with GapR peaks corresponding to higher spatial contacts than TopoI peaks. Additionally, we found that Hi-C contact frequencies alone could identify GapR, but not TopoI, binding sites with high accuracy. Finally, we found that majority of positive and negative supercoils coincides with highly active transcription units, with a minor group likely associated with replication. Our results suggest that both positive and negative supercoiling enhance chromosomal interactions, but the positive supercoils contribute significantly more than the negative supercoils to bring distant chromosomal loci closer in space.

Introduction

All living cells have DNA, which instructs the cell to develop, grow, and reproduce. Genomic DNA topologies are known to be a key contributing factor to its spatial organization and gene regulation [1]. Unlike in eukaryotes, where the nuclear envelope surrounds the nucleus and histones organize the chromosomes, bacterial DNA packaging is mainly achieved by DNA supercoiling and the binding of nucleoid-associated proteins [2], as genetic materials in bacteria are not bounded by a nuclear membrane. The existence of supercoiling in the *Escherichia coli* DNA has long been described [3], and it is known that supercoils are associated with transcription [4]. Furthermore, supercoiling can be categorized as negative or positive, depending on whether it twists in the opposite or in the same direction of the double helix, and both types of supercoiling form coils (twists) or plectonemes (writhes) [5]. However, the location and the distribution of supercoiling along the *E. coli* genome remained unknown until the recent release of GapR-seq [6 7 8] and EcTopoI-seq [9]. These two high-resolution chromatin immunoprecipitating sequencing (ChIP-seq) methods probe the genomic location at which *Caulobacter crescentus* GapR and *E. coli* Topoisomerase I preferentially bind, respectively. It has been shown that GapR has a high affinity only for positively supercoiled DNA, making it a potential sensor for positive supercoils in *E. coli* [6]. Similarly, EcTopoI (TopoI) binds and removes excessive negative supercoils [10], hence its binding sites reveal the genomic regions that are negatively supercoiled. Recently, advances in chromosome conformation capture based technologies such as Hi-C allow one to interrogate genome organization by measuring the physical interaction frequency of two DNA fragments in the nuclear space with high resolutions [7 8 11 12]. Despite providing an overall genomic structure of the chromosome and offering topological insights such as DNA looping and topological associated domains (TAD), Hi-C experiments typically have resolution of 0.5-5 Kb and do not reveal precise locations of supercoiling, which is thought to take place on shorter length scales [13]. Therefore, GapR-seq and EcTopoI-seq (TopoI-seq) augment our understanding of the *E. coli* genome topology.

In this study, we integrate GapR-seq and TopoI-seq data with Hi-C data to investigate how supercoiling contributes to the *E. coli* genome spatial organization using traditional statistical methods and modern machine-learning approaches. Our statistical analyses show that genomic loci containing supercoiling correlate with high spatial interactions. Most importantly, we find that positive supercoiling corresponds to higher spatial contact than negative supercoiling. Both findings have not been previously reported. Additionally, using a machine learning model, we find that Hi-C data can be used to identify the locations of positive, but not

negative, supercoiling accurately. Finally, to explore the relationship between supercoiling distribution and transcription activities, we integrate GapR-seq and EcTopoI-seq data with RNA-seq data. We found that most supercoiling sites colocalize with active transcription units, with a minor group likely associated with replication. Our results illustrate the prevalence of positive supercoiling and important role of transcription in organizing the genome.

Results

GapR-binding sites correlate with higher-than-average Hi-C contact frequencies in *E. coli*

To examine whether positive supercoiling contributes to enhanced chromosomal interactions, we compared Hi-C contact frequencies of genomic loci that do or do not contain the GapR-binding sites. Using the *E. coli* GapR-seq data (LB, 37°C) [6], we identified 609 reproducible GapR-binding peaks (in **Methods**, ChIP-seq data processing and analysis) along the *E. coli* genome. Because GapR preferentially binds to overtwisted double-stranded DNA (dsDNA), these peak positions correspond to positively supercoiled sites [6 14]. We then divided the *E. coli* genome of 4.6 million base pairs (Mbp) into either 928 (5-Kbp per bin) or 9284 (0.5-Kbp per bin) bins based on the maximum resolution of two publicly available 5C and Hi-C datasets (referred to as Hi-C data in the paragraphs below)[7 8], in addition to one independent Hi-C experiment we performed in this study. Each 5-Kbp bin is assigned a tag, gP (GapR-peak) or gNP (non-GapR-peak), depending on whether it contains at least one of the GapR ChIP-seq peak centers or not, and each 0.5-Kbp bin follows similar assignment depends on whether it overlaps with any GapR ChIP-seq peaks (see **Supplement Fig. S1B, Table 1** for summary). When we aligned GapR peak-containing bins (black lines) alongside with the corresponding Hi-C heatmaps of the *E. coli* genome [7], we observed that genomic locations containing a large number of GapR peaks have an overall higher Hi-C contact frequencies (**Fig. 1C, 1D, 1E** dense black bar) with other genomic loci a few hundred Kbp away than the ones with sparse GapR peaks (**Fig. 1C, 1D, 1E** sparse black bar). To quantify this observation, we summed up each locus' total contact in the 100Kbp to 1Mbp window and observed that the total contacts are high in the genomic regions with a large number of GapR peaks (at ~3 Mbp in the genome) and low in the region with sparse GapR peaks (at ~1Mbp in the genome) (**Fig. 1F, 1G, 1H**). This observation indicates that GapR peak-containing bins have closer spatial arrangements than non-peak containing bins in the 100 Kbp – 1Mbp distance separation range. We are also able to validate this observation in three other experimental replicates (**Supplement Fig. S1A**).

To quantitatively verify the difference in Hi-C contact frequencies of regions enriched or depleted of GapR peaks, we calculated the inverse of the normalized Hi-C contact frequency between any two bins and plotted the averages against the linear genomic distance between the loci pairs. We used the inverse of Hi-C contact frequency as the apparent indicator for the spatial separation because it was shown to be proportional to the spatial distance between two chromosomal sites[15]. We found that the spatial separation between two bins both containing GapR peaks (**Fig. 1I, 1J, 1K,** gP-gP, blue) was smaller than that between all bins in all distance separation ranges (all-all, red). On the contrary, the spatial separation between two bins not containing any GapR peaks (**Fig. 1I, 1J, 1K**, gNP-gNP, green) was larger than that between all bins. We also plotted the difference between the gP-gP and gNP-gNP curves and verified that, for all three independent datasets, such a difference reaches a maximum near the 500 Kbp loci separation distance and gradually decays (**Fig. 1L, 1M, 1N**). The other replicates also demonstrate the same pattern (**Supplement Fig. S1A, S1C**). Therefore, our observations are highly reproducible and robust to different high-throughput chromosome conformation capture technologies.

Interestingly, although the 5 Kbp and 0.5 Kbp Hi-C datasets were collected in different growth conditions (LB for the 5 Kbp datasets and minimal medium for the 0.5 Kbp dataset), the trend of higher Hi-C contacts between GapR-containing bins remained, except that the 0.5Kbp dataset has an overall smaller difference between the gP-gP and the gNP-gNP curves compared to the 5Kbp datasets. These observations suggest that although overall configuration and compaction of the *E. coli* chromosome might be different under independent experiments (**Fig. 1I, 1J, 1K,** all-all, red, and **Supplement Fig. S1C**), positive supercoils marked by GapR-binding peaks have strong correlation with closer chromosomal contacts. To verify that the pattern we observed is indeed related to specific binding of GapR to positive supercoils, we analyzed the Hi-C data using a $GapR^{1-76}$-seq dataset. $GapR^{1-76}$ lacks the dimerization domain required for tetramerizing on overstretched, positively supercoiled DNA, and hence binds to DNA nonspecifically[6]. $GapR^{1-76}$ binding sites are distributed uniformly along the *E. coli* genome (**Fig. 1C, 1D, 1E**, green bars), unlike GapR and TopoI peaks that preferentially bind to the surroundings of the *oriC* half of the chromosome (black and blue bars in **Fig. 1A, 1B, 1C**; see **Supplemental Table 1A** for definitions of *ori*C half and *ter* half). Using this $GapR^{1-76}$-seq dataset, the spatial separation difference between gP-gP and gNP-gNP containing bins is substantially diminished, suggesting that the pattern we observed is unique to GapR's binding to positive supercoils (**Supplement Fig. S1D**).

TopoI-binding sites correlate with higher-than-average Hi-C contact frequencies in *E. coli,* and positive supercoiling has a larger effect in enhancing chromosomal DNA interaction than negative supercoiling

Bacterial genomes are generally negatively supercoiled [15], and previous studies have shown that negative supercoiling is a main contributor for genome compaction [16]. To investigate the role of negative supercoiling in organizing the genome, we used an *E. coli* Topo-seq data set in which all *E. coli* TopoI binding sites were identified using immunoprecipitation under the same growth condition as the 5 Kbp Hi-C dataset (LB, 37°C)[9]. *E. coli* TopoI removes excessive negative supercoiling and can also associate with RNA polymerase during transcription [10]. Hence, TopoI's peak positions on the chromosome excluding the ones contained in an active transcription unit may reveal negative supercoiled sites [9] (see **Methods**, ChIP-seq processing and analysis). Using the TopoI-seq dataset, we identified 446 reproducible TopoI-binding sites and assigned the *E. coli* genome bins as either tP (Topo-peak) containing bins or tNP (non-Topo-peak) containing bins, akin to the GapR peak analysis. We then aligned TopoI peak-containing bins (blue lines) with the corresponding Hi-C heatmap [7 8] along the *E. coli* chromosome (**Fig. 1C, 1D, 1E**). TopoI peaks along the chromosome have a largely similar distribution as GapR peaks in that they are densely distributed surrounding the *ori*C half of the chromosome (**Supplemental Table 1A**, *ori*C half and *ter* half) which contains multiple *rrn* operons, and sparsely surrounding the *ter* half of the chromosome. As such, genomic locations containing a large number of TopoI peaks (**Fig. 1C, 1D, 1E,** dense blue bars) also showed an overall higher Hi-C contact frequencies with other genomic loci a few hundred Kbp away than the ones with sparse TopoI peaks (sparse blue bars).

To quantitatively examine this observation, we plotted the apparent spatial separation between two bins along their linear separation along the chromosome (**Fig. 2A, 2B, 2C**). Similar to the observation in the GapR peak analysis, for all datasets and despite their difference in resolution, growth media, and experiment technologies, we found that the spatial separation between two bins both containing TopoI peaks (tP-tP, blue) was the smallest, followed by that between all genomic location (all-all, red) and that between two bins both containing no TopoI peak (tNP-tNP, green). In all three Hi-C datasets, despite their different resolution, the difference between tP-tP and tNP-tNP follows the same trend as observed in the GapR analysis – such a difference reaches a maximum around loci linear separation distance of 500 Kbp and gradually decays (**Fig 2D, 2E, 2F**). However, the numerical differences between tP-tP and tNP-tNP, which maximizes at ~0.1 in the magnitudes, appeared to be less than that of the corresponding GapR

peak analysis, which maximizes at ~0.2. These observations are also reproducible across all experimental replicates (**Supplement Fig. S2A**).

