Computational design of probes to detect bacterial genomes by multivalent binding: Supporting Information

Tine Curk[a,b], **Chris A. Brackley**[c], **James D. Farrell**[a], **Zhongyang Xing**[d,e], **Darshana Joshi**[d], **Susana Direito**[c], **Urban Bren**[b], **Stefano Angioletti-Uberti**[f], **Jure Dobnikar**[a,g,h], **Erika Eiser**[d], **Daan Frenkel**[g], **and Rosalind J. Allen**[c,1]

[a]Chinese Academy of Sciences, Institute of Physics, Beijing 100190, China; [b]University of Maribor, Faculty of Chemistry and Chemical Engineering, Maribor 2000, Slovenia; [c]School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, UK; [d]University of Cambridge, Cavendish Laboratory, Cambridge CB3 0HE, UK; [e]Current address: College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, Hunan, 400073, China; [f]Imperial College London, Department of Materials, London SW7 2AZ, UK; [g]University of Cambridge, Department of Chemistry, Cambridge CB2 1EW, UK; [h]Songshan Lake Materials Laboratory, Dongguan, Guangdong 523808, China

1. Monovalent vs multivalent probe-target binding

Let us consider a surface grafted with oligonucleotide probes, in contact with a sample solution that contains single-stranded target DNA molecules. We first suppose that each target molecule can bind to only a single probe (see Figure 1 (blue) and Figure 5 in the main text). In this monovalent binding scenario, each probe can be treated independently and the surface density of probes that are bound by the target, ρ_b^{mono}, is given by the standard Langmuir isotherm:

$$\rho_b^{mono} = \rho \left[\frac{c_t e^{-\beta \Delta G}/c_0}{1 + c_t e^{-\beta \Delta G}/c_0} \right] , \qquad \text{[S1]}$$

where ρ denotes the surface density of oligomer probes, c_t is the molar concentration of targets in the sample, ΔG is the free energy of probe-target hybridization, $\beta \equiv 1/(k_{\mathrm{B}}T)$ and $c_0 = 1$ M is a standard reference concentration. For low surface coverage of bound targets, $c_t e^{-\beta\Delta G}/c_0 < 1$, and the above expression reduces to:

$$\rho_b^{mono}/\rho \approx c_t e^{-\beta\Delta G}/c_0 = c_t K_{\mathrm{A}} , \qquad \text{[S2]}$$

as given in the introduction to the main text. Here $K_{\mathrm{A}} \equiv e^{-\beta\Delta G}/c_0$ is the equilibrium association constant.

Now let us consider the case where a single target DNA fragment can bind multiple probes simultaneously (Figure 1 (red) and Figure 5 in the main text). Let us suppose there are k probe binding sites on a single target DNA strand. In Refs (1–4) we have shown that the adsorption isotherm for a flexible polymer which can bind multiple sites on the surface also follows a Langmuir form:

$$\rho_b^{multi} l_t^2 = \frac{c_t l_t^3 N_{\mathrm{A}} q_b(\rho, k, \beta\Delta G)}{1 + c_t l_t^3 N_{\mathrm{A}} q_b(\rho, k, \beta\Delta G)} , \qquad \text{[S3]}$$

where N_{A} is Avogadro's number and the surface is assumed to be discretised into lattice sites of the size of the target l_t which, for a flexible polymer target, is determined by the radius of gyration of the target polymer R_g: $l_t \sim R_g$.

A key quantity in Eq. (S3) is the partition function q_b for the surface-bound polymer target. This function enumerates all possible binding configurations, and is given by

$$q_b(\rho, k, \beta\Delta G) = \left(1 + \rho e^{-\beta\Delta G}/(l_t c_0)\right)^k - 1 . \qquad \text{[S4]}$$

The form of the partition function (S4) arises because each of the k sites on the target can be either free (with statistical weight 1) or bound to any of the $n_p = l_t^2 \rho$ probes within an area l_t^2; ΔG is the single probe hybridisation free energy as above. $q_{1u} = l_t^3 c_0 N_{\mathrm{A}}$ is the unbound partition function or a free configurational "volume" of an unbound target site on the DNA strand, while the bound partition function of a single site is: $q_{1b} = n_p e^{-\beta\Delta G}$ as the site can be attached to any of the n_p probes. The ratio of bound/unbound partition functions of a single target site is thus: $q_{1b}/q_{1u} = \rho e^{-\beta\Delta G}/(l_t c_0)$. In bulk solution all sites on a polymer are unbound, the unbound partition function is: $q_u^* = (q_{1u})^k$ because all k sites are assumed independent. For a polymer next to the surface each site can be either free or attached: $q_b^* = (q_{1u} + q_{1b})^k - (q_{1u})^k$, where the subtraction of $-(q_{1u})^k$ arises because we consider only states in which the target polymer has at least one site bound to the surface - thus the weight of the completely unbound polymer state needs to be subtracted. Hence, the expression $q_b = \frac{q_b^*}{q_u^*} = (1 + (q_{1b}/q_{1u}))^k - 1$ is the normalised partition function for binding of k independent target sites, relative to the solution state where each of the k sites is unbound; i.e. $q_b = e^{-\beta \Delta G_{\text{polymer}}}$ where $\Delta G_{\text{polymer}}$ is the bound contribution to the free energy difference between the polymer being at a specific location on the surface vs. a specific location in bulk solution. We have previously shown that this extension of the Langmuir model gives a good description of multivalent binding of a polymer to a surface (1, 2, 4). For the case of low oligonucleotide probe surface coverage, $c_t l_t^3 N_{\mathrm{A}} q_b(\rho, k, \beta\Delta G) < 1$, the denominator in Eq (S3) is approximated as unity and the expression for the binding curve simplifies to

$$\rho_b^{multi} \approx c_t l_t \left[\left(1 + \frac{\rho}{c_0 l_t} e^{-\beta \Delta G} \right)^k - 1 \right] . \qquad \text{[S5]}$$

For target strands that bind only a single probe, $k = 1$, the above expression reduces to the monovalent isotherm, Eq. (S1), in the low surface coverage limit. The full expression, Eq. (S3) and Eq. (S4), does not directly reduce to the Langmuir isotherm, Eq. (S1), because the full expression includes a lateral polymer excluded area of size l_t^2 which is not captured in the standard Langmuir isotherm (see refs (1, 2, 4) for further discussion).

2. Design of oligonucleotide probes

Score function method. Our in-house algorithm chooses oligonucleotide probes based on a score function that measures the number of regions of complementarity between the probe sequence and the target DNA (considering both the forward and reverse strands of the pathogen genome). We first choose the length l, in nucleotide bases, of the desired probes. For short probes, $l \leq 10$, our algorithm generates and evaluates all possible test probe sequences (e.g. there exist 4^{10} different sequences of length 10nt). If $l > 10$, the algorithm instead considers all distinct sequences of length l that occur within the target pathogen genome.

A test probe sequence i of length l is compared to all length l subsequences j in the genome and its reverse complement, and the numbers n_{ija} of exact matches of length $1 < a < l$ between i and the j are tallied. For example, if $l = 5$, $i =$AAAAA and $j =$ATAAA, then $n_{ij1} = 4, n_{ij1} = 2, n_{ij3} = 1$ and $n_{ij2} = n_{ij4} = n_{ij5} = 0$. Probe sequence i is then assigned a score S_i, evaluated according to

$$S_i = \log \left[\sum_{a=1}^{l} 4^a n_{ia} \right] . \qquad \text{[S6]}$$

This score function sums the numbers of matches $n_{ia} = \sum_j n_{ija}$ over all subsequence lengths a. Matches of length a are weighted by a factor 4^a to account for the fact that longer matches are less likely to happen by chance (the probability of finding a match of length a in a random target DNA sequence is $(1/4)^a$). The logarithm ensures that the score values remain manageable even for long probes (larger values of l), and that score values for genomes of very different lengths remain comparable.

As an aside, the score function, Eq. (S6), can be thought of as an estimate of the interaction free energy between the probe and the target. Briefly, the factor of 4^a can be seen as a Boltzmann factor $e^{-E/(k_{\mathrm{B}}T)}$, where the "energy" E is $-k_{\mathrm{B}}T \log(4)$ per matching nucleotide. The term in the square brackets in Eq. (S6) would then correspond to a partition function.

For a random DNA target interacting with a random oligonucleotide probe, we can compute the expected value of the score function. For the random target DNA, all subsequences of length a within the probe sequence are equally likely to be exact matches with any given part of the target sequence. Since the probability of obtaining any given random DNA sequence of length a is $(1/4)^a$, we expect a particular subsequence of length a to appear, on average, $(L - a + 1) \times (1/4)^a \approx L \times (1/4)^a$ times in a random DNA target of length L (assuming $L \gg a$). Therefore, remembering that there are $(l - a + 1)$ distinct subsequences of length a within the probe, the expected value of n_{ia} in this random scenario is $L \times (l - a + 1) \times (1/4)^a$, and the expected score S_i for random probe i is $S_{i,\text{random}} = \log \left[L \times \sum_{a=1}^{l} (l - a + 1) \right] = \log \left[L \times \left(l(l + 1) - \sum_{a=1}^{l} a \right) \right] = \log \left[L \times (l(l + 1)/2) \right]$.

To select the desired probes, the score function is computed for all test probe sequences of length l. The test probe sequences are then ordered based on their scores. For the *E. coli* bl21-de3 genome, the ten top-scoring probe sequences of length 10nt, and their scores S, are provided in Table 1, and the equivalent data for probe sequences of length 20nt is provided in Table 2.

