The rugged DNA-binding sequence-fitness landscape of the LacI/GalR Family is a product of asymmetry in the operator:repressor complex

Anthony T. Meger[1,‡], Matthew A. Spence[2,‡], Mahakaran Sandhu[2], Colin J. Jackson[2,3,4,*], Srivatsan Raman[1,5,6,*]

[1]Department of Biochemistry, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[2]Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia

[3]ARC Centre of Excellence for Innovations in Peptide & Protein Science, Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia

[4]ARC Centre of Excellence for Innovations in Synthetic Biology, Research School of Chemistry, Australian National University, Canberra, ACT 2601, Australia

[5]Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[6]Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, Wisconsin 53706, United States

[‡]These authors contributed equally to this work

Contact Information

*Corresponding authors: Colin Jackson (colin.jackson@anu.edu.au) and Srivatsan Raman (sraman4@wisc.edu)

ABSTRACT

Our understanding of how proteins evolve over time and across fitness landscapes of variable ruggedness remains incomplete. Our ability to generate and characterize expansive and sufficiently diverse datasets to examine this question has been limited, which has hampered efforts to understand why some fitness landscapes are more rugged than others and how they are a product of protein structure and function. Here we describe the use of ancestral sequence reconstruction to generate a 1158 sequence phylogeny of the DNA-binding domain (DBD) of the LacI/GalR transcriptional repressor family. This revealed a rugged fitness landscape, which we rationalize through biophysical analysis to be intrinsic to the structure and function of DBDs: the asymmetry of the operator-DBD complex means that the fitness landscape is in fact a composite of two separate fitness landscapes. This ruggedness underlies the rapid evolution of orthogonal specificity within DBDs, which is an essential part of their function.

KEYWORDS: ancestral sequence reconstruction (ASR); protein evolution; deep mutational scanning (DMS); epistasis; sequence space; fitness landscape

INTRODUCTION

A central question in molecular evolution is how proteins acquire novel functions, such as new binding specificities, via mutation and selection. The sequence-fitness landscape is a valuable construct that has allowed us to understand and visualize the complex relationship between evolutionary sequence changes and protein function.[1–5] The topology of a fitness landscape largely reflects the nature of the evolutionary process: “smooth” landscapes, with small numbers of discrete but connected peaks across which new activities can gradually evolve via additive and predictable mutational steps, are relatively well understood, with many enzymes and binding proteins as examples.[2,6–10] The molecular basis for the smoothness of these transitions can be partly explained by the modulation of the conformational dynamics of proteins, in which mutations can gradually shift the conformational equilibrium towards conformations better suited to the new activity.[11–13] In contrast, our understanding and characterization of “rugged” fitness landscapes, in which mutations tend to have unpredictable epistatic effects that result in fitness landscapes consisting of multiple peaks separated by many non-functional sequences (“valleys”), is less well developed.[14–16] One intriguing hypothesis is that the sequence-fitness landscapes of certain proteins are inherently defined by the function and fold.[17] This partially deterministic view has important implications for evolution of new activities (such as drug[17,18] or vaccine resistance[19]) and protein engineering. [20,21]

To understand and characterize the role of rugged fitness landscapes in molecular evolution, we need to comprehensively map sequence-function relationships over large and diverse sequence spaces that span the full evolutionary history of protein families. While directed evolution and deep mutational scanning (DMS) experiments have provided profound insight into molecular evolution at the level of single mutational steps, they seldom explore the large evolutionary timescales and the full diversity of sequence space within protein families.[4,7] In contrast, ancestral sequence reconstruction (ASR) can sample large spans of sequence space through the computational

67 reconstruction of ancestral species from a phylogenetic tree and a sequence evolution model.[22,23]
68 However, ASR studies generally characterize a relatively small number of divergent ancestral
69 sequences, which makes it difficult to deconvolute adaptive and neutral mutations and understand
70 the stepwise additive and context-specific (epistatic) effects of mutations.[24]
71
72 In contrast to enzymes, which have been shown to often evolve across relatively smooth fitness
73 landscapes in which promiscuous activities can be gradually optimized to become the primary
74 function,[11–13] transcription factors likely display a more rugged fitness landscape as they are
75 generally defined by high DNA specificity, and incremental mutations leading to promiscuity could
76 be evolutionarily disfavored. The lac repressor (LacI), which belongs to the LacI/GalR family
77 (LGF) of prokaryotic gene regulators,[25] is a well-studied model for DNA recognition and allostery.[26]
78 Proteins of the LGF show remarkable diversity in their amino acid sequences and their DNA
79 recognition. While much is known about the operator specificity of *Escherichia coli* LacI
80 (EcLacI),[27–31] our understanding of the sequence-fitness landscape of DNA-binding specificity in
81 the wider LGF, the historical evolutionary trajectory of operator recognition and the molecular
82 mechanisms that underpin the evolution of DNA-specificity in proteins remains incomplete. The
83 LGF is thus a compelling system to study ruggedness in sequence-fitness landscapes and how
84 the combination of structure and function dictate the evolutionary dynamics and selective
85 pressures exerted on these regulators.
86
87 In this study, we experimentally characterize a complete phylogenetic tree (1158 extant and
88 ancestral sequences) of the LGF and perform DMS on extant EcLacI to reveal the sequence-
89 fitness landscape of DNA specificity for the *E. coli* Lac operator. We find the landscape to be
90 extremely rugged due to high levels of epistasis, with most sequences having no affinity for the
91 *E. coli* lac operator sequence. However, the screen unearthed dozens of functional repressors
92 from distinct phylogenetic clades, with as many as 32/60 amino acid substitutions in the DNA

93 binding domain (DBD) compared to the EcLacI DBD. Analysis of the local evolutionary landscape
94 within clades shows gain/loss of function between adjacent nodes, indicating rapid switches of
95 specificity, which may be beneficial for developing orthogonal genetic regulation. The ruggedness
96 of this fitness landscape may therefore be a characteristic feature of regulators and an essential
97 functional requirement of high specificity and orthogonality. The molecular basis for this
98 observation is revealed through *in vitro* binding assays and simulations, which show that
99 ruggedness of the LGF fitness landscape arises in part due to the necessity for regulators to
100 simultaneously evolve specificity for either DNA half-site in asymmetric operators and explains
101 why this fold and operator structure has been evolutionarily selected for genetic regulation. This
102 study provides new fundamental insight into nature's design of genetic regulation, evolutionary
103 rules of protein-DNA recognition, the biophysical mechanism of transcription factor allostery, and
104 how structure and function combine to determine the sequence-fitness landscapes of proteins.
105

106 **RESULTS**

107 **Synthesis and characterization of a complete phylogeny of LGF DNA-binding domains.**
108 The *E. coli* lac operator sequence has emerged over a ~3-billion-year evolutionary history. To
109 generate a set of DBDs that spans the full evolutionarily accessible sequence space of the LGF,
110 we performed phylogenetic inference to reconstruct 577 ancestral sequences from a dataset of
111 581 extant EcLacI homologs (**Fig. 1a**). Consistent with previous phylogenetic studies,[32] our
112 analysis reconstructs the family as three major lineages. When rooted at the phylogenetic
113 midpoint, these include the LacI-clade as the most ancestral, a single common ancestor giving
114 rise to descendent clades comprising the catabolite control protein (CcpA/RegA), ribose, maltose,
115 sucrose, galactose, arbutin/salicin, purine and cytosine regulators (RbsR, MalR, SacR/ScrR,
116 GalR/GalS, AscG, PurR and CytR, respectively), and an uncharacterized lineage of LacI-like
117 proteins (**Fig. 1a**). On average, the mean posterior probability of each ancestral DBD was 93%,
118 indicating strong statistical support across the full dataset (**Supplementary Fig. 1a**). Likewise,

119 phylogenetic branch supports were consistently high across the topology (**Supplementary Fig.**
120 **1b**), and the approximately unbiased test failed to reject the topology we present in Fig. 1 among
121 9 other tree-search replicates (P-value=0.682)[33] (**Supplementary Fig. 2**).
122
123 One limitation of ASR is that, while it is possible to generate hundreds of ancestral sequences
124 computationally, only a handful (<10) of evolutionarily distant nodes are typically characterized
125 experimentally. Thus, the incremental functional adaptations associated with the full evolutionary
126 trajectory can remain obscured by neutral variation. Here, we synthesized the DBDs of all 577
127 reconstructed ancestors, as well as all 581 extant sequences used in our phylogeny, generating
128 a dataset of 1158 diverse DBDs that fully covers the phylogenetic tree constructed in this work
129 (**Fig. 1a**). We used chip-based oligonucleotides to synthesize the DBD sequences and cloned
130 them into a plasmid library to encode chimeric variants consisting of DBDs fused to the ligand-
131 binding domain (LBD) of EcLacI, which is induced by allolactose/IPTG (**Fig. 1b**). We constructed
132 chimeras with an invariant LBD for several reasons: (i) by studying the ancestral DBD in the
133 context of the EcLacI LBD, we can be confident that the regulatory mode of the chimeras will
134 match that of the extant EcLacI and will repress in the absence of an allosteric effector;[34] (ii) since
135 the inducer of the EcLacI LBD is known, we can study the allosteric communication between LBD
136 and DBD in the chimeras; (iii) we can ensure that the experimental conditions we used were
137 conducive to repression, giving accurate insight on exclusively DNA specificity in the DBD. Deep
138 sequencing the plasmid library confirmed 100% coverage of the DBD library with minimal skew
139 i.e., all 1158 DBDs were present (**Supplementary Fig. 3**).
140
141

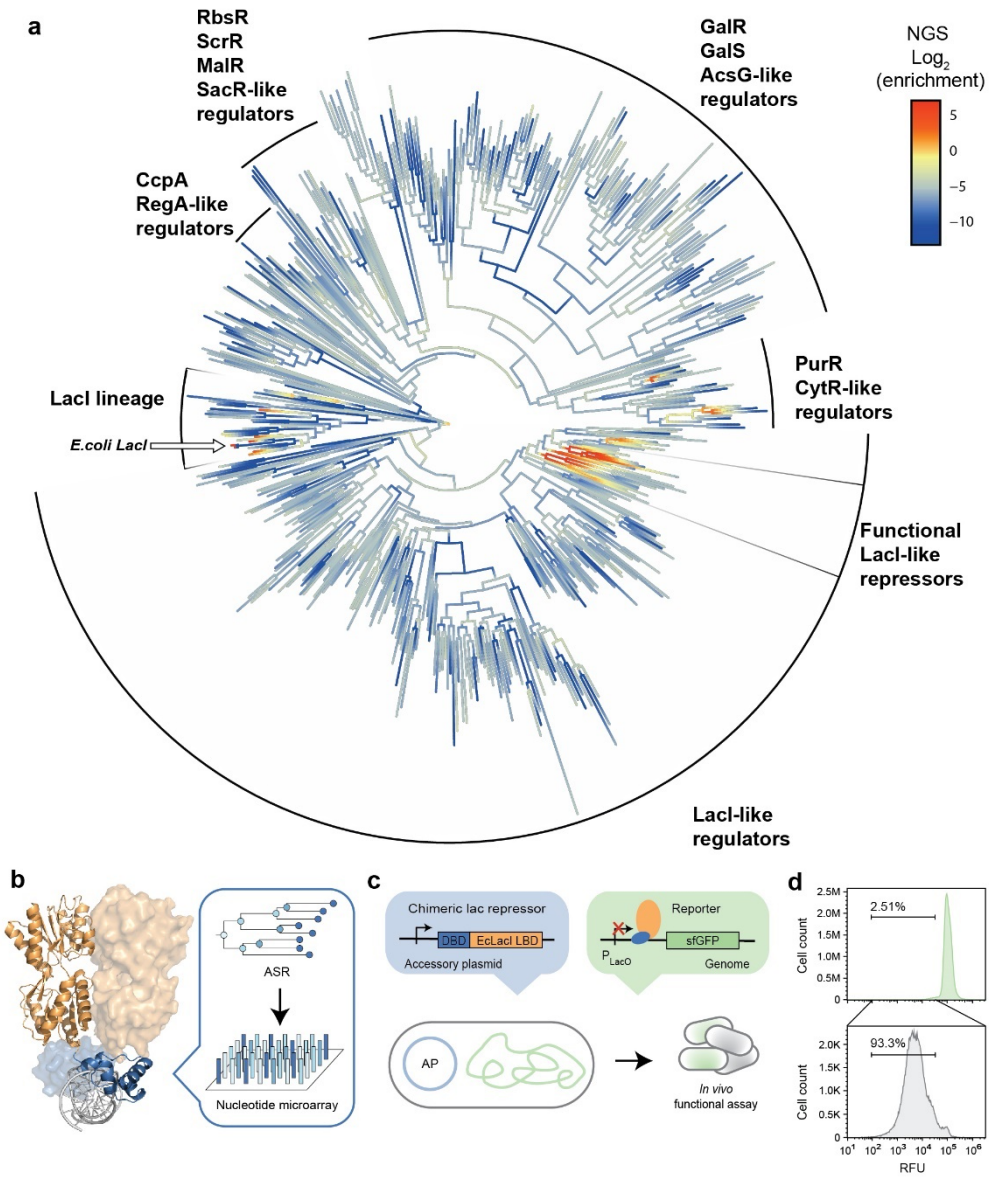


Fig. 1. Evolution of LacO recognition. (**a**), Phylogenetic inference of the LacI/GalR family of prokaryotic transcription factors. 577 sequences were ancestrally reconstructed using from a dataset of 581 extant LacI homologs (**Supplementary Fig. 1,** see "methods"). Color gradient shows average repression scores from replicate (*n*=2) high-throughput screens (**Supplementary Fig. 4**). (**b**), Structure (PDB ID: 1EFA) of dimeric extant *E. coli* lac repressor bound to LacO (left). LBD (orange), DBD (blue), and LacO (gray) are shown in cartoon representation. Each extant and ancestral DBD was generated using multiplex micro-array DNA synthesis (right). (**c**), *In vivo* characterization of repressor function. DBDs were fused to the *E. coli* LacI LBD and constitutively expressed on an accessory plasmid. A reporter construct consisting of the gene encoding sfGFP under transcriptional control of LacO was integrated into the genome of E. coli to assay repressor function. (**d**), Flow cytometry of the pre-selected phylogenetic DBD library (top) and after two sequential sorts to isolate repression competent DBDs (bottom). The indicated gate was used for selection by FACS (**Supplementary Fig. 4**). RFU, relative fluorescence units.