Our analyses so far showed that both positive supercoiling (indicated by GapR binding peaks) and negative supercoiling (indicated by TopoI binding sites) contribute to chromosome compaction, while positive supercoiling appears to have a larger effect in compaction (**Fig. 1L, 1M, 1N, 2D, 2E, 2F**). To investigate this effect quantitatively, we re-assigned each 5Kbp-bin a tag, gOnly (contains GapR peaks only but no TopoI peak), tOnly (contains Topo peaks only but no GapR peak), or NN (contains no peak of either type). This mutually exclusive bin assignment avoids the collective influence on spatial contacts contributed by both supercoiling types. As a result of this assignment, bins containing both GapR and TopoI peaks were not used in the analysis (162 5-Kbp bins and 539 0.5-Kbp bins). Plotting the apparent spatial separation within these three groups confirmed the observation that bins between GapR-only peaks (**Fig. 2G, 2H, 2I,** gOnly-gOnly, red) was the closest, those between TopoI-only peaks (tOnly-tOnly, blue) further, and those between two non-peak-containing loci the furthest (NN-NN, green). Interestingly, at linear separations > 1.25 Mbp, the spatial separation of non-peak-containing bins (NN-NN) gradually decreased and finally became slightly smaller than the corresponding tOnly-tOnly bins, but the spatial separation between two loci containing only GapR peaks is always lower than that between two loci containing only TopoI peaks through all linear distances, as demonstrated by the subtracting the gOnly from the tOnly spatial separation (**Fig. 2J, 2K, 2L**). From these comparisons, we conclude that while both types of supercoils contribute to genome interaction, positive supercoiling is a stronger driving force than negative supercoiling, where the difference in magnitude is the most pronounced at linear separation distance from ~500 Kbp to ~1.5Mbp depending on the chromosome conformation in different Hi-C experiments (**Fig. 2J, 2K, 2L**). Similar patterns were found in other independent Hi-C replicates (**Supplement Fig. S2B**).

Machine learning model confirms that contact maps could capture positive but not negative supercoiling information

Our analyses so far suggest that positive and negative supercoiling may contribute to unique spatial arrangements captured by contact maps generated from Hi-C/5C experiments. To investigate whether the contact matrix itself encodes such supercoiling information, we trained a series of random forest models to predict positive and negative supercoiling locations based on Hi-C contact frequency for each supercoiling type, respectively. The input matrix is obtained by

rearranging the contact matrix (5C, resolution = 5 Kbp), generated by Lioy et al. [7] by aligning its diagonal with the middle column, so that each column (or termed as a feature) represents a signed genomic separation distance from a genomic locus (**Fig. 3A**). In the training set, we used the contact frequencies of each bin with all other bins across the genome as the input (**Fig 3A**, rows) and the corresponding information on whether this bin contains GapR (gP/gNP) or TopoI (tP/tNP) peaks as the ground truth (**Fig. 3A**, colored bars) to train the random forest model. To evaluate the predication accuracy of the trained model, we compared the predicated g/tP or g/tNP assignment of each genomic bin with the underlying true assignment derived from the GapR-seq and TopoI-seq data in the testing set. Two metrics of evaluation are the area under the Receiver Operating Characteristic Curve (AUROC) and the area under the Precision Recall Curve (AUPRC), both computed as the average from 10 independent experiments based on 5-fold cross-validation (see **Methods**, random forest model).

We observed that the random forest model performed well for predicting GapR peaks with a AUROC value of ~0.74 when using as little as 25% of data for training. The prediction performance remains similar when using more data for training. This prediction model exceeds the baseline model trained using randomly assigned gP or gNP to a genomic bin which has an AUROC value of 0.43 (**Fig. 3B**). Similar performance patterns are also shown by PR curves with AUPRC ~0.68 (**Fig. 3C**). Similar performance across models with different training-to-testing data ratios suggest that all GapR peak-containing bins have uniform, strong interaction patterns with all other genomic loci, and such patterns are absent in non-GapR peak-containing bins. To further examine which feature (linear distance between bins) in the model contributes the most to the correct assignment of gP or gNP to each genomic bin in all data, we calculated the value of Mean Decrease in Accuracy (MDA) for all loci with different linear distances. We found that the MDA plot peaked at ~500 Kb, indicating that the contact between loci separate by this distance contribute the most to the correct categorization of gP and gNP bins (**Fig. 3D**). It confirms our finding that gP and gNP differ the most in their interactions with genomic loci at a linear distance ~ 500 Kbp. In summary, the model performance indicates that P and NP bins are largely distinguishable based on their Hi-C contacts with other genomic loci, suggesting that positively supercoiled sites, compared to sites without positive supercoiling, may have a different interaction pattern with their neighboring genomic segments on the chromosome.

Next, to test whether Hi-C contact frequency can be used to predict negative supercoiling in a similar fashion, we trained the random forest model following the same procedure except that TopoI peak data was used instead of GapR peak data. Surprisingly, ROC and PR curves

generated at different training-to-testing data ratios have estimated AUROC values cap at 0.58 and AUPRC at 0.44, indistinguishable from or slightly worse than that from the baseline. (**Fig. 3E, 3F**). Furthermore, the MDA suggests that the random forest makes a classification not relying on any particular genomic regions (**Fig. 3G**). These results demonstrate that Hi-C contact frequencies of a genomic locus with others alone cannot effectively determine whether it contains a TopoI peak, in contrast to what we observed on GapR (**Fig. 3B, 3C, 3D**). The poor performance of TopoI peak prediction can be explained by the lack of difference between the interaction of tP and P with all genomic loci (**Fig. 3H, 3I, Supplement Fig. S3**). Although there is a significant difference between tP-tP and tNP-tNP interaction (**Fig. 2A**), our machine learning models have no access to this information because the features are not labeled as peak or non-peak. Instead, the models rely on the difference between peak and non-peak with all genomic loci to make a classification. Since such a difference is significant in GapR peaks but not TopoI peaks, we conclude that recovering positive supercoiling signals, but not negative supercoiling, is possible from using HiC datasets alone, which suggests that positive supercoiling has a characteristic signature that is reflected in the Hi-C contact map.

Transcription-independent and -dependent supercoiling

Previous studies have established that transcription is a major supercoil-generator in bacteria [17] [18]. During transcription, an elongating RNA polymerase (RNAP) molecule under-twists upstream DNA and over-twists downstream DNA, generating negative and positive supercoiling, respectively [19]. By quantifying where genomic supercoils locate with respect to the transcription activities of their nearby transcription units, we can assess the ability of transcription to generate sustained supercoils and consequently its contribution to compact the chromosome.

We used a set of RNA-seq data (LB, 37°C) with 2640 transcription units (TU) annotated on the *E. coli* chromosome, whose transcriptional activities are measured in Reads Per Kilobase Million (RPKM) [20]. A TU is considered highly active if its expression falls in the top quartile (**Supplement Fig. S4A**). Next, we calculated the distance between a transcription end site (TES) and the nearest GapR peak downstream, and that between a transcription start site (TSS) and its nearest TopoI peak upstream through all the annotated TUs. The distance distributions of TopoI and GapR peaks to their nearest TU do not differ from each other significantly (p=0.13). When we plotted the TES–GapR or TopoI–TSS distance against the corresponding transcription activity of the TU, we observed that majority (94.7% and 94.8%) of GapR and TopoI peaks are within 2.5 Kbp of its nearest TES and TSS, respectively. For comparison, on average 90.0% and 77.0% of

random chromosomal sites are within the same 2.5 Kbp distance from the nearest TES or TSS, respectively (**Fig. 4A, 4B, 4C, 4D, Supplement Fig. S4B, S4C**). This comparison demonstrates that the majority of supercoils are associated with transcription activities, and that it makes up to ~96% of total sustained genomic supercoils, hence contributing significantly to genome compaction.

Given that the intergenic regions only make up to 8% of the *E. coli* genome with an average length of ~ 80 bp [20], over 75% of GapR and TopoI peaks have at least 85% of their peak length overlapping within adjacent gene bodies. Furthermore, 44.9% of all TUs overlap with either GapR or TopoI peaks, and 73.1% of these overlapping regions are contributed by only GapR peaks. Previous studies have shown that DNA supercoils generated by ongoing transcription activities could modulate the cooperativity of adjacent RNA polymerase molecules on the same gene and also impact the transcription of adjacent genes [20 21]. Therefore, the observation that more GapR than TopoI peaks reside in adjacent genes suggest that positive supercoiling also plays a more pervasive role than negative supercoiling in regulating *E. coli* transcription.

We also observed a minor group of GapR and TopoI peaks (3.8% and 4.1% of all peaks, respectively) locating more than 2.5 Kbp from any TUs (hex bins above the orange line, **Fig. 4A, 4B**). About ~ 72% and 79% of them associate with TUs with low transcription activities, respectively, (purple highlighted quadrant, **Fig. 4A, 4B**). This observation suggests that these supercoils may not be introduced by transcription because they are far away. Interestingly, the genomic locations of these supercoil sites are near the *Ori* and *Ter* regions but sparse at in between (**Fig. 4E, 4F**), suggesting that they may be introduced by replication, which is another major supercoiling contributor among genomic processes [22]. These patterns are absent in the controls of random chromosomal sites, which are uniformly distributed across the genome as expected (**Fig. 4G, 4H**). In summary, this analysis demonstrates that the location of supercoils are mostly associated with transcription activities, aligns with previous findings that transcription activity is a main driving force for genome compaction.

Discussion

In this work, we explored the role of positive and negative supercoiling in the *E. coli* genome organization and quantified the difference in their effects (**Fig. 1C-N**). We observed a consistent, robust pattern across independent Hi-C data sets that the spatial separation between two genomic loci both containing GapR peaks are the smallest compared to the ones containing only one or not containing GapR peaks, or the ones containing TopoI binding sites indicative of

negative supercoiling (**Fig. 2A-F, 2G-L**). These observations add to the idea that supercoiling leads to chromosome compaction and suggest that positive supercoiling enhances the spatial interactions of chromosomal loci more significantly than negative supercoiling. Building upon these observations, we developed a supervised machine learning model and found that Hi-C data can accurately predicate GapR-binding sites but not TopoI binding sites, suggesting that the spatial interactions contributed by positive supercoiling dominate Hi-C measurement (**Fig. 3A-I**). Finally, we found over 95% of GapR and TopoI sites reside next to a highly expressed TU, and that more GapR sites than TopoI sites are within the body of an adjacent gene. These observations suggest that transcription is a major driver to generate supercoiling to compact the genome, and that supercoiling may indeed play an important role in regulating transcription of genes within the same topological domain. Taken together, our results suggest that positive supercoiling contributes more than negative supercoiling to genome compaction and transcription, which has not been proposed by previous studies (**Fig. 4A-F**).