We note that the sequences come in complementary pairs; this is a consequence of the fact that we consider both the forward and reverse strands of the pathogen genome. We also note that many of the sequences are very similar to one another, and probably correspond to overlapping parts of the pathogen genome. Should this not be desirable, one could easily modify the selection criteria to prevent overlapping probe sequences being chosen. For the simulation results presented in the main text only the top two sequences in this list were used: because these are a complementary pair, this amounts to using a single probe sequence, plus its reverse complement.

sequence i	score S_i
CGCCAGCGCC	21.262
GGCGCTGGCG	21.262
CCGCCAGCGC	21.24
GCGCTGGCGG	21.24
CCAGCGCCAG	21.231
CTGGCGCTGG	21.231
GCGCCAGCGC	21.221
GCGCTGGCGC	21.221
CAGCGCCAGC	21.203
GCTGGCGCTG	21.203

Table 1. Top-scoring 10 nucleotide sequences for the *E. coli* bl21-de3 strain (including both forward and reverse genome strands)

sequence i	score S_i
AGGCGTTCACGCCGCATCCG	32.686
CGGATGCGGCGTGAACGCCT	32.686
GATGCGGCGTGAACGCCTTA	32.671
TAAGGCGTTCACGCCGCATC	32.671
AAGGCGTTCACGCCGCATCC	32.668
GGATGCGGCGTGAACGCCTT	32.668
ATAAGGCGTTCACGCCGCAT	32.664
ATGCGGCGTGAACGCCTTAT	32.664
GATAAGGCGTTCACGCCGCA	32.663
TGCGGCGTGAACGCCTTATC	32.663

Table 2. Top-scoring 20 nucleotide sequences for the *E. coli* bl21-de3 strain (including both forward and reverse genome strands)

Targeted method for distinguishing similar genomes. In some cases, it is important to be able to detect the target genome in the presence of other genomic DNA that is closely related to it. For example, one might need to distinguish between strains of the same bacterial species, such as the O157 Sakai strain of *E. coli*, which causes food poisoning, in the presence of harmless strains (represented here by the wild-type lab strain bl21-de3). In this case, it is likely that the top-scoring oligonucleotide probe sequences for both target genomes will be very similar, making it hard to achieve selective binding.

To differentiate between similar bacterial genomes (here denoted A and B) we propose a modified method of probe selection. Rather than simply scoring probe sequences according to their number of regions of complementarity with the target genome, we propose instead to rank them by the *difference* in their score for genomes A and B:

$$\Delta S_i = S_i(A) - S_i(B)\ . \qquad [S7]$$

The probe sequences of length 10nt and 20nt that maximise the difference ΔS_i between the O157 Sakai and bl21-de3 strains of *E. coli* are shown in Tables 3 and 4 respectively, along with their score difference ΔS. ΔS is much larger for the 20nt strands, therefore we used the first two 20nt sequences in Table 4 (corresponding to a single sequence and its reverse complement) in the simulations in the main text (Fig. 3)).

Alternative oligonucleotide selection method using BLAST. An alternative approach to scoring a probe sequence is simply to count how many times it appears in the target genome. Here we describe an algorithm to do this, which uses the popular

sequence i	score ΔS_i
GGTGTATGAC	0.335
GTCATACACC	0.335
ATCCGGATGA	0.323
TCATCCGGAT	0.323
CATCCGGATA	0.323
TATCCGGATG	0.323
GGTGACGGAC	0.319
GTCCGTCACC	0.319
GGGTGACGGA	0.319
TCCGTCACCC	0.319

Table 3. Top-scoring 10 nucleotide sequences maximising the score difference, Eq. (S7)**, between the O157 and bl21-de3 strains of *E. coli***

sequence i	score ΔS
GGAGACTAAACTCCCTGAGA	10.463
TCTCAGGGAGTTTAGTCTCC	10.463
CTCAGGGAGTTTAGTCTCCA	10.452
TGGAGACTAAACTCCCTGAG	10.452
AGGGAGTTTAGTCTCCAGGA	10.443
TCCTGGAGACTAAACTCCCT	10.443
GAGACTAAACTCCCTGAGAA	10.436
TTCTCAGGGAGTTTAGTCTC	10.436
CAGGGAGTTTAGTCTCCAGG	10.429
CCTGGAGACTAAACTCCCTG	10.429

Table 4. Top-scoring 20 nucleotide sequences maximising the score difference, Eq. (S7)**, between the O157 and bl21-de3 strains of *E. coli***

sequence	BLAST matches
GCGCTGGCGG	9657
CCGCCAGCGC	9657
GCGCTGGCGA	9379
TCGCCAGCGC	9379
ACGCCAGCGC	9233
GCGCTGGCGT	9233
ACGCTGGCGG	9112
CCGCCAGCGT	9112
GCGCCAGCGT	8995
ACGCTGGCGC	8995

Table 5. Top-scoring 10 nucleotide sequences for the bl21-de3 strain of *E. coli* obtained using the BLAST method.

BLAST+ software suite (5).

First, we determine a set of candidate probe sequences from (both strands of) the target genome – specifically for probes of length 10nt, we slide a 10 base window along the genome in steps of 5 bases, taking the sequence within each window as a candidate (after removing any duplicate sequences). Next, we remove from the candidates any sequences of low complexity, e.g. those containing repetitive sequences such as “TAAAAAAAGA” or “TCGCGCGCGC”, since these tend to appear frequently within genomes and can lead to probes which are more likely to self-hybridize. This is done using the “dustmasker” software (which is part of the BLAST+ suite (5)).

For the remaining candidate sequences, we perform a BLAST search for matches on the target genome (using the “blastn” tool), counting the number of matches of 7 nt or longer. Sequences are then ranked in descending order of the number of matches, and the list further refined by removing any sequences which have 7 nt or longer identical regions with a sequence which is higher in the list. The top-scoring probe sequences are then taken from the top of the ranked list. Table 5 shows the top-scoring sequences found using the BLAST method for probes of length 10 nt targeting the bl21-de3 *E. coli* genome.

Reassuringly, the top four probe sequences obtained from the BLAST method match the top four obtained from our in-house algorithm, allowing for single-base shifts. The BLAST and score function methods find the same two top probe sequences. This BLAST method for probe selection is likely to be more computationally efficient for longer genomes, since it utilizes a highly optimised heuristic sequence alignment algorithm. However, the score function method outlined above is likely to rank candidate strands more robustly, since it considers matches of any length.

3. Coarse-grained polymer model for genomic DNA

Our coarse-grained polymer model for genomic DNA is outlined in the main text: here we discuss some of its aspects in more detail.

Blob radius of gyration. In our coarse-grained model, the single-stranded DNA of the target bacterial pathogen is treated as a chain of “blobs”, each of which represents ~ 400 nucleotides. At sufficiently high temperature, we can assume that any base-pairing interactions between the blobs can be neglected and the genomic DNA can be treated as a self-avoiding walk (SAW) polymer (6). The radius of gyration of a SAW polymer is given by

$$R_g = \frac{b}{\sqrt{6}} N_{\mathrm{Kuhn}}^{\nu} \quad , \qquad \text{[S8]}$$

where b is the Kuhn segment length of the polymer, $\nu = 0.588$ is the scaling exponent and N_{Kuhn} is the number of Kuhn segments in the polymer (7). If N_m is the number of monomers and a is the contour length per monomer, N_{Kuhn} is given by aN_m/b. For single-stranded DNA, $a = 0.65\,\mathrm{nm}$ (6, 8), and at a physiological salt concentration of 0.1M, $b \approx 2\,\mathrm{nm}$ (6, 8, 9). This leads to a prediction for the radius of gyration, R_g, of an $N_m = 400$ nucleotide blob, of 10 nm. At lower temperatures the radius of gyration will be affected by self-hybridisation, but recent results using a more detailed simulation model show that the macroscopic properties of ssDNA (e.g the radius of gyration) are not significantly affected by self-hybridisation for temperatures above 40°C (10). This insensitivity arises due to the opposing effects of rigidification (the persistence length of dsDNA is ≈ 50nm compared to ≈ 1nm for ssDNA) which increases R_g, and hybridisation between distant parts of the strand, which decreases R_g. All of our calculations are performed at temperatures above 40°C, where R_g can be assumed constant. For the same reason we do not include any specific blob–blob attractive interactions arising due to base pairing in our model, as these would reduce the radius of gyration of the simulated DNA polymer.

Implementation of Langevin Dynamics simulations in LAMMPS. Our Langevin dynamics simulations used the open source molecular dynamics simulation package LAMMPS (11). Specifically, the equations of motion for a set of particles representing the "blobs" as they interact with each other and with the surface are solved in the NVT ensemble (constant particle number, volume and temperature) using a velocity Verlet algorithm. This amounts to a numerical solution of the Langevin equation

$$m\frac{d^2\mathbf{r}_i}{dt^2} = -\nabla U_i(\mathbf{r}_i) - \xi\frac{d\mathbf{r}_i}{dt} + \sqrt{2k_{\mathrm{B}}T\xi}\boldsymbol{\eta}_i \qquad , \tag{S9}$$

where $\mathbf{r}_i$ is the position of particle i, $U_i(\mathbf{r}_i)$ is the potential energy of the particle i, the second term on the right hand side captures viscous drag and the components of $\boldsymbol{\eta}_i$ are independent δ-correlated white noise with unit variance and zero mean. m and ξ are the particle mass and the friction due to the implied solvent respectively, and these lead to a velocity decorrelation time $\tau_0 = m/\xi$. In the simulations we use length units of blob radius r_b, energy units of $k_{\mathrm{B}}T$, and mass units where m is the mass of the single blob. This leads to a simulation time unit $\tau = \sqrt{mr_b^2/k_{\mathrm{B}}T}$, and the integration is performed using a time step of 0.02τ. In reality this system will be over-damped ($m \ll \xi$), but this would lead to infeasibly long simulation run times; instead we use a reduced friction $\xi = m/100$, though since we run our simulations until they reach an equilibrium state we do not expect this to affect our results. Typically we run simulations for a total of $10^6\tau$ while counting the blob-surface contacts only in the last third of the total simulation time. Where simulation results are reported they correspond to a single simulation run, averaged over the last third of the run. Figure S1 shows that our results for the number of blob-surface contacts typically converge after $\approx 5\times10^5\tau$.

