Supplementary Notes and Figures

Sequence-encoded DNA mechanics regulates binding and catalysis by DNA gyrase

Bailey G. Forbes[1], Rosella Pinckney[1], Aditi Biswas[1], Sisi Liu[1,2], Heather Potter[1], Elizabeth R. Morris[1,*], and Aakash Basu[1,*]

[1]Department of Biosciences, Durham University, Durham, DH1 3LE, United Kingdom.

[2]College of Chemistry, Huazhong Agricultural University, Wuhan, Hubei, 430070, China.

[*]Correspondence should be addressed to Aakash Basu (aakash.basu@durham.ac.uk) and Elizabeth R. Morris (liz.morris@durham.ac.uk).

Contents:

Supplementary note 1: methods and datasets

Supplementary note 1A: gyrase activity assay

To verify that the reconstituted gyrase tetramer assembled from purified GyrA and GyrB subunits was catalytically active, we performed *in vitro* DNA supercoiling. The initial relaxed DNA substrate was the plasmid pSG483, nicked with Nt.BbvCI and ligated to produce a relaxed thermal distribution of topoisomers[1].

Six reactions (27 µl each) were prepared containing relaxed plasmid DNA (~300 ng total) and ATP (2.22 mM), in a final reaction buffer comprising 0.1 mg ml^{-1} BSA**,** 73.3 mM potassium glutamate**,** 30.6 mM Tris–HCl (pH 7.5), 2 mM DTT, 5.56 mM $MgCl_2$, and 10.1% (v/v) glycerol. Gyrase tetramer was included at final concentrations of 55.6 nM, 27.8 nM, 13.9 nM, 7.0 nM, 3.4 nM, or 0 nM (negative control).

Reactions were incubated at 37 °C for 1 h and stopped by the addition of EDTA and SDS (final concentrations of 10 mM EDTA and 1% (w/v) SDS). Stopped reactions were resolved directly on a 1% agarose gel in TAE at 18 V for 16 h, stained with SYBR Gold, and imaged (Fig. S1a). In absence of gyrase, we see the thermal distribution of topoisomers present in the initial relaxed substrate. Gyrase collapses the topoisomers into a single fast-running supercoiled band.

Supplementary note 1B: Processing of raw reads, and supplementary datasets

SELEX

For rounds 1-7 of SELEX, all unique sequences detected in each round, the number of times they occur in each round, and their counts per million reads in each round, can be obtained from the raw reads (NCBI SRA accession number PRJNA1406499). SELEX reads reported here are saved under accession SRX31898394 ("Gyrase Binding SELEX 3 (modified primer)"). For every round, there are two forward read files (for example, for Round 1, these filenames end in "…Round1_forward_reads_part1.fq" and "…Round1_forward_reads_part2.fq") and two reverse read files (again, part 1 and part 2). The "part1" forward read should be merged with the "part1" reverse read, and likewise for part 2. Finally, the merged part 1 and part 2 reads can just be combined. Ultimately, this exercise will yield one combined set of reads for every round 1-7. Individual reads which (i) do not have the left adapter, or (ii) do not have the right adapter, or (iii) have both adapters but do not have exactly 133 bp between the two are discarded. From this set of reads, unique sequences, the number of times they occur, and their counts per million, can be obtained.

SELEX round 0

The initial pre-selection SELEX library (round 0; 133-bp variable region) was archived but was not sequenced. As a surrogate baseline, we sequenced a separately prepared starting library produced using the same oligonucleotide synthesis, amplification, and adapter design workflow, except that it contained a 191-bp variable region. Raw reads are available under SRA accession

SRX31962802 ("Round 0 Gyrase Binding SELEX 3 (Modified Primer)") and comprise one forward and one reverse read file, which were merged prior to processing. Reads were filtered as for rounds 1–7 using the same adapter sequences; the only difference was that a 191-bp variable region was required between adapters. For all round 0 analyses presented in figures, reads were truncated to retain the adapter sequences and the central 133 bp of the 191-bp variable region, thereby matching the length and positional context of the SELEX variable region used in rounds 1–7.

The set of 1,253 sequences in the 167 bp library, the 225 bp library, and the minicircle library

Supplementary Dataset 1 (filename "Scores_1253_sequences.csv") lists all 1,253 sequences in set A (methods) (central 133 bp only), their binding scores in the 167 and 225 bp libraries, and their supercoiling scores. The column names are 'Sequence', 'Binding Score 167 bp', 'Binding Score 225 bp', 'Supercoiling Score 225 bp'. The first row serves as header, storing these column names.

The set of 4,280 sequences spanning the ORF start sites of 40 *E. Coli* genes

Supplementary Dataset 2 (filename 'Scores_4280_sequences.csv') lists the 4,280 additional sequences in the minicircle library (central 191 bp only) and their supercoiling scores. The column names are 'Sequence' and 'Supercoiling Score'. The first row serves as header, storing these column names.

Supplementary note 1C: predicting intrinsic cyclisability from sequence

Reported models in the literature:

Our loop-seq measurements[2,3] have been used by us and several other independent groups to develop and train machine-learning models for predicting intrinsic cyclisability of any given 50 bp DNA sequence[3–8]. In this study, we used the latest reported model (the Park et. al. model[8]), where the reported Pearson's r between predicted and measured values is similar to that between multiple technical repeats of loop-seq measurements itself. Despite this, we verify its predictions against our previously published model (the Basu et. al. model[3]) – see below.

Predicting intrinsic cyclisability:

Prediction models take as input a 50 bp DNA sequence and predicts its intrinsic cyclisability. If the sequence in question is longer than 50 bp, (for example, 167 bp in several cases considered in this study), then each sequence is computationally split into overlapping 50 bp fragments, each fragment offset from the previous by 1 bp. A sequence of length L can be split into L-50+1 such fragments. Intrinsic cyclisability of each such fragment is predicted independently to obtain L-50+1 numbers, which we refer to as the intrinsic cyclisability profile of the sequence. 'Mean intrinsic cyclisability' of the sequence, is defined as the mean of these L-50+1 numbers. When the profile is plotted as a function of position, we typically plot the x-axis starting at 25 rather than 1. This is because we nominally assign the intrinsic cyclisability of the first fragment (which extends from position 1 through 50 of the original L bp sequence) as the intrinsic cyclisability at the middle

of the fragment (position 25 along the original long sequence). The intrinsic cyclisability of the second fragment, likewise is assigned to the position 26, and so on.

Verifying the model against our previous published model:

The Park et. al. model[8] is the best reported thus far: the Pearson's r between measured and predicted values is comparable to the Pearson's r between technical repeats of the loop-seq measurement itself. Our earlier published model (Basu et. al. model[3]) performs worse. However, we decided to compare the Park et. al. with the Basu et. al. model predictions. We observe that the two are correlated, both for totally random 50 bp sequences, and for 50 bp DNA fragments that span the 1,253 sequences in our 167 bp library (pearson's r = 0.6 and 0.7 respectively) (Fig. S1b-c). Finally, we verify that the overall shape and trend in the intrinsic cyclisability profiles of the 10 groups of synthetic DNA sequences (groups 1-10 in set A sequences – see methods and Fig. 4a) are captured in both models (Fig. S1d). All further analysis in this study is done using the Park el. al. model.

Supplementary note 1D: Detailed Methods

Protein expression

GyrA and GyrB proteins were produced and the GyrA-GyrB complex assembled according to a protocol adapted from Vanden Broeck et al 2019[9] and Michaelczyk et al 2023[10].

Plasmids pET28b-EcGyrATWS and pET28b-EcGyrBTWS, produced by V. Lamour (IGBMC and Strasbourg University Hospitals) and gifted by J. Heddle (Durham University), comprised a modified pET28b backbone with the coding sequences for *E. coli* GyrA (2-875) or GyrB (2-875), respectively, inserted between an N-terminal His_{10}-tag and a C-terminal Twin-Strep-tag. GyrA and GyrB proteins were expressed in the T1 phage-resistant *E. coli* BL21 (DE3) derivative strain ER2566 grown in Terrific Broth at 37 °C with shaking. Protein expression was induced by addition of 0.5 mM IPTG to log phase cultures (A600 = 0.6) and the cells incubated for a further 18 h at 18 °C. Cells were harvested by centrifugation and resuspended in 50 mL lysis buffer (20 mM HEPES-NaOH pH 7.5, 300 mM NaCl, 1 mM $MgCl_2$, 10% (v/v) glycerol, 20 mM imidazole, 0.5 mM TCEP, 1:10,000 Benzonase, and 1× EDTA-free mini complete protease inhibitors (Roche) per 10 g of cell pellet).

Protein purification

The cell suspension was lysed by sonication and clarified by centrifugation for 1 h at 50,000×g and 4 °C. The supernatant was subsequently applied to a 5 mL HisTrap FF (Cytiva) column on an AKTA Pure 25 M (Cytiva) held at 2-4 °C in a refrigerated cabinet. The column was washed with 10 column volumes (CVs) HisTrap Wash Buffer (20 mM HEPES-NaOH pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 20 mM imidazole, 0.5 mM TCEP) and step elution applied with HisTrap Elution Buffer (20 mM HEPES-NaOH pH 7.5, 300 mM NaCl, 10% (v/v) glycerol, 375 mM imidazole).

Peak fractions were pooled and loaded at a slow flow rate onto a 20 mL column containing StrepTactin (IBA) resin, followed by washing with 15 CVs StrepTactin wash buffer (20 mM HEPES-NaOH pH 8, 60 mM NaCl, 10% (v/v) glycerol, 1 mM EDTA-NaOH pH 8, 0.5 mM TCEP) and a step elution with StrepTactin Elution Buffer (20 mM HEPES-NaOH pH 8, 60 mM NaCl, 10% (v/v) glycerol, 1 mM EDTA, 2.5 mM desthiobiotin, 0.5 mM TCEP).