The differential roles of positive and negative supercoiling

Our results show that positive supercoiling marked by GapR peaks is a stronger driving force for genome compaction compared with negative supercoiling identified by TopoI peaks (**Fig. 2G-L**). This marked difference has not been reported before. The *E. coli* chromosome is naturally negatively supercoiled [15], which is regarded necessary to allow genomic processes such as replication or transcription to proceed without being energetically costly to open the double-stranded chromosomal DNA. As such, it is reasonable to expect that the role of positive supercoiling is indeed to compact the genome and compensate for the opening of dsDNAs due to naturally existing negative supercoiling, which may destabilize the chromosome by randomly introducing ssDNA regions due to stochastic dsDNA opening in negatively supercoiled DNAs.

Why would positive and negative supercoiling have differential effects on chromosome organization? The mechanisms that regulate positive and negative supercoiling in *E. coli* are complex. However, we note that there are several abundant DNA structuring proteins in *E. coli* (e.g., HU, H-NS, Fis) that are known to constrain negative supercoiling, suggesting that the effects of negative supercoiling may be buffered by these DNA binding proteins [23 24 25]. Hence, the role of negative supercoiling may be more pertinent to genomic processes while that of positive supercoiling may be more pertinent to genome compaction Further investigation will be needed to address this question.

Genomic locations of positive and negative supercoiling

When we arranged GapR and TopoI peaks along the *E. coli* chromosome, we found that they share similar patterns. Both GapR and TopoI peaks are enriched in *Ori* and *ter* regions, but less in between (**Fig. 1C-E**). As the *Ori* macrodomain contains most of the active genes, and both *Ori* and *Teri* are involved in DNA replication [16 26], the dense location of GapR and TopoI peaks in these two regions is consistent with the expectation that transcription and replication are major supercoil generators. Between the two, transcription appears to have a more significant role in generating supercoils, as over 95% of GapR and TopoI peaks are associated with highly active TUs. Of note, there are more GapR peaks (609) compared to TopoI peaks (446) (see **Supplementary Table 1** for statistics), which could be due to different capturing efficiency of the corresponding antibody in the ChIP-Seq experiments.

Short-distance and long-distance compaction

We show that the spatial distance separation between two bins containing GapR-peak or TopoI-peak is smaller than those without these peaks, with the largest effects on compaction occurring for loci 0.5-1 Mbp apart in a linear chromosome (**Fig. 1L-N, 2D-F**). We hypothesize that the fact that supercoiling-dependent compaction peaks at 0.5-1 Mbp is a result of the chromosome conformation and is dictated by the macrodomain organization of the *E. coli* genome. Short-distance interactions such as TADs occur on only ~100 Kbp scales [26], which are too short to explain our observations. In contrast, the observed supercoiling-dependent compaction matches the size of macrodomains, which are estimated to be ~1 Mb [26]. While the actual peak is slightly different for each experiment, they all fall in the 0.5-1 Mbp genomic distance range. Such a difference may be due to experimental variations since it has an association with the number of valid paired reads from the Hi-C/5C experiment. As such, we hypothesize that positive and negative supercoiling contributes to organizing the chromosome by enhancing the interaction within the macrodomains.

Materials and Methods

Hi-C data processing, contact matrices computation, and contact heatmap generation. All Hi-C/5C sequence reads were aligned to the *E. coli* reference genome NC_000913.3 and processed using the Hi-C-Pro 3.1.0 pipeline [27]. Contact count matrices were normalized using the iterative correction and eigenvector decomposition (ICED) technique standard to the Hi-C-Pro pipeline. All ICED contact counts are log2 transformed with a pseudo count added, and the heatmaps are generated by the ComplexHeatmap [28] package in R.

ChIP-seq data processing and analysis. GapR ChIP-seq and TopoI-seq. Quality check was performed using FastQC [29], followed by Trimmomatic [30] to remove adapters in TopoI-seq sequence reads. All sequence reads were aligned to the *E. coli* reference genome NC_000913.3 using Bowtie 2 [31], followed by Samtools [32] which converts SAM files to BAM. Genomic peaks were identified using macs2's callpeak [32] function, and peaks from experiment replicates were processed to generate a set of reproducible peaks using a customized R script (available on the GitHub repo). A genomic bin (either 5 or 0.5 Kb) is labeled as gP or tP only if it contains the center of at least one GapR or TopoI peak, respectively. gNP and tNP are the respective complementary set of bins to the set of gP and tP bins. gOnly bins are defined as the bins that contain only GapR peak centers but no Topo peak center. tOnly bins are bins that contain only Topo peak centers but no GapR peak center. Bins that do not contain any peak center are labeled as NN.

Random forest model. Data set preprocessing: The raw Hi-C counts are normalized within each data set with the ICED method using the HiC-Pro software [27]. The reciprocals of these counts will be used to construct the Hi-C frequency matrix. A transformation is applied to this matrix to get the training and testing data matrix: the diagonal of the original is aligned with the middle column, representing the self-contacts. Each column, represents a genomic separation distance from a given locus (row), is a feature to the model. Training and testing data: The 928 by 928 matrix (bin size is 5k determined by the Hi-C data resolution), after the above transformation, is partitioned into four non-overlapping set. We built four series of models using these four quarters, respectively, as the testing data. For each testing set, we built 10 random forests by randomly selecting 25%, 50%, 75%, respectively, from its complement data set as the training set. Random forest training using the R package caret [33]: Each random forest is trained using 5-fold cross-validation with 10 repeats. mTry (from 2 to 42 by 5) and nTrees (from 200 to 1200 by 100) are

optimized using the ROC metric. Prediction and evaluation: After each model is trained, we made prediction using the optimized parameters on its designated testing set, which remained unseen by the model in our experiment design. For each testing data set and each prediction percentage (25%, 50%, and 75%), there are 10 predictions being made. We then computed their average AUROC and AUPRC. The baseline models are trained in the same way, except that the input data set is labeled using a set of randomly generated peak locations.

Growth conditions and chemical treatments for *E. coli* used in Hi-C *E. coli* was grown in LB (10 g/L NaCl, 10 g/L tryptone, 5 g/L yeast extract) at 37 °C with shaking at 200 rpm. Optical density was measured at 600 nm using a Genesys 10 Bio Spectrophotometer.

Chromosome conformation capture deep-sequencing (Hi-C) Hi-C experiments were performed as described previously [12] (Le et al, 2013) with the following modifications. Wild-type *E. coli* K-12 MG1655 was grown to OD600 ~0.2 and then fixed with 3% formaldehyde for 10 min at room temperature before quenching with 0.375 M final concentration glycine for 15 min. Fixed cells were pelleted, washed twice with 1X M2 salts (6.1 mM Na_2HPO_4, 3.9 mM KH_2PO_4, 9.3 mM NH_4Cl), then washed 2 more times with 1X TE buffer (10 mM Tris-HCl [pH 8.0] and 1 mM EDTA), before resuspension in 1X TE at ~10^7 cells / µL. Resuspended cells were divided into 25 µL aliquots before storage at -80 °C. Each Hi-C experiment was performed with 2 x 25 µL aliquots. To each 25 µL aliquot, 20 000 U of Ready-Lyse Lysozyme (Lucigen) was added, and the mixture was incubated for 30 min before addition of SDS to a final concentration of 0.25% for 15 additional minutes to completely dissolve cell membranes and release chromosomal DNA.

Each aliquot of DNA was then digested with HpaII (NEB) as follows. A 50 uL reaction was assembled containing lysed cells, rCutSmart buffer, and 5 µL of 10% Triton-X 100, which was allowed to sit for 15 min at RT to inactivate SDS. Subsequently, 100 U of HpaII was added and the DNA was digested at 37 °C for 2 hr. The reaction was cooled on ice before labelling cleavage overhangs with Biotin-14-dCTP (ThermoFisher) in a 60 uL reaction (digestion reaction, 0.9 µL dGTP, dTTP and dATP at 2 mM each, 4.5 µL 0.4 mM biotin-14-dCTP, 1.2 µL of Klenow Large fragment [5000 units/ml]) for 45 min at room temperature. Reactions were quenched by adding 0.125% final concentration of SDS and incubated for 5 min at room temperature. Blunted DNAs were ligated in dilute conditions as follows: A master mix containing 75 µL 10% Trition-X 100, 100 µL 10X T4 Ligase Buffer (NEB), 2.5 µL of 20 mg / mL BSA (NEB), and 800 µL water per

blunting reaction was assembled and chilled on ice. Chilled ligase mix was then added to the blunted DNAs and incubated for 15 min at room temperature to inactivate SDS. Ligation reactions were then chilled on ice for 15 min before addition of 3 µL T4 DNA ligase (NEB, 2,000,000 U / mL). Reactions were mixed well and incubated at 10-16°C for 4 hr.

After ligation, 10 mM final concentration EDTA and 2.5 µL of 20 mg / mL Proteinase K (NEB) was added, and reactions incubated at 65 °C overnight to digest proteins and reverse crosslinks. DNA was then extracted twice with phenol:chloroform:isoamyl alcohol, isopropanol precipitated, and then the entirely of the DNAs from the two original aliquots was resuspended in 60 µL of water. Next, non-ligated biotinylated DNAs were eliminated in a 100 µL reaction (60 µL purified DNA, 10 µL Buffer #2 (NEB), 0.5 µL 20 mg / mL BSA (NEB), 5 µL each of 2 mM dATP and dTTP). DNA was sheared using a Bioruptor (Diagenode) and size-selected by gel extraction to be between 200 and 500 bp.

Paired-end Illumina libraries were generated as in (Guo et al, 2021) as follows using DNA LoBind tubes (Eppendorf). First, DNA was end-repaired with 4 µL T4 DNA polymerase (NEB), 4 µL T4 PNK (NEB), and 0.75 µL Klenow large fragment (NEB) in 100 µL T4 DNA ligase buffer with 0.25 mM final concentration dNTPs for 30 min at room temperature. Repaired DNA was recovered by Ampure XP (Beckman Coulter) bead purification, using 100 µL beads in 300 µL 20% PEG/NaCl solution. Beads were washed twice with 80% ethanol, dried, and resuspended in 32 µL EB (10 mM Tris-HCl [pH 8.0]). The bead slurry was directly treated with 3 µL Klenow (3'→5' exo-) (NEB) in 40 µL reaction with 0.2 mM ATP at 37 °C for 30 min to add 3' overhangs to DNA. Repaired DNA was recovered by Ampure XP capture and resuspended in 15 µL EB without beads. Y-shaped adaptors were prepared by annealing Illumina PE adapter 1 and Illumina PE adapter 2. Y-shaped adapters were added to bead slurry, and the mix was ligated in 25 µL total volume with 5 µL 15 µM ligated adapters and 1.5 µL T4 DNA Ligase (NEB, 400,000 U / mL) for 30 hr at room temperature, with occasional mixing. Biotin-labeled junctions were purified from non-labelled junctions with Dynabeads MyOne Streptavidin C1 (ThermoFisher). Dynabeads were first washed twice with 1X NTB (5 mM Tris-HCl [pH 8.0], 0.5 mM EDTA and 1 M NaCl) then added to ligation and mixed well. Biotin-labeled DNAs were allowed bind for 30 min at room temperature with occasional mixing before pulldown with a magnet. The beads were washed three times with 100 µL 1X NTB, then washed with 100 µL water, resuspended in 10 µL water, transferred a new tube, and stored at -20 °C. DNA libraries were amplified in 50 µL final volume with 2X KAPA HiFi Master Mix (Roche) supplemented with 5 % final concentration DMSO (Sigma) and appropriate barcoded primers. The total number of cycles was optimized for

each sample to minimize the number of cycles required for library generation. Libraries were purified by two-step Ampure XP capture to recover 200-500 bp amplified libraries. DNA was recovered from Ampure beads by resuspending in 20 µL 10 mM Tris-HCl (pH 8.0). Paired-end sequencing was performed with a 91nt kit on a NextSeq500 at the MIT Bio Micro Center.