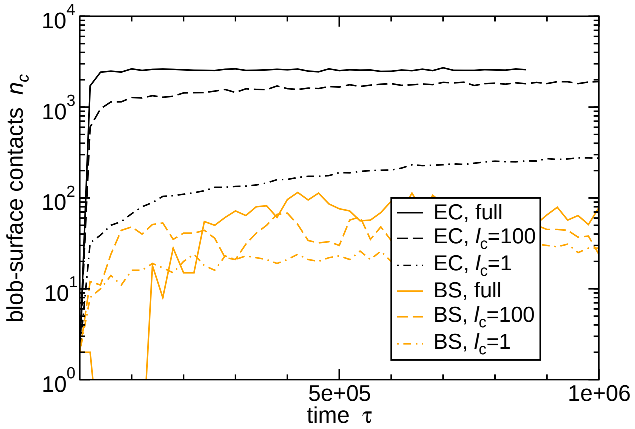


Fig. S1. Adsorption kinetics (number of blob-surface contacts as a function of time) in our simulations of the *E. coli* bl21-de3 (EC) and *B. subtilis* (BS) genomes binding to a surface coated with *E. coli* probes of length 10nt. Results are shown for the full genome (solid line; "full"), for genomes fragmented into fragments of length 100 blobs (40,000nt; dashed lines, "$l_c = 100$") and genomes fragmented into fragments of length 1 blob (400nt; dot-dashed lines, "$l_c = 1$"). The probe surface density is $\rho = 0.003r_b^{-2}$.

The potential energy function $U(\mathbf{r}_i)$ in Eq. (S9) is a sum of terms taking into account chain connectivity, blob-blob repulsion, and the blob-surface interaction (a short-range repulsion and longer-ranged sequence-dependent attraction) as given in Eqs.(4)-(6) in the main text. Some of these Gaussian potentials have previously been implemented in LAMMPS (11); the additional Gaussian potential describing specific base-pairing interactions between the target DNA and the DNA-grafted surface (second term in Eq. (6) of the main text) was implemented in LAMMPS as gaussian/cut. We mixed the different potential types using the hybrid/overlay option in LAMMPS.

4. Nearest-neighbour model for DNA hybridisation free energies

The nearest neighbour (NN) model for DNA hybridisation (12, 13) assumes that the hybridisation free energy can be written as a sum over base pairs, and that the contribution for a given base-pair depends on its identity and that of its immediate neighbours. The total hybridisation free energy between two single-stranded pieces of DNA is then written as

$$\Delta G^{NN} = \Delta G_{\text{stack}} + \Delta G_{\text{boundary}} + \Delta G_{\text{symmetry}} \ , \qquad \text{[S10]}$$

where $\Delta G_{\text{stack}} = \sum_{<i,j>} \Delta G_{<i,j>}$ is the sum is over all nearest neighbour base pairs (eg CG/GC), $\Delta G_{\text{boundary}}$ accounts for the strand ends and $\Delta G_{\text{symmetry}}$ is an entropic penalty that is applied to self-complementary duplexes (i.e. those that can form internal structure such as hairpins), to account for the fact that these duplexes have C2 symmetry (14).

In this work, we use the SantaLucia parameterisation of the NN model (14, 15), which specifies the various nearest neighbour base pair contributions to ΔG_{stack}, as well as $\Delta G_{\text{boundary}}$ and $\Delta G_{\text{symmetry}}$. All of these terms are assumed to be composed of enthalpic, ΔH^{NN}, and entropic, ΔS^{NN}, parts which are independent of temperature T, such that the free energy terms ΔG have a linear temperature dependence: $\Delta G^{NN} = \Delta H^{NN} - T\Delta S^{NN}$. The salt concentration dependance is incorporated as a linear correction to the enthalpy and entropy terms. Additional penalties for misalignment, bulges, hairpins and loops are also applied (see Ref. (15) for a detailed explanation of the model). The SantaLucia model has been shown to faithfully reproduce experimental data for DNA hybridisation free energies over a wide range of temperatures and salt concentrations.

Calculating SantaLucia DNA-binding free energies with NuPack. NuPack (16–19) is an open source program for calculating SantaLucia free energies for DNA strand hybridisation. It is available at http://www.nupack.org. The user specifies the sequences of n interacting DNA strands, the temperature (in this work T=50°C) and salt concentration (we have used the default value, 1M NaCl) and the program calculates the hybridization free energy $\Delta G^{NN}(s_1, s_2, ...s_n)$, considering all possible binding combinations, including all possible partial matches, self–hybridisation and three–way junctions, as well as defects such as bulges and loops. In this work, we use NuPack to obtain the hybridisation free energy between two strands, one of which is a 400nt-long "blob" in our coarse-grained DNA polymer, and the other of which is an oligonucleotide probe of length 10-20nt. We also use NuPack to compute the self-hybridisation free energies of the blob sequence and of the probe sequence. Denoting the blob sequence as s_j and the probe sequence as s_k, we then calculate the free energy change due to their interaction, $\Delta\tilde{G}_{j,k}$, as:

$$\Delta\tilde{G}_{j,k} = \Delta G^{NN}(s_j, s_k) - \Delta G^{NN}(s_j) - \Delta G^{NN}(s_k) \ . \qquad \text{[S11]}$$

In other words, we compute the difference between the free energy of the blob and probe DNA strands in contact with each other and that of the isolated blob and probe strands s_j and s_k.

It is important to note that, in NuPack, the SantaLucia free energies are obtained from partition functions that include all possible configurations, including the fully unbound state (which has free energy zero). Correctly accounting for these unbound states leads to the factors of -1 and +1 in Eq. (5) of the main text. We also note that the calculation of $\Delta G^{NN}(s_j, s_k)$ includes all the states considered in both $\Delta G^{NN}(s_j)$ and $\Delta G^{NN}(s_k)$ plus additional states where the two strands bind each other. Hence, $\Delta\tilde{G}_{j,k}$ must be negative.

The free energies provided by NuPack are given with respect to a reference concentration which is taken to be that of water: $c_w = 55$M. This leads to the factor of $c_w N_{\text{A}}$ in Eq. (6) of the main text.

Interaction free energy between a blob and the probe-coated surface. $\Delta\tilde{G}_{j,k}$ as computed using NuPack is the interaction free energy between a 400nt strand corresponding to a blob, and a single oligonucleotide probe. However, we require the interaction free energy $\Delta\tilde{G}_{j,\text{surf}}$ between a blob and the entire probe-coated surface. This is obtained using the following formula (3):

$$\Delta\tilde{G}_{j,\text{surf}} = -k_B T \log\left[1 + \rho r_b^2 \sum_k f_k \left(e^{-\beta\Delta\tilde{G}_{j,k}} - 1\right)\right] \ , \qquad \text{[S12]}$$

where ρ is the surface density (number per area) of oligonucleotide probes on the surface and r_b is the radius of gyration of the blob, such that r_b^2 is assumed to be the area of the surface over which the blob interacts, and ρr_b^2 is the number of probes that interact with the blob. In Eq. (S12) we have, for generality, supposed that there can be a mixture of distinct probe types on the surface, such that probe type k has fractional abundance f_k. In the main text, Eq. (5), we presented a simplified form of this equation, for only a single probe type. For our simulations, we used two probe types, one of which was the reverse complement of the other. Since the densities of the two probe types were equal, in our calculations we used $f_1 = f_2 = 0.5$.

The -1 in the round brackets in Eq. (S12) arises because, as discussed earlier, the SantaLucia hybridisation free energy, Eq. (S11), is calculated using a partition function that includes the state where the probe and blob are not bound. To avoid double-counting, this state needs to be subtracted from each of the terms describing a particular probe type. It is then added back, via the +1 in the square bracket, to include the state where the blob is not bound to any of the probes.

Eq. (S12) is a mean–field formula, which assumes that the blob binds independently to each probe. This assumption is valid if the probe grafting density is sufficiently low: $\rho \leq 1/r_b^2$. The mean-field formula also assumes that the probes are homogeneously mixed on the surface.

5. Derivation of the interaction potential prefactor for the Langevin dynamics simulations

To obtain the prefactor H in the coarse-grained interaction potential that we use in our simulations (Eq. (6) in the main text), we match coarse-grained and microscopic definitions of the partition function for blob-surface binding.

From a microscopic perspective, a blob j can be bound to any of the oligonucleotide probes on the surface, with binding free energy $\Delta\tilde{G}_{j,k}$ for a probe of type k, or the blob can be free in solution (not bound to the surface). The statistical weight of the state in which blob j is bound to probe k is $\left[e^{-\beta\Delta\tilde{G}_{j,k}} - 1\right]/(c_w N_\text{A})$. Here, the -1 takes into account that the SantaLucia calculation performed by NuPack includes the unbound state as discussed above, and the factor $1/(c_w N_\text{A})$ (where $c_w = 55\text{mol/l}$ and N_A is Avogadro's number) arises because $\Delta\tilde{G}_{j,k}$ is determined in NuPack with respect to the reference concentration of liquid water; i.e. $\Delta\tilde{G}_{j,k}$ is the hybridisation free energy with respect to the chemical potential of an ideal gas with concentration c_w. The statistical weight of the unbound state is given by the integral over the system volume V. Therefore, the microscopic configurational partition function for the blob with the surface, summing over all states, is

$$Q_{j,\text{micro}} = \frac{1}{c_w N_\text{A}} N_p \sum_k f_k \left(e^{-\beta\Delta\tilde{G}_{j,k}} - 1\right) + \int_V d\mathbf{r}\ , \qquad \text{[S13]}$$

where N_p is the total number of probes on the surface. The first term in Eq. (S13) accounts for configurations in which the blob is bound to the surface and the second term accounts for configurations where it is free in the solution.