156 To characterize the affinity of the DBDs within the phylogeny to the *E. coli* lac operator and their
157 allosteric activation by IPTG, we used a cell-based pooled screening strategy (**Fig. 1c**). We
158 inserted a single-copy GFP cassette under the control of the lac operator (LacO) into the *E. coli*
159 genome. Variants that can bind to LacO with sufficient affinity repress GFP expression, while
160 those that cannot produce high GFP expression. Using fluorescence-activated cell sorting
161 (FACS), repression-competent variants are enriched by sorting low GFP cells. Only a small
162 fraction (2.5-2.6%) of low-fluorescence cells were observed via flow cytometry, indicating that
163 relatively few extant and ancestral DBDs in this library repress LacO (**Fig. 1d**). The subset of
164 repression competent variants was enriched to 92.7-93.3% of the total population after two rounds
165 of sorting (**Supplementary Fig. 4**). The pre- and post-sorted libraries were sequenced to estimate
166 the enrichment ratios. A higher enrichment ratio implies greater GFP repression and is a proxy
167 for protein-DNA affinity. The enrichment ratios were highly correlated (R^2=0.99) between
168 independent replicates (**Supplementary Fig. 5**). We found 15 ancestral and 7 extant DBDs that
169 can repress GFP from the 1158-member library. These 22 enriched DBDs comprise 1.9% of the
170 1158-member library, consistent with the fraction of the low-fluorescence population observed via
171 flow cytometry.
172
173 **The sequence-fitness landscape of the DNA binding domains is rugged.** A fundamental
174 question in molecular evolution is whether the emergence of new functions is mediated by the
175 gradual partitioning of functions (smooth) or by discrete switches (rugged). On a smooth fitness
176 landscape, we would expect a single lineage to exhibit gradual changes in activity and maintain
177 significant promiscuity. For example, the binding specificities of proteins within the periplasmic
178 amino acid binding protein family (same fold as the LBD of the LGF family) were shown to
179 gradually alter over time along evolutionary lineages through the selection of promiscuous
180 functions.[24,35] In contrast, a rugged fitness landscape involving discrete specificity switches, would
181 result in different functions being sparsely dispersed across the phylogeny. Our results reveal an

182 extraordinarily rugged fitness landscape for the DBDs, in which functional repressors are
183 dispersed within multiple lineages across the phylogenetic tree (**Fig. 1a, Fig. 2a-c**). Indeed, *LacO*
184 repression appears to have emerged in three evolutionarily distinct lineages; the EcLacI clade,
185 PurR/CytR clade and a third clade that we dub the functional LacI-like regulators. Notably, these
186 clades are evolutionarily distinct and share an ancestor only at the LCA of the full LGF, which
187 diverged ~3Gya.
188
189 We next traced the evolution of function within the clade in which *LacO* repression was observed,
190 showing that recognition of LacO is sporadic within each group. For instance, in the EcLacI clade,
191 out of the 28 ancestral sequences, only seven were repression competent. Additionally, we
192 identify 32 instances where LacO recognition is either gained or lost between adjacent nodes
193 (**Fig. 2a-c**). Thus, while evolutionary selection and optimization of ancestral function exist to some
194 extent, we observe that rapid gain or loss of that same function is more dominant across the LGF
195 phylogeny. In summary, the DBDs of the LGF family appear to be evolving across a rugged fitness
196 landscape, leading to rapid gain/loss of function transitions.

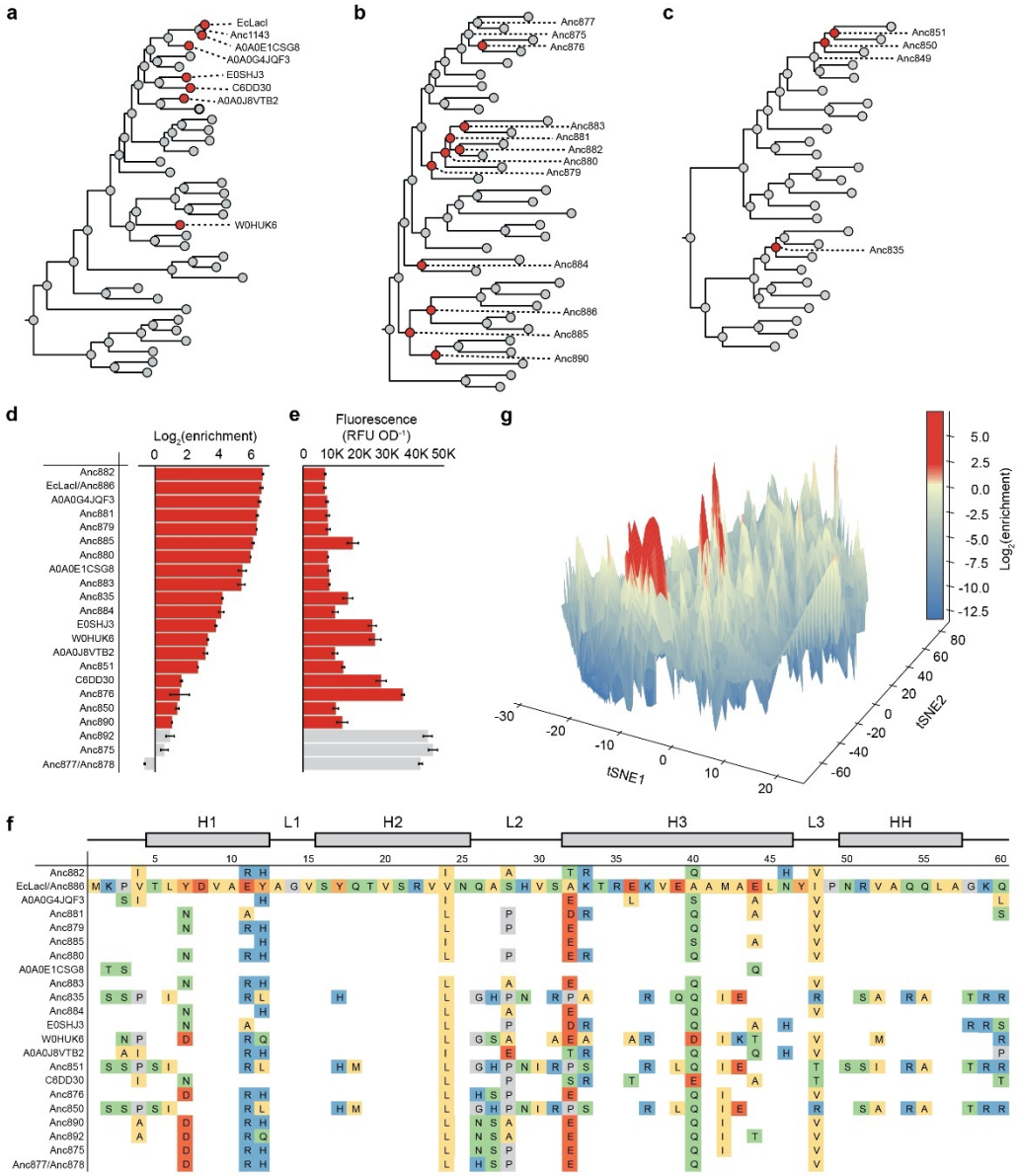


Fig. 2. Sequence-fitness landscape and phylogenetic mapping of LacO recognition. (**a-c**), Lineages containing functional repressors of LacO. Ancestrally reconstructed sequences are numbered by tree node and extant sequences are indicated with UniProt IDs. Red nodes indicate clonally validated functional repressors of LacO within the LacI lineage (**a**), functional LacI-like regulators (**b**), and PurR/CytR-like regulators (**c**), and gray nodes indicate non-functional repressors. (**d**)**,** Extant and ancestral DBD variants with the greatest enrichment from our high-throughput screens. Enrichment values were determined by comparing NGS distributions of the pre- and post-sorted populations. Error bars denote the standard deviation of replicate (n=2) sorting and NGS experiments. (**e**), Clonally assayed mean fluorescence values (n≥2) normalized to OD_{600}. (**f**), Primary sequences of characterized ancestral and extant DBDs and overall structural topology of the DBD (top). (**g**), Sequence-fitness landscape of repressor function. Enrichment values were mapped onto the t-SNE landscape of the UniRep represented phylogenetic DBD sequences (see “methods”) to assay ruggedness. RFU, relative fluorescence units. OD, optical density.

212 To understand how and why sequence changes altered LacO recognition, we investigated the
213 enrichment of different sequences, which showed a strong correlation between the enrichment
214 ratio and the ability of a variant to repress (**Fig. 2d-f, Supplementary Fig. 6)**. Clonal screening
215 revealed a near-binary functional switch in repression competency separating enriched and
216 depleted variants (**Fig. 2d,e**). Of the 1158-member variant library, 7 extant and 13 ancestral
217 sequences were clonally validated as repression competent. Analysis of the sequence
218 conservation patterns revealed among the four helices of the DBD, H2 and HH are most
219 conserved as these are essential for LacO recognition and allosteric activation, respectively (**Fig.
220 2f; Supplementary Fig. 7**).[36] Indeed, within the recognition helix (H2) residues Tyr17, Gln18 and
221 Arg22, which directly interact with the lac operator and control DNA specificity[37], are rarely
222 mutated in the functionally repressive DBDs. The only exceptions include Y17H and Q18M, which
223 have both been demonstrated to maintain lac operator binding in the genetic background of
224 *Ec*LacI.[38,39] Notably, ancestors 835, 850, and 851 have multiple substitutions in the HH; these
225 variants repress GFP but are not allosterically activated by 1 mM IPTG (**Supplementary Fig. 8**),
226 consistent with the role of HH in allosteric signaling.[36]
227
228 To obtain a global view of the sequence-function landscape, we encoded the DBD sequences as
229 high dimensional vectors with unified representation (UniRep), a deep learning model that allows
230 sequences with similar structural, functional, and evolutionary properties to be clustered together
231 in space (**Fig. 2g**).[40] This highlights the extreme ruggedness of the sequence-function landscape,
232 with three regions that correspond to the LacI lineage, functional LacI-like regulators, and
233 PurR/CytR-like regulators, containing most of the functional peaks. The rugged surface and
234 isolated functional peaks suggest that there is little promiscuous binding across the family, that
235 there is extreme epistasis with a small and finite number of functional sequences, and that there
236 is low robustness or tolerance to mutations. In summary, our broad analysis of LacO repressor
237 function across the 1158 sequences of this phylogenetic tree suggests that (a) LacO recognition

238 is an easily accessible evolutionary state (appeared independently multiple times) and (b) it
239 simultaneously exists within narrowly confined solution spaces (the fitness peaks that do exist are
240 very small and surrounded by inactive sequences).
241
242 **Diverse substitutions shape local ruggedness among related DBDs.** To better understand
243 how the sequence space surrounding the functional repressors affects function, we identified
244 specific mutations that affect function. This is challenging in genetically diverse datasets because
245 neutral genetic drift and epistasis can both mask and confound sequence-function relationships.
246 In this dataset, the context-dependence of three amino acid positions, 18, 22, and 26, illustrate
247 how epistasis creates ruggedness (**Fig. 2a-c**). First: Gln18 in the non-functional ancestor Anc849
248 is mutated to Met18 (previously shown to increase affinity (2.5-fold) to LacO in EcLacI[39]) in
249 Anc850 (alongside Glu39Leu), resulting in gain of function, followed by the introduction of
250 Arg48Thr, Asn50Ser, and Ala52Ile in Anc851, which retains function even though these mutations
251 are highly represented in non-functional variants (**Fig. 2c; Supplementary Fig. 9**). Thus, the
252 deleterious effects of Arg48Thr, Asn50Ser, and Ala52Ile are mitigated by the presence of
253 Gln18Met.[41] Second: Arg22 is essential for LacO binding as it forms a critical base-specific
254 hydrogen bond (**Supplementary Fig. 7b**);[42]while it is found in every functional variant in the
255 screen (**Fig. 2f**), 1018 non-functional variants also contain Arg22. Arg22His is among the
256 substitutions found in Anc1143, which is the closest non-functional ancestor of EcLacI (**Fig. 2a**).
257 Third: Asn26 in the non-functional ancestor Anc875 is mutated to His26 in Anc876 and introduces
258 function *de novo*. However, when Asn26His is introduced alongside the Ile42Met in Anc877, there
259 is no gain of function (**Fig. 2b**). In summary, we find subtle mutational perturbations have large
260 effects on function, indicating a metastable state, and that epistasis contributes to the ruggedness
261 of the landscape.
262

263 **Historical contingency and genetic drift shape evolution.** While evolutionary trajectories of
264 proteins are constrained by biological function, stochastically sampled permissive substitutions
265 also shape the sequence landscape and contribute to historical contingency i.e., the dependence
266 of evolutionary trajectories on historical mutations.[43–45] From ASR alone, it is difficult to parse
267 whether mutations are functionally essential, permissive, or neutral because ancestral and extant
268 sequences are often separated by many substitutions. Therefore, to investigate the role of each
269 DBD position, we used deep mutational scanning (DMS), to systematically assay all single-amino
270 acid substitutions in EcLacI. By combining ASR with a DMS screen of EcLacI, we sought to
271 disentangle conserved residues that have been fixed by adaptive evolution from those that arose
272 through neutral drift.
273
274 To functionally characterize all single amino acid substitutions of the *Ec*LacI DBD (1121 variants
275 in total), we used the same technologies (oligonucleotide chip synthesis, one-pot library cloning,
276 FACS, and deep sequencing) as described previously for our high-throughput phylogenetic
277 screen (**Supplementary Fig. 10**). We sorted the low fluorescence population to enrich the
278 repression-competent variants. The number of repression-competent DMS variants (25% of the
279 population) was 10-fold higher than from the phylogenetic library (2.5%) (**Supplementary Fig.
280 11a**). These were enriched to ~95% of the population after two rounds of sorting (**Supplementary
281 Fig. 11b,c**). Sequencing of pre- and post-sorted libraries yielded enrichment ratios, normalized to
282 native EcLacI, that were highly correlated (R^2=0.93) between independent replicates
283 (**Supplementary Figs. 12**)[31].