Acknowledgments

The authors thank all members of the Xiao, Ji laboratories for valuables discussions and advice on this study. All computations were performed on the Joint High Performance Computing Exchange (JHPCE), organized by the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health, and the Maryland Advanced Research Computing Center (MARCC) Rockfish cluster. Z. F. wants to thank Jiafang Song for marking the illustration cartoons. The project is supported by NIGMS 1R35GM136436 and NSF MCB1817551 (to J. X.), and R00GM134153 to (M. S. G.)

Declaration of Competing Interests

The authors declare no competing interests.

Data and Code Availability

Two raw, publicly available Hi-C datasets were downloaded from the National Center of Biotechnology Information's Gene Expression Omnibus under the accession number GSE: 107301 and Accession number PRJNA587586. Datasets generated during this study are deposited at the Gene Expression Omnibus (GEO): GSE214856. All custom-made codes used in this study are available from the Xiao Laboratory GitHub repository (XiaoLab JHU: https://github.com/XiaoLabJHU).

Author Information

Authors and Affiliations

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA

Ziqi Fu, Weiqiang Zhou

Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA

Jie Xiao

Department of Microbiology, University of Washington School of Medicine, Seattle, WA 98195, USA

Monica S. Guo

Contributions

Z.F., J.X., W.Z., and M.S.G. designed the statistical experiments. M.S.G. conducted the Hi-C experiments. Z.F. analyzed and integrated all genomics data sets used in this study. Z.F. and W.Z. designed the machine learning experiments. Z.F. built the machine learning models and wrote analysis codes. Z.F., J.X. wrote the original draft. All authors reviewed and edited the manuscript.

References

1. Cavalli, G. & Misteli, T. Functional implications of genome topology. *Nat Struct Mol Biol* **20**, 290-299 (2013).
2. Ali Azam, T., Iwata, A., Nishimura, A., Ueda, S. & Ishihama, A. Growth phase-dependent variation in protein composition of the Escherichia coli nucleoid. *J. Bacteriol.* **181**, 6361-6370 (1999).
3. Lal, A. *et al*. Genome scale patterns of supercoiling in a bacterial chromosome. *Nat Commun* **7**, 11055 (2016).
4. Botchan, P., Wang, J. C. & Echols, H. Effect of circularity and superhelicity on transcription from bacteriophagelambda DNA. *Proc Natl Acad Sci U S A* **70**, 3077-3081 (1973).
5. White, J. H. & Bauer, W. R. Calculation of the twist and the writhe for representative models of DNA. *J. Mol. Biol.* **189**, 329-341 (1986).
6. Guo, M. S., Kawamura, R., Littlehale, M. L., Marko, J. F. & Laub, M. T. High-resolution, genome-wide mapping of positive supercoiling in chromosomes. *Elife* **10**, 10.7554/eLife.67236 (2021).
7. Lioy, V. S. *et al*. Multiscale Structuring of the E. coli Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell* **172**, 771-783.e18 (2018).
8. Cockram, C., Thierry, A. & Koszul, R. Generation of gene-level resolution chromosome contact maps in bacteria and archaea. *STAR Protoc* **2**, 100512 (2021).
9. Sutormin, D. *et al*. Interaction between transcribing RNA polymerase and topoisomerase I prevents R-loop formation in E. coli. *Nat. Commun.* **13**, 4524-5 (2022).
10. Tan, K. *et al*. Structural basis for suppression of hypernegative DNA supercoiling by E. coli topoisomerase I. *Nucleic Acids Res.* **43**, 11031-11046 (2015).
11. van Berkum, N. L. *et al*. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **(39):1869. doi**, 10.3791/1869 (2010).
12. Le, T. B. K., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731-734 (2013).
13. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* **21**, 207-226 (2020).
14. Guo, M. S., Haakonsen, D. L., Zeng, W., Schumacher, M. A. & Laub, M. T. A Bacterial Chromosome Structuring Protein Binds Overtwisted DNA to Stimulate Type II Topoisomerases and Enable DNA Replication. *Cell* **175**, 583-597.e23 (2018).

15. Wang, S. *et al*. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598-602 (2016).
16. Verma, S. C., Qian, Z. & Adhya, S. L. Architecture of the Escherichia coli nucleoid. *PLoS Genet.* **15**, e1008456 (2019).
17. Racko, D., Benedetti, F., Dorier, J. & Stasiak, A. Are TADs supercoiled? *Nucleic Acids Res.* **47**, 521-532 (2018).
18. Ma, J. & Wang, M. D. DNA supercoiling during transcription. *Biophys. Rev.* **8**, 75-87 (2016).
19. Liu, L. F. & Wang, J. C. Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 7024-7027 (1987).
20. Herzel, L., Stanley, J. A., Yao, C. & Li, G. Ubiquitous mRNA decay fragments in E. coli redefine the functional transcriptome. *Nucleic Acids Res* **50**, 5029-5046 (2022).
21. Johnstone, C. P. & Galloway, K. E. Supercoiling-mediated feedback rapidly couples and tunes transcription. *Cell Reports* **41** (2022).
22. Bates, A. D. & Maxwell, A. in *DNA Topology* (OUP Oxford, 2005).
23. Rouvière-Yaniv, J., Yaniv, M. & Germond, J. E. E. coli DNA binding protein HU forms nucleosomelike structure with circular double-stranded DNA. *Cell* **17**, 265-274 (1979).
24. Broyles, S. S. & Pettijohn, D. E. Interaction of the Escherichia coli HU protein with DNA. Evidence for formation of nucleosome-like structures with altered DNA helical pitch. *J Mol Biol* **187**, 47-60 (1986).
25. Schneider, R., Travers, A. & Muskhelishvili, G. FIS modulates growth phase-dependent topological transitions of DNA in Escherichia coli. *Mol Microbiol* **26**, 519-530 (1997).
26. Valens, M., Penaud, S., Rossignol, M., Cornet, F. & Boccard, F. Macrodomain organization of the Escherichia coli chromosome. *EMBO J.* **23**, 4330-4341 (2004).
27. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).
28. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).
29. Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
31. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).

32. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
33. Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Soft.* **28**, 1 (2008).

Figures and Tables

Figure 1.

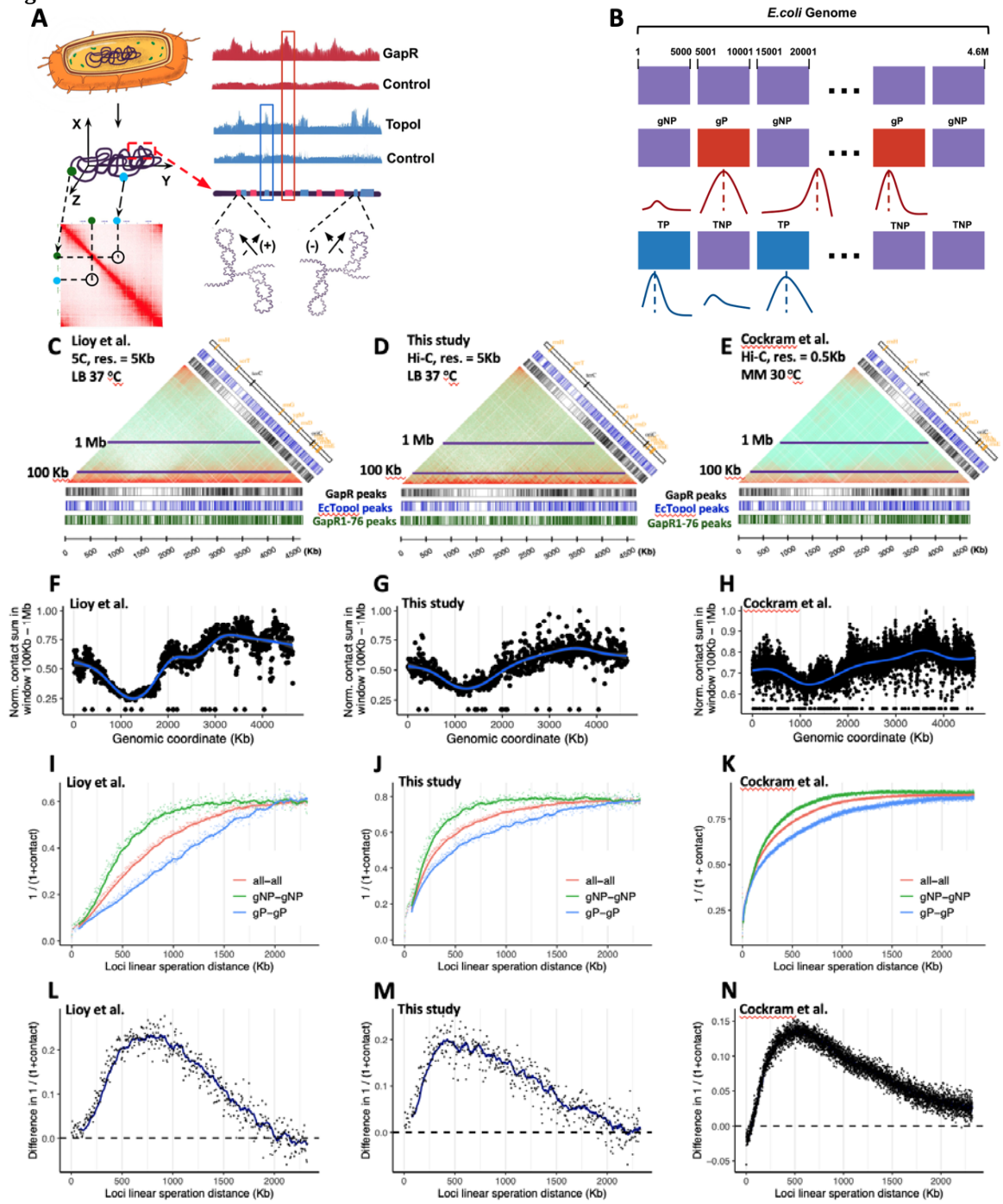


Figure 1. Supercoiling locations revealed by GapR and TopoI ChIP-seq demonstrate high-than-average spatial contacts.

(**A**) A cartoon illustration of the entire study. Isolated *E. coli* chromosomes (wiggling lines) are subjected to measure spatial contacts between any two genomic loci (represented by the green and cyan dots) using Hi-C (bottom contact matrix), and to measure protein binding sites using chromatin-immunoprecipitation

coupled sequencing (ChIp-Seq, right) for GapR (red, marking positive supercoiling) and TopoI (blue, marking negative supercoiling).