From a coarse-grained perspective, the interaction due to hybridisation between the blob and the surface is described, in our model, with a Gaussian attraction potential

$$U_j(z) = \frac{H_j}{\sqrt{2\pi}} e^{-\frac{z^2}{2r_b^2}}\ . \qquad \text{[S14]}$$

This potential depends on the height z of the blob above the surface; H_j is the blob-dependent prefactor that we are aiming to determine. In the coarse-grained representation, the partition function for blob-surface binding, considering only attractive interactions, can then be written as

$$Q_{j,\text{cg}} = \int_V e^{-\beta U_j(z)} d\mathbf{r}\ . \qquad \text{[S15]}$$

The partition functions obtained for the microscopic and coarse-grained representations of our system need to be equal, since they describe the same system. Therefore we equate Eq. (S13) and Eq. (S15) to obtain

$$\int_V \left[e^{-\beta U_j(z)} - 1\right] d\mathbf{r} = \frac{1}{c_w N_\text{A}} N_p \sum_k f_k \left(e^{-\beta\Delta\tilde{G}_{j,k}} - 1\right)\ . \qquad \text{[S16]}$$

Using Eq. (S12) this reduces to

$$\int_V \left[e^{-\beta U(r)} - 1\right] d\mathbf{r} = N_p \left[\frac{e^{-\beta\Delta\tilde{G}_{j,\text{surf}}} - 1}{c_w N_\text{A} \rho r_b^2}\right]\ . \qquad \text{[S17]}$$

Since the Gaussian form (S14) for $U(z)$ depends only on the height above the surface, integration over the lateral coordinates x, y gives a factor of the surface area S on the left-hand side of Eq. (S17). Since $N_p/S = \rho$ (the probe density), ρ cancels and we get

$$\int_{-\infty}^{-\infty} \exp\left[-\frac{H_j}{k_B T\sqrt{2\pi}} e^{\frac{-z^2}{2}}\right] - 1\ dz = \frac{e^{-\beta\Delta\tilde{G}_{j,\text{surf}}} - 1}{c_w N_\text{A} r_b^2}\ . \qquad \text{[S18]}$$

This relation determines the mapping between the attractive interaction prefactor H_j (from Eq. (S14)) and the blob-surface interaction free energy $\Delta G_{j,\text{surf}}$ that we obtain from our SantaLucia calculations.

Figure S2 (solid line) shows this mapping. We note that the integrand in Eq. (S18) is a double exponential which is highly peaked around $z = 0$. We can hence approximate this as $\int_{-\infty}^{-\infty} \exp\left[-\frac{H_j}{k_\text{B}T\sqrt{2\pi}} e^{\frac{-z^2}{2}}\right] - 1\ dz \approx e^{-\frac{H_j}{k_\text{B}T\sqrt{2\pi}}}$. Moreover, it turns out that in all our simulations, $\beta\Delta G_{j,\text{surf}} \leq -5$, therefore, we can approximate $e^{-\beta\Delta G_{j,\text{surf}}} - 1 \approx e^{-\beta\Delta G_{j,\text{surf}}}$, and a simple linear relation follows, as given in the main text:

$$H_j = \sqrt{2\pi}\left[\Delta G_{j,\text{surf}} + k_\text{B}T \ln[r_b^3 c_w N_A]\right]\ . \qquad \text{[S19]}$$

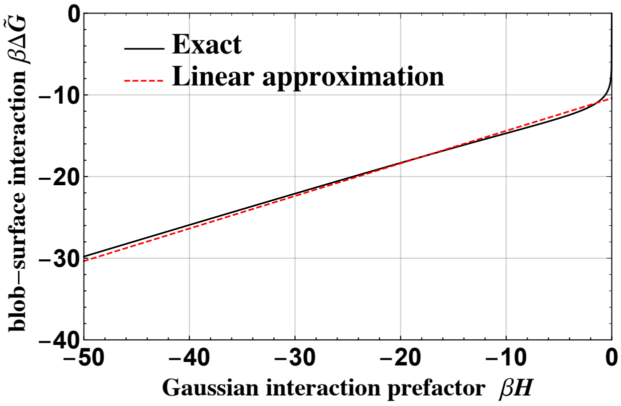


Fig. S2. Blob-surface interaction: mapping between the prefactor H in the Gaussian coarse-grained attractive interaction, and the blob-surface interaction free energy $\Delta\tilde{G}_{j,\mathrm{surf}}$. The mapping is obtained from Eqs. (S14) and (S18). Here, $r_b = 10$nm (which corresponds to 400 nucleotides per blob), and $\beta = 1/(k_\mathrm{B}T)$. The dashed red line shows the approximate linear relationship Eq. (S19).

We note that in our coarse-grained simulations, we also include an additional exponential repulsion between the blob and the surface, which represents the entropic penalty of confining the polymer close to the surface. Because our aim here was to match the partition functions for the attractive part of the polymer-surface interaction between the microscopic and coarse-grained representations, the repulsive term in the coarse-grained potential was not included in the mapping procedure.

6. Blob-surface interaction free energy and binding selectivity

Figure S3 plots the blob-surface interaction free energy values $\Delta\tilde{G}_{j,\mathrm{surf}}$ for each of the blobs that make up the genomic DNA polymer, for the *E. coli* wild-type (bl21-de3) and *B. subtilis* genomes binding to a surface designed to target *E. coli.* For probes of length 10nt (left panels in Figure S3) the difference in blob-surface binding free energy is, on average, $\sim 1k_\mathrm{B}T$ between *E. coli* and *B. subtilis.* Interestingly, coating the surface with more strand types (using the 20 highest scoring probes - 10 distinct sequences plus their reverse complements) does not seem to appreciably change the distribution of interactions (Figure S3 (e) and (f), comparing $n_t = 2$ for a single distinct probe - forward and reverse sequences, with $n_t = 20$ for a mixture of 10 probes, including forward and reverse sequences).

Perhaps counterintuitively, increasing the oligonucleotide probe length to 20nt (Figure S3b and c) reduces both the average interaction strength ($\Delta\tilde{G}_{j,\mathrm{surf}}$ is on average less negative) and the average interaction difference between the *E. coli* and *B. subtilis* genomes; in other words, the specificity with which the surface binds *E. coli* is reduced for longer probes! This effect is due to longer probe strands having, on average, stronger self-interaction, which passivates them, decreasing their binding affinity for the target genome. However, the distribution of interaction free energies $\Delta\tilde{G}_{j,\mathrm{surf}}$ is very different for long versus short probes (compare Figure S3(e) and (f)). For the longer, 20nt probes (Figure S3(f)), the interaction free energy distribution shows a prominent shoulder at low free energy values for the targeted *E. coli* genome, with a significant number of blobs with $\Delta\tilde{G}_{j,\mathrm{surf}} < 20k_\mathrm{B}T$ (even though the average interaction is weaker for 20nt probes than for 10nt probes). In contrast, few strong blob-surface interactions are observed for the non-target *B. subtilis* genome.

As one might expect, differentiating between very similar genomes, for example different strains of *E. coli*, is more challenging. Here, as discussed in section 2, we suggest choosing probe sequences based on the *difference* ΔS in score function values between the targeted and non-targeted genome. We tested this approach by comparing binding of the *E. coli* O157 Sakai strain vs. the wild-type *E. coli* bl21-de3 strain, for a surface coated in probes that target *E. coli* O157 Sakai, selected by ranking of the ΔS values between Sakai and wild-type (Figure S4). For probes of length 10nt the difference in the average blob-surface interaction free energy between the two genomes is negligible (Figure S4(a)), but for longer 20nt probes (Figure S4(b)) the targeted Sakai strain shows a strongly interacting "shoulder" in the distribution of blob-surface interaction free energies, while the nontargeted bl21-de3 strain does not show any strong blob-surface interactions. Thus, the "ΔS" method does provide a way to design probes that discriminate between similar strains.

Interestingly, if one needs to discriminate between two strains that are not very similar, choosing probes based on the score difference ΔS has little advantage. Figure S5 compares histograms of blob-surface binding free energies for the *E. coli* and *B. subtilis* genomes, for probes targeting *E. coli*, chosen either using the basic score function method or the "ΔS" method. Using the "ΔS" method makes the overall interaction weaker (the histogram is shifted towards higher $\Delta\tilde{G}$), but the specificity, *i.e.* the difference between the target (EC) and non-target (BS) histograms, does not improve.

In summary, to differentiate between sufficiently dissimilar genomes, such as those of different species, it appears to be sufficient to choose oligonucleotide probes based solely on the multiplicity of binding to the target genome, Eq. (S6). However, if strains of the same species must be differentiated, one should choose the oligonucleotide probes by maximising the target-to-non-target score difference, Eq. (S7).

7. Effect of blob size

In this work, we have coarse-grained the single-stranded genomic DNA into a polymer of "blobs", with each blob representing 400nt. Within each blob, it is assumed that the genome behaves as a self-avoiding walk polymer. A key tenet of coarse-grained polymer theory is that the results are independent of the chosen blob size (20, 21). Here, we test this by repeating some of our calculations using different blob sizes, i.e. different numbers of nucleotides per blob.