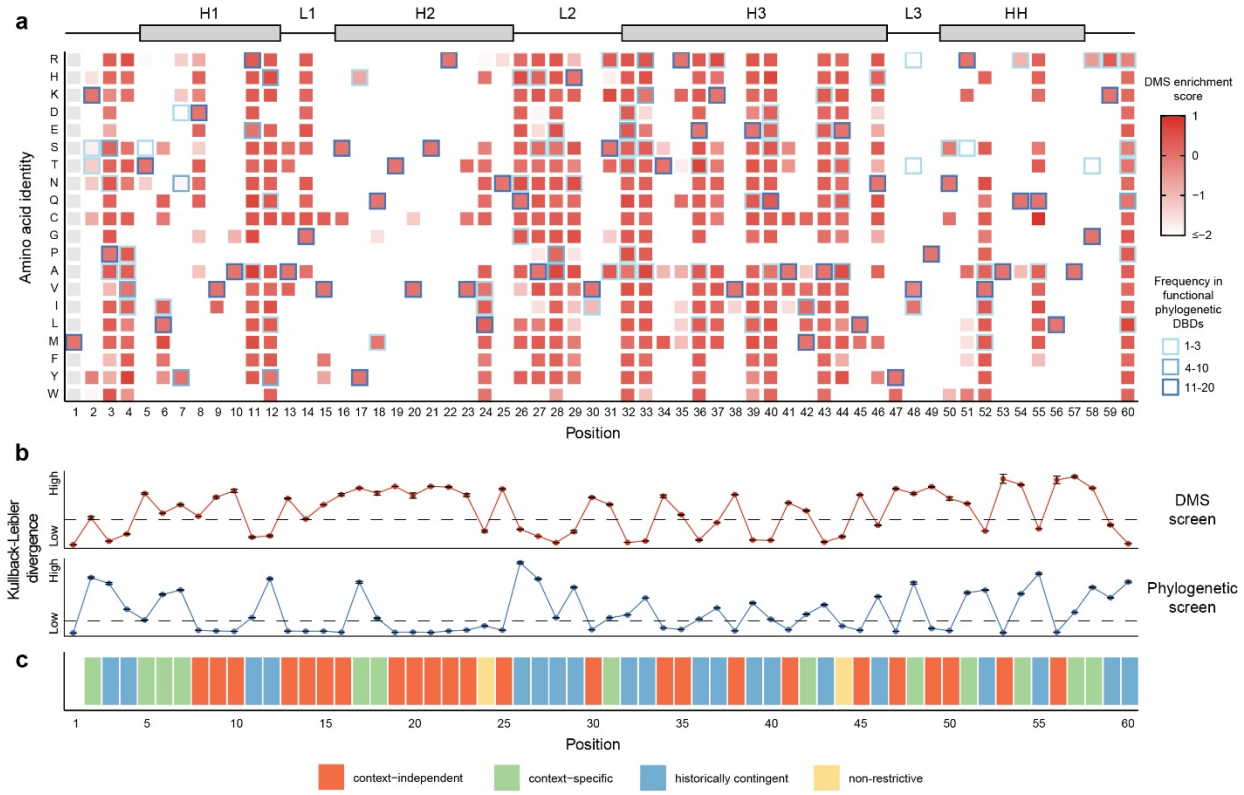


Fig. 3. DMS to assign evolutionary roles to each position of the DBD. (**a**), Heat map showing the normalized enrichment scores (red gradient) of each EcLacI substitution after selection of repression competent DBDs. Secondary structure topology of the DBD (top) and substitution frequency (blue gradient) among functional extant and ancestral DBDs are shown. (**b**), Kullback-Liebler divergence (KLD) describes how much the amino acid probability distribution changes at each position in response to functional selection in both the DMS and phylogenetic screens. For the DMS screen, positions with high KLD are functionally restrictive, whereas low KLD positions are more tolerant to substitutions. (**c**), Assigned evolutionary roles to each DBD position based on KLD analysis of the DMS and phylogenetic screens. Context-independent positions (low phylogenetic KLD and high DMS KLD) contribute to stability and general HTH fold. Context-specific positions (high phylogenetic and DMS KLD) modulate LacO recognition. Historically contingent positions (high phylogenetic KLD and low DMS KLD) are not functionally constrained in the genetic background of *E. coli* LacI, despite sequence convergence among functional phylogenetic variants. Non-restrictive positions (low phylogenetic and DMS KLD) are generally tolerant to substitutions.

For this DMS experiment, each DBD position has a probability distribution of the 20 canonical amino acids pre- and post-selection. We used Kullback-Lieber divergence (KLD[46]; a statistical method for comparing two probability distributions) to discern which residues are functionally restrictive (high score) and non-restrictive (low score). This showed that the restrictive and non-

305 restrictive positions are non-uniformly distributed across the DBD, with H2 and HH largely
306 intolerant to substitutions (**Fig. 3a, b**), consistent with the sequencing results from the phylogeny
307 (**Fig. 2**), while the C-terminus, loop 1, loop 2 and most of H3 are robust to mutation.
308
309 While DMS provided key insights into the functional role of each residue, it lacks an evolutionary
310 perspective. For instance, using DMS alone we cannot distinguish between residues essential for
311 LacO recognition ("context-specific") and those required for stability and general fold ("context-
312 independent"). By combining DMS with our extensive phylogenetic screen, it is possible to define
313 an evolutionary role for each residue (**Fig. 3b**). There are four possible combinations of DMS
314 (high/low) and phylogenetic (high/low) KLD scores, each representing a unique evolutionary role
315 of a residue (**Fig. 3c**). Context-independent functionally restrictive residues, such as those that
316 are structurally essential, have high DMS KLD scores (functionally restrictive) and low
317 phylogenetic KLD scores - because positions that are conserved across all phylogenetic variants
318 cannot undergo a shift in probability distribution upon selection and thus have low KLD scores
319 despite potentially being functionally relevant (**Fig. 4a,b**). These context-independent residues
320 are heavily concentrated within the core of the HTH fold and across the dimerization interface of
321 the HH, and include essential LacO-binding residues, such as Arg22 (**Fig. 4a,b**). Context-
322 dependent functionally restrictive positions have high KLD scores for both DMS and phylogenetic
323 screens. These residues are critical for LacO recognition, either making direct contact with the
324 operator DNA in both the major and minor grooves, helping to orient the DBD for LacO binding,
325 or interacting with the LBD for allosteric communication. Context-dependent residues include
326 Tyr7, Tyr17 and Gln18 and Met42 – all of which contact LacO or are associated with an epistatic
327 change-of-function, such as in Anc877 (**Fig. 4c,d**). Historically contingent positions have low DMS
328 and high phylogenetic KLD scores i.e., while mutationally robust in terms of EcLacI function, they
329 are restricted in phylogenetic analysis owning to their deterministic role in evolution (**Fig. 4e,f**).
330 Many of these residues are located adjacent to context-dependent functionally restrictive

positions, consistent with a role in modulating their effect, such as Asn26 (historically contingent) and Met42 (context-specific) (**Fig 4f**). Functionally non-restrictive positions have low KLD scores for both DMS and phylogenetic screens. Only two solvent-exposed residues, Val24 and Glu44, fall within this category. Altogether, this analysis demonstrates that the evolutionary role of all residues within the DBD of EcLacI can be revealed through a statistical analysis of DMS and ASR data.

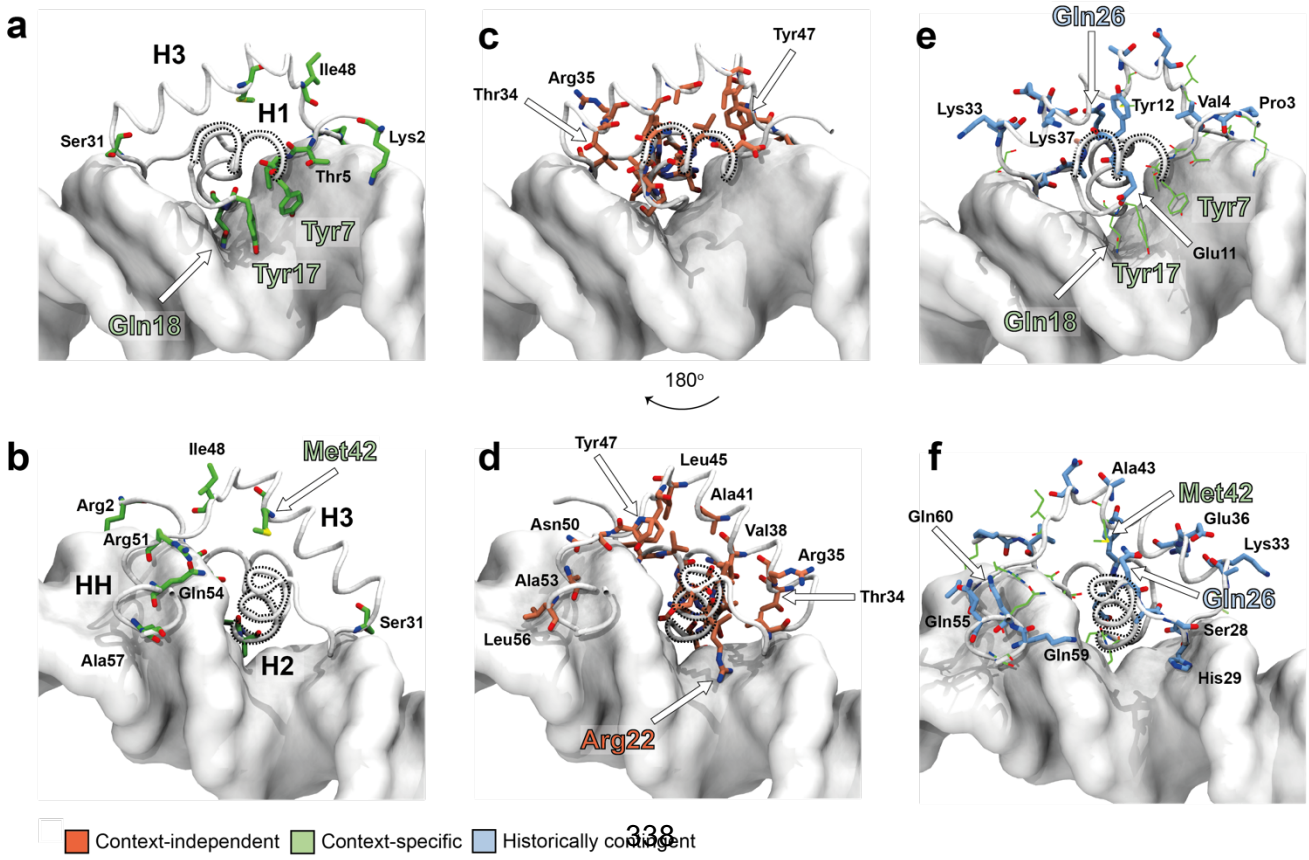


Fig. 4. Structural role of DBD mutations. (a) Context-specific mutations in the LacO major groove, where Tyr7, Tyr17 and Gln18 make essential contacts with the DNA. (b) context-specific mutations in the minor groove. Context-specific mutations are dominated by locations that contact the DNA, particularly in the major groove. (c) Context-independent mutations in the major groove. Context-independent mutations cluster around the core packing of the DBD. (d) Context-independent mutations in the minor groove. Context-independent mutations also include those essential for DNA-recognition, such as Arg22. (e) Historically contingent mutations in the major groove. Context specific mutations are represented as lines (green), historically contingent are represented as sticks (cyan). (f) Historically contingent mutations in the minor groove. Historically contingent mutations cluster around the context specific mutations that they set the background for. For example, Met42 and Gln26 (epistatically linked) are oriented towards one another on H3 and H1, respectively.