(**B**) A cartoon illustration of Hi-C and ChIP-seq data integration. The 4.6 Mbp *E. coli* genome is divided into a number of bins, with each bin assigned to gP (GapR-peak containing), gNP (not containing GapR peak), tP (Topo I-peak containing), or tNP (not containing Topo I peak), based on GapR-Seq and Topo I-seq data. A bin can contain either one, both, or none.

(**C-E**) Visualization of GapR and TopoI peak genomic location with respect to their spatial contacts with other genomic locations in the contact heatmaps. Normalized (log2) contact counts are plotted as symmetric halves from 5C data set (LB, 37 ºC, 5 Kbp resolution) generated by Lioy et. al (left), Hi-C data set (LB, 37ºC, 5 Kbp resolution) generated in this study (middle), and the Hi-C data set (MM, 30ºC, 0.5 Kbp resolution) generated by Cockram et al (right), with a 45% counterclockwise rotation. The indicator bars mark the peaks identified in the GapR (black), $GapR^{1-76}$ (green), and TopoI (blue) ChIP-seq experiments. The region bounded by the two horizontal lines in purple (10Kb, 1Mbp) represents the window used to compute the contact sum (in **F-H**) at each genomic locus (5 Kbp bin). The two reference rules provide the genomic coordinates (bottom, increment = 500 Kbp) and locations of *Ori*C, *ter*C, and seven rrn operons (top-right, *E. coli* MG 1655, 4.6 Mb).

(**F-H**) At each locus (black dots), we sum up its total contact in the 100 Kbp and 1 Mbp window. Normalized (log2) sum (y-axis) is plotted along the *E. coli* genomic coordinates (x-axis). Simple moving average (blue curve, window size = 15) of the normalized contact sum peaks near ~3-4 Mbp.

(**I-K**) The average contact frequencies (y-axis, inversed contact counts representing the apparent spatial separation between two loci) at each genomic separation distance (x-axis) are plotted (colored dots) with simple moving averages (colored curves, window size = 15). Different colors indicate the interaction between loci pairs both containing GapR peaks (gP-gP, blue) and that with no GapR peaks (gNP-gNP, green). The average contact frequencies at each genomic separation are plotted in red as a reference.

(**L-N**) The numeric difference between the averaged contact frequencies (y-axis) at each genomic separation distance (x-axis) by subtracting the gP-gP from gNP-gNP (dots in **I-K**).

Figure 2.

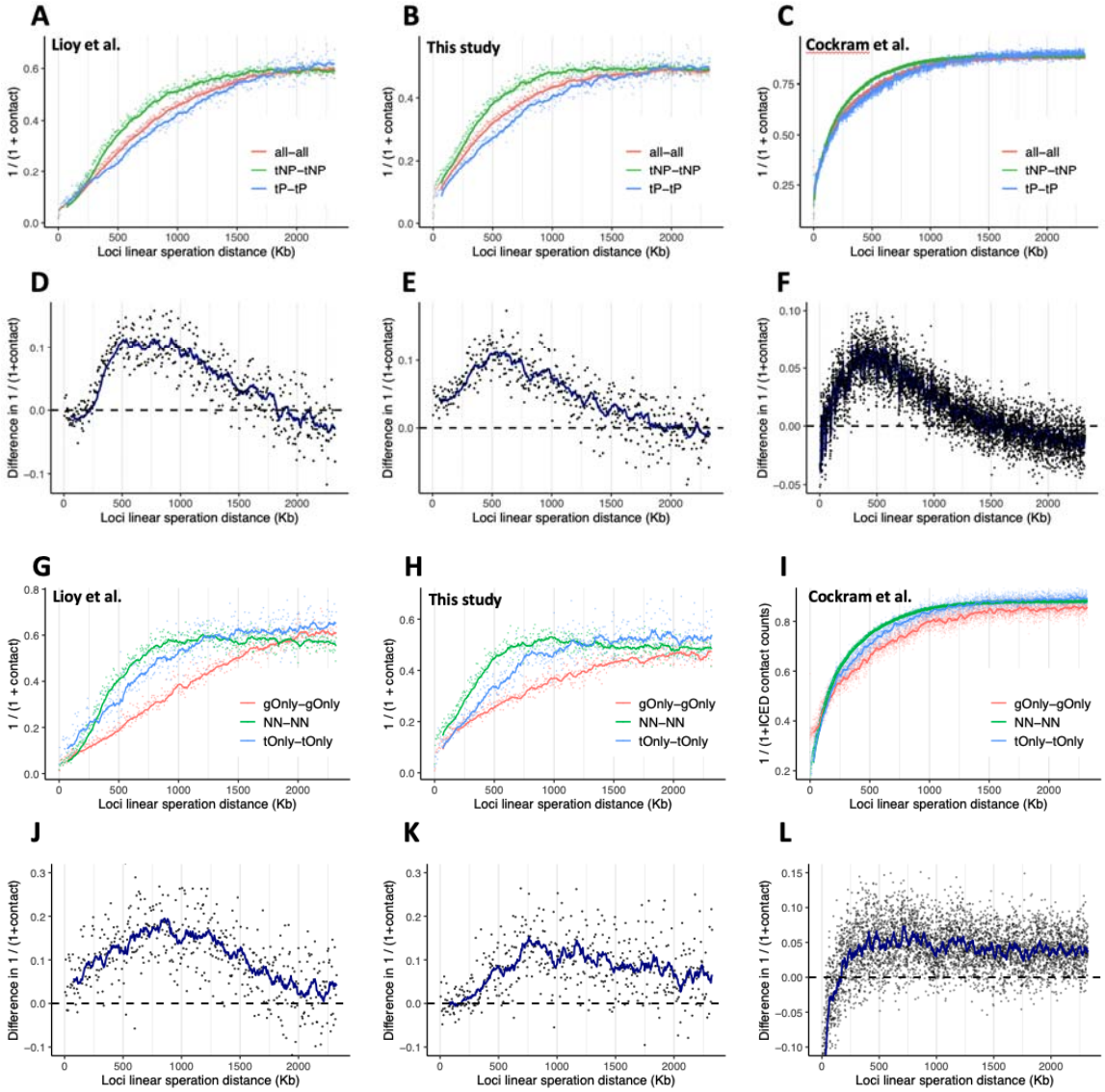


Figure 2. TopoI-binding sites correlate with higher-than-average Hi-C contact frequencies, and positive supercoiling has a larger effect in enhancing chromosomal interaction.
(**A-C**) Average contact frequencies (y-axis, inversed contact counts representing the apparent spatial separation between two loci) at each genomic separation distances (x-axis) are plotted (colored dots) with simple moving averages (colored curves, window size = 15). Different colors indicate the interaction between loci pairs both containing TopoI peaks (tP-tP, blue) and that with no TopoI peaks (tNP-tNP, green). The average contact frequencies at each genomic separation are plotted in red as a reference.
(**D-F**) The numeric difference between the averaged contact frequencies (y-axis) at each genomic separation distance (x-axis) by subtracting the tP-tP from tNP-tNP (dots in **A-C**).
(**G-I**) The average contact frequencies (y-axis, inversed contact counts representing the apparent spatial separation between two loci) at each genomic separation distance (x-axis) are plotted (colored dots) with simple moving averages (colored curves, window size = 15). Different colors indicate the interaction between loci pairs both containing only GapR peaks (gOnly-gOnly, red) and that containing only TopoI peaks (tOnly-tOnly, green). The average contact frequencies at each genomic separation of loci pairs both containing no peaks of either type are plotted in blue as a reference.

(**J-L**) The numeric difference between the averaged contact frequencies (y-axis) at each genomic separation distance (x-axis) by subtracting the gOnly-gOnly from tOnly-tOnly (dots in **G-I**). Overall, the positive values indicate that loci both containing only GapR peaks have smaller spatial separation distance than that containing only TopoI peaks.

Figure 3.

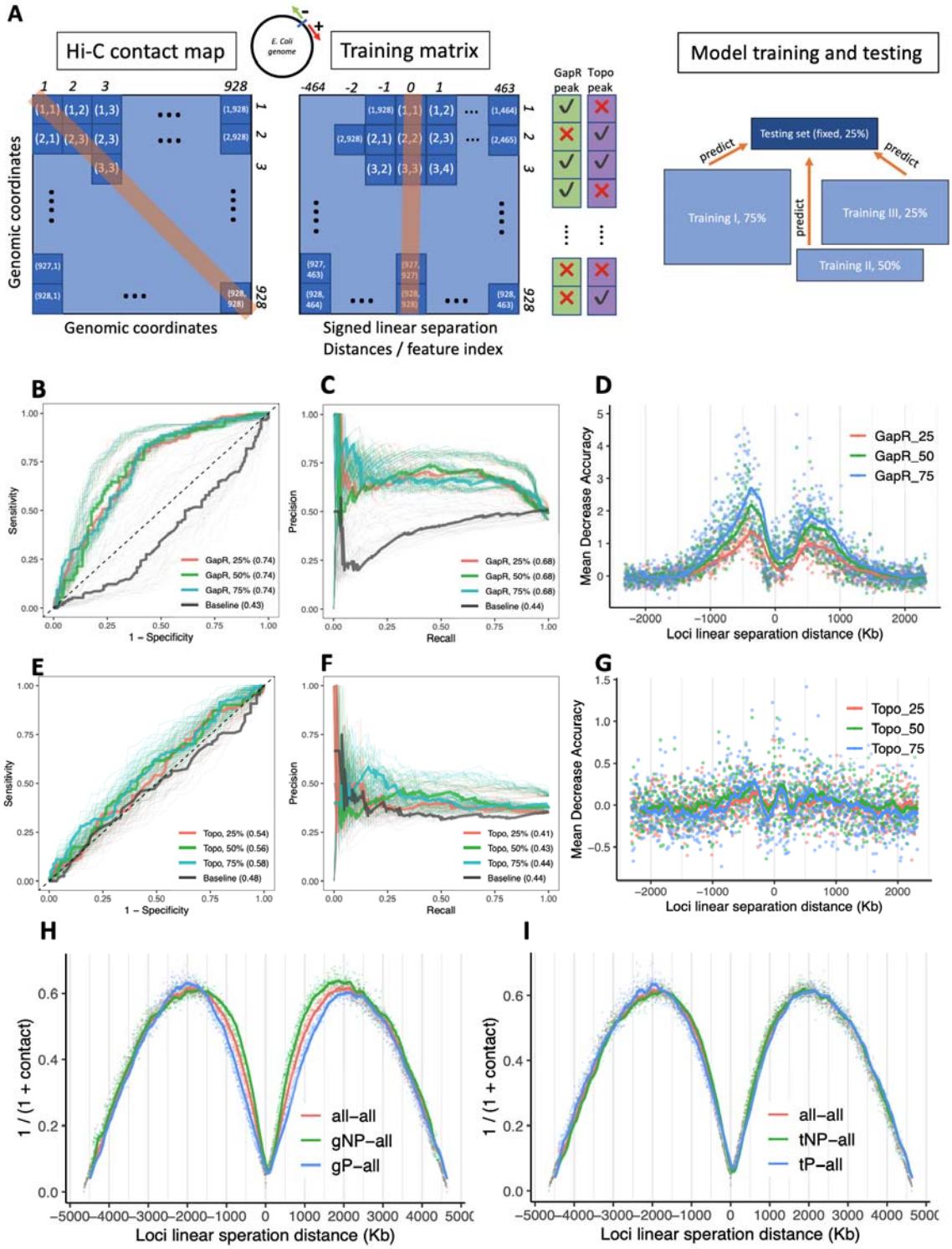


Figure 3. Machine learning models able to recover positive but not negative supercoiling signal from contact matrix.