Peak fractions were pooled and the His_{10}- and Twin-Strep-tags were cleaved overnight at 4 °C using in-house produced His_6-TEV and GST-3C proteases in a 1:100 protease:protein ratio. Protein was loaded onto two 5 mL HiTrapQ HP (Cytiva) anion exchange columns in series, washed with 20 CVs HGED buffer (50 mM Na-HEPES pH 8, 10% glycerol (v/v), 1 mM EDTA, 0.5 mM TCEP) and a series of increasing 0.1 M NaCl steps were used to elute protein. Peak protein eluted at 0.2 M NaCl was pooled and concentrated to 3-10 mg/mL, snap frozen and stored at -80 °C.

GyrA-GyrB complex assembly and activity assay

GyrA was diluted to 2 μM and GyrB was diluted to 10 μM in Storage Buffer (50 mM Tris-HCl pH 7.5, 100 mM potassium glutamate, 10% (v/v) glycerol, 2 mM DTT) and the two were combined in 1:1 volumetric ratio to achieve a 1:5 stoichiometric ratio of GyrA:GyrB. After incubation for 1 h at 4 °C, the 1 μM GyrA-GyrB complex was snap frozen in liquid nitrogen and stored at -80 °C. We verified that the complex was active by performing *in vitro* DNA supercoiling (supplementary note 1A).

SELEX

We used SELEX to select, over 7 rounds of selection and amplification, the best sequences for gyrase binding from a large pool of approximately 10^{13} unique 133 bp random double-stranded DNA fragments, each flanked by constant 17 bp adapters for PCR amplification. Although this represents only a small subset of the theoretical sequence space for 133 bp DNA, it provides extensive diversity for selection. In each SELEX round, gyrase tetramers were incubated with the DNA library under conditions of limiting enzyme, allowing for the selection of the best bending DNA sequences. DNA bound and unbound to gyrase were separated *via* Electrophoretic Mobility Shift Assay (EMSA) (Fig. 2b). Bound DNA was extracted; a small aliquot was sequenced, and the remainder was PCR-amplified to generate input for the next round of selection (7 rounds in total).

The initial dsDNA random pool was prepared from purchased ssDNA (IDT) with a central random 133 nt region (during synthesis, all bases were present in roughly equal proportions), flanked by constant 17 nt adapters (left adapter: 5'-ATTGCCGTCCGTACCGT-3'; right adapter: 5'-TTGGGTGCGCGAAACGA -3'). ssDNA was chemically synthesised and PAGE purified (IDT). A single cycle of PCR-based strand elongation with Phusion High Fidelity polymerase (NEB) in presence of just the reverse primer was carried out to produce a dsDNA library, which served as the initial library for SELEX rounds. (4 50 ml PCR reactions, each containing 75 pmol of template DNA and 8 mM reverse primer, were subject to 3 min at 98 ºC (melting), 2 min at 60 ºC (annealing), and 8 mins at 72 ºC (elongation). All 4 reactions were purified and DNA eluted in a

total of 100 ml of Elution Buffer (10 mM tris-Cl pH 8.5), yielding a final concentration of ~160 ng/ml).

For each SELEX round, gyrase tetramer (in tetramer storage buffer) was incubated with the PCRed dsDNA library (amount per round of SELEX (1-7): 35, 17, 17, 10, 10, 4, 4 pmol) at a DNA:tetramer ration of 10:1 for rounds 1-5, and 5:1 for subsequent rounds, in a total reaction volume of 100 µl for rounds 1-3, 30 ml for rounds 4-5, and 10 ml for rounds 6-7, in presence of 1x gyrase binding buffer (50 mM tris-Cl, pH 7.5, 55 mM KCl, 10 mM $MgCl_2$,5 mM DTT, 5% glycerol), for 1 hour at room temperature. Excess DNA forces the DNA sequences to compete among each other for enzyme binding. Electrophoretic Mobility Shift Assay (EMSA) was performed to separate bound and unbound fractions using a 6% TBE gel in 1X TAE buffer supplemented with 10 mM $MgCl_2$ for 2 h at 45 V (Fig. 2b). Gels were stained with SYBR gold, the bands corresponding to the bound fraction were excised, and DNA extracted from it using crush and soak. Gel slices were crushed and soaked in 300µl of elution buffer for rounds 1-3 and 200ml for rounds 4-7. The gel slices were left shaking at 200 rpm at 37 ºC overnight. The tubes were centrifuged for 1 min and the supernatant extracted. This supernatant was used as template for PCR using KAPA HiFi polymerase (Roche) to amplify the DNA for the next round of SELEX. Each such PCR reaction utilised 23 µl of this supernatant as template and the number of cycles used was 12 (rounds 1, 4) or 13 (rounds 2,5,6 and 7). The number of such PCR reactions carried out was empirically determined based on the amount of product from one test reaction (it turned out to be 4 for round 1, 2 for round 2, and 1 for rounds 3-7). Each PCR reaction was purified into 1 spin column (Qiagen) and eluted in 50 ml of Elution Buffer. The PCR conditions for these reactions were 95°C 20 s (denaturation), 52°C for 15 s (annealing), 72°C for 19 s (elongation). The amplified DNA was used for the next round of SELEX and 20 nano grams of amplified DNA was used in subsequent PCR reactions to prepare the same for Illumina Sequencing using the Nextera XT index kit, following library preparation sequencing methods for 16S metagenomic sequencing (Illumina).

Library design and preparation for competitive gyrase binding and supercoiling assays

A core set of 1,253 133 bp DNA sequences (set A) representing various categories were flanked by the constant 46 nt adapter sequences 5'-GCCGGCTATCCATTCCAGGAGCTAACACAATTGCCGTCCGTACCGT-3' and 5'-TTGGGTGCGCGAAACGACATCGGTACGTACCCAGTGAATCCCGTAA-3'. This yielded 1,253 different 225 nt long molecules, with the central 133 bp variable (but specified) while the flanking 46 nt constant.

When set A sequences were PCR amplified using primers that target the inner 17 nt of the 46 nt adapter sequences for the 1,253 molecules, it yields a "167 bp library" of 1,253 molecules. These molecules had the central 133 bp variable, and constant outer 17 bp. This resultant library was used for competitive gyrase binding assays.

When set A sequences were amplified using primers that target the outer 17 nt (which are common to all 1,253 + 4,280 sequences), it results in a "225 bp library". These molecules again had the central 133 bp region variable, and the flanking 46 bp regions constant. This resultant library was also used for competitive gyrase binding assays.

An additional 4,280 191 nt DNA sequences (set B) taken from around the ORFs of 40 genes in *E. coli* were each flanked by the constant 17 nt adapter sequences representing the outer 17 nt of the previous set of 225 bp 1,253 DNA sequences (5'-GCCGGCTATCCATTCCA-3' and 5'-CCCAGTGAATCCCGTAA-3'). This set was mixed with setA, yielding a total of 5,533 225 nt long molecules. Of these, 1,253 molecules (i.e. set A) had the central 133 nt variable and the flanking regions constant, while 4,280 molecules had the central 191 nt region variable and the flanking regions constant. This combined set A plus set B sequences were PCR amplified with primers that target the outer 17 nt (common to set A and set B), along with additional overhangs (see below) to generate a minicircle library for competitive supercoiling assays.

Competitive gyrase binding assay

To measure binding propensity of individual DNA sequences, we performed competitive gyrase binding assays separately on the 167 bp library and the 225 bp library. Gyrase was incubated with the dsDNA library under low enzyme concentrations, allowing the different DNA sequences to compete for binding. Bound and unbound DNA fractions were separated by gel electrophoresis, and both pools were deep-sequenced. We define binding score of a sequence as the log of the ratio of its counts in the bound to that in the unbound pool.

The initial ssDNA library (set A) was PCR amplified (10 ng template, 2 mM of each primer, using KAPA HiFi polymerase (Roche), for 14 cycles (same conditions as SELEX PCR), followed by elution in 50 ml Elution Buffer, yielding a typical final product concentration of ~100 ng/µl). ~4 pmols of PCR amplified DNA was mixed with 0.8 pmol of gyrase tetramer to achieve a 5:1 DNA:gyrase ratio during binding. Binding was carried out in a final volume of 10 µl (50 mM Tris-HCl pH 7.5, 55 mM KCL, 10 mM $MgCl_2$, 5 mM DTT, 5% glycerol) and incubated for 1 hour at room temperature. DNA-gyrase binding was allowed to proceed under low enzyme concentrations to promote competitive interactions among the library sequences. After incubation, bound and unbound DNA fraction bands were separately excised from a gel, followed by crush and soak of each in 200 µl Elution Buffer, as was done in the SELEX round 7. Two 50 µl PCR reactions were set up for the bound and unbound fractions, each utilizing 23 µl of the extracted supernatant from crush and soak, and all other conditions being identical to the PCR in SELEX, except for cycle number (9 for free DNA band and 13 for bound). Each was eluted in 50 µl Elution Buffer, and yielded a final concentration of ~60 ng/µl in each case. Library preparation for sequencing was done as for SELEX.

Competitive gyrase supercoiling assay

To measure supercoiling propensity, we first converted the linear DNA library into minicircles, incubated it with gyrase in absence of ATP, added ATP briefly to allow supercoiling, and separated the supercoiled form relaxed DNA via EMSA. Deep-sequencing of both pools allowed us to calculate supercoiling score of every sequence, defined as the log of the ratio of the relative population of the sequence in the supercoiled pool to that in the relaxed pool.

To produce minicircles, following a previously published protocol[11]. We started with the combined set A plus set B library and PCR amplified it using KAPA HiFi polymerase (Roche) using primers (5'-ATCGCAGGTCTCAGTACGGCCGGCTATCCATTCCA-3' and 5'-ATAAGCGGTCTCAGTACGTTACGGGATTCACTGGG-3') that bind to the outer 17 nt, and additionally add overhangs for subsequent restriction digestion and generation of sticky overhangs as described[11]. 15 50 µl PCR reactions were set up using 0.15 µl of template library per reaction, for 16 cycles. PCR conditions were the same as SELEX, except for 28 s of elongation. Each PCR reaction was purified and eluted in 50 µl Elution Buffer, yielding ~750 µl of total amplified library at a final concentration of ~170 ng/µl. The linear DNA was digested with BsaI (NEB) to produce 4 nt sticky ends, which were subsequently ligated using T4 DNA ligase in the same reaction, as described[11]. This creates 231 bp DNA minicircles. T5 exonuclease was used to digest non-circular DNA. Minicicles were gel-purified on a 1% agarose gel.