Figure S6(a) confirms that changing the blob size does not appreciably affect the binding specificity, i.e. the difference in blob-surface interaction free energy between the *E. coli* and *B. subtilis* genomes, for an *E. coli*-targeting surface. The absolute values of the blob-surface interaction free energies $\Delta\tilde{G}_{j,\text{surf}}$ do change with the blob size, as we would expect. However, the difference between the target (EC) and non-target (BS) genomes (*i.e.* the binding specificity) stays constant. In these calculations, we require the radius of gyration r_b of a blob, which changes with blob size. Here, we have assumed that r_b follows the standard scaling law for a self-avoiding walk: $r_b = 10\text{nm}\times(l_b/400\text{nt})^{\nu}$ with the scaling exponent $\nu = 0.588$ (for $l_b = 400$nt we recover $r_c = 10$nm as in our other calculations). The mass of the blob was kept constant.

Using these blob-surface interaction free energies, we also performed Langevin dynamics simulations of genome-surface binding, for a range of blob sizes. Figure S6(b) shows that the results, in terms of genome-surface binding (number of nucleotides within 30nm of the surface) are essentially the same for blob sizes $l_b = 400$nt and $l_b = 800$nt. This is because, although larger blobs interact more strongly with the surface (Figure S6(a)), there are fewer of them per genome. For the smallest blob size $l_b = 200$nt our simulations are no longer independent of blob size (the overall genome-surface binding is predicted to be somewhat stronger). We attribute this to the fact that our model neglects blob-blob hybridisation interactions. For large blobs these are negligible because the vast majority of binding configurations are due to intra-blob hybridisation interactions. For smaller blobs, however, neglecting blob-blob hybridisation tends to push the binding equilibrium towards genome–surface probe binding.

While we could simulate blobs that are even larger than 800nt, our theory relies on the assumption that a single blob can bind at most a single surface probe. This assumption limits the probe density ρ that can be simulated, via the condition $r_b < \rho^{-0.5}$. Our chosen blob size of $l_b = 400$nt is large enough to neglect the blob-blob hybridisation term (Figure S6(b)) while being small enough to be able to model a wide range of probe densities.

8. Equivalence of changing probe density, temperature and salt concentration

In this work, we have mostly investigated genome-surface binding as a function of the probe surface density ρ, keeping the temperature and salt concentration fixed at $T = 50°$C and 1M NaCl. This was done for convenience, since the SantaLucia interactions need to be recalculated if the temperature or salt concentration change. However,it turns out that varying ρ is equivalent to varying either temperature or salt concentration.

Figure S7 plots blob-surface interaction free energies for the *E. coli* wild-type genome binding to an *E. coli*-targeting surface (20nt probes), for a range of probe densities ρ, temperatures T and salt concentrations c. Reducing the salt concentration to $c = 0.1$M weakens the interaction, while increasing the temperature also uniformly weakens the interaction. If both the temperature and salt concentration are reduced the interactions remain virtually unchanged (compare the data for $T = 50, c = 1$M and $T = 40, c = 0.1$M in Figure S7). Thus, to a first approximation changing temperature or salt concentration simply translates the blob-surface binding interactions towards stronger/weaker binding, as does changing the probe density (Eq. (S12), and compare red and violet curves in Figure S7). Therefore the results that we plot as a function of probe surface density would look essentially equivalent if plotted as a function of either temperature or salt concentration.

9. Genome similarity calculation for the two *E. coli* strains

In this work, we have used two *E. coli* strains, O157 Sakai and bl21-de3, as examples of closely related genomes. It is useful to be able to quantify the similarity between these genomes and to identify which parts of the genome are similar between the strains.

To compare two genomes A and B, we first identify a subsequence j of genome B – this might be the sequence corresponding to one of the blobs in our coarse-grained polymer model. We then consider all subsequences of length l_s from genome A, and count the number of times they appear exactly within subsequence j of genome B. This quantity is denoted n_{A,B_j}. We then normalise by the number of distinct sequences of length l_s that exist in subsequence j, which is $l_b - l_s + 1$, where l_b is the length of subsequence j, to get a similarity measure $S^{\text{sim}}_{\text{A},\text{B}_j}$:

$$S^{\text{sim}}_{\text{A},\text{B}_j} = n_{\text{A},\text{B}_j}/(l_b - l_s + 1)\ . \qquad \text{[S20]}$$

If genome A and genome B were identical, the measure $S^{\mathrm{sim}}_{\mathrm{A},\mathrm{B}_j}$ would be 1 or greater (it would be 1 if every subsequence of length l_s were unique but will be increased if the genome contains repeat subsequences). If, in contrast, there were no matching regions of length l_s in subsequence j on genome, the similarity measure would be zero.

This similarity measure clearly depends on the chosen length l_s of the subsequences in genome A. For practical purposes, it is useful to choose a length l_s comparable to that of the oligonucleotide probes, since this allows us to estimate how well two genomes can be discriminated by the probes.

Figure S8(a) shows the similarity measure $S^{\mathrm{sim}}_{\mathrm{A},\mathrm{B}_j}$, computed between the *E. coli* 0157 Sakai and *E. coli* bl21-de3 wild type strains, taking $l_s = 20$nt, and using the 400nt blobs of the bl21-de3 as the subsequences j. Indeed, the two genomes are very similar, but there do exist regions of dissimilarity that are dispersed throughout the genome (i.e. they happen for multiple blobs). Reassuringly it is these regions of dissimilarity which are selected by our probe selection algorithm, Eq. (S7); this is evident when we plot (in Figure S8(b)) the blob-surface hybridisation free energies $\Delta\tilde{G}_{j,\mathrm{surf}}$ for a surface coated in 20nt probes designed by the "ΔS" method of Eq. (S7) to discriminate between the strains.

Although other measures for quantifying genome similarity exist and are widely used, for our purposes the measure $S^{\mathrm{sim}}_{\mathrm{A},\mathrm{B}_j}$ provides a useful way to estimate similarities along the coarse-grained polymer chain on a "blob to blob" basis, taking account of the chosen length of the oligonucleotide probes.

10. Performance of our approach compared to that of existing DNA microarray probes

To assess the performance of our proposed multivalent probe design strategy, we compared our results to simulations of target-probe binding for several published probe sequences designed for DNA detection via DNA microarrays. These were:

- Probe A (40nt, Wiesinger et al. (2007) (22)) GTAACGTCAATGAGCAAAGGTATTAACTTTACTCCCTTCC. This targets the 16S ribosomal RNA gene of *E. coli* (therefore it should bind *E. coli* DNA from all strains).
- Probe B (70nt, Vora et al. (2004) (23)) GGTTGTCACGAATGACAAAACACTTTATGACCGTTGTTTACATTT-TAAAGGCCAAGGATTAGCTGTACAT. This targets the *rfbE* gene which is specific to the O157 *E. coli* strain.
- Probe C (27nt, Jin et al. (2005) (24)) GGTGGAATGGTTGTCACGAATGACAAA. This probe also targets the *rfbE* gene which is specific to the O157 *E. coli* strain.

Using our theoretical approach, we investigated the binding free energies for probe A (and its reverse complement) interacting with both *E coli* wild-type and *B. subtilis* genomic DNA, and for probes B and C interactions with the *E coli* wild-type and O157 Sakai strain genomic DNA. The results are shown in Figure S9 (left panels). Figure S9(a) shows that probe A has several strong binding sites along the *E coli* wild-type genome: comparing with Figure S3 (b) for our 10nt probes we see that probe A binds more strongly than our probes, but with far fewer binding sites on the genome. The picture is similar for probes B and C (Figure S9(c) and (e)): both of these probes have only two binding sites on the O157 Sakai genome, corresponding to the locations of the *rfbE* gene in the forward and reverse strands. In contrast, Figure S8 (b) shows that our Sakai-targeting 20nt probe has many more binding sites along the O157 Sakai genome, although each one is weaker.