(**A**) A cartoon illustration of the machine learning model design. The input matrix is a rearrangement of the conventional contact matrix[7], and the input features (x-axis or columns) are signed genomic separation

distance from a given loci (y-axis or rows). The responses, or whether a 5 Kb bin contains any GapR peaks or TopoI peaks are indicated by the check and error marks. A set of randomly chosen loci (25%) is preserved as a fixed testing set for all random forests built.

(**B-C**) ROC and PR curves of 10 independent experiments with 25% (red), 50% (green), and 75% (blue) training data predicting positive supercoiling (marked by GapR peaks) locations up to the Hi-C resolution (5 Kb in this case). The performance of the baseline model, trained with randomized peak labels, is plotted in black. Highlighted curves represent experiments with the median AUROC and AUPRC, respectively. Experiments with difference training percentages have similar AUROC (0.74) and AUPRC (0.68) values, indicating that positive supercoils form strong and uniform spatial patterns along the *E. coli* genome captured by contact maps.

(**D**) Feature importance plot showing the average feature importance (over 10 independent experiments) measured in mean decrease accuracy (y-axis) of each feature (x-axis) as colored dots. The simple moving averages (window size = 15) from experiments with 25% (red), 50% (green), and 75% (blue) training data peak at ~ 500 Kb genomic separation distance, indicating that these features contribute the most to the effective classification of GapR peak containing bins.

(**E-F**) ROC and PR curves of 10 independent experiments with 25% (red), 50% (green), and 75% (blue) training data predicting negative supercoiling locations (marked by TopoI peaks) up to the Hi-C resolution (5 Kb in this case). The performance of the baseline model, trained with randomized peak labels, is plotted in black. Highlighted curves represent experiments with the median AUROC and AUPRC, respectively. Experiments with difference training percentages have similar AUROC (0.54-0.58) and AUPRC (0.41-0.44) values, which indicate that our models cannot effectively recover negative supercoiling spatial patterns from contact maps.

(**G**) Feature importance plot showing the average feature importance (over 10 independent experiments) measured in mean decrease accuracy (y-axis) of each feature (x-axis) as colored dots. The simple moving averages (window size = 15) from experiments with 25% (red), 50% (green), and 75% (blue) training data show uniform patterns across different genomic separation distances.

(**H**) The average contact frequencies (y-axis, inversed contact counts representing the apparent spatial separation between two loci) at each signed genomic separation distance (x-axis) are plotted (colored dots) with simple moving averages (colored curves, window size = 15). Different colors indicate the interaction between loci pairs with at least one containing GapR peaks (gP-all, blue) and that with at least one containing no GapR peaks (gNP-all, green). The average contact frequencies at each genomic separation are plotted in red as a reference. The significant difference between gP-all and gNP-all explains the effective prediction of GapR peak locations on *E. coli* genome.

(**I**) The TopoI peak counterpart of (**H**). The absence of a significant difference between tP-all and tNP-all explains the ineffective prediction of TopoI peaks on the *E. coli* genome.

Figure 4.

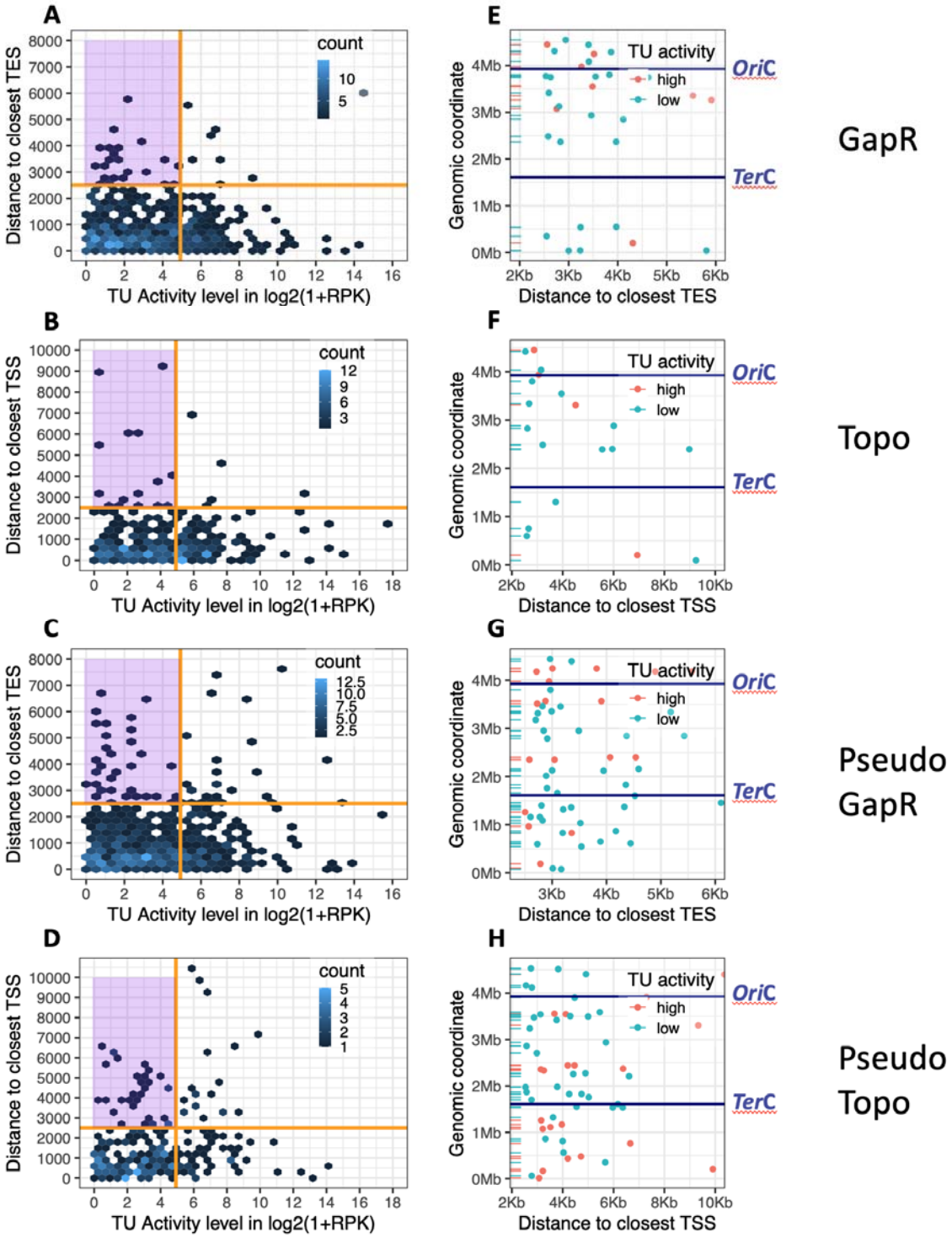


Figure 4. Transcription is a strong driving force for supercoiling generation.
(**A**) 2D-histrogram showing the distribution of GapR peaks (hexagon bins colored by density) with respect to the distance (y-axis) to their nearest upstream TES and the normalized (log2) transcription activity (x-axis, RPK) of that associated TU. GapR peaks on the right to the vertical line in orange (x-intercept =

29.34, the 75th percentile of all TU activities) associate with highly active TUs. GapR peaks on top of the horizontal line in orange (y-intercept = 2500) are more than 2.5 Kb distant from any TUs. The GapR peaks highlighted by a purple block are GapR peaks distant from any TUs and do not associate with highly transcribed TUs, suggesting that they may not be introduced by *E. coli* transcription activities.

(**B**) The TopoI peak counterpart of (**A**).

(**C-D**) Control experiments repeating the analysis in (**A**) by randomly generating a set of pseudo-GapR and pseudo-TopoI peaks. There are more pseudo-peaks lie above the orange horizontal lines compared with peaks in (**A-B**).

(**E**) The group of GapR peaks (colored dots) above the orange horizontal line in (**A**) are plotted to visualize their location along the *E. coli* genome (y-axis) and distance to their nearest TES (x-axis). Two horizontal lines in blue indicate the position of *Ori*C and *Ter*C. GapR peaks are colored in red if they associate with a highly transcribed TU and in green otherwise.

(**F**) The TopoI peak counterpart of (**E**).

(**G-H**) Control experiments repeating the analysis in (**E**) by randomly generating a set of pseudo-GapR and pseudo-TopoI peaks.

Supplemental figures and tables for

Differential Roles of positive and negative supercoiling in compacting the *E. coli* genome

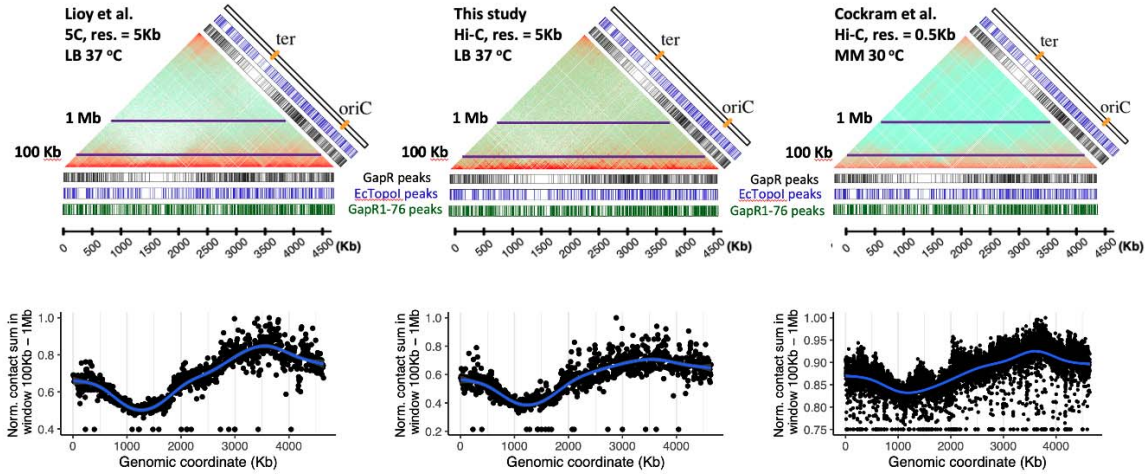
Ziqi Fu[1], Monica S. Guo[2#], Weiqiang Zhou[1#], Jie Xiao[3#]

1. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St., Baltimore, MD 21205
2. Department of Microbiology, University of Washington School of Medicine, 959 NE Pacific St., Seattle, WA 98195
3. Department of Biophysics and Biophysical Chemistry, Johns Hopkins School of Medicine, 725 N. Wolfe St., Baltimore, MD 21205, U.S.A.