Purified formed relaxed circles (~1.3 pmol) were incubated with ~1.3 pmol gyrase tetramer in 1X gyrase supercoiling assay buffer without ATP (25 mM Tris-HCl pH 7.5, 40 mM potassium glutamate, 5 mM $MgCl_2$, 2 mM DTT, 0.1 mg/ml BSA, and 10% glycerol) in a total volume of 18 µl for 1 hour at room temperature. 2 µl of 20 mM ATP was added and the reaction transferred to 37°C for 10 s to allow supercoiling, followed by the addition of 2 µl of stop buffer (final concentration after adding stop buffer: 10 mM EDTA and 1% SDS). DNA was purified via ethanol precipitation, eluted in 20 µl of Elution Buffer, and resolved on a 5% polyacrylamide gel prepared in TB buffer (90 mM Tris base, 90 mM boric acid, and 10 mM $MgCl_2$). The gel was run at 80 V for 90 min using a matching running buffer and stained with SYBR gold. The bands corresponding to the relaxed and supercoiled fractions were separately excised, the DNA in each was extracted via crush-and-soak in 200 µl Elution Buffer as done for SELEX. 23 µl of this (in each case) was used to set up a 50 µl PCR reaction using KAPA HiFi polymerase (Roche) and primers that matched the 17 nt left and right adapters of the original 225 nt linear DNA library (not the primers used initially to prepare the DNA for circularisation), for 15 cycles, with conditions matching SELEX except for 28 s of elongation. Purification and elution in 50 µl Elution Buffer yielded typical concentrations of 90-100 ng/µl for both the supercoiled and relaxed bands. Each DNA sample was separately prepared for Illumina sequencing as done for SELEX. Sequence-specific supercoiling scores were calculated analogously to the binding scores, using the natural logarithm of the ratio of sequence counts in the supercoiled fraction to those in the relaxed fraction, providing a quantitative measure of gyrase supercoiling across the library.

<u>Sequencing and data analysis</u>

Each round of SELEX was sequences at a depth of 6 gigabases. Bound and unbound or supercoiled and relaxed fractions in competitive binding/ supercoiling assays were each sequenced at a depth of 0.5 gigabases. Bidirectional sequencing for 150 cycles was carried out on an Illumina NovaSeq X Plus machine (Novogene). Reads were merged using fastp. For each round of SELEX, merged reads were first filtered for only those reads that contained the left adapter and the right adapter, and exactly 133 or 191 nt between them. Following this, unique reads were identified from each round of SELEX and the counts per million (CPM) of each unique read was calculated. Reads were sorted according to decreasing CPM.

Predicting intrinsic cyclisability

To predict intrinsic cyclisability of any given 50 bp sequences, we use the latest reported machine-learning model trained on the basis of millions of loop-seq measurements[8]. Nevertheless, we confirm that the model predictions correlate with our previously published model[3] (supplementary note 1C). See also supplementary note 1C for explanations of how the predictive models are applied to predict intrinsic cyclisability along given long sequences.

Supplementary note 2: intrinsic cyclisability in SELEX reads

Supplementary note 2A: plotting details for Fig. 2a

The sequences of pSC101, the Mu phage genome, and PBR322, were obtained from NCBI. The reported gyrase binding sites were present between coordinates 4580 – 4864, 17,765 – 18050, and 846 – 1126. The DNA sequence from 25 bp upstream of this window to 25 bp downstream was extracted and split into overlapping 50 bp fragments, each offset from the previous by 1 bp. Intrinsic cyclisabilities of each of these fragments were obtained via the Park et. al. model[8] (supplementary note 1C). Plotted are these intrinsic cyclisabilities. For any 50 bp fragment that extends from position n to n+49 along the genome or the plasmid sequence, its intrinsic cyclisability was assigned as the intrinsic cyclisability of DNA at position n+25 along the genome or plasmid sequence.

Supplementary note 2B: plotting details for Fig. 2c

SELEX reads were all 167 bp long, comprising the left and right adapters (see methods) flanking the central variable 133 bp regions. SELEX round 0 was archived but not sequenced; enrichment analyses for round 0 used a separately prepared starting library processed identically (see supplementary note 1B). For every round, each 167 bp sequence read was split into 118 50 bp DNA fragments offset from each other by 1 bp. Intrinsic cyclisability values of each of these fragments was obtained using the Park et. al. model (supplementary note 1C). The predicted values were organised into a matrix of $N_{reads} \times 118$, where N_{reads} is the number of reads for that round (Table 1 in the main text). Each row corresponds to one read and each column to one 50 bp fragment position along that read. For each SELEX round, the column-wise mean of this matrix was calculated and plotted to generate the average flexibility profile across all reads for that round. The x-axis positions for each of the 118 columns were assigned numerical values beginning with 25, increasing by 1 bp per column. In other words, for any 167 bp DNA sequence, the intrinsic cyclisability of its first 50 bp fragment (positions 1–50) is assigned as the intrinsic cyclisability of this overall longer DNA sequence at position 25, and so on.

The assumed footprint of the two CTD is obtained from the recent cryo-EM structure of the gyrasr holoenzyme (PDB: 9GBV)[12]. Based on the structure, we determined the centre of the DNA fragment (approximate centre between the two GyrA active Y122 tyrosines) at position 20 along chain E. Likewise, we determined visually that DNA along chain E between positions 4 and -37 was associated with the CTD.

Supplementary note 2C: control related to Fig. 2c

To check whether the bimodal pattern of intrinsic cyclisability profile along SELEX-enriched DNA reads (Fig. 2c) were dominated by the profiles of a few high-copy number sequences in the reads, we repeated the analysis for unique sequences only for every round of SELEX. This time, for every round, we used only the set of unique sequences from among our reads to carry out the analysis in supplementary note 2B (results in Fig. S2a). The near-identical pattern of flexibility

profiles to that in Fig. 2c suggest that the dual-peaked profile it is not dominated by the intrinsic cyclisability profiles of just a few high-copy number sequences, but that unrelated families of selected strong gyrase binders share this mechanical property.

Supplementary note 2D: plotting details for figure 2d

Each of the 1,585,271 unique 167 bp sequence from SELEX round 7 was divided into 118 overlapping 50 bp fragments, offset by 1 bp relative to the previous fragment. The intrinsic cyclisability of each fragment was predicted using the Park et al model. For each sequence, the mean intrinsic cyclisability of the first 59 fragments ("left half") and that of the last 59 fragments ("right half") were calculated and the difference (right minus left) taken. Sequences were sorted according to this difference. A matrix of intrinsic cyclisability values of these sorted sequences was constructed, where each of 1,585,271 rows corresponds to a sequence, and each of 118 columns stores the intrinsic cyclisability of that 50 bp fragment along that sequence. This matrix was binned along the row dimension into 2,000 equal-sized bins by column-wise averaging consecutive rows within each bin. Fig. 2d shows a heatmap representation of this binned matrix.

Supplementary note 2E: control related to figure 2d

When obtaining Fig. 2d, we have defined a metric for a sequence (difference between mean intrinsic cyclisability on either half of the sequence) and sorted the sequences according to this metric (see supplementary note 2D). To verify whether the apparent asymmetry see in Fig. 2d is a sorting artifact, we created a dummy matrix of intrinsic cyclisability values by taking the matrix used to generate the heatmap in Fig. 2d and randomly permuting the values along rows at every column. The column-wise mean of this permuted matrix is unaltered from that of the original matrix, but its heatmap (Fig. S2b) when sorted in the same way, shown no apparent sequence-level asymmetry as in Fig. 2d.

Supplementary note 2F: plotting details for Fig. 2e

To analyse the symmetry of DNA flexibility profiles among gyrase-binding sequences, we used all the 1,585,271 unique sequences from the 7th round of SELEX. Each sequence was 167 bp long, comprising the constant left and right adapters flanking a 133 bp variable region. Each 167 bp sequence was divided into 118 overlapping 50 bp fragments, offset by 1 bp relative to the previous fragment. The intrinsic cyclisability of each fragment was predicted using our neural-network–based model. The resulting vector of 118 intrinsic cyclisability values was smoothed using a Gaussian filter with a standard deviation $\sigma = 2$ bp to suppress noise while preserving the broad shape of the profile. Smoothing was performed using python scripts using scipy.ndimage.gaussian_filter1d from SciPy 1.10.0. Peaks were identified in each smoothed profile using the python scripts that call the function scipy.signal.find_peaks from SciPy 1.10.0, with the height threshold set to the mean + one standard deviation of that profile. This algorithm can detect any number of peaks along the profile that satisfy the threshold—there was no constraint

on the number of peaks per side. Detected peak positions were then partitioned into those lying before the midpoint, and those after the midpoint, of the 167 bp DNA region. Sequences containing at least one left and one right peak were labeled *"both"*. Sequences with only left-side peaks were labeled *"left"*. Sequences with only right-side peaks were labeled *"right"*. Sequences with no detected peaks were labeled *"none"*.

Profiles were grouped by these four categories. For each class, the mean and standard deviation of intrinsic cyclisability were computed at every position across the 118 columns, producing an average ± SD profile for that class. These 4 profiles for the 4 classes are plotted in Fig. 2e

Supplementary note 2G: control related to Fig. 2e

Fig. 2e focused on classifying the different unique sequences in SELEX round 7. We confirmed that similar profiles are obtained if the entire set of reads (all 17,211,837 reads, not just the 1,585,271 unique sequences, see table 1 of the main text) were used for the classification process. See Fig. S2c.

Supplementary note 3: other sequence features in SELEX reads

Supplementary note 3A: GC content oscillations

Fig. 3a presents mean GC content as a function of position, averaged over all reads in every round. To ensure that the pattern is not dominated by a few highly-enriched sequences (which have high copy numbers, especially in later rounds), we re-calculated the plot, but this time, averaging only over the unique sequences in each round (Fig. S3a).