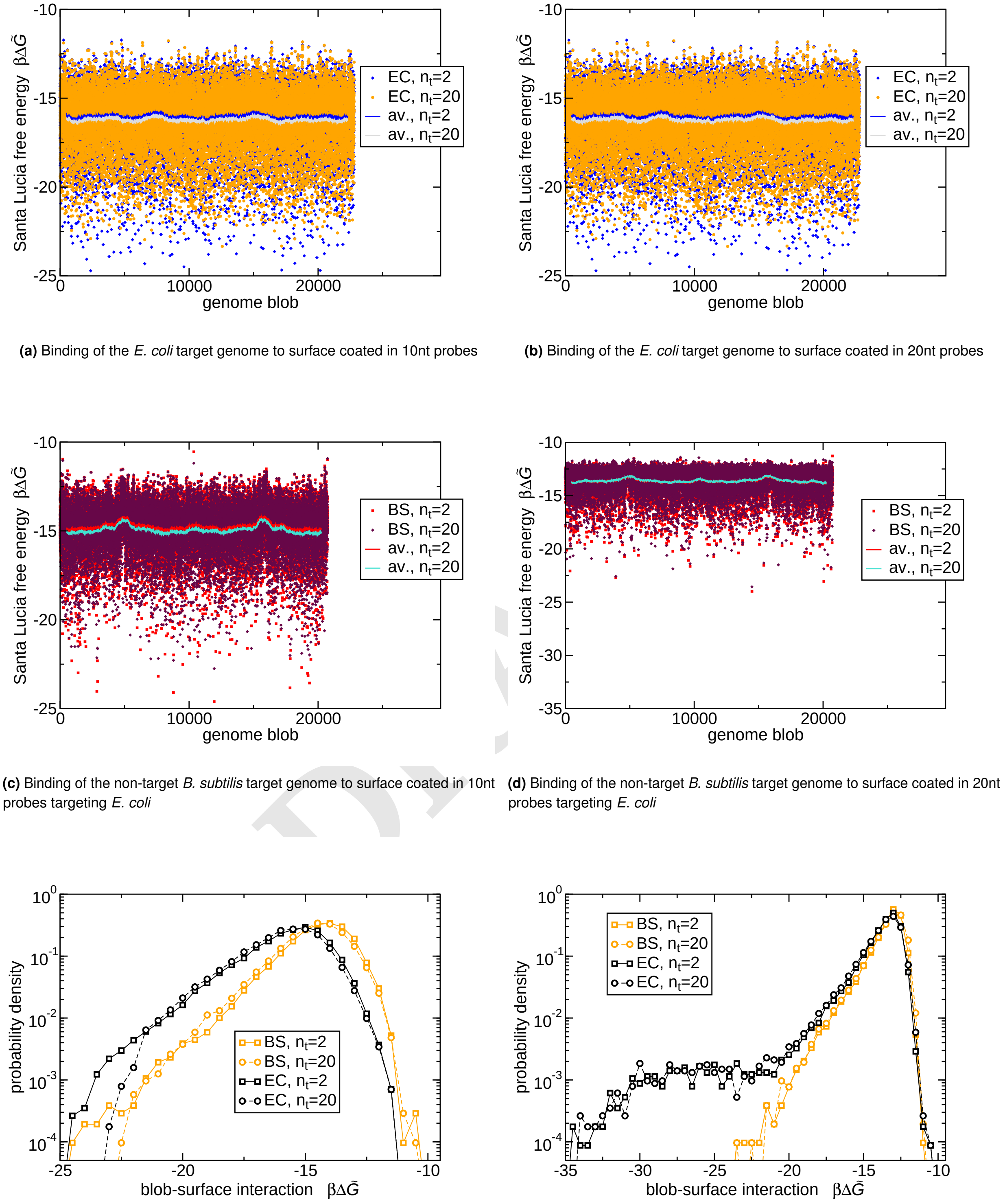
We also performed Langevin dynamics simulations of the binding of target and non-target DNA to surfaces coated in probes A, B and C (and their reverse complements), compared to surfaces coated in our top-scoring probes. The results are shown in the right-hand panels of Figure S9. In each case, we assumed that the genomic DNA was fragmented into 400nt fragments for binding to the literature probes (A, B and C) (since short fragments are typically used in DNA microarray methods (23)). For the simulations with our 10nt probes we assumed unfragmented genomic DNA.

Figure S9 (b) compares the simulated binding of *E. coli* wild-type DNA and *B. subtilis* DNA to a probe A-coated surface, and to a surface coated in our 10nt *E. coli*-targeting probes. The data shows increased selectivity (difference in binding between the target and non-target DNA), and increased sensitivity (strong binding at low probe density) for our multivalency approach compared to the existing probe A.

Similar results were obtained when we simulated the binding of *E. coli* O157 Sakai and *E. coli* wild-type DNA to surfaces coated in either probe B/C, or our 20nt probes targeting O157 Sakai (designed using the ΔS method). Figure S9 (d) and (f) show that the sensitivity and selectivity of binding are improved for our multivalent binding approach compared to either probe B or C.

We also checked the performance of our method compared to probes A, B and C for the same level of genome fragmentation. Figure S10 shows equivalent results to those of the right panels of Figure S9, but in the case where the genomic DNA is fragmented into 4000nt fragments (for all probes). Our multiple target probe design approach still shows improved sensitivity and selectivity in this case, even though the overall performance is slightly diminished by the fragmentation of the DNA.

1. Dubacheva GV, et al. (2014) Superselective targeting using multivalent polymers. *J. Am. Chem. Soc.* 136(5):1722–1725.
2. Dubacheva GV, Curk T, Auzély-Velty R, Frenkel D, Richter RP (2015) Designing multivalent probes for tunable superselective targeting. *Proc. Natl. Acad. Sci. USA* 112(18):5579–5584.
3. Curk T, Dobnikar J, Frenkel D (2017) Optimal multivalent targeting of membranes with many distinct receptors. *Proc. Natl. Acad. Sci. USA* pp. 7210–7215.
4. Dubacheva GV, Curk T, Frenkel D, Richter RP (2019) Multivalent recognition at fluid surfaces: The interplay of receptor clustering and superselectivity. *J. Am. Chem. Soc.* 141(6):2577–2588.
5. Camacho C, et al. (2009) Blast+: architecture and applications. *BMC Bioinformatics* 10:421.
6. Sim AYL, Lipfert J, Herschlag D, Doniach S (2012) Salt dependence of the radius of gyration and flexibility of single-stranded dna in solution probed by small-angle x-ray scattering. *Phys. Rev. E* 86(2):021901.
7. de Gennes PG (1979) *Scaling concepts in polymer physics.* (Cornell University Press).
8. Toan NM, Micheletti C (2006) Inferring the effective thickness of polyelectrolytes from stretching measurements at various ionic strengths: applications to dna and rna. *J. Phys: Condensed Matter* 18(14):S269–S281.
9. Chen H, et al. (2012) Ionic strength-dependent persistence lengths of single-stranded rna and dna. *Proc. Natl. Acad. Sci. USA* 109(3):799–804.
10. Henrich O, Gutiérrez Fosado YA, Curk T, Ouldridge TE (2018) Coarse-grained simulation of dna using lammps. *Eur. Phys. J. E* 41(5):57.
11. Plimpton S (1995) Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Phys.* 117(1):1–19.
12. Crothers DM, Zimm BH (1964) Theory of melting transition of synthetic polynucleotides – evaluation of stacking free energy. *J. Mol. Biol.* 9:1–9.
13. DeVoe H, Tinoco I (1962) Stability of helical polynucleotides – base contributions. *J. Mol. Biol.* 4:500–517.
14. SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide dna nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* 95(4):1460–1465.
15. SantaLucia Jr J, Hicks D (2004). *Annu. Rev. Biophys. Biomol. Struct.* 33:415.
16. Zadeh JN, et al. (2011) NUPACK: analysis and design of nucleic acid systems. *J. Comput. Chem.* 32:170–173.
17. Dirks RM, Pierce NA (2003) A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24:1664–1677.
18. Dirks RM, Pierce NA (2004) An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comput. Chem.* 25:1295–1304.
19. Dirks RM, Bois JS, Schaeffer JM, Winfree E, Pierce NA (2007) Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev.* 49:65–88.
20. Pierleoni C, Capone B, Hansen JP (2007) A soft effective segment representation of semidilute polymer solutions. *J. Chem. Phys.* 127(17):171102.
21. Louis AA, Bolhuis PG, Hansen JP, Meijer EJ (2000) Can polymer coils be modeled as “soft colloids”? *Phys. Rev. Lett.* 85(12):2522–2525.
22. Wiesinger-Mayr H, et al. (2007) Identification of human pathogens isolated from blood using microarray hybridisation and signal pattern recognition. *BMC Microbiology* 7(1):78.
23. Vora GJ, Meador CE, Stenger DA, Andreadis JD (2004) Nucleic acid amplification strategies for dna microarray-based pathogen detection. *Appl. Environ. Microbiol.* 70(5):3047–3054.
24. Jin HY, Tao KH, Li YX, Li FQ, Li SQ (2005). *World J. Gastroenterol.* 11(37):5811–5815.



(a) Binding of the *E. coli* target genome to surface coated in 10nt probes

(b) Binding of the *E. coli* target genome to surface coated in 20nt probes

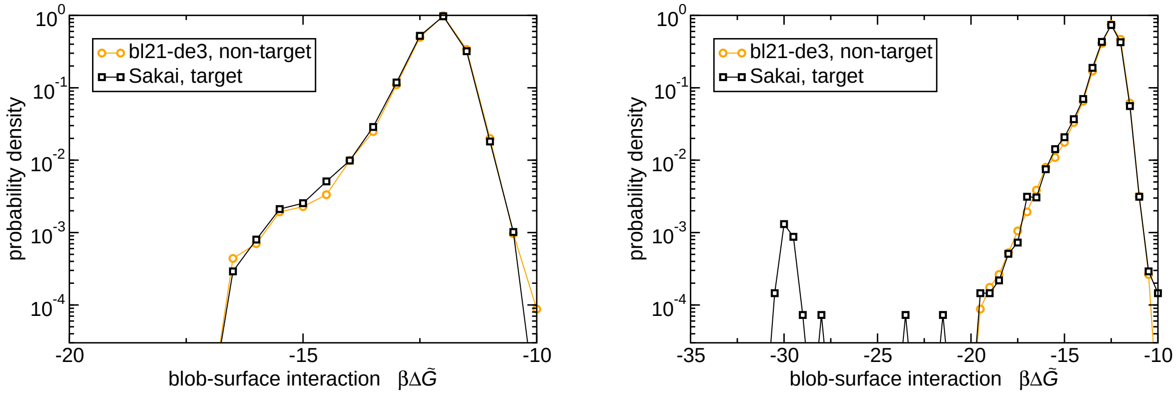
(c) Binding of the non-target *B. subtilis* target genome to surface coated in 10nt probes targeting *E. coli*

(d) Binding of the non-target *B. subtilis* target genome to surface coated in 20nt probes targeting *E. coli*

(e) Histogram of binding free energies for a surface coated in 10nt probes targeting *E. coli*

(f) Histogram of binding free energies for a surface coated in 20nt probes targeting *E. coli*