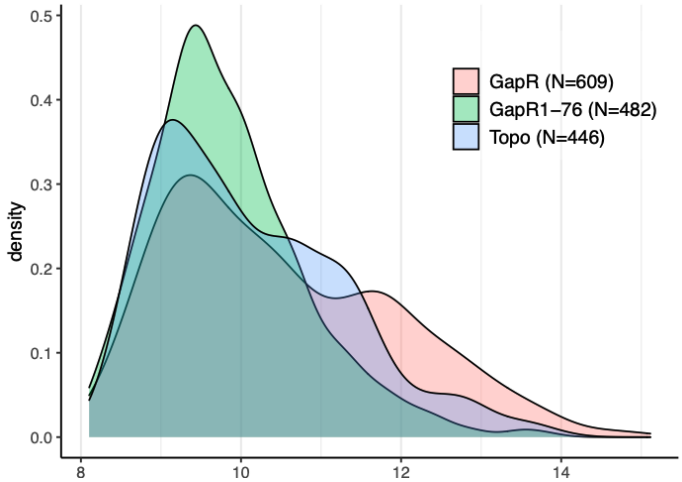
[#]: Correspondences should be addressed to, msguo@uw.edu, wzhou14@jhu.edu, and xiao@jhmi.edu

Figure S1.

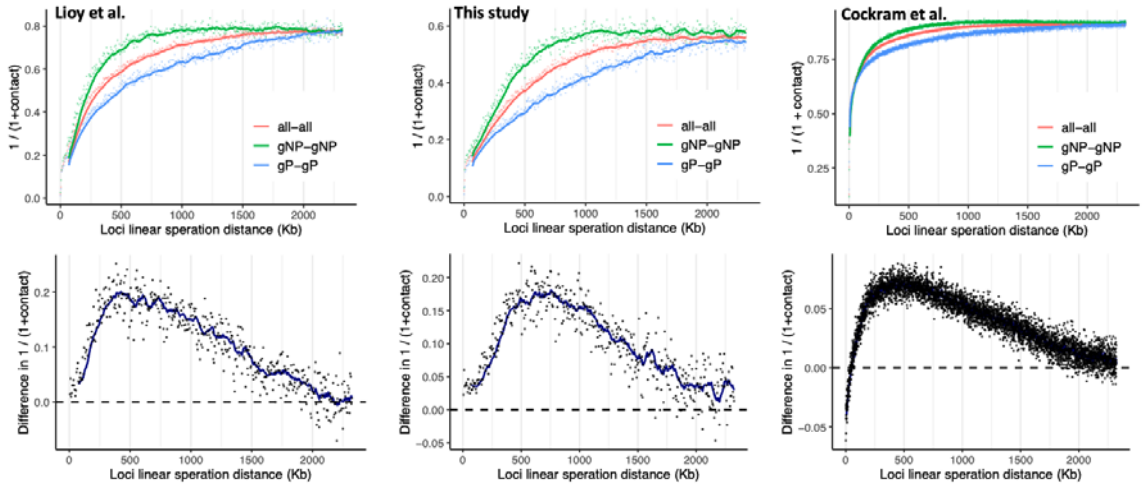
A



B



C



D

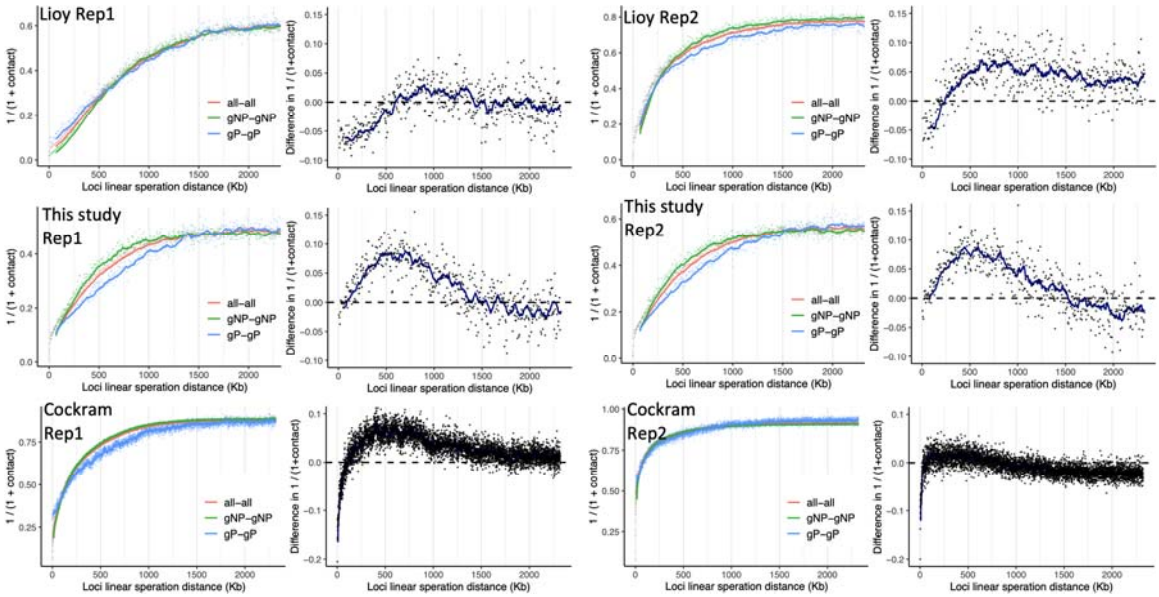


Figure S1.

(A) Replicates of the corresponding Hi-C and 5C experiments illustrated in (**Fig. 1C-E, F-H**) verified the reproducibility of the patterns described in the article.

(B) Density plots for lengths (log2 normalized, x-axis) of identified reproducible GapR peaks (salmon, N=609), GapR$^{1-76}$ peaks (green, N=482), and TopoI peaks (blue, N=446).

(C) Replicates of the corresponding Hi-C and 5C experiments illustrated in (**Fig. 1I-K, L-N**) verified the reproducibility of the patterns described in the article.

(D) GapR$^{1-76}$ counterpart of the analysis illustrated in (**Fig. 1I-K, L-N**).

Table S1.

1A. The *E. coli* genome is divided into 928 5Kb bins and 4642 0.5Kb bins. We identified 609 reproducible GapR peaks and 446 reproducible TopoI peaks (see methods, ChIP-seq analysis). The *OriC* half is defined as the half of the *E. coli* genome centering at the 5Kb or 0.5Kb bin which contains *OriC* (coordinate = 3,925,744, EcoCyc). The *ter* half is defined by the set of bins complement to the *OriC* half.

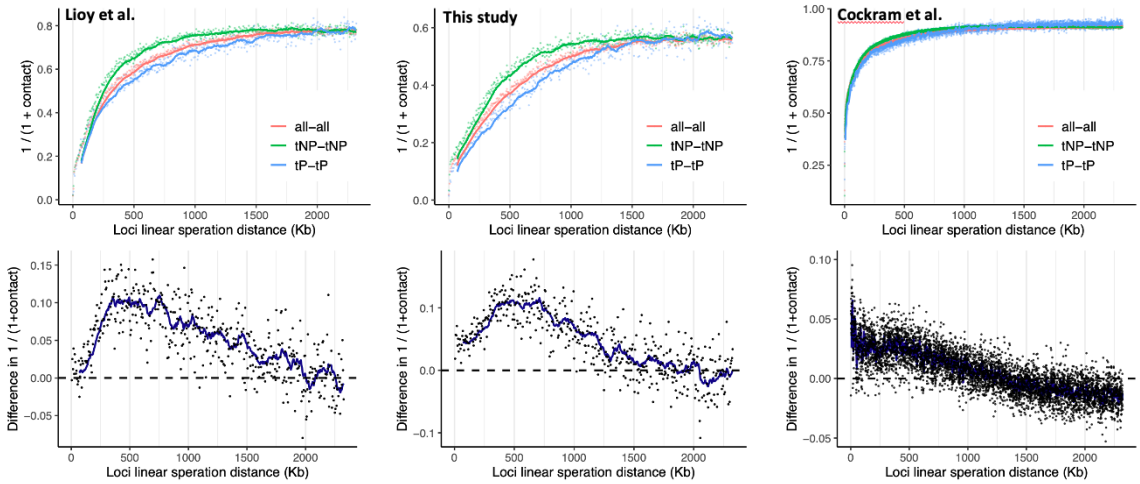
	GapR	Topo
Total peak number	609	446
5Kb bins (928)	454	352
0.5Kb bins (9284)	3672	1890
5Kb bins in *ter* half (464)	147	150
5Kb bins in *OriC* half (464)	307	201
0.5Kb bins in *ter* half (4642)	714	716
0.5Kb bins in *OriC* half (4642)	2958	1174

1B: **Statistics of three sets of identified, reproducible peak widths**

	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
GapR	278	654	1193	2532	3112	**35457**
TopoI	330	558.2	946.5	1647.2	1942	**15616**
GapR$^{1-76}$	257	583.8	820	1201.8	1302.5	**14331**

Figure S2.

A



B

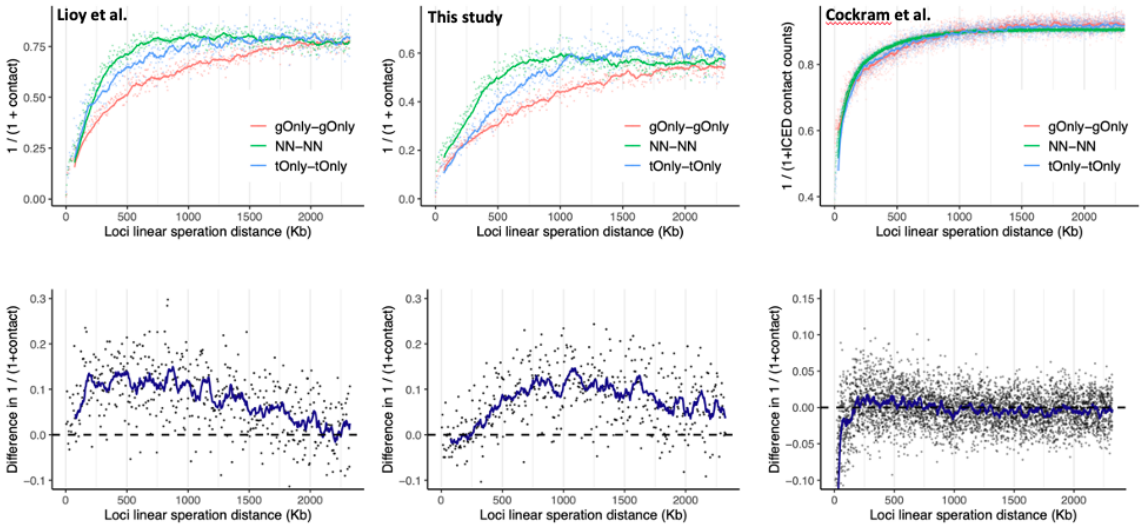


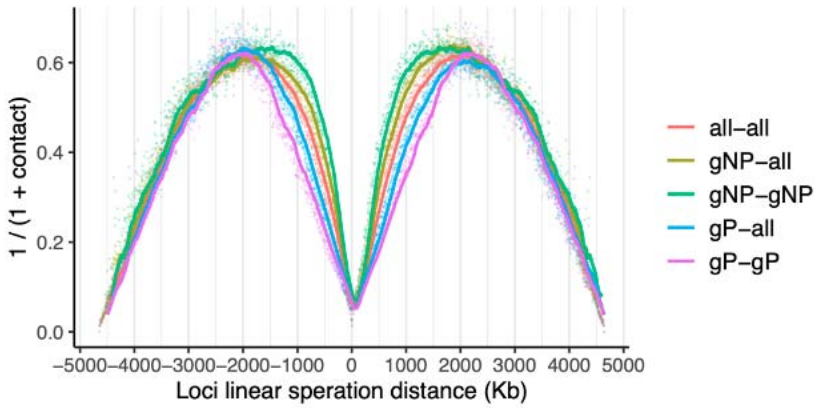
Figure S2.