Having confirmed that presence of oscillations in the mean GC content as a function of position, we sought to determine whether the extent of oscillations are expected on the basis of random noise. We scrambled each unique sequence, maintaining every sequence's overall base composition, and plotted the mean GC content as a function of position, averaged over the scrambled unique sequences of round 7. The oscillations significantly reduce in magnitude, establishing the baseline oscillations for the given overall base composition of the individual sequences (Fig. S3b), and indicating that the oscillations seen in Fig. 3a reflect an underlying preference for individual sequences in later rounds for G/C vs A/T are specific positions.

Supplementary note 3B: Details of how power spectrum in Fig. 3b was calculated

For each SELEX round, we first computed mean GC content as a function of position across the 133 bp variable region of each read. This yielded a single one-dimensional trace $g(i)$ of length $L = 133$, where $g(i) \in [0,1]$ is the fraction of reads in that round which have a G or C at position i (i.e. the mean GC content at that position). We then quantified periodic structure in this mean profile using a simple Fourier power spectrum as follows:

The GC-fraction trace $g(i)$ was mapped to a symmetric $[-1, +1]$ scale via $y(i) = 2g(i) - 1$,

and then mean-centred to remove the DC component: $y(i) \leftarrow y(i) - \frac{1}{L}\sum_{i=0}^{L-1} y(i)$

To reduce spectral leakage due to finite-length trace, we multiplied by a Hann window $w(i)$:

$$y_w(i) = y(i)w(i)$$

with

$$w(i) = \frac{1}{2}\left(1 - cos\left(\frac{2\pi i}{L-1}\right)\right)$$

We also computed the window normalisation constant

$$U = \frac{1}{L}\sum_{i=0}^{L-1} w(i)^2$$

which corrects the expected reduction in energy signal due to windowing.

We then computed the one-sided real Fast Fourier Transform ($rFFT$) of the windowed signal, $X(k) = rFFT(y_w)$ for frequency indices $k = 0, 1, 2, \ldots, \left\lfloor \frac{L}{2} \right\rfloor$. The nomalised power spectrum was then

$$P(k) = \frac{|X(k)|^2}{L^2 U}$$

We excluded the 0-frequency bin (k=0) from downstream analysis.

The x-axis in Fig 3b indicated period, not frequency. Using a sample spacing of 1 bp, the Fourier frequencies are $f(k) = \frac{k}{L}$ cycles per bp. We convert to periodicity in bp as

$$T(k) = \frac{1}{f(k)} = \frac{L}{k}$$

again omitting $k = 0$ (for which $T = \infty$).

Supplementary note 3C: phasing and categorical spectral analysis

The average of individual non periodic sequences can be periodic:

We present below a simplified explanation for why phased but non-periodic sequences can be periodic upon averaging. The following sequences S_1 – S_4 have no periodicity individually:

S_1 = [0,0,4,0,0,0,0,0,0,0,0,0,0,0,0,0]

S_2 = [0,0,0,0,0,0,0,4,0,0,0,0,0,0,0,0,0]

S_3 = [0,0,0,0,0,0,0,0,0,0,0,0,4,0,0,0,0,0]

S_4 = [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,4]

However, the average sequences S_{mean} (column-wise average at each position) is:

S_{mean} = [0,0,1,0,0,0,0,1,0,0,0,0,1,0,0,0,0,1]

S_{mean} has a periodic structure.

As a biologically relevant example, when the average GC content as a function of position for the identified 147 bp nucleosomal DNA sequences from *S. cerevisiae* is plotted, clear periodic oscillations in GC content as a function of position is seen and confirmed *via* Fourier analysis of the averaged trace[13]. However, recent analysis has confirmed that individual 147 bp nucleosomal DNA sequences do not possess any statistically significant periodicity, and that the periodicity of the average signal is purely due to proper phasing.[14]

Categorical spectral analysis

To analyse whether individual reads from SELEX have a periodic oscillation in GC content as a function of position, we instead must take the Fast Fourier Transforms of individual sequences first, before averaging the individual power vs periodicity plots of the individual reads. This is distinct from averaging the individual sequences first, and the obtaining a single power vs frequency plot. Mathematically, the two operations of Fourier transform and averaging do not commute, but this has been the source of much confusion in the literature[14].

To analyse periodic patterns in the base composition of DNA sequences, we performed Fourier analysis. Taking the Fourier transform of a DNA sequence inherently requires some strategy for converting base composition into numeric data. The exact numeric representation (for example, A/T = 0 and G/C = 1) affects the calculated power spectral density[14]. We therefore performed categorical spectral analysis which treats individual sequences as a sequence of various 'categories' (A/T or G/T), and numerically represents categories in an unbiased manner.

We performed categorical (Voss-type) spectral analysis separately for the AT and GC base categories for every given sequence. Every sequence is first converted into two indicator binary traces of the same length: an AT indicator trace, where positions corresponding to A or T are assigned a value of 1 and all others 0, and a GC indicator trace, where positions corresponding to G or C are assigned a value of 1 and all others 0. All analyses were performed in Python 3.10, using NumPy (v1.24.3), SciPy (v1.10.0)**,** and Matplotlib (v3.7.1).

Each indicator trace was detrended by subtracting its mean value, ensuring that spectral power reflects oscillatory variation rather than base composition bias. To suppress edge artifacts, a Hann window was applied prior to the Fourier transform (as described in Supplementary Note 3B).

After detrending and applying a Hann window (supplementary note 3B), the indicator traces for AT and GC bases were converted from the spatial domain to the frequency domain using the Fast Fourier Transform (FFT). The squared magnitudes of the resulting complex Fourier coefficients were used to compute the power spectrum for each sequence, reflecting the relative strength of periodic A/T or G/C patterns along the DNA.

For every sequence, the total categorical power spectrum was calculated as the sum of the individual A/T and G/C power spectra at each frequency. The total categorical power spectra from all sequences were averaged frequency-wise to obtain the quantity plotted as 'Power' in Fig. S3c. It indicates that individual sequences also possess 10 bp periodic structure, which is enriched in subsequent SELEX rounds.

Supplementary note 3D: periodicity in WW dinucleotides

Earlier analysis of nucleosomal DNA sequences revealed enrichment of WW dinucleotides (AA/TT/AT dinucleotides) at 10 bp periodicity (which leads to an overall periodic GC content structure)[13,15]. We also observe phased periodic WW dinucleotides in the average trace of mean WW nucleotide content as a function of position, averaged over all reads in every round (Fig. S3d).

Supplementary note 3E: GC oscillations on either half of the molecules

In the average trace in Fig. 3a, the oscillations in GC content are more prominent visually on the "left" half of the molecule. We first verified that this is not owing to individual sequences being periodic only on one half of the 133 bp region.

Plotted in Fig. S3e is power vs periodicity (obtained via categorical spectral analysis over individual sequences followed by averaging individual power spectra) for SELEX round 7, but only either the left 60 bp (positions 1-60 along the 133 bp region) and right 60 bp (positions 74 - 133 along the 133 bp region) were considered (separately). We do find strong ~10 bp periodic oscillations in either set.

As discussed in the main text, the protein:DNA interaction is expected to be symmetric. Therefore, the only effect that could break the left-right symmetry and lead to visually less prominent oscillations in the mean trace in Fig. 3a is the fact that the two constant primer-binding sequences flanking the 133 bp variable region are not the same. It is possible that the phase imposed by the "left" adapter better matches the phase imposed by the periodic lysines/ argenines, than the phase imposed by the right adapter.

Supplementary note 3F: asymmetries in GC oscillations and intrinsic cyclisability

We find that sequences classified to have flexibility peaks on the "left" half also had periodic GC content oscillations restricted to the first 60 bp only. Likewise for sequences classified to have flexibility peaks on the right half. See figure S3f.

Supplementary note 3G: analysis of GC-matched random sequences and cleavage sites

GC-matched random sequences were generated from the reads in round 7 of SELEX. For every read, we constructed a counterpart read where we replaced Gs or Cs randomly with Gs or Cs,

likewise for As and Ts. Thus, every counterpart read matches the G/C distribution of the SELEX read used to generate it, but was otherwise random. These counterpart reads were again flanked by the same 17 nt left and right adapters. Mean intrinsic cyclisability along these sequences was plotted in Fig. 3d (plotting details similar to what was done in Fig. 2c).

For the cleavage sequences, all previously reported 4,627 Cfx-induced gyrase cleavage sites along E. *coli*[16] were considered, and the 133 bp DNA sequence flanking these sites were extracted. These 133 bp sequences were then flanked by the same 17 nt left and right adapters to generate 167 bp sequences. Mean intrinsic cyclisability as a function of position was plotted for these sequences in Fig. 3d (plotting details similar to what was done in Fig. 2c).

Supplementary note 3H: *in silico* evolution of DNA sequences from a random seed

We performed *in silico* evolution to generate a DNA sequence whose intrinsic cyclisability profiles matches some given profile. We start with a totally random 133 bp DNA sequence, flanked by the same 17 bp adapters as used for SELEX (total length = 167 bp). We consider all possible single point mutation variants in the central 133 bp region of this sequence (there are 399 such: 3 possible point mutations at each of 133 positions), calculate the intrinsic cyclisability profiles of each of these variant sequences (by splitting the sequence into overlapping 50-mers and predicting their intrinsic cyclisabilities), and select the one with the least RMS deviation from the profile we are trying to match. We then consider all possible single point mutations of this selected sequence and repeat the process. This iteration, comprising all possible single point mutation followed by selection of the best, was carried out *in silico* 15 times.

Supplementary note 3I: power spectra of *in silico* DNA sequences

Figure 3g shows the power spectrum of mean GC content as a function of position, averaged over SELEX sequences and over the two groups of evolved sequences. Here we confirm that even individual sequences within the two groups of evolved sequences (as opposed to the mean sequence) show far reduced periodic oscillations. We performed categorical spectral analysis on each of the 9,900 *in silico* evolved DNA sequences from each group (‘evolved mean’ or ‘evolved indv’, as explained in the main text). The power spectra (power vs periodicity plot) of each individual sequences were averaged across all sequences in each group. See Fig. S3g. This average power spectra of each evolved group shows far reduced peak at 10 bp periodicity when compared to SELEX sequences.