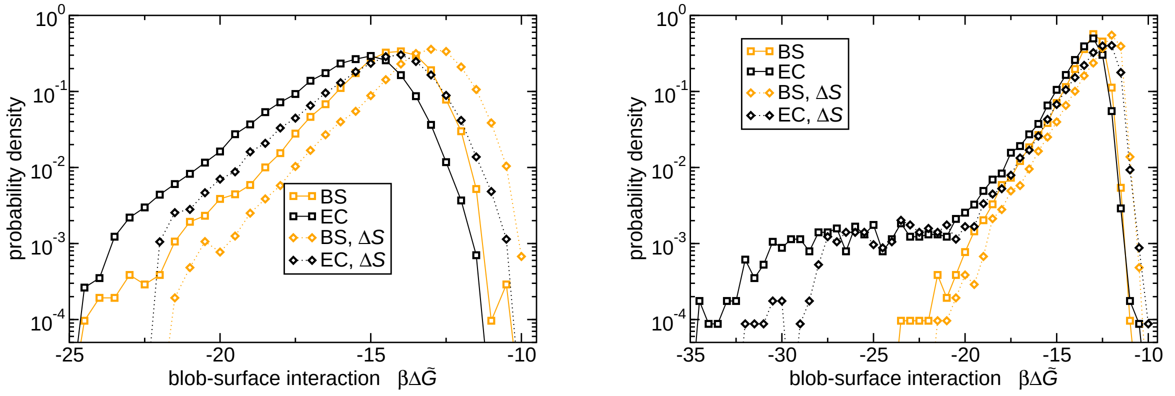
Fig. S3. Effects of changing probe length, and of using a mixture of probe sequences, for probes targeting the *E. coli* wild-type genome. Panels (a)-(d) show individual blob-surface binding free energy values $\Delta\tilde{G}_{j,\mathrm{surf}}$, obtained using Eq. (S12), together with moving averages over 1000 consecutive blobs. Panels (e) and (f) show histograms of blob-surface binding free energy values. The left panels correspond to short oligonucleotide probes (10nt), while the right panels are for longer probes (20nt). In the legends, 'BS" stands for the *B. subtilis* QB928 genome and "EC" stands for the *E. coli* bl21-de3 wild-type genome. $n_t = 2$ refers to a surface coated in the highest-scoring probe sequence, plus its reverse complement, while $n_t = 20$ refers to a surface coated in a mixture of the 10 highest-scoring probe sequences, plus their reverse complements. The other parameters are: $\rho = 0.1 r_b^{-2}$, $l_b = 400$nt, $T = 50^\circ$C.



(a) $l_s = 10$

(b) $l_s = 20$

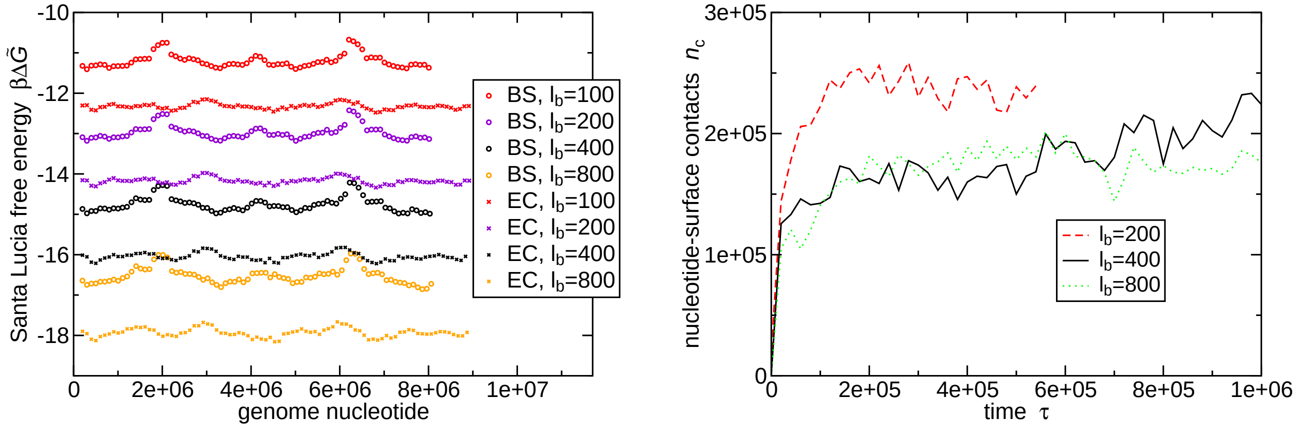
Fig. S4. Histogram of SantaLucia blob-surface interaction free energies $\Delta\tilde{G}_{j,\mathrm{surf}}$, for the two *E. coli* strains 0157 Sakai and bl21-de3, interacting with a surface that is coated in probes designed to obtain the largest score difference between the strains, $\Delta S = S_{Sakai} - S_{bl21-de3}$. The calculations are performed with the single top-scoring probe sequence, plus its reverse complement. In panel (a) the probe length is 10nt; in panel b) it is 20nt. The black lines correspond to binnding of the target Sakai strain; the orange lines correspond to binding of the non-target wild-type strain. The other parameters are $\rho = 0.1 r_b^2$, $l_b = 400$nt, $T = 50^\circ$C.



(a) $l_s = 10$

(b) $l_s = 20$

Fig. S5. The "ΔS" probe selection method has little effect on selectivity between different bacterial species. Histograms of blob-surface interaction free energies $\Delta\tilde{G}_{j,\mathrm{surf}}$ are shown for a surface targeting *E. coli* wild-type, for probes selected by maximising the score S (square symbols), or using the ΔS method (diamond symbols). Using the ΔS method shifts the interaction histograms but does not affect the selectivity of binding between the targeted *E. coli* wild-type genome and the non-targeted *B. subtilis* genome. In panel a) the probe length is 10nt; panel b) corresponds to 20nt probes. In each case a single probe sequence is used, together with its reverse complement. The other parameters are $\rho = 0.1 r_b^{-2}$, $l_b = 400$nt, $T = 50^\circ$C.



(a)

(b)

Fig. S6. Effect of changing the blob size in our coarse-grained simulations. In panel (a) the 40000 nucleotide moving average of the blob-surface interaction free energy $\Delta\tilde{G}_{j,\mathrm{surf}}$ is plotted for different blob sizes $l_b = 100, 200, 400, 800$nt, for a surface coated in 10nt probes designed to target *E. coli* bl21-de3. In the legend, EC denotes binding of the target *E. coli* genome and BS denotes binding of the non-target *B. subtilis* genome. Panel b) shows results of our Langevin dynamics simulations, for three different blob sizes. The number of nucleotides within a 30 nm distance of the surface is plotted as a function of time for simulations of the *E. coli* genome binding to a surface coated in 10nt *E. coli*-targeting probes. The parameters are $\rho = 0.1r_b^2, T = 50^\circ$C.

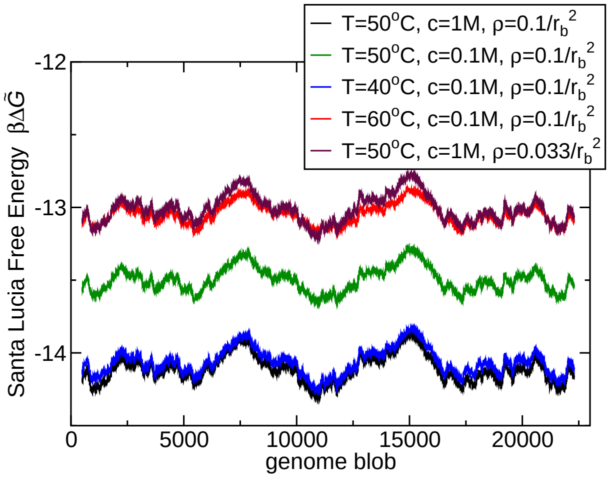
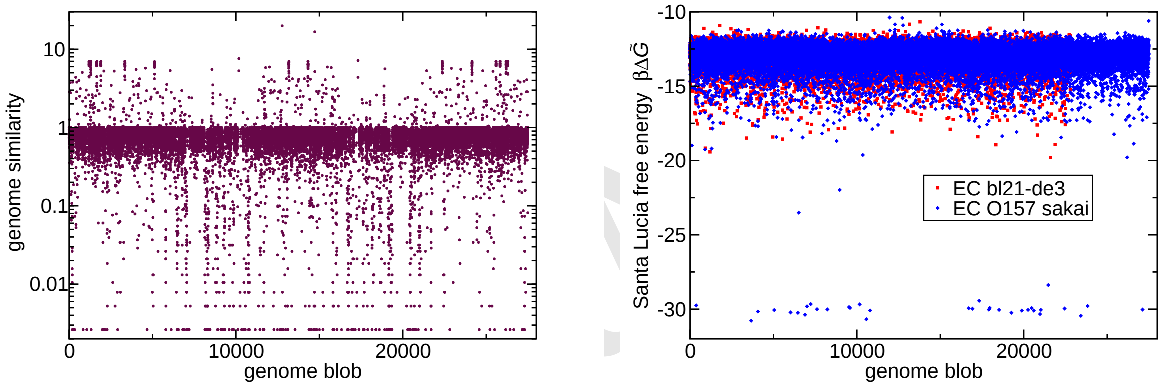