352

353 **Operator asymmetry contributes to the rugged fitness landscape.** Despite asymmetric DNA
354 operator sequences having lower affinity for DBDs than symmetrical sequences,[47] they are
355 overwhelmingly dominant throughout evolution and must therefore confer some selective benefit.
356 To investigate this, we performed *in vitro* DNA-binding assays using surface plasmon resonance
357 (SPR; **Fig. 5**). We find that all functional variants tested (Anc880, Anc881 and Anc882) bind the
358 symmetrical $LacO_{sym}$ with K_d constants greater than an order of magnitude lower than their
359 asymmetric operator counterparts (LacO and $LacO_1$), confirming that operator asymmetry
360 significantly reduces a repressor's DNA-binding capacity in ancestral, as well as extant DBDs
361 (**Fig. 5a-c**; **Supplementary Fig. 13**).

362

363 To investigate the molecular basis for binding of asymmetrical operator sequences, we performed
364 all-atom molecular dynamics (MD) simulations on models of Anc880 and 882, as well as EcLacI,
365 complexed with LacO (**Fig. 5**). Each system was simulated with 10 independent 60 ns replicates
366 (600 ns total sampling time) (**Supplementary Fig. 14**). Previous studies have identified motions
367 related to DNA-binding specificity in *Ec*LacI occur over ns – μs timescales, consistent with the
368 timescale of our simulations.[48,49] We observe asymmetry in the protein-DNA interactions between
369 the two monomers that constitute the functional dimer (**Fig. 5d,e**). Although both monomers
370 interact with the DNA via the same residues, the relative contributions these residues make to the
371 overall binding energy and their DNA contacts are distinct. In Anc882, which is the variant with
372 the greatest LacO affinity, we identified Tyr7, Tyr17, Gln18 and Arg22 as being essential for DNA
373 recognition and binding. However, the conformations of each of these residues differ between
374 monomers to accommodate the respective operator half-sites and principal component analysis
375 (PCA) reveals that backbone motions differ between either monomer (**Fig. 5d-f**). Identical
376 analyses of apo-Anc882 trajectories found that in the absence of DNA, backbone motions are
377 nearly indistinguishable between monomers (**Supplementary Fig.14**). This analysis is consistent

with NMR structures of the EcLacI DBD complexed with the natural $LacO_1$ operator[50], and alternate DNA binding modes have been alluded to by analyses of LGF repressors and EcLacI mutants.[51] Thus, asymmetrical operator sequences impose a complex selective pressure: the DBDs must be able to undergo a conformational change to bind two distinct half-sequences with physiologically relevant affinity. Being subject to multiple selection pressures associated with either half-site to maintain complete operator binding is consistent with the extreme ruggedness we observe (**Fig. 2**).

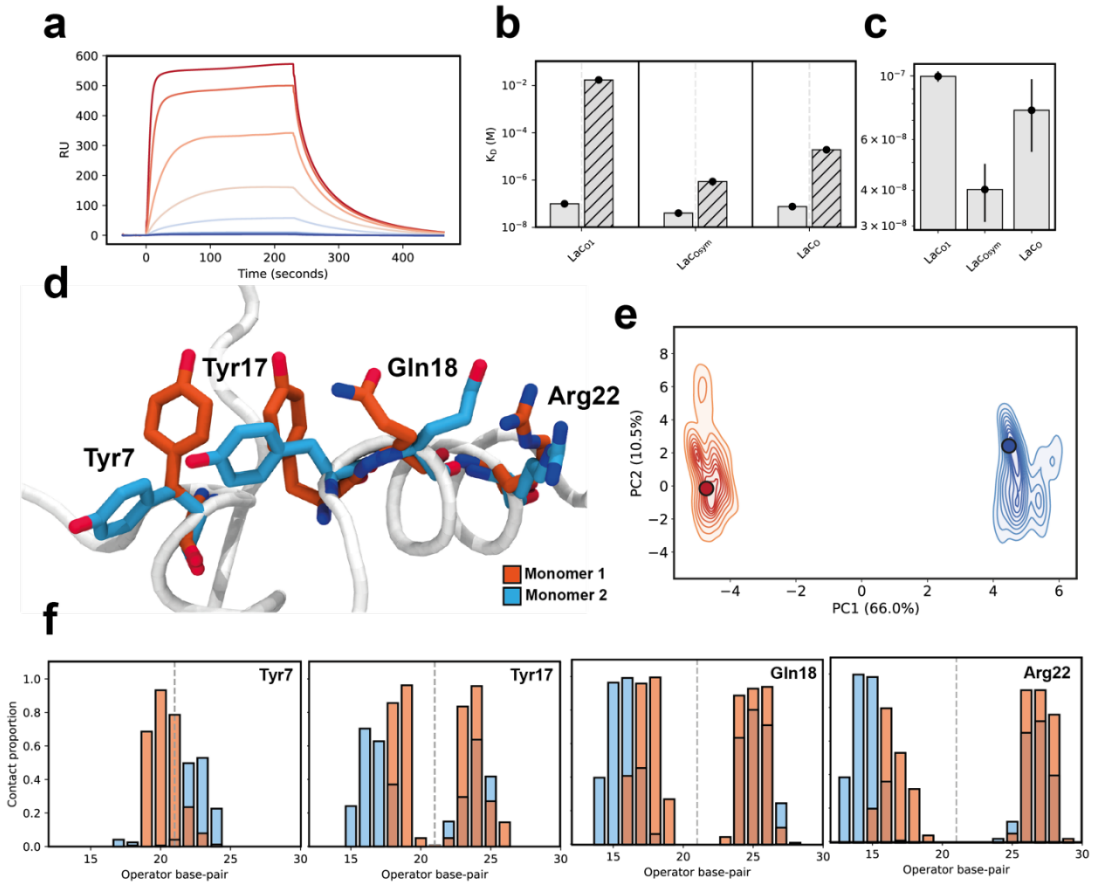


Fig. 5 Transcriptional regulator asymmetry and fitness landscape ruggedness. (a) Multi-cycle kinetics SPR sensogram for Anc882 LacO binding. (b) Anc882 K_d values, determined by SPR for $LacO_1$, $LacO_{sym}$ and LacO DNA binding when repressed (solid) and induced with 10 mM IPTG (hashed). (c) Magnified view of repressed K_d values. (d) Conformational snapshots of the most sampled conformations for monomers 1 (red) and 2 (cyan). Dominant conformations of Tyr7, Tyr17, Gln18 and Arg22 are shown from principal component analysis (PCA; panel e). (e) PCA of Anc882, showing monomers A (red) and B (blue). Conformational snapshots presented in (d) are shown as circles. (f) Distribution of key protein-DNA contacts in Anc882 and LacO. DNA residues (X-axis) have been aligned for visualization.

395 The demand on the DBD to simultaneously evolve high affinity for two distinct half-site DNA
396 sequences using a single binding site (as well as allosteric regulation with the LBD) means that
397 the fitness landscape we observe across this phylogeny can be viewed as a composite of two or
398 more fitness landscapes. To test this, we performed *in silico* evolutionary simulations on small
399 model systems using theoretical elementary landscapes as a model (**Fig. 6**)[52–54]. Elementary
400 landscapes derive from a robust and well-established graph-theoretic approach to combinatorial
401 landscape optimization problems[54]; specifically, they are the orthogonal eigenvectors of the
402 graph-Laplacian of a sequence space graph. For our purposes, they provide a theoretically
403 rigorous way of testing hypotheses that pertain to the ruggedness of a fitness landscape. Using
404 a linear combination of elementary landscapes, each representing a different hypothetical fitness
405 function, we find that the combination of two fitness landscapes can indeed be more rugged than
406 either constitutive fitness landscape is alone, akin to constructive/deconstructive wave
407 interference (**Fig. 6**). This demonstrates that ruggedness can emerge from the composition of
408 otherwise smooth fitness landscapes and predicts that proteins with a single function, such as
409 enzymes or solute binding proteins, may evolve over smoother fitness landscapes than proteins
410 such as LGF repressors, where multiple and complex fitness constraints are imposed
411 simultaneously.
412

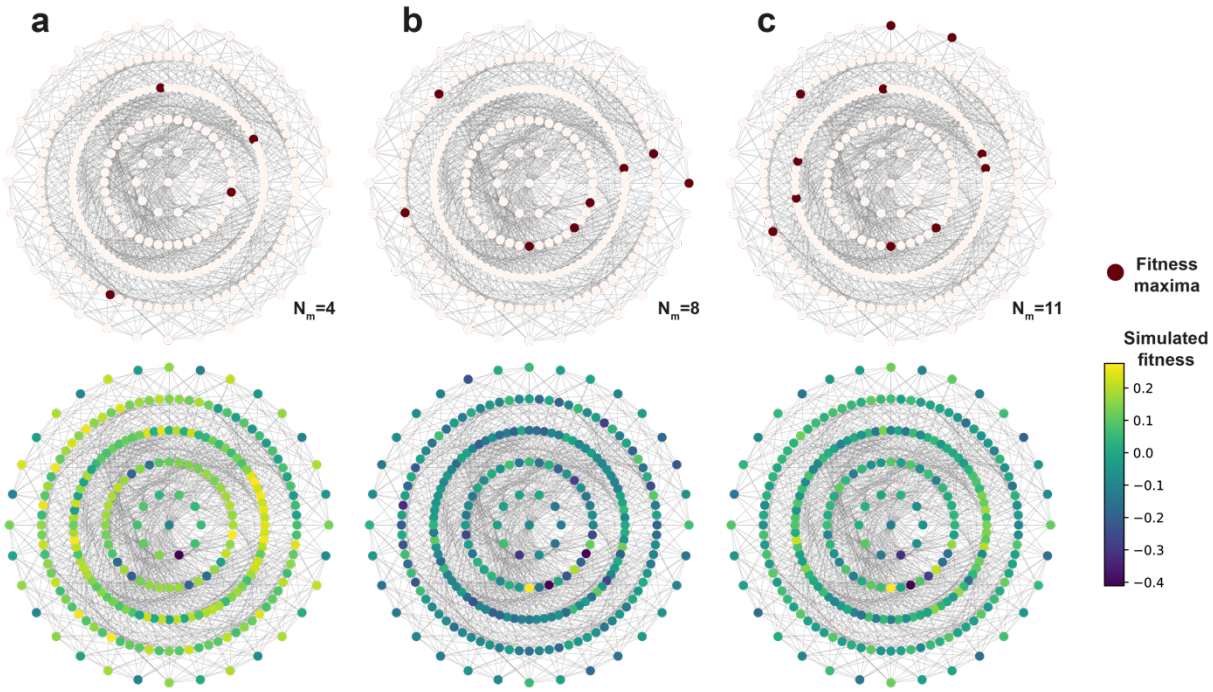


Fig 6. Elementary landscape model simulations. 2nd and 3rd elementary landscapes in panels (a), (b) respectively. Each vertex represents a unique genotype, colored fitness maxima (top) and quantitative simulated fitness (bottom). Edges connect genotypes accessible by a single mutation. (a) and (b) have 4 and 8 fitness maxima (N_m), respectively (where N_m is the number of fitness peaks where all neighboring genotypes have a lower fitness). (c) A linear combination of 2nd and 3rd elementary landscapes. The composite landscape is characterized by 11 fitness maxima, more than either constituent landscape, and is more rugged.