Supplementary note 4: (lack of) SELEX-enriched motifs

Supplementary note 4A: Attempts at identification of SELEX-enriched motifs

To identify sequence motifs enriched during SELEX, we performed a comprehensive enrichment analysis of all 10 bp sequences comparing the final selected pool (round 7) against at early pool (round 1). For each read, we extracted all overlapping 10-mers using a sliding window (124 windows per 133 bp sequence). Separately for round 7 and round 1, this yielded a vector of length 4^{10} containing the counts of all possible 10-mers, from which the frequency of each of the possible 4^{10} 10-mers in round 7 and round 1 were calculated. For each 10-mer observed in both rounds, we calculated the ratio of its occurrence frequency in round 7 to round 1, and defined it as the enrichment ratio of that 10-mer. Supplementary Fig. S4 shows a histogram of enrichment ratios.

To test whether the most enriched 10-mers correspond to any classical short sequence-motiff, we submitted the most enriched 0.1% of 10-mers to "The MEME suite"[17] and were unable to find any statistically significant motif (best e-value ~ 0.26), indicating that the enrichment is not driven by simple position-specific consensus sequences. This is consistent with a model where selection favours broader compositional or structural sequence properties rather than a discrete binding motif.

Supplementary note 4B: mechanical properties of enriched and de-enriched 10-mers

From among all possible 4^{10} 10-mers, we identified the 1,000 10-mers with the highest and the 1,000 10-mers with the lowest enrichment ratios between round 7 and round 1 of SELEX (from supplementary note 4A). Additionally, we constructed 1,000 10-mers totally at random. This yielded a total of 3,000 10-mers across the three groups (most enriched, random, and most de-enriched). For each 10-mer, we generated 1,000 random 50 bp DNA sequences with the constraint that each sequence had that 10-mer in the middle. This resulted in 3 million 50 bp DNA sequences (1 million is each group). We predicted the intrinsic cyclisabilities of these 3 million molecules, and displayed the result in Fig. 3h. Black horizontal lines represent the mean intrinsic cyclisabilities of the 1 million molecules in each group.

Supplementary note 5: competitive binding

Supplementary note 5A: synthetic DNA sequences

The 400 synthetic sequences, comprising 20 groups of 20 sequences each were designed to have specific intrinsic cyclisability variations along their lengths. We first defined 10 intrinsic cyclisability profiles (intrinsic cyclisability as a function of position along the 167 bp region) possessing a single peak in intrinsic cyclisability on the left half of the molecule, with peak location and width matching the top 100 SELEX-selected sequences with the highest copy number from round 7 and also classified as having a single peak on the left side (Fig. 4a, bold line). We varied the peak heights across the 10 profiles, with the shallowest peak heights dipping below the baseline intrinsic cyclisability and peak heigh being variable across the 10 profiles. For each profile, we generate, at random, through *in-silico* evolution (see supplementary note 3H), a group of 20 DNA sequences across which the mean intrinsic cyclisability as a function of position closely matches the profile. Fig. 4a shows the mean flexibility as a function of position, averaged over the 20 sequences in each group. Similarly, 10 additional profiles with two flexibility peaks in each profile were used to generate 10 more groups (groups 11-20) of 20 sequences each, whose mean flexibility as a function of position, averaged over sequences within each group, is shown in Fig. 4a.

Supplementary note 5B: metric to capture phased periodic dinucleotide repeats

The 200 most-enriched SELEX sequences in the 167 bp library were significantly better gyrase binders than any other sequence in any other category (Fig. 4b). However, intrinsic cyclisability alone cannot explain their binding strength. Sequences within group 8 for instance each had an intrinsic cyclisability profile that matched the mean profile for the 100 top SELEX sequences with peaks on the left side only (SELEX1 in Fig. 4b), yet the SELEX1 sequences are significantly better gyrase binders than group 8 sequences.

We have already observed that SELEX enriches for sequences with periodic and phased GC content and specific dinucleotide contents, which is unrelated to the selection of intrinsic cyclisability peaks. To assess the relevance of phased periodic GC content and dinucleotide content to gyrase binding score, we sought to construct a quantitative metric to determine how similar any individual 133 bp sequence is to the phased periodic patterns that emerge when all reads in round 7 of SELEX are averaged. Subsequently, we sought to calculate this metric for all sequences in the library to see if it could distinguish SELEX sequences from the other sequences.

As a first minimal model, we considered only positional preferences for GC vs AT content. We defined an ideal 133 bp sequence in which positions where the mean GC content in round 7 exceeds the mean GC content in round 1 (as calculated in Fig. 3a) were assigned as G and all other positions were assigned as A. We further assigned a positional strength equal to the absolute difference between the round 7 and round 1 mean GC content at each of the 133 locations. To score any given 133 bp sequence, each position was classified as either matching or mismatching the ideal sequence (treating G and C, and A and T equivalently). Matches contributed positively to the

total score by the corresponding positional strength, whereas mismatches contributed negatively. While the scores thus obtained of the SELEX sequences were on average higher than sequence from other groups (Fig. S5a), we sought to refine the model accounting for positional strength of all dinucleotides.

We extended the model to distinguish all four bases and their nearest-neighbour context by constructing a dinucleotide enrichment model. For each of the 132 overlapping dinucleotide positions along the 133 bp random region, we computed the empirical frequencies of all 16 dinucleotides (ApA, ApC, …, TpT) from among all reads in round 1 and round 7 of SELEX. From these, we defined a position-specific log-likelihood ratio (LLR) weight for each dinucleotide:

$$W(i, xy) = log\left(\frac{P_7(i, XpY)}{P_1(i, XpY)}\right)$$

Where $P_7(i, XpY)$ and $P_1(i, XpY)$ denote the frequencies of dinucleotide XpY at position i in rounds 7 and 1 respectively. Positive weights therefore indicate dinucleotides enriched by selection, whereas negative weights indicate dinucleotides depleted by selection.

This yields a 132 x 16 weight matrix that assigns a score to each possible dinucleotide at each position. This weight matrix is visualised in Fig. S5b, which shows clear periodic oscillatory trends in the enrichment extent (LLR weights) of various dinucleotides.

Any given 133 bp sequence can then be scored by summing the corresponding weights across its 132 overlapping dinucleotides. Fig. S5c shows the obtained scores of all categories of sequences in the 1,253 member 167 bp library. It shows better separation of scores between the SELEX sequences and all other categories of sequences.

Importantly, this model retains the additive structure of the previous model, but is more expressive, allowing discrimination between G and C, A and T, and between distinct dinucleotide steps such as GG, CC, CG, and TA. This is important because several earlier models for DNA binding to the surface of proteins aided by lysines and argenines inserting into the minor groove show local 10 bp preference for specific dinucleotides [13,18].

Supplementary note 5C: SELEX enrichment of overall GC content

Here we investigate the effect of overall GC content on gyrase binding in SELEX as well as competitive gyrase binding data. It is suggested that the reader had read the main text at least till the section on competitive gyrase binding assays to properly appreciate this supplementary note.

Fig. 3a suggests that with increasing rounds of SELEX, overall GC content also increases, especially in the flanking regions. We quantify this in Fig. S5d. However, we note that even among completely random sequences, increase in average GC content also automatically leads to increase in dinucleotide periodic score (Fig. S5e), suggesting that the effects may not be unrelated.

Consistent with SELEX selecting for sequences with increasing overall GC content, we find that there is overall positive correlation between mean GC content and gyrase binding score within sequences in the 167 bp library. However, this correlation is driven largely by the sequences splitting into two clouds: the 200 top SELEX sequences, and all other categories (Fig. S5f). Within the SELEX sequences, and within all other sequences, there is no significant correlation (Fig. S5f). Likewise, when the scores are used from the 225 bp library, we observe a similar effect: an overall weak positive correlation driven by the sequences splitting into SELEX and non-SELEX groups, with no correlations within individual categories (Fig. S5f). These results suggest that, overall, GC content is likely not a significant determinant of gyrase binding score, and that SELEX sequences have slightly higher GC content likely because of selection of specific phased periodic patterns.

Supplementary note 5D: detailed and further analysis related to Fig. 4g

For any given 167 bp sequence, 'left peak' in cyclisability is defined as the mean intrinsic cyclisability from positions 35 to 55 bp, and 'right peak' in cyclisability is defined as the mean intrinsic cyclisability from positions 115 to 135 bp. This is irrespective of whether there is an actual peak in these locations (groups 5-10 or 15-20), or dips in these locations (groups 1-4, or 11-20). The definition of 'right peak' holds even for sequences in groups 1-10, where intrinsic cyclisability on the 'right' side is essentially flat – we simply average what that flat intrinsic cyclisability is.

Binding scores from the 225 bp library (Fig. 4g, S5g):

Using binding scores as calculated from the 225 bp library, we plot, for each of the 20 groups of synthetic sequences, binding score averaged over all 20 sequences within a group vs mean height of 'left' intrinsic cyclisability peak of all 20 molecules within that group (Fig. S5g left panel). We connect the points for groups 1-10 in blue and the points for groups 11-20 in orange, and find that the orange curve starts off lower than the blue crosses over, and end up higher than the blue. This suggests that if we compare a molecule with a certain cyclisability peak on the left only, with another with a similar peak on the left but also an identical peak on the right, the latter molecule with have a higher (or lower) binding score, depending on whether the peak is above or below baseline.

To further quantitatively verify if the effect is additive, we re-plot the figure, except this time, along the x-axis, we plot the sum of the peak heights at both the left and side, for all groups. Groups 1-10 each contribute only a near constant baseline for 'right' peak height. We now see that the two plots better coincide (Fig. S5g right panel). This implies that sequences with both a 'left' and 'right' peak in intrinsic cyclisability have similar binding scores as compared to sequences with only a 'left' peak but whose value equals the sum of the left and right peak heights of the first sequence.

Binding scores from the 167 bp library:

When the above calculations are repeated for binding scores obtained from the 167 bp library, the effects, including the additive nature, are largely the same when comparing between groups where intrinsic cyclisability peaks range from baseline through above baseline values (groups ~5-10 or

~15-20). See Fig. S5h). However, the dependence of binding score on intrinsic cyclisability peak heights when the peak dips below the baseline, is flatter when compared to the 225 bp library.