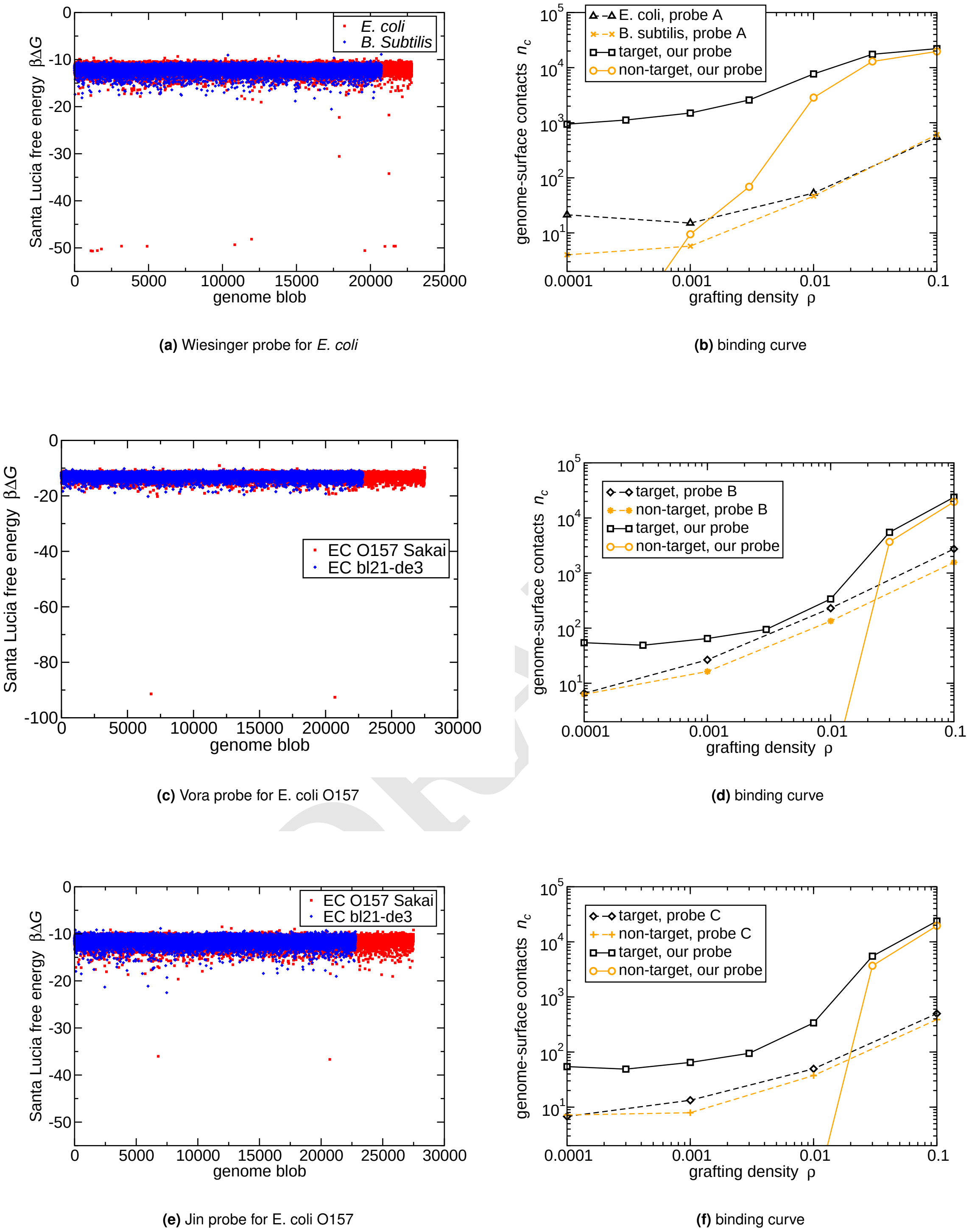
Fig. S7. Effect of changing the temperature T, salt concentration c and the probe surface density ρ on genome-surface binding free energies. The plot shows 400000 nucleotide moving averages of the SantaLucia blob-surface binding free energy $\Delta\tilde{G}$ for an *E. coli* bl21-de3 genome binding to a surface grafted with the top scoring 20nt probe and its reverse complement. The black curve shows the same data as in Figure S3(b) at $T = 50^\circ$C and $c = 1$M, which are also the values used in all previous calculations.



(a)

(b)

Fig. S8. Distinguishing between similar genomes: *E. coli* 0157 Sakai and *E. coli* bl21-de3 wild type. (a) Genome similarity calculated between bl21-de3 and 400 nucleotide segments (blobs) of the 0157 Sakai strain, applying Eq. (S20). (b) SantaLucia blob-surface interaction energies $\Delta\tilde{G}_{j,\mathrm{surf}}$ for 0157 Sakai and bl21-de3 genomic DNA binding to a surface coated with the top scoring 20nt probe that maximises the difference function $\Delta S = S_{Sakai} - S_{bl21-de3}$, and its reverse complement. This data corresponds to the genome-surface binding plots in Fig 3 in the main text.



(a) Wiesinger probe for *E. coli*

(b) binding curve

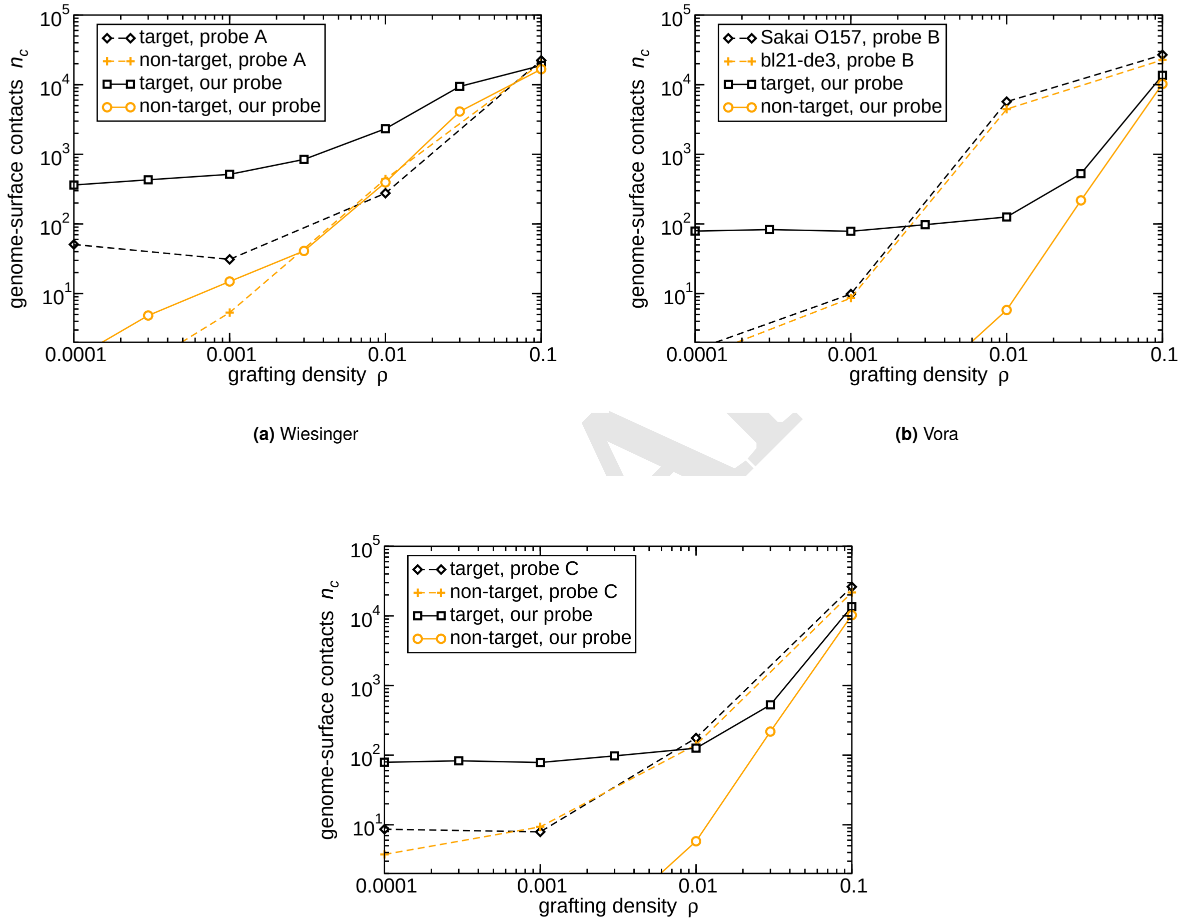
(c) Vora probe for E. coli O157

(d) binding curve

(e) Jin probe for E. coli O157

(f) binding curve

Fig. S9. Comparing the performance of our "multivalent binding" approach with probes A, B and C taken from the literature. The left panels show blob-surface SantaLucia interaction free energies $\Delta\tilde{G}_{j,\mathrm{surf}}$ along the target and non-target genomes, for probe density $\rho = 0.01$, for (a) probe A, target = *E. coli* wild-type, non-target = *B. subtilis*; (b): probe B, target = *E. coli* O157 Sakai, non-target = *E. coli* wild-type; and (c): probe C, target = *E. coli* O157 Sakai, non-target = *E. coli* wild-type. In all cases the forward and reverse genomic DNA strands are included. The right panels show the results of Langevin dynamics simulations of binding of target and non-target DNA binding to surfaces coated in either the literature probe or our top-scoring probe (including both forward and reverse probe strands). For simulations with the literature probes the genome was assumed to be fragmented into 400nt (1 blob) fragments; for simulations with our probes the genome was assumed to be unfragmented. Panel (b) shows results for probe A versus our top-scoring 10nt *E. coli* probe, for *E. coli* wild-type or *B. subtilis* DNA. Panel (d) shows results for probe B versus our top-scoring 20nt O157 Sakai probe, for *E. coli* O157 Sakai or wild-type. Panel (e) is equivalent to panel (d) but for probe C. In all cases, $T = 50^\circ$C. In the simulation plots, we note that the number of genome-surface contacts does not go to zero as the probe density decreases to zero. This is because, in our simulations, a single genome (forward and reverse strands) is confined in a simulation box of size $10 \times 10 \times 50\mu$m. Even in the absence of any genome-surface binding, we expect on average 10 of the genome blobs to be close enough to the surface to be classified as being in "surface contact".



(a) Wiesinger

(b) Vora

(c) Jin

Fig. S10. Equivalent simulation results to those of Figure S9 (right panels), but for the case where the genomic DNA is fragmented into 4000nt (10 blob) fragments for both the literature probes and our probes.