(A) Replicates of the corresponding Hi-C and 5C experiments illustrated in (**Fig. 2A-C, D-F**) verified the reproducibility of the patterns described in the article.

(B) Replicates of the corresponding Hi-C and 5C experiments illustrated in (**Fig. 2G-I, J-L**) verified the reproducibility of the patterns described in the article.

Figure S3.

A



B

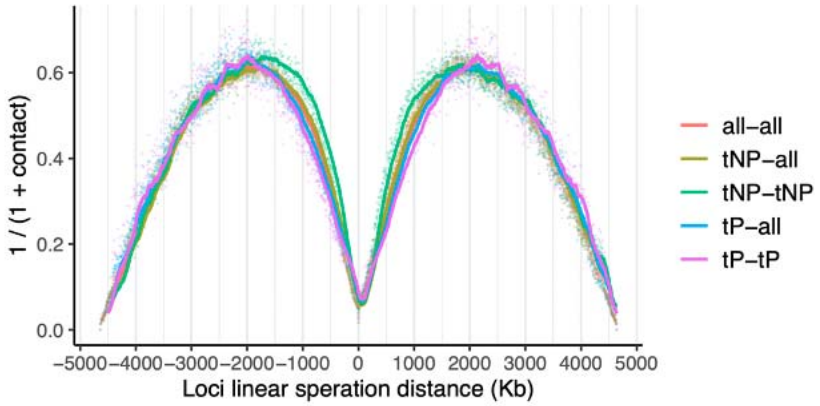
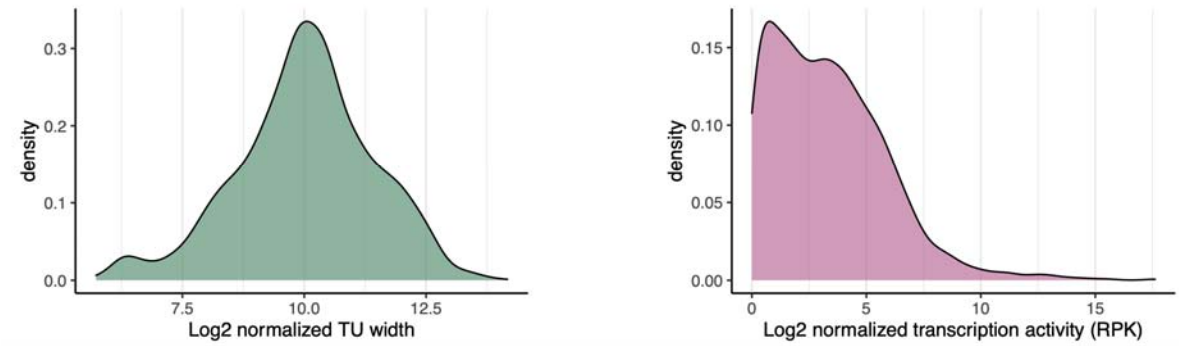


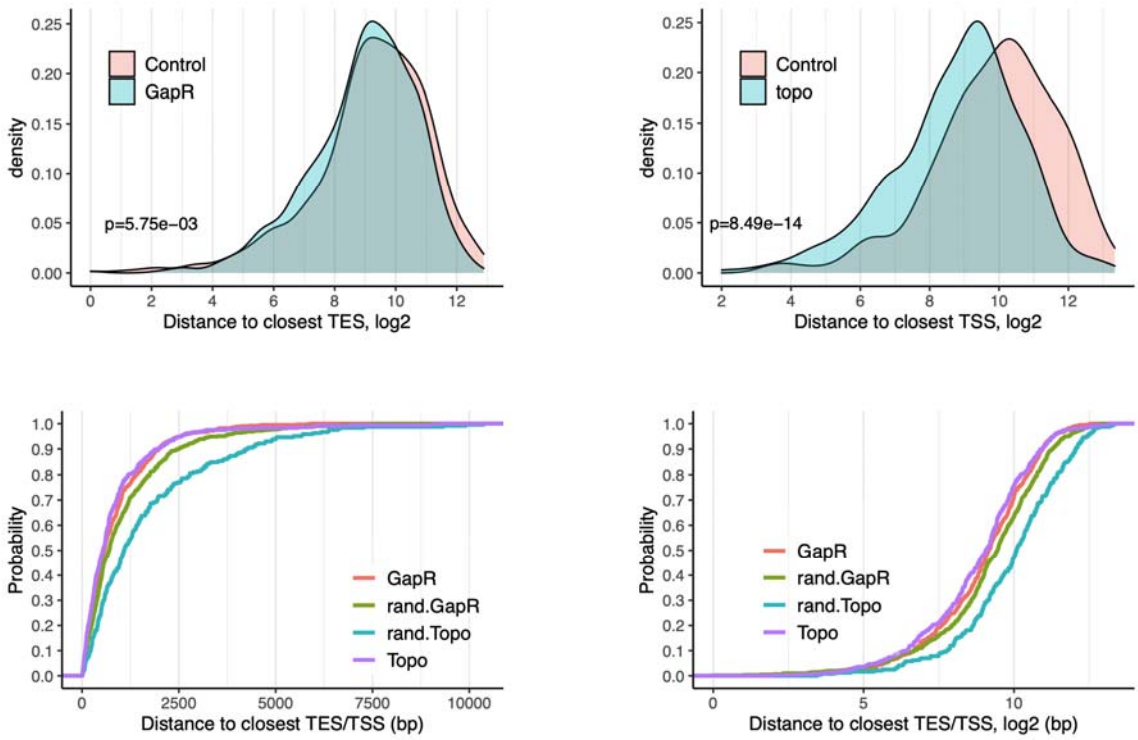
Figure S3.
(A) Average contact frequencies (y-axis, inversed contact counts representing the apparent spatial separation between two loci) at each signed genomic separation distances (x-axis) are plotted (colored dots) with moving averages (colored curves, window size = 15). Different colors indicate the interaction between all loci pairs (all-all, red), when at least one locus does not contain GapR peak (gNP-all, yellow), when both loci do not contain GapR peak (gNP-gNP, green), when at least one locus contains GapR peaks (gP-all, blue), or when both loci contain GapR peaks (gP-gP, pink).
(B) The TopoI peak counterpart of (**A**).

Figure S4.

A


density
Log2 normalized TU width
Log2 normalized transcription activity (RPK)

B


Control
GapR
p=5.75e−03
Distance to closest TES, log2
Control
topo
p=8.49e−14
Distance to closest TSS, log2
Probability
GapR
rand.GapR
rand.Topo
Topo
Distance to closest TES/TSS (bp)
Distance to closest TES/TSS, log2 (bp)

C

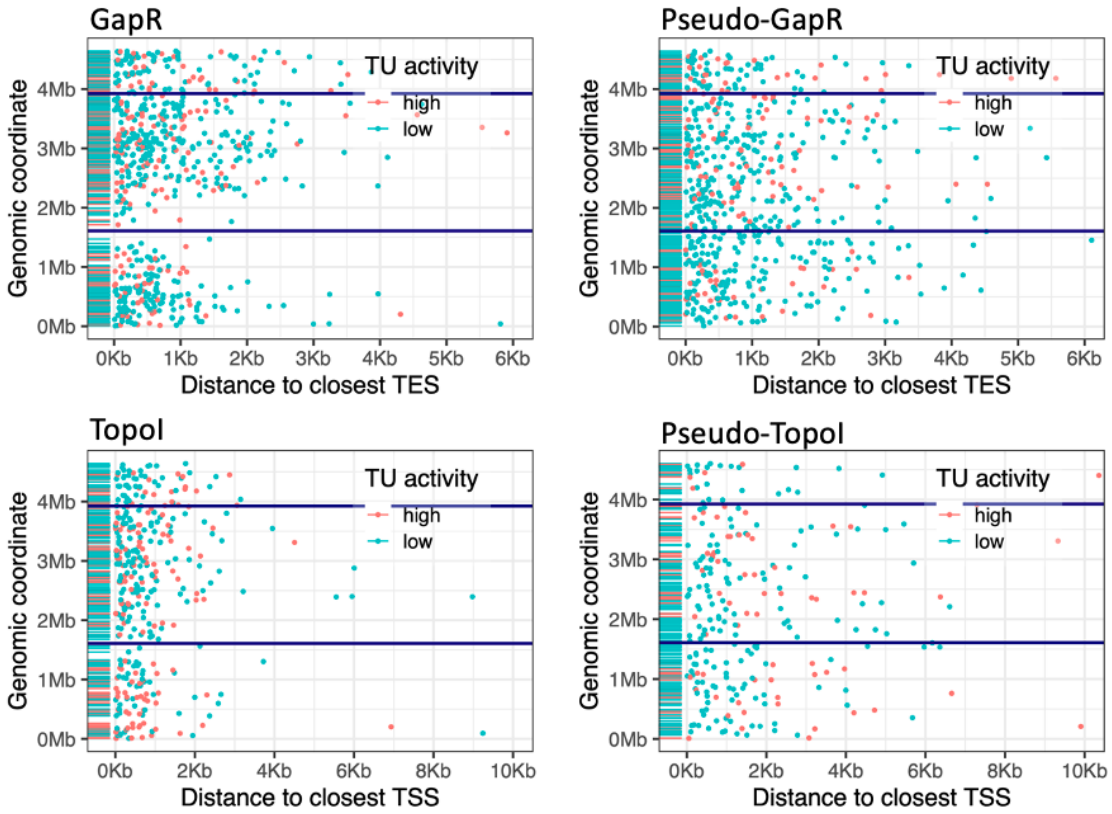


Figure S4.
(A) Density plots of the log2 normalized transcription unit length (left) and activity (right2, measured in Reads Per Kilobase).
(B) Density plots of the distance from GapR v.s. pseudo-GapR peaks to their nearest TES (top left) and from TopoI v.s. pseudo-TopoI peaks to their nearest TSS (top right). The cumulative distribution functions (bottom left) of the four distributions in the above two figures and with log2 normalization (bottom right).
(C) Four sets of peaks are plotted to visualize their location along the *E. coli* genome (y-axis) and distance to their nearest TU (x-axis). Two horizontal lines in blue indicate the position of *ori*C and *ter*. GapR peaks are colored in red if they associate with a highly transcribed TU and in green otherwise.