Supplementary Note 6: competitive supercoiling

Supplementary note 6A: sequence dependence of supercoiling score

To analyse sequence-dependence of supercoiling scores, we first examined the pairwise relationships between binding score, supercoiling score, and GC content using scatter plots (Fig. S6a). For each scatter plot, we also visualise the relationship by binning sequences into deciles of one of the observables and plotting the mean ± s.e.m of the other variable within each bin (Fig. S6a).

We observe that supercoiling score is weakly correlated with overall GC content, and shows a non-monotonic relationship with binding score. Moreover, this non-monotonic relationship cannot be trivially explained by GC content because the relationship between GC content and binding score, and between GC content and supercoiling score, are both approximately monotonic (Fig. S6a).

Further, to assess whether the non-monotonic dependence of supercoiling on binding score could arise by chance, we performed 200 random permutations of the binding scores, and in each case, obtained the binned plot of supercoiling vs binding scores (Fig. S6b). When compared to the un-permuted case, none of these show any similar non-monotonic trend.

As binding score is highly correlated with mean intrinsic cyclisability (defined as the mean of the 118 intrinsic cyclisabilities of the 118 50 bp DNA fragments that a 167 bp DNA fragment can be tiled into), we anticipated a similar non-monotonic relationship between these two variables. We confirm this via a binned plot – Fig. S6c, while Fig. S6d shows the corresponding scatter plot.

Supplementary Note 7: binding and supercoiling around *E. coli* ORFs

Supplementary note 7A: selection of the 40 *E. coli* genes

We selected 40 genes from *E. coli* and measured their gyrase binding and supercoiling scores in the genomic region surrounding their ORF start sites. This supplementary note describes the criteria used to select these genes. The full list of genes, together with their classification and relevant metadata, is provided below, at the end of this supplementary note.

All genes met the criteria that they were the first gene downstream from their respective promoters[19], ensuring that the analysed promoter-proximal DNA is not influenced by upstream transcriptional units.

A previous study measured the change in transcriptional level of the entire *E. Coli* transcriptome during global genomic supercoil relaxation induced *via* gyrase inhibition[20]. The study classes genes as SSG+ if they were upregulated during gyrase inhibition (supercoil relaxation) and SSG- if downregulated. From this dataset, we identified the 20 most strongly relaxation induced and the 20 most strongly relaxation supressed genes. From each of these two sets 5 were selected that met the above promoter-position criteria.

To provide comparison groups independent of whether genes were SSGs, we additionally selected five highly expressed housekeeping genes at random from previously reported lists of *E. Coli* housekeeping genes[21], again restricting selection to genes that are the first transcriptional unit downstream of the promoter.

Five poorly expressed genes were selected at random from among those genes not detectably expressed in the earlier genome-wide transcriptional study[20], and that also satisfied the promoter-position criteria.

Finally, 20 non-SGS genes were selected at random that all met the promoter-position criterion.

Gene	Category
b1330	SSG+
dcuC	SSG+
yjeS	SSG+
lrhA	SSG+
ydeH	SSG+
polA	SSG-
pepN	SSG-
yidQ	SSG-
pncB	SSG-
sbcB	SSG-
recA	Highly expressed
rpoB	Highly expressed

rpoA	Highly expressed
gapA	Highly expressed
rho	Highly expressed
sbcC	Poorly expressed
betT	Poorly expressed
mrcB	Poorly expressed
fhuD	Poorly expressed
dinJ	Poorly expressed
glyA	Random gene
ydeB	Random gene
dnaX	Random gene
metE	Random gene
tyrB	Random gene
mrp	Random gene
ychH	Random gene
dacC	Random gene
b1731	Random gene
b1452	Random gene
bioA	Random gene
yhfA	Random gene
yiaF	Random gene
galS	Random gene
ndh	Random gene
yhbY	Random gene
pyrD	Random gene
galR	Random gene
dsdC	Random gene
trxB	Random gene

Supplementary Figures

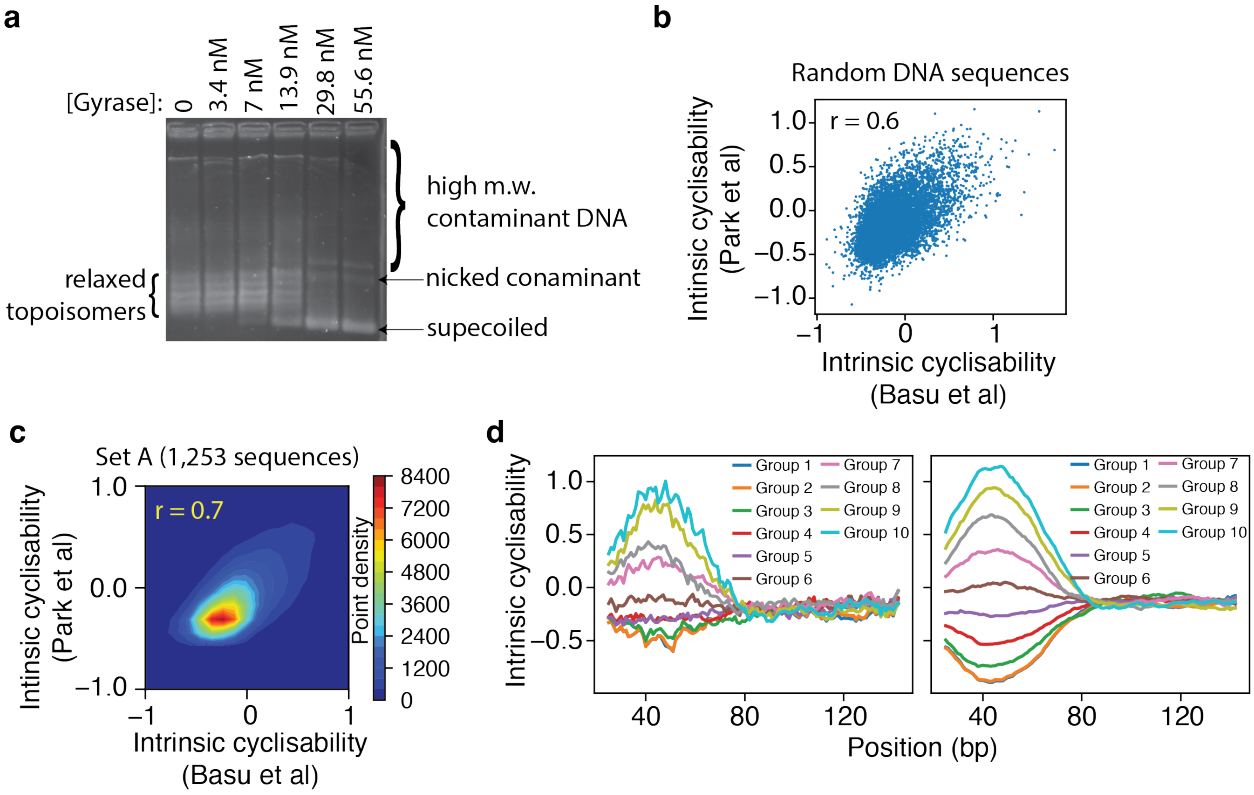

Figure S1: (a) *Gyrase activity assay, as described in supplementary note 1A. Reactions were carried out in presence of 2.2 mM ATP, indicated [gyrase] and other component for 1 hour at 37°C, quenched with SDS and EDTA, run on a 1% agarose gel at 18V for 16 hours, and stained with SYBR gold. The substrate plasmid has a higher molecular weight contaminant.* ***(b)*** *Scatter plot of predicted intrinsic cyclisabilities of 10,000 randomly generated 50 bp DNA sequences, obtained via the Park et al vs. the Basu et al models. Pearson's r = 0.6.* ***(c)*** *Contour plot of the Park vs Basu predicted intrinsic cyclisabilities of all the (167-50+1)x1253 = 147,854 50 bp DNA sequences that span the 167 bp length of all the 1,253 sequences in the 167 bp library. Pearson's r = 0.7. The very large number of points makes it difficult to directly plot a scatter.* ***(c)*** *Basu et al (left panel) and Park et al (right panel) model predictions of the mean intrinsic cyclisability as a function of position, averaged of the 20 sequences in each group 1 – 10. See Fig. 4a.*

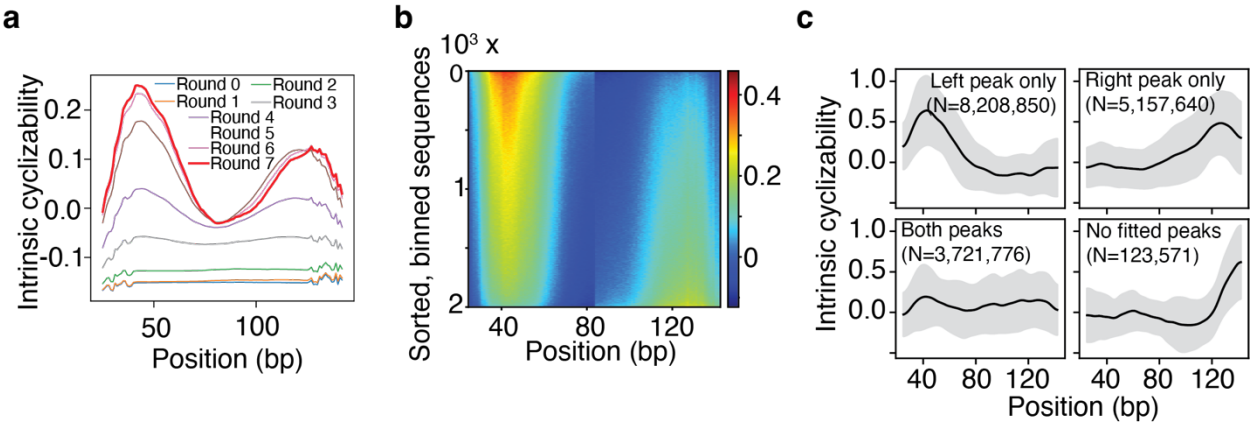

Figure S2: ***(a)*** *Intrinsic cyclisability of DNA as a function of position averaged over all unique sequences in each round of SELEX.* ***(b)*** *Same heatmap as in Fig. 2d, after randomly permuting the individual rows of the intrinsic cyclisability matix, as described in supplementary note 2E.* ***(c)*** *Same as Fig. 2e, except all reads from round 7 of SELEX (not just the unique sequences) were used, as described in supplementay note 2G.*