DISCUSSION

These results show that the fitness landscape for operator binding by DBDs of the LGF is extremely rugged i.e., the viable fitness space for LacO recognition and repression is narrow and highly localized. Interestingly, this ruggedness has long been alluded to by mutational studies, but never comprehensively observed or studied prior to our work.[5] In contrast, most of the protein families studied by ASR to date demonstrate protein evolution over smooth fitness landscapes where gradual changes in function can be observed.[24,35,45,55–58] Even those characterized by extensive epistasis have well-defined mutational trajectories from semi- or non-functional intermediate sequences to functional ones[44,59]. The profound difference in evolutionary dynamics between these studies and the LGF may reflect fundamentally different biological and physical

432 constraints. While any complex molecular trait can be described as a combination of arbitrarily
433 smooth fitness functions (i.e. folding, substrate-recognition, electrostatic preorganization,
434 immunogenicity, among many catalytic and structural parameters[60]) DNA recognition among LGF
435 regulators is particularly rugged due to the compounding of two DNA-binding landscapes that
436 likely have complexity in isolation; if either half-site requires the correct folding, dynamical
437 ensemble, electrostatic and shape complementarity (among likely many other traits), a protein
438 sequence that binds both half-sites must compound the molecular requirements for both of the
439 half-sites.
440
441 One notable aspect of the rugged DBD landscape is the evolutionary metastability of repression:
442 descendants of a repression competent ancestor are seldom functional with the same operator
443 sequence and mutations have extreme epistatic relationships. This indicates that both selection
444 pressure for LacO specificity has been variable and that LacO specificity has emerged
445 independently numerous times throughout evolutionary history, each time in different sequence
446 contexts. Like the notion of an epistatic ratchet, where a phenotype becomes completely
447 inaccessible within a specific background[43,61], the metastability we observe occurs through non-
448 specific mutations that counteract or reciprocate the fitness effects of previously fixed mutations
449 in backgrounds that are on the edge of a binary phenotype, making a few genetic changes
450 sufficient to rewire specificity. Using a combination of ASR and DMS, we have comprehensively
451 mapped historical contingency within the DBD showing that permissive sites, which are essential
452 in an evolutionary context and underlie the epistasis and metastability, can be identified.
453
454 The rugged landscape, the asymmetry of the DBD:DNA complex and the evolutionary
455 metastability of the repressor function are all interesting observations, but how do they relate to
456 the physiological role of these proteins? Since effective genetic regulation is paramount to
457 organismal fitness, precise and binary interactions are essential. To impart a selective advantage,

458 regulators must be highly specific for their cognate operator sequences. Unlike promiscuous
459 enzymatic activity, promiscuous DNA-binding causes metabolic cross-regulation and severe
460 disruption of cellular function.[62,63] Thus, we hypothesize that this ruggedness fitness landscape is
461 an intrinsic aspect of fitness for the LGF: asymmetric operator half-sites have evolved as a
462 mechanism to minimize metabolic crosstalk between regulators and create an evolutionary
463 dynamic in which activity is essentially binary. Over a rugged fitness landscape, as we observe,
464 DNA-binding specificity is metastable, with rapid gain/loss of function, essentially eliminating the
465 risk of non-specific DNA-binding and metabolic cross-regulation between diverging regulators.
466 This suggests that inherently epistatic protein folds, such as that of the DBD, have likely been
467 evolutionarily selected for regulatory purposes. Indeed, previous ASR studies have shown that
468 epistasis is a hallmark of several classes of eukaryotic transcription factors,[59,64–67] and recent
469 results have highlighted the importance of molecular unpredictability in transcription factor
470 divergence.[68] Interestingly, similar observations of rapid *de novo* promoter emergence have also
471 recently been made, indicating that the evolutionary dynamics we observe in regulatory proteins
472 may also be present in their DNA binding partners.[69]
473
474 Altogether, our analysis of the sequence-function landscape of the LGF DBDs has led to several
475 valuable insights. Most importantly, we prove a molecular level explanation for the high fidelity
476 that has long been observed in gene regulation by proteins of the LGF family, showing that it is
477 an intrinsic property of the rugged fitness landscape that is itself a function of the biophysical
478 properties of DBD structure and asymmetry of DNA operator sequences. Future work could
479 investigate whether rugged fitness landscapes, and functional partitioning through asymmetry,
480 are common when the physiological function of the protein family makes promiscuous activity
481 deleterious, as for these transcriptional regulators.
482
483

484 **METHODS**

485 **Phylogenetic inference**

486 To collate an expansive dataset of the full LGF, EcLacI, RbsR, MalR, SacR/ScrR, GalR, GalS,
487 AscG, PurR, CytR, and CcpA were each individually queried against the NCBI non-redundant
488 protein database using pBLAST. ~250 sequences were retrieved from each individual BLAST
489 search, using an e-value significance threshold of 1.0E-10. Redundancy in this dataset was
490 removed to 90% sequence identity using CD-HIT[70] and incomplete sequences, or sequences with
491 poorly conserved (<1% of sequences) indels were manually removed. Sequences were aligned
492 using the ESPRESSO protocol of T-COFFEE. [71] This alignment was manually refined to remove
493 poorly conserved sequences or insertions and benchmarked against available LGF X-ray crystal
494 structures. Phylogenetic inference was performed in IQ-TREE.[72] The ML model, LG+R9,[73,74] was
495 fitted using ModelFinder (as implemented in IQ-TREE).[74] Tree-search was performed using
496 default parameters and branch supports were computed as ultra-fast bootstrap approximations.[75]
497 Tree search was repeated for 10 independent replicates, which were tested for statistical
498 equivalency by the approximately unbiased (AU) test conducted to 10000 replicates.[33] An
499 additional inference was performed using the same protocol, however, only including the N-
500 terminal 60 columns of the LGF dataset to determine how much evolutionary signal the DBD
501 provides to the full-length sequences (Supplementary Fig 2). Ancestral sequences were
502 reconstructed on the single topology presented in Fig. 1 using CodeML from the PAML software
503 package.[76] The sequence evolution model for ASR was manually set to LG+G4[73] and indel events
504 were processed in the ancestral sequences according to the principle of parsimony.

505

506 **Construction of GFP reporter strain**

507 The *E. coli* strain DH10β was modified by lambda Red recombineering to insert a superfolder
508 GFP gene driven by the pLlacO[77] promoter and a kanamycin resistance cassette at the LacI locus
509 to generate the GFP "reporter" cell line. The temperature sensitive plasmid pKD46 (Genbank

510 AY048746) was used for arabinose-inducible expression of Red recombinase.[78] Linear double-
511 stranded donor DNA was amplified off a pSC101 plasmid using overhang primers to add 50bp
512 homology arms directed to the 5' end of the LacI gene. A frozen glycerol stock of DH10β
513 transformed with the pKD46 plasmid was struck out on LB carbenicillin (100µg/mL) and incubated
514 at 30°C overnight. A colony was selected and inoculated into 5mL LB carbenicillin and grown for
515 16h in a shaking incubator at 30°C. Cells were diluted 50X into 25mL of LB carbenicillin and grown
516 to an OD of 0.1. Red recombinase was induced with 100mM L-arabinose and cells were grown
517 to an OD of 0.6. The cells were harvested and prepared for electroporation. 25µL of cells were
518 transformed with 500ng of donor DNA. Cells were recovered for 2h in SOC media in a shaking
519 incubator at 37°C and plated on LB kanamycin (30 µg/mL). The transformants were incubated at
520 37°C overnight to cure the cells of the temperature sensitive pKD46 plasmid. A visibly fluorescent
521 colony was selected and inoculated into 5mL LB kanamycin to be grown for 16h in a shaking
522 incubator at 37°C. The genome modification was confirmed via sequencing and a glycerol stock
523 was stored at -80°C.
524
525 **General library DNA assembly**
526 Plasmids were constructed using standard molecular biology techniques of PCR and Golden Gate
527 assembly with Kapa HiFi DNA Polymerase (KAPA Biosystems), restriction enzymes (NEB), T4
528 DNA Ligase (NEB), and Antarctic Phosphatase (NEB).
529
530 Oligonucleotides encoding residues 2-60 of all LGF DBDs (1158 variants in total) or all single
531 amino acid substitutions of the *E. coli* LacI DBD (1121 variants in total) were synthesized as
532 single-stranded oligonucleotide pools (Agilent Technologies). Extant DBDs exceeding 60aa long
533 were truncated from the N-terminus to 60 residues in length to comply with DNA synthesis
534 restrictions. Oligonucleotides were converted to double-stranded DNA using 15 cycles of PCR
535 amplification and purified on a spin column (EZNA Cycle Pure kit from Omega BioTek). A pSC101

536 backbone containing extant *E. coli* LacI gene under control of the strong pLtetO promoter and a
537 spectinomycin resistance gene was amplified using a primer pair encoding BsaI cut sites that
538 matched the DBD insertion location of both oligonucleotide libraries. The amplified backbone was
539 treated with Dpn1 for 2h at 37°C and purified using a spin column. The backbone was treated
540 with BsaI-HF v2 for 2h at 37°C, Antarctic phosphatase for 1h at 37°C, and subsequently purified
541 using a spin column. A Golden Gate assembly reaction (30 cycles of 37°C for 5min and 16°C for
542 5min) was performed using 0.042 pmol pSC101 backbone and 0.21 pmol amplified DBD library
543 in a 1:5 molar ratio. The assembled product was dialyzed on a semi-permeable membrane
544 (Millipore) for 1h at 25°C against dH_2O. A 25µL aliquot of electrocompetent DH10β *E. coli* cells
545 were transformed with 2µL of the dialyzed product using electroporation. Cells were recovered
546 for 1h in SOC media in a shaking incubator at 37°C and dilutions were plated on LB Spectinomycin
547 (50 µg/mL) to calculate transformation efficiency (>10^6 CFU/mL). Remaining recovered cells were
548 diluted 5x, incubated for 6h, and then diluted 50X for overnight growth in a shaking at 37°C for
549 16h. Library plasmids were extracted using a DNA miniprep kit and 1µL (~100ng) was used to
550 transform 25µL of electrocompetent reporter cells following the transformation protocol described
551 above. Glycerol stocks of the reporter cells transformed by the preselected phylogenetic and DMS
552 plasmid libraries were stored at -80°C.
553
554 **Fluorescence-activated cell sorting**
555 Thawed glycerol stocks of reporter cells containing the LGF DBD or DMS libraries were used to
556 inoculate (50µL each) 5mL of LB kanamycin (30µg/mL) / spectinomycin (50µg/mL) in duplicate.
557 Cells were grown in a shaking incubator at 37°C for 16h and subsequently diluted 50X in PBS
558 (137mM NaCl, 2.7mM KCl, 10mM Na2HPO4, 1.8mM KH2PO4) for sorting. Sorting was performed
559 with a Sony MA900 cell sorter. Cells were excited with a 488nm laser and GFP fluorescence was
560 monitored through a 525/50 filter. A sorting gate was drawn to isolate the low-fluorescence
561 populations of the LGF DBD (2.5-2.6% of total population, 250,000 cells collected) and DMS

562 (25.4-25.6% of total population, 500,000 cells collected) libraries. Cells were sorted into 1mL of
563 LB, and the total volume was adjusted to 5mL after sorting. Cells recovered for 1h at 37°C in a
564 shaking incubator before antibiotics were added (kanamycin and spectinomycin), and incubation
565 was continued for another 15h. Glycerol stocks were made and stored at -80°C. Plasmids were
566 harvested using a DNA miniprep kit and 1μL (~100ng) was used to transform 25μL of fresh
567 electrocompetent reporter cells. The procedure described above was repeated for a second round
568 of low-fluorescence sorting using the same sorting gates to further enrich for repression
569 competent variants.
570
571 **Next-generation sequencing**
572 We used deep sequencing to evaluate presorted and sorted LGF and DMS populations. The DBD
573 was amplified in two stages with Kapa HiFi DNA Polymerase (KAPA Biosystems) for tailed
574 amplicon sequencing. The first PCR reaction was performed with overhang primers that anneal
575 to 5' and 3' constant regions surrounding the DBD. The overhangs add a variable N region (to
576 enhance nucleotide diversity), and a portion of the universal Illumina adapter. The first reaction
577 was performed using 14 cycles with 1ng of extracted plasmid DNA (DMS or LGF library, presorted
578 or sorted) as template. The product was purified on a spin column (EZNA Cycle Pure kit from
579 Omega BioTek). A second PCR reaction was used to add the index (for pooled sequencing) and
580 the Illumina 'stem'. This amplification was performed using 10 cycles with 10ng of the purified
581 product from the first reaction as template. The amplicons were purified and used for deep
582 sequencing. Samples were sequenced on an Illumina MiSeq System, using a MiSeq Reagent Kit
583 v2 (2x250 cycles) following the manufacturer's documentation.
584
585 Paired-end Illumina sequencing reads were merged with PEAR (Paired-end read merger).[79]
586 Phred scores (Q scores) were used for quality filtering. Reads with an expected number of errors
587 exceeding 1 were removed and total read counts for each DBD variant were computed

588 (**Supplementary Figs. 3,10**). Raw read counts were highly correlated ($R^2 \geq 0.97$) for all replicate
589 samples. Enrichment for each variant was computed using

590 $$Enrichment = \frac{f_{sorted}}{f_{presorted}}, \qquad (eq. 1)$$

591 where $f_{presorted}$ and f_{sorted} are relative read count frequencies before and after enrichment of
592 repression competent variants, respectively. Higher enrichment indicates tighter repression of
593 GFP and higher affinity to LacO. For the DMS library, enrichment scores (Escore) were computed
594 by normalizing enrichment to WT and applying a log2 transformation. Thus, the WT DMS Escore
595 is set to 0, Escore > 0 indicates improved function, and Escore < 0 indicates reduced function.
596
597 Kullback-Liebler divergence of a specific position is given by the following:

598 $$KLD = \sum_{a} f_{sorted,a} \log_2 \frac{f_{sorted,a}}{f_{presorted,a}}, \qquad (eq. 2)$$

599 where the sum is over all 20 canonical amino acid identities, and $f_{sorted,a}$ and $f_{presorted,a}$ are the
600 relative frequencies of an amino acid a in the repressed sorted or presorted distributions,
601 respectively. If the relative frequency of an amino acid at a position is zero, then the value was
602 excluded from the summation.
603
604 **Microplate fluorescence assay for clonal characterization**
605 The presorted and repressed sorted LGF DBD libraries were struck out on LB kanamycin
606 (30μg/mL) / spectinomycin (50μg/mL) plates. Colonies were selected and inoculated into 150μL
607 LB kanamycin/spectinomycin in a 96-well plate. Cells were incubated at 37°C in a microplate
608 shaker for 8h. The cultures were diluted 50X in fresh media in a 96-well plate with varying
609 concentrations of IPTG (0, 0.1, 0.5, 1, 5, 25, 100 and 1000μM). Diluted cultures were incubated
610 at 37°C in a microplate shaker for 14h. GFP fluorescence and OD_{600} were measured in a BioTek
611 Synergy HTX Multi-Mode 96-well plate reader. Fluorescence was normalized to OD to account

612 for differences in cell density. Assayed colonies were sequenced to identify the DBD variant. The
613 mean and standard deviation of normalized fluorescence for replicates (n ≥ 2) for each
614 concentration of IPTG were used to fit these dose-response curves to the Hill–Langmuir equation:

615 $$fluoresence = b + \frac{(F_{max} - b)[x]^n}{EC_{50}^n + [x]^n}, \qquad (eq. 3)$$

616 where b is the basal fluorescence (measure of LacO affinity in the absence of inducer), F_{max} is
617 the maximum fluorescence signal achieved at 1mM IPTG, EC_{50} is the half maximal effective
618 concentration, x is the concentration of IPTG, and n is the Hill coefficient.
619

620 **Protein expression and purification**

621 Anc880, 881 and 882 were synthesized and cloned into the NdeI/XhoI multiple cloning site of
622 pET-28a(+) with a C-terminal 6x histidine tag. All three ancestral proteins were recombinantly
623 expressed in BL21(DE3) lab strain *E. coli* at 25 °C for 16 hours in Luria-Bertani (LB) media. Protein
624 expression was induced with 10 mM IPTG once media OD_{600} reached approximately 1.5 AU. Cells
625 were pelleted by centrifugation at 5,000 RPM for 15 minutes and were lysed by sonication once
626 resuspended in buffer A [20 mM PBS, 150 mM NaCl, 20 mM imidazole (pH 7.4)] containing the
627 recommended amount of *Serratia marscecens* Turbonuclease (Sigma-Aldrich). The lysate was
628 pelleted by centrifugation at 12,000 RPM, 4 °C for 1 hour before being filtered (0.22 μM) and
629 loaded onto a prepacked 5 mL HisTrap FF immobilized metal ion affinity chromatography column
630 (GE lifescience) equilibrated in buffer A. The loaded column was washed for 5 column volumes
631 in buffer A and 3 column volumes in 94% buffer A/ 6% buffer B [20 mM PBS, 150 mM NaCl, 250
632 mM imidazole (pH 7.4)], before being eluted with 100% buffer B on an AKTA Start FPLC
633 instrument (GE lifescience). IMAC purified protein was then filtered (0.22 μM) and purified by size
634 exclusion chromatography (SEC) on a prepacked HiLoad superdex 200 16/60 SEC column (GE
635 lifescience) equilibrated in SEC buffer [20 mM PBS, 150 mM NaCl (pH 7.4)] and run on an AKTA
636 FPLC system (GE lifescience).

637

638 **Surface Plasmon resonance**

639 Proteins purified to homogeneity by IMAC and SEC were buffer exchanged into SPR running
640 buffer [10 mM HEPES, 300 mM NaCl, 3 mM EDTA, 0.05% (v/v) tween-20 (pH 7.4)] and diluted
641 into a $\log_2$ dilution series from 250 nM - 3.91 nM for repressed binding samples and 10 μM – 78.1
642 nM for samples induced with 70 μM IPTG. $_{\text{Double}}$-stranded DNA oligonucleotides for LacO,(
643 TGTGTGGAATTGTTATCCGCTCACAATTTCACACA) $LacO_1$
644 (TGTGTGGAATTGTGAGCGGATAACAATTTCACACA), $LacO_{sym}$
645 (TGTGTGGAATTGTGAGCGCTCACAATTTCACACA), and random DNA
646 (AGGTCAAAAAGCCAGTGGTTATTTTAAGATGTCGC) were synthesized and 5'-biotinylated
647 (IDT). All SPR experiments were performed on a Biacore 8K instrument (GE lifescience). An SA
648 streptavidin Biacore chip (GE lifescience) was activated with 1 M NaCl and 20 mM NaOH before
649 four channels were charged with approximately 400 RU of each respective DNA ligand. Binding
650 affinity was determined through multi-cycle kinetic analysis at the 7 serial $\log_2$ protein dilutions,
651 and one buffer only blank. All protein analytes were run at 50 μL/min for a ligand contact time of
652 120 seconds. After each analyte cycle, the ligand was regenerated with 3 M NaCl run at 20 μL/min
653 for a contact time of 60 seconds. The K_d of DNA binding was determined from a multi-cycle model
654 fitted to the reference subtracted sensorgrams in the Biacore 8K evaluation software.

655

656 **Molecular dynamics simulations**

657 Models for Anc880 and Anc882 were generated from X-ray crystallography coordinates of EcLacI
658 bound to $LacO_{sym}$ (PDB: 1EFA)[80] using FoldX[81]. The $LacO_{sym}$ DNA ligand in 1EFA was mutated
659 to the LacO sequence that was experimentally tested using X3DNA. [82] To avoid DNA terminal
660 fraying, 5' and 3' termini were extended with sequences GGAAT and TCC respectively in X3DNA
661 for all models (Full DNA sequence: GGAATTGTTATCCGCTCACAATTCC). Simulations were

662 performed in the GROMACS MD engine using the Amber ff14SB[83] force field with PARMBSC1[84]
663 nucleic acid parameter set. MD integration used a timestep of 2 fs. Columb interactions were
664 modelled using particle mesh Ewald (PME) with a cut-off radius of 1.0 nm. Leonard-Jones
665 interactions used a cut-off scheme with a radius of 1.0 nm. Hydrogen bonds were constrained
666 using the LINCS algorithm. [85] For each system, 10 independent replicates of equilibration (from
667 the NVT ensemble onwards) and production MD were performed. Each protein-DNA complex
668 was put in a rhombic dodecahedral box extending at least 1.0 nm from the closest atom to the
669 box boundary in all directions. Simulation boxes were filled with TIP3P water molecules, and the
670 charge was neutralized with sodium and chloride ions. To replicate physiological conditions, each
671 box was additionally filled with 150 mM NaCl. All systems were energy minimized by steepest
672 descent with a step size of 0.01 nm until the greatest force was less than 1000 KJ/mol/nm. The
673 systems were equilibrated to 300 K in the NVT ensemble with harmonic restraints on all heavy
674 atoms for 100 ps using a velocity rescale thermostat and brought to 1 Bar pressure in the NPT
675 ensemble for 100 ps using a Berendsen barostat. Following temperature and pressure
676 equilibration, unrestrained production MD was performed for 100 ns in the NPT ensemble using
677 a velocity rescale thermostat and a Parrinello-Rahman barostat with references of 300 K and 1
678 Bar, respectively. Following periodic boundary corrections, all trajectories were analyzed in
679 MDTraj and every 10th frame was sampled. All production MD was performed on the Australian
680 National Computing Infrastructure GADI supercomputer.
681
682 **Elementary landscape simulations**
683 Elementary landscapes were calculated via eigendecomposition of the graph adjacency matrix
684 corresponding to a combinatorial sequence space over an alphabet of 3 amino acids and 5
685 number of positions. Elementary landscapes were indexed by order (e.g. first-order, second-order
686 and so on), and composition of landscapes performed combinatorically between 1st and 2nd
687 order, 1st and 3rd order, and 1st and 3rd order landscapes. The composition operation was

defined as a simple linear combination of landscapes (in analogy to superposition of waves in physics). All graph construction and matrix operations were performed in NumPy and Networkx. The number of local maxima were calculated using custom functions. Graph plotting was performed in Networkx using custom functions.

ACKNOWLEDGEMENTS

This work was supported in part by the ARC Centre of Excellence in Synthetic Biology (CE200100029), the ARC Centre of Excellence in Peptide and Protein Science (CE200100012), the Great Lakes Bioenergy Research Center, U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Number DE-SC0018409 (S.R and A.T.M) and the NIH Director's New Innovator Award DP2GM132682 (S.R). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, Department of Defense, Department of Energy, or other federal agencies.

AUTHOR CONTRIBUTIONS

A.T.M., M.A.S., C.J.J, and S.R. conceptualized the project. M.A.S. performed the phylogenetic analysis and ancestral protein reconstruction. A.T.M. created designs, performed sorting and NGS experiments, and Rosetta stability calculations. M.A.S. performed binding experiments and operator half-site analyses. M.S. simulated the NK model landscapes. A.T.M., M.A.S., C.J.J, and S.R. drafted the manuscript, A.T.M., and M.A.S. created the figures. C.J.J. and S.R. supervised the work. All authors edited and approved the final manuscript.

710 **REFERENCES**

711 1. Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225**, 563–564
712 (1970).
713 2. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of
714 evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).
715 3. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**,
716 397–401 (2016).
717 4. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat.*
718 *Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
719 5. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes
720 reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).
721 6. Otwinowski, J., McCandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis.
722 *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7550–E7558 (2018).
723 7. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat.*
724 *Methods* **11**, 801–807 (2014).
725 8. Zhou, J. *et al.* Higher-order epistasis and phenotypic prediction. *Proc. Natl. Acad. Sci. U. S.*
726 *A.* **119**, e2204233119 (2022).
727 9. Bloom, J. D. & Arnold, F. H. In the light of directed evolution: pathways of adaptive protein
728 evolution. *Proc. Natl. Acad. Sci. U. S. A.* **106 Suppl 1**, 9995–10000 (2009).
729 10. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow
730 only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
731 11. Campbell, E. *et al.* The role of protein dynamics in the evolution of new enzyme function.
732 *Nature Chemical Biology* **12**, 944–950 (2016).
733 12. Jackson, C. J. *et al.* Conformational sampling, catalysis, and evolution of the bacterial
734 phosphotriesterase. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 21631–21636 (2009).

735 13. Kaczmarski, J. A. *et al.* Altered conformational sampling along an evolutionary trajectory
736 changes the catalytic activity of an enzyme. *Nat. Commun.* **11**, 5945 (2020).
737 14. Nahum, J. R. *et al.* A tortoise-hare pattern seen in adapting structured and unstructured
738 populations suggests a rugged fitness landscape in bacteria. *Proc. Natl. Acad. Sci. U. S. A.*
739 **112**, 7530–7535 (2015).
740 15. Hayashi, Y. *et al.* Experimental rugged fitness landscape in protein sequence space. *PLoS*
741 *One* **1**, e96 (2006).
742 16. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness
743 landscapes is facilitated by indirect paths. *Elife* **5**, (2016).
744 17. Rodrigues, J. V. *et al.* Biophysical principles predict fitness landscapes of drug resistance.
745 *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1470-8 (2016).
746 18. Flynn, J. M. *et al.* Comprehensive fitness landscape of SARS-CoV-2 Mpro reveals insights
747 into viral resistance mechanisms. *Elife* **11**, (2022).
748 19. Louie, R. H. Y., Kaczorowski, K. J., Barton, J. P., Chakraborty, A. K. & McKay, M. R. Fitness
749 landscape of the human immunodeficiency virus envelope protein that is targeted by
750 antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E564–E573 (2018).
751 20. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with
752 Gaussian processes. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E193–E201 (2013).
753 21. Gonzalez Somermeyer, L. *et al.* Heterogeneity of the GFP fitness landscape and data-driven
754 protein design. *Elife* **11**, (2022).
755 22. Spence, M. A., Kaczmarski, J. A., Saunders, J. W. & Jackson, C. J. Ancestral sequence
756 reconstruction for protein engineers. *Curr. Opin. Struct. Biol.* **69**, 131–141 (2021).
757 23. Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene
758 reconstruction. *Curr. Opin. Struct. Biol.* **20**, 360–366 (2010).
759 24. Clifton, B. E. *et al.* Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding
760 protein. *Nat. Chem. Biol.* **14**, 542–547 (2018).

761 25. Nguyen, C. C. & Saier, M. H., Jr. Phylogenetic, structural and functional analyses of the LacI-
762 GalR family of bacterial transcription factors. *FEBS Lett.* **377**, 98–102 (1995).
763 26. Monod, J., Wyman, J. & Changeux, J.-P. On the nature of allosteric transitions: A plausible
764 model. *J. Mol. Biol.* **12**, 88–118 (1965).
765 27. Marklund, E. *et al.* Sequence specificity in DNA binding is mainly governed by association.
766 *Science* **375**, 442–445 (2022).
767 28. Otwinowski, J. & Nemenman, I. Genotype to phenotype mapping and the fitness landscape
768 of the E. coli lac promoter. *PLoS One* **8**, e61570 (2013).
769 29. Zuo, Z., Chang, Y. & Stormo, G. D. A quantitative understanding of lac repressor's binding
770 specificity and flexibility. *Quant. Biol.* **3**, 69–80 (2015).
771 30. Barnes, S. L., Belliveau, N. M., Ireland, W. T., Kinney, J. B. & Phillips, R. Mapping DNA
772 sequence to transcription factor binding energy in vivo. *PLoS Comput. Biol.* **15**, e1006226
773 (2019).
774 31. Garruss Alexander, S., Collins Katherine, M. & Church George, M. Deep representation
775 learning improves prediction of LacI-mediated transcriptional repression. *Proceedings of the*
776 *National Academy of Sciences* **118**, e2022838118 (2021).
777 32. Fukami-Kobayashi, K., Tateno, Y. & Nishikawa, K. Parallel evolution of ligand specificity
778 between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Mol. Biol. Evol.*
779 **20**, 267–277 (2003).
780 33. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**,
781 492–508 (2002).
782 34. Meinhardt, S. *et al.* Novel insights from hybrid LacI/GalR proteins: family-wide functional
783 attributes and biologically significant variation in transcription repression. *Nucleic Acids Res.*
784 **40**, 11139–11154 (2012).

785 35. Clifton, B. E. & Jackson, C. J. Ancestral Protein Reconstruction Yields Insights into Adaptive
786 Evolution of Binding Specificity in Solute-Binding Proteins. *Cell Chemical Biology* **23**, 236–
787 245 (2016).
788 36. Swint-Kruse, L. & Matthews, K. S. Allostery in the LacI/GalR family: variations on a theme.
789 *Curr. Opin. Microbiol.* **12**, 129–137 (2009).
790 37. Spronk, C. A. *et al.* The solution structure of Lac repressor headpiece 62 complexed to a
791 symmetrical lac operator. *Structure* **7**, 1483–1492 (1999).
792 38. Lehming, N. *et al.* The interaction of the recognition helix of lac repressor with lac operator.
793 *EMBO J.* **6**, 3145–3153 (1987).
794 39. Daber, R. & Lewis, M. Towards evolving a better repressor. *Protein Engineering, Design and*
795 *Selection* **22**, 673–683 (2009).
796 40. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein
797 engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–
798 1322 (2019).
799 41. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218
800 (2016).
801 42. Milk, L., Daber, R. & Lewis, M. Functional rules for lac repressor–operator associations and
802 implications for protein–DNA interactions. *Protein Science* **19**, 1162–1172 (2010).
803 43. Bridgham, J. T., Ortlund, E. A. & Thornton, J. W. An epistatic ratchet constrains the direction
804 of glucocorticoid receptor evolution. *Nature* **461**, 515–519 (2009).
805 44. Pillai, A. S. *et al.* Origin of complexity in haemoglobin evolution. *Nature* **581**, 480–485 (2020).
806 45. Yang, G. *et al.* Higher-order epistasis shapes the fitness landscape of a xenobiotic-degrading
807 enzyme. *Nat. Chem. Biol.* **15**, 1120–1128 (2019).
808 46. Kullback, S. & Leibler, R. A. On Information and Sufficiency. *ann. math. stat.* **22**, 79–86
809 (1951).

810 47. Sadler, J. R., Sasmor, H. & Betz, J. L. A perfectly symmetric lac operator binds the lac
811 repressor very tightly. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 6785–6789 (1983).
812 48. Liao, Q. *et al.* Long time-scale atomistic simulations of the structure and dynamics of
813 transcription factor-DNA recognition. *J. Phys. Chem. B* **123**, 3576–3590 (2019).
814 49. Glasgow, A., Hobbs, H. T., Perry, Z. R., Marqusee, S. & Kortemme, T. Ligand-induced
815 changes in dynamics mediate long-range allostery in the lac repressor. *bioRxiv* (2021)
816 doi:10.1101/2021.11.30.470682.
817 50. Kalodimos, C. G. *et al.* Plasticity in protein-DNA recognition: lac repressor interacts with its
818 natural operator 01 through alternative conformations of its DNA-binding domain. *EMBO J.*
819 **21**, 2866–2876 (2002).
820 51. Zuo, Z. & Stormo, G. D. High-resolution specificity from DNA sequencing highlights
821 alternative modes of Lac repressor binding. *Genetics* **198**, 1329–1343 (2014).
822 52. Stadler, P. F. Landscapes and their correlation functions. *J. Math. Chem.* **20**, 1–45 (1996).
823 53. Stadler, P. F. Fitness landscapes. in *Biological Evolution and Statistical Physics* 183–204
824 (Springer Berlin Heidelberg, 2007).
825 54. Chicano, F., Whitley, L. D. & Alba, E. A methodology to find the elementary landscape
826 decomposition of combinatorial optimization problems. *Evol. Comput.* **19**, 597–637 (2011).
827 55. Hadzipasic, A. *et al.* Ancient origins of allosteric activation in a Ser-Thr kinase. *Science* **367**,
828 912–917 (2020).
829 56. Kaltenbach, M. *et al.* Evolution of chalcone isomerase from a noncatalytic ancestor. *Nat.*
830 *Chem. Biol.* **14**, 548–555 (2018).
831 57. Bar-Rogovsky, H., Hugenmatter, A. & Tawfik, D. S. The evolutionary origins of detoxifying
832 enzymes. *J. Biol. Chem.* **288**, 23914–23927 (2013).
833 58. Castro-Fernandez, V. *et al.* Reconstructed ancestral enzymes reveal that negative selection
834 drove the evolution of substrate specificity in ADP-dependent kinases. *J. Biol. Chem.* **292**,
835 15598–15610 (2017).

836 59. Anderson, D. W., McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the
837 function and evolution of an ancient transcription factor and its DNA binding sites. *Elife* **4**,
838 e07864 (2015).
839 60. Markin, C. J. *et al.* Revealing enzyme functional architecture via high-throughput microfluidic
840 enzyme kinetics. *Science* **373**, eabf8761 (2021).
841 61. Ben-David, M. *et al.* Enzyme evolution: An epistatic ratchet versus a smooth reversible
842 transition. *Mol. Biol. Evol.* **37**, 1133–1147 (2020).
843 62. Capra, E. J., Perchuk, B. S., Skerker, J. M. & Laub, M. T. Adaptive mutations that prevent
844 crosstalk enable the expansion of paralogous signaling protein families. *Cell* **150**, 222–232
845 (2012).
846 63. Zarrinpar, A., Park, S.-H. & Lim, W. A. Optimization of specificity in a cellular protein
847 interaction network by negative selection. *Nature* **426**, 676–680 (2003).
848 64. Liu, Q. *et al.* Ancient mechanisms for the evolution of the bicoid homeodomain's function in
849 fly development. *Elife* **7**, (2018).
850 65. Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence
851 space of an ancient protein. *Nature* **549**, 409–413 (2017).
852 66. McKeown, A. N. *et al.* Evolution of DNA specificity in a transcription factor family produced a
853 new gene regulatory module. *Cell* **159**, 58–68 (2014).
854 67. Srivastava, M. & Payne, J. L. On the incongruence of genotype-phenotype and fitness
855 landscapes. *PLoS Comput. Biol.* **18**, e1010524 (2022).
856 68. Park, Y., Metzger Brian, P. H. & Thornton Joseph, W. Epistatic drift causes gradual decay of
857 predictability in protein evolution. *Science* **376**, 823–830 (2022).
858 69. Yona, A. H., Alm, E. J. & Gore, J. Random sequences rapidly evolve into de novo promoters.
859 *Nat. Commun.* **9**, (2018).
860 70. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation
861 sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

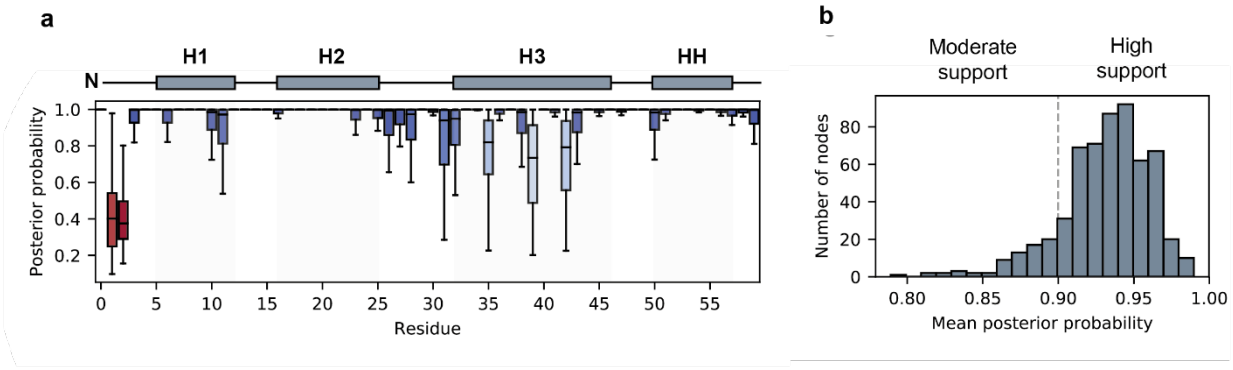
862 71. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate
863 multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
864 72. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective
865 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–
866 274 (2015).
867 73. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.*
868 **25**, 1307–1320 (2008).
869 74. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
870 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**,
871 587–589 (2017).
872 75. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving
873 the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* vol. 35 518–522
874 Preprint at https://doi.org/10.1093/molbev/msx281 (2018).
875 76. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and*
876 *Evolution* vol. 24 1586–1591 Preprint at https://doi.org/10.1093/molbev/msm088 (2007).
877 77. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in Escherichia
878 coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements. *Nucleic Acids Res.* **25**,
879 1203–1210 (1997).
880 78. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in Escherichia
881 coli K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6640–6645 (2000).
882 79. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-
883 End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
884 80. Bell, C. E. & Lewis, M. A closer view of the conformation of the Lac repressor bound to
885 operator. *Nat. Struct. Biol.* **7**, 209–214 (2000).
886 81. Delgado, J., Radusky, L. G., Cianferoni, D. & Serrano, L. FoldX 5.0: working with RNA, small
887 molecules and a new graphical interface. *Bioinformatics* **35**, 4168–4169 (2019).

82. Lu, X.-J. & Olson, W. K. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat. Protoc.* **3**, 1213–1227 (2008).

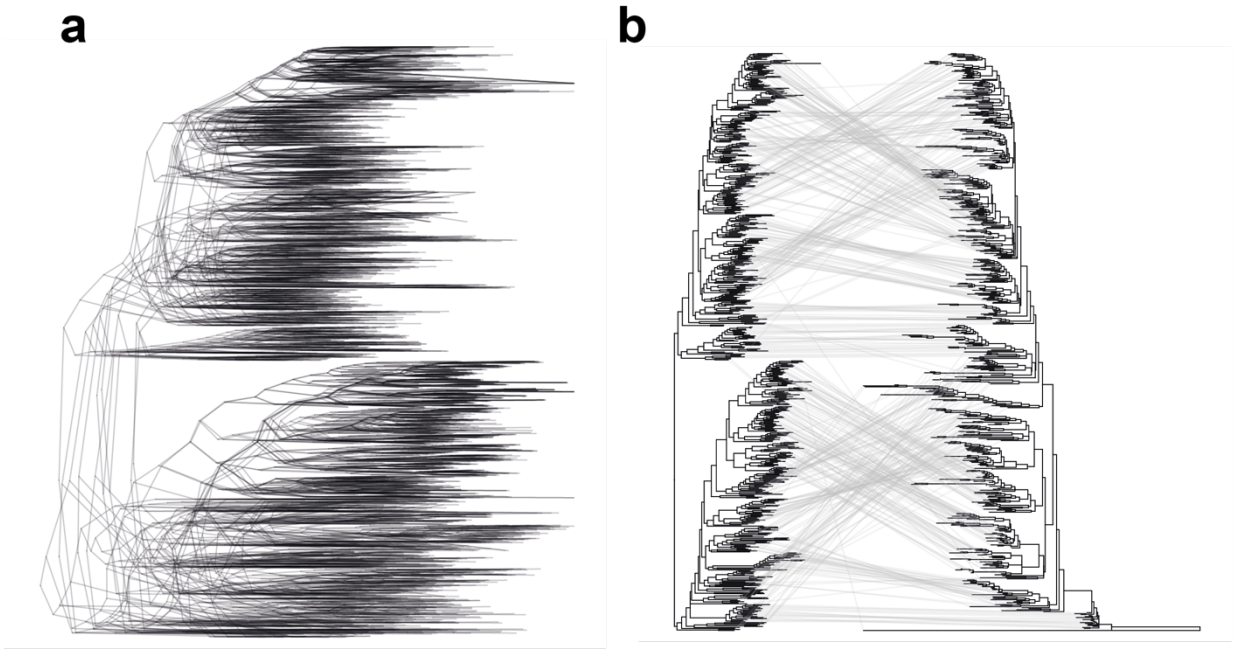
83. Maier, J. A. *et al.* Ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).

84. Ivani, I. *et al.* Parmbsc1: a refined force field for DNA simulations. *Nat. Methods* **13**, 55–58 (2016).

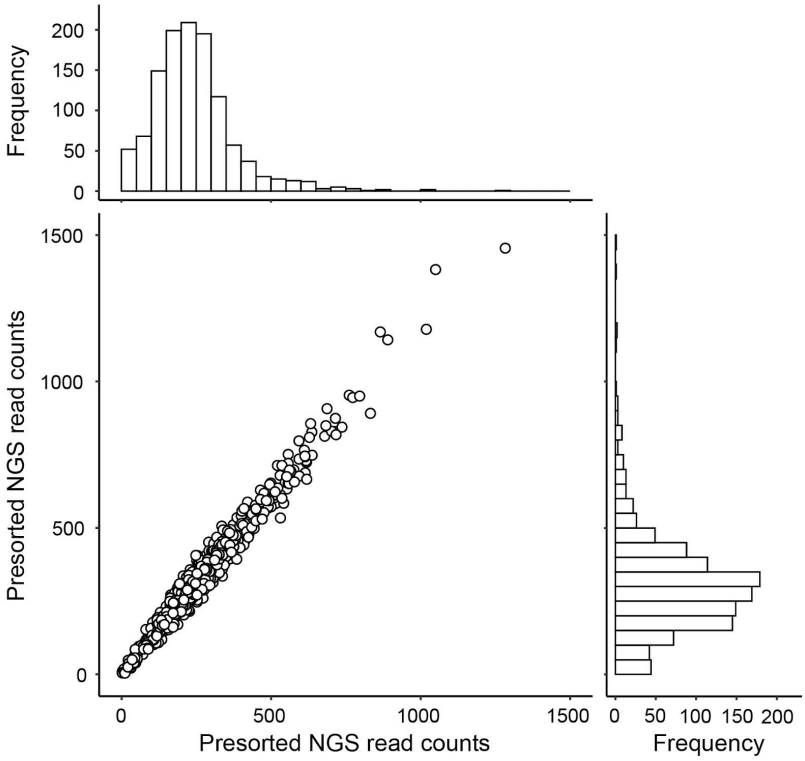
85. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).