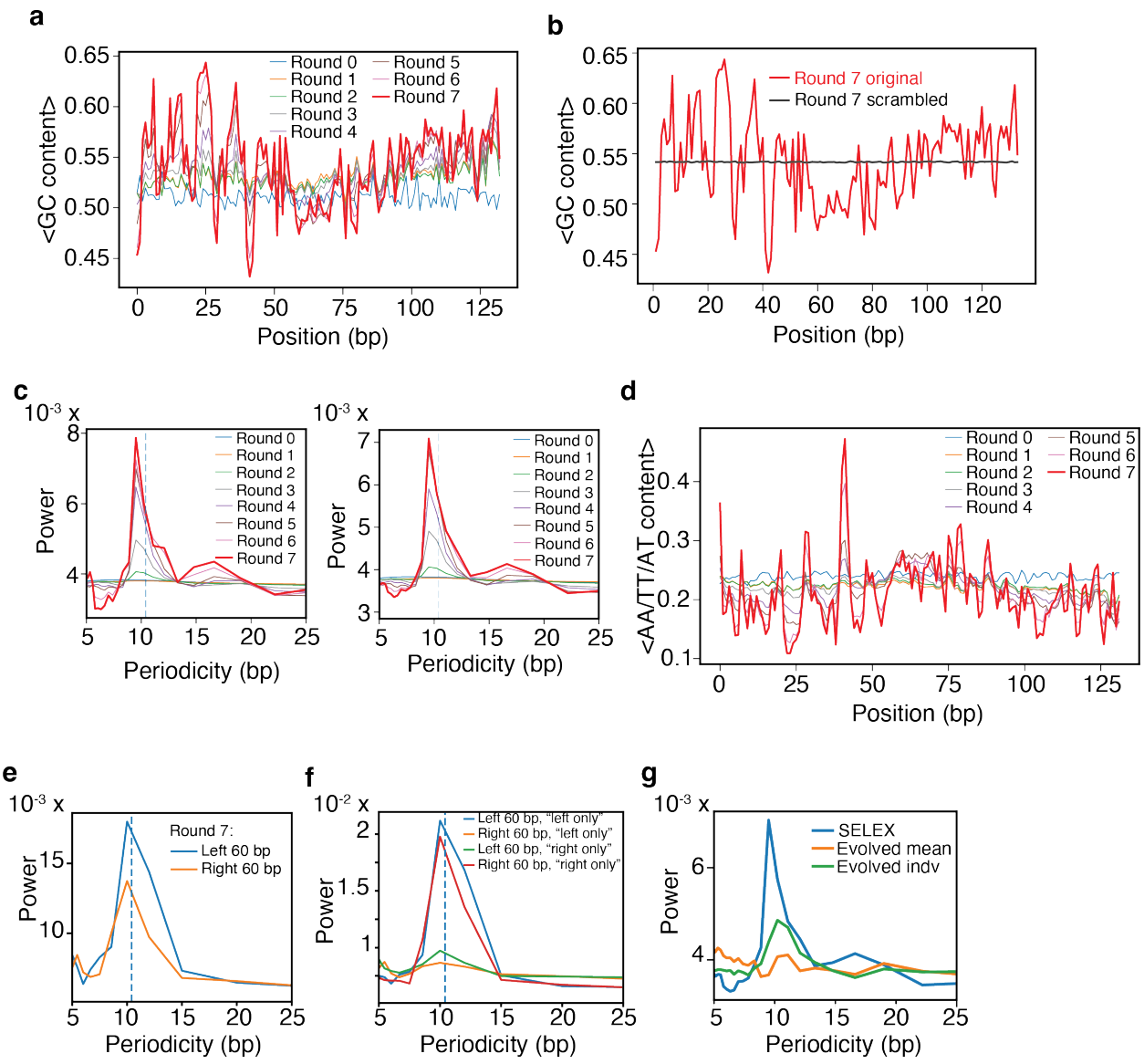

Figure S3: ***(a)*** *Mean GC content as a function of position, averaged over all the unique sequences only in each round of SELEX.* ***(b)*** *Mean GC content as a function of position averaged over all unique sequences in SELEX round 7 (red), and over the same set of sequences but after each sequence has been randomly permuted within itself (still preserving the original base compositions) (black).* ***(c)*** *Left panel: Average power spectrum as a function of position, obtained by performing categorical spectral analysis on each read in each round of SELEX, followed by averaging the spectra across all reads within a round. Right panel: same as left panel, except instead of all reads within a round, only unique sequences within each round were considered.* ***(d)*** *Mean content of WW dinucleotides (AA/TT/AT) as a function of position, averaged over all unique sequences in each round of SELEX.* ***(e)*** *All unique sequences in round 7 of SELEX were split into two sub-sequences, comprising the first and last 60 bp from the parent sequence. Categorical spectral analysis was carried out on all sub-sequences to obtain a power spectral density. This was averaged separately among the sub-sequences representing the "left" and "right" ends of the*

original sequences. ***(f)*** *Unique sequences from round 7 of SELEX were categorised as "left only" or "right only", depending on whether their intrinsic cyclisability profile has peaks only on the left or right half of the sequence, as done in Fig. 2e (see also Supplementary Note 2F) Within each goup, sequence were further split into sub-sequences, representing the first 60 bp or last 60 bp, as done in panel e. Categorical spectral analysis was performed on all sub-sequences from either groups, and the individual power spectra averaged over all left 60 bp sub-sequences separately from the "left only" and "right only" groups, and then over all right 60 bp sub-sequences separately from the "left only" and "right only" groups.* ***(g)*** *Same as Fig. 3g, except categorical spectral analysis was performed on individual sequences before averaging the individual power spectra.*

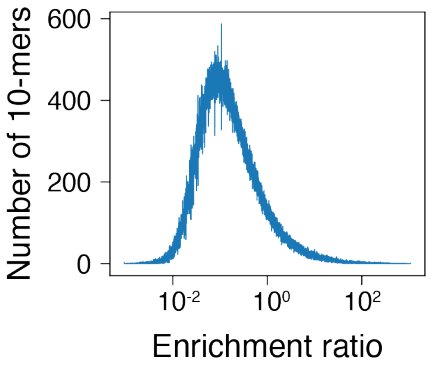

Figure S4: *Histogram of enrichment ratios of all 10-mers that occur within the reads of both round 7 and round 1 (see supplementary note 4A).*

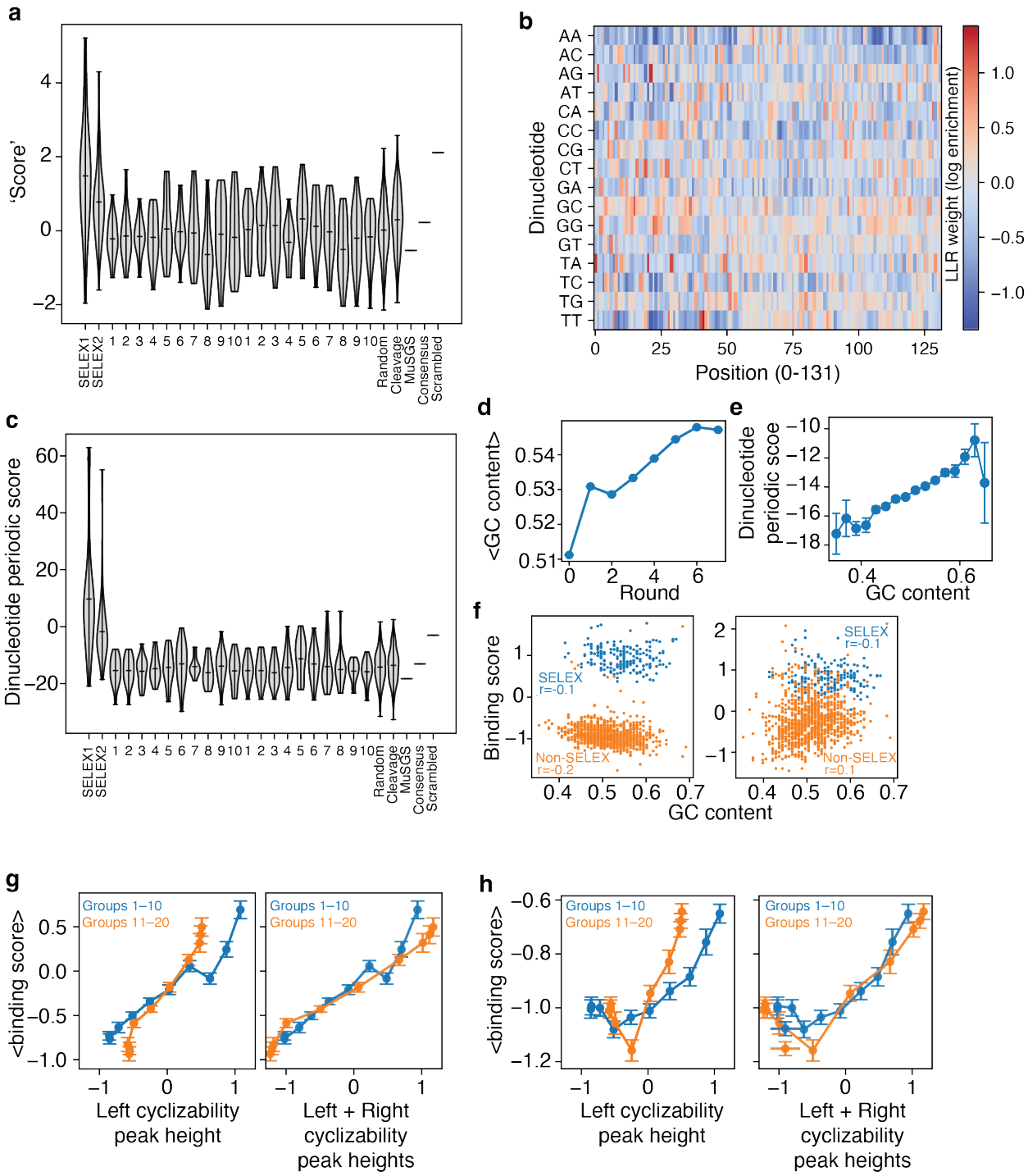

Figure S5: (a) *Violin plot of 'score' from the minimal model that consideres only positional AT or GC preferencing (Supplementary note 5B), across all categories within the 1,253 sequences in the 167 bp library.* ***(b)*** *The weight matrix of the position-specific log-likelihood ratio of enrichment of every dinucleotide at all positions along the 133 bp molecule, as explained in supplementary note 5B.* ***(c)*** *Dinucleotide periodic score calculated for all categories within the 1,253 sequences in the 167 bp library.* ***(d)*** *Mean GC content, averaged over all reads in each round of SELEX (errorbars*