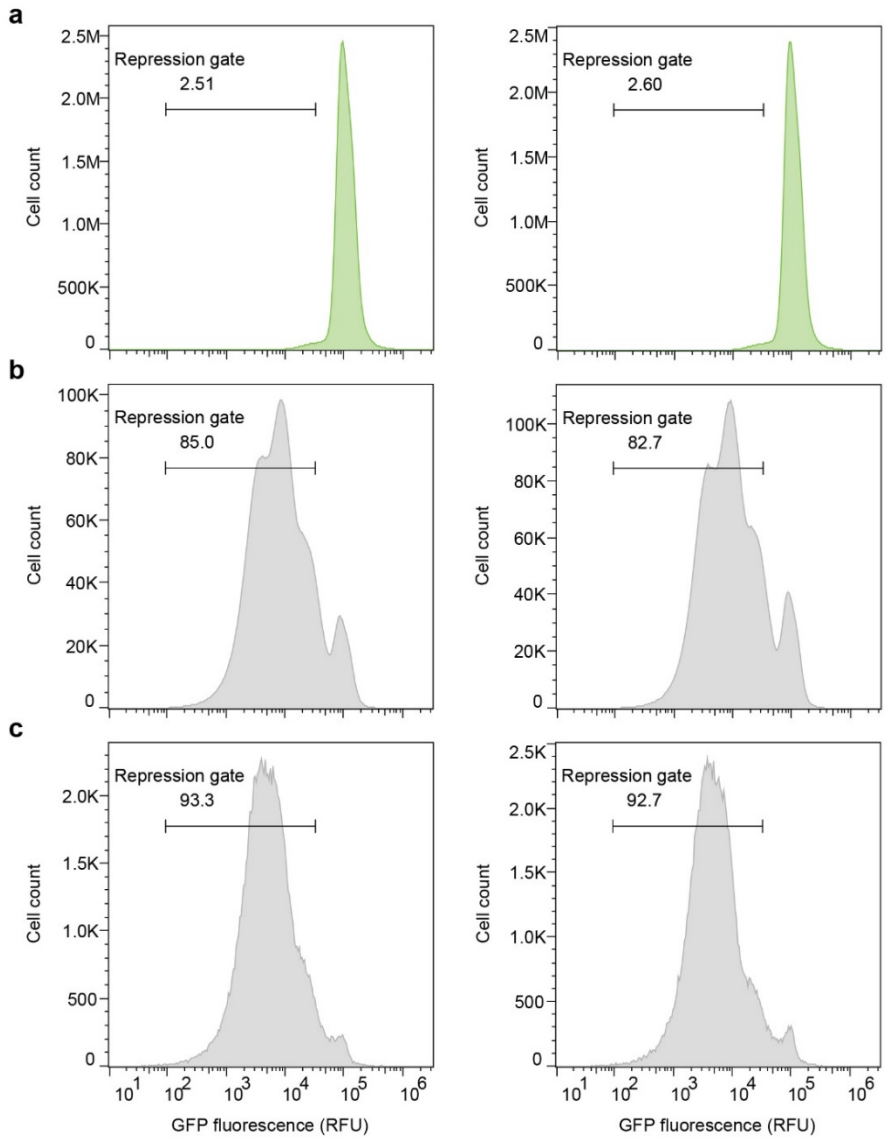
Supplementary Fig. 1. Ancestral sequence reconstruction of prokaryotic transcriptional regulators. (**a**), Site-wise posterior probability distribution averaged over all reconstructed ancestral sequences. (**b**), Distribution of posterior probabilities averaged over the DBD sequence. Most ancestral sequences are reconstructed with high posterior probability (>0.9 mean posterior probability) and very few are reconstructed with less than moderate (<0.8 mean posterior probability). Helix 2 and the hinge-helix, which are most important for DNA-recognition, are reconstructed unambiguously.



Supplementary Fig. 2. Tree topology testing. (**a**), All 10 reconstructed topologies that passed the AU test conducted to 10000 replicates. Each independent tree search converges on a similar topology. (**b**), Comparison the full DBD and LBD topology presented in main text Fig. 1 (left) and a tree-search performed under the same criteria using only the DBD. Most of the phylogenetic signal in the inference comes from the large LBD sequence, as the DBD-only topology is incongruent with the full-length topology.



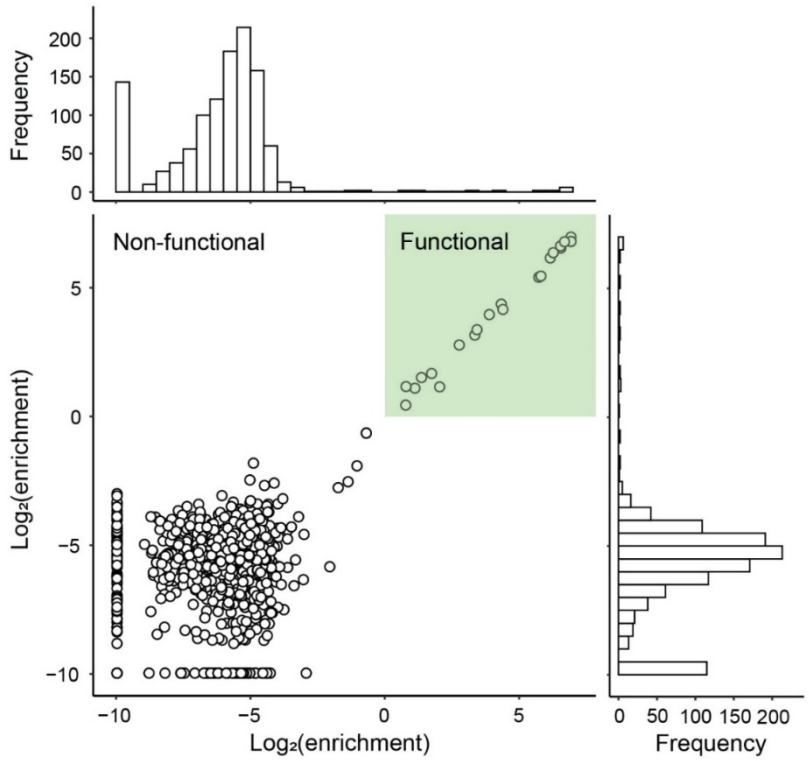
Supplementary Fig. 3. Presorted library distribution of extant and ancestral DBD variants. Raw read counts of all 1158 DBD variants from replicate next-generation sequencing runs (*n*=2). All variants were present in the pooled library (median=250±40 reads) prior to selection of functional repressors. Histograms of presorted read counts show that DBD variants were approximately normally distributed with a moderate right-shift (skewness=1.6±0.1).



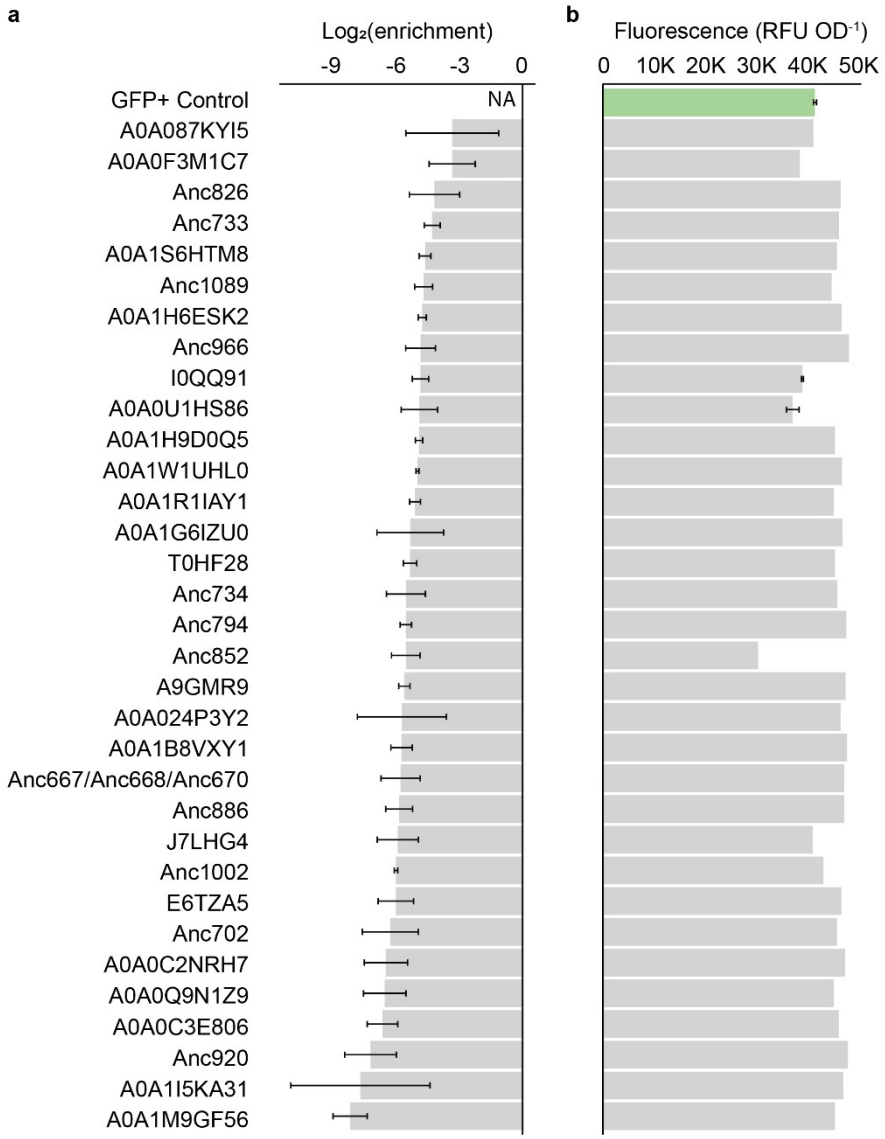
925

926 **Supplementary Fig. 4. Isolation of repression competent extant and ancestral DBDs using**
927 **FACS.** (**a**), Distribution of GFP fluorescence of the pre-selected phylogenetic DBD library using
928 flow cytometry. The small fraction (2.5-2.6%) of low-fluorescence cells shows that relatively few
929 extant and ancestral DBDs library members repress LacO. (**b-c**), Distribution of GFP fluorescence
930 after one (**b**) or two (**c**) rounds of FACS-based selection to isolate the low fluorescence
931 subpopulation. The indicated repression gate was used to enrich for repression competent
932 variants. RFU, relative fluorescence units.

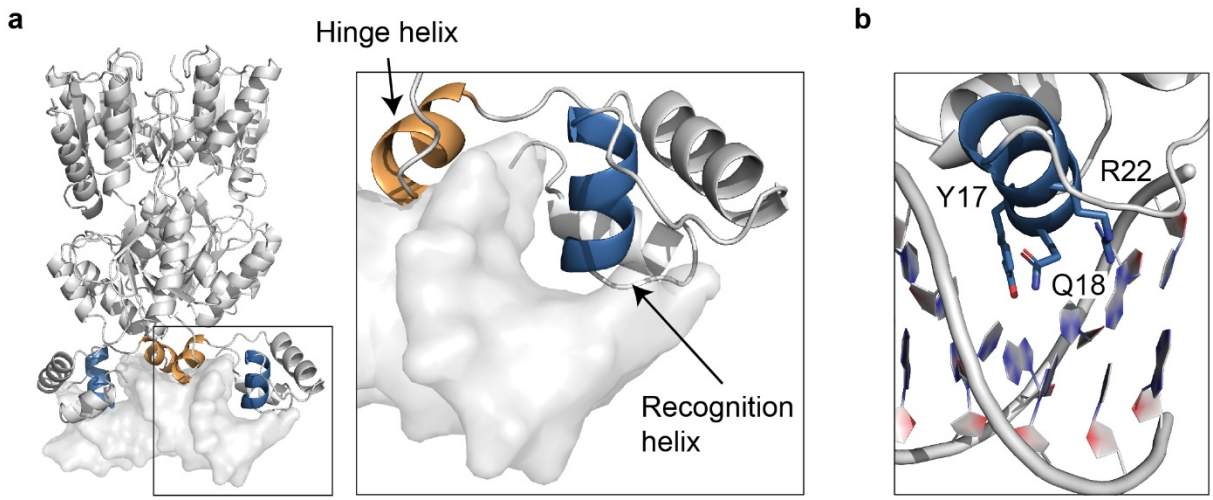
933