are s.e.m.) ***(e)*** *Dinucleotide periodic score vs GC content, each calculated from 10,000 randomly-generated 133 bp DNA sequences. Sequences were binned according to GC content. Errorbars are s.e.m.* ***(f)*** *Scatter plot of binding score from the 167 bp (left panel) and 225 bp (right panel) libraries vs the overall GC content of all 1,253 sequences. 'SELEX' refers to the 200 SELEX-derived sequences.* ***(g)*** *left panel: identical to Fig. 4g. Right panel: the x-axis plots the sum of the left and right intrinsic cyclisability peaks. See supplementary note 5D for details.* ***(h)*** *Same as panel g, except binding scores from the 167 bp library are used.*

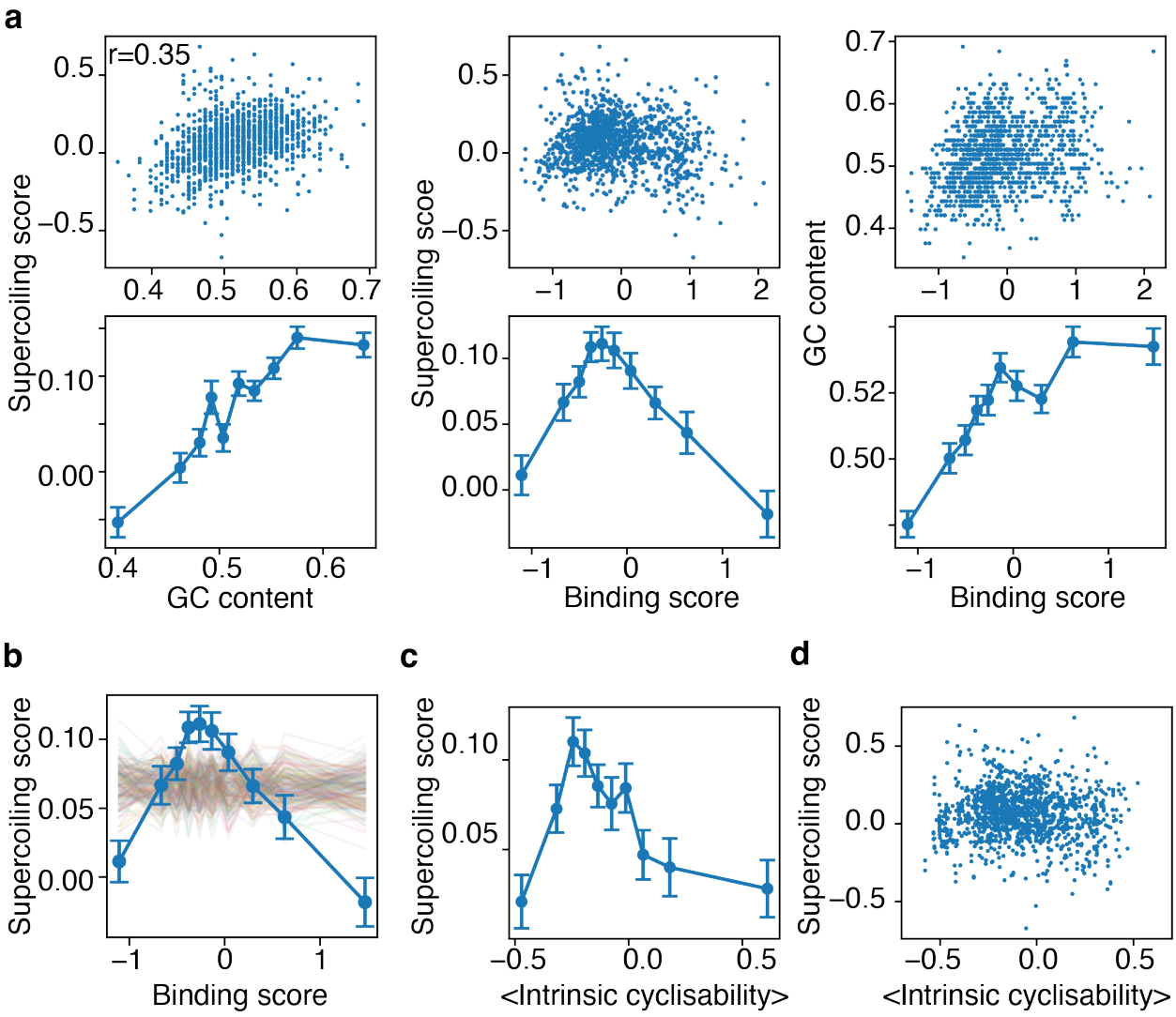

Figure S6: ***(a)*** *Top row: scatter plots showing pairwise relationships between (left) supercoiling score and GC content (Pearson's r = 0.35), (middle) supercoiling score and binding score, and (right) GC content and binding score for all 1,253 sequences. Bottom row: the same relationships shown after binning sequences into deciles of the x-axis variable (points show mean ± SEM within each bin).* ***(b)*** *Faint colours: 200 random permutations of the binding scores of the 1,253 sequences were taken, and the binned analysis for supercoiling vs binding score (as in panel b) repeated. Solid colours: reproduction of the corresponding plot from panel a).* ***(c)*** *Sequences within the 225 bp library were ordered according to mean intrinsic value along the sequence (mean of the individual intrinsic cyclisability values of the 50 bp overlapping fragments that tile any sequence) and split into deciles of ~125 sequences. Plotted is the mean ± s.e.m. of supercoiling score vs mean intrinsic cyclisability of all sequences within each decile.* ***(d)*** *Scatter plot of supercoiling score vs mean intrinsic cyclisability of all 1,253 167 bp sequences.*

References

1. Tretter, E. M. & Berger, J. M. Mechanisms for Defining Supercoiling Set Point of DNA Gyrase Orthologs: I. A NONCONSERVED ACIDIC C-TERMINAL TAIL MODULATES ESCHERICHIA COLI GYRASE ACTIVITY. *J. Biol. Chem.* **287**, 18636 (2012).

2. Basu, A. *et al.* Measuring DNA mechanics on the genome scale. *Nature 2020 589:7842* **589**, 462–467 (2020).

3. Basu, A. *et al.* Deciphering the mechanical code of the genome and epigenome. *Nature Structural & Molecular Biology 2022 29:12* **29**, 1178–1187 (2022).

4. Li, K., Carroll, M., Vafabakhsh, R., Wang, X. A. & Wang, J. P. DNAcycP: a deep learning tool for DNA cyclizability prediction. *Nucleic Acids Res.* **50**, 3142 (2022).

5. Khan, S. R., Sakib, S., Rahman, M. S. & Samee, M. A. H. DeepBend: An interpretable model of DNA bendability. *iScience* **26**, 105945 (2023).

6. Jiang, W. J. *et al.* Assessing base-resolution DNA mechanics on the genome scale. *Nucleic Acids Res.* **51**, 9552–9566 (2023).

7. Back, G. & Walther, D. Predictions of DNA mechanical properties at a genomic scale reveal potentially new functional roles of DNA flexibility. *NAR Genom. Bioinform.* **5**, (2023).

8. Park, J. *et al.* Probing mechanical selection in diverse eukaryotic genomes through accurate prediction of 3D DNA mechanics. *bioRxiv* 2024.12.22.629997 (2024) doi:10.1101/2024.12.22.629997.

9. Vanden Broeck, A., Lotz, C., Ortiz, J. & Lamour, V. Cryo-EM structure of the complete E. coli DNA gyrase nucleoprotein complex. *Nat. Commun.* **10**, 1–12 (2019).

10. Michalczyk, E. *et al.* Molecular mechanism of topoisomerase poisoning by the peptide antibiotic albicidin. *Nat. Catal.* **6**, 52–67 (2023).

11. Oliynyk, R. T. & Church, G. M. Efficient modification and preparation of circular DNA for expression in cell culture. *Commun. Biol.* **5**, (2022).

12. Michalczyk, E. *et al.* Structural basis of chiral wrap and T-segment capture by Escherichia coli DNA gyrase. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2407398121 (2024).

13. Brogaard, K., Xi, L., Wang, J. P. & Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496–501 (2012).

14. Jin, H., Rube, H. T. & Song, J. S. Categorical spectral analysis of periodicity in nucleosomal DNA. *Nucleic Acids Res.* **44**, 2047–2057 (2016).

15. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).

16. Sutormin, D., Rubanova, N., Logacheva, M., Ghilarov, D. & Severinov, K. Single-nucleotide-resolution mapping of DNA gyrase cleavage sites across the Escherichia coli genome. *Nucleic Acids Res.* **47**, 1373–1388 (2019).

17. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).

18. Rosanio, G., Widom, J. & Uhlenbeck, O. C. In vitro selection of DNAs with an increased propensity to form small circles. *Biopolymers* **103**, 303–320 (2015).

19. Salgado, H. *et al.* RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic Acids Res.* **52**, D255–D264 (2024).

20. Peter, B. J. *et al.* Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli. *Genome Biol.* **5**, (2004).

21. Rocha, D. J. P., Santos, C. S. & Pacheco, L. G. C. Bacterial reference genes for gene expression studies by RT-qPCR: survey and analysis. *Antonie Van Leeuwenhoek* **108**, 685–693 (2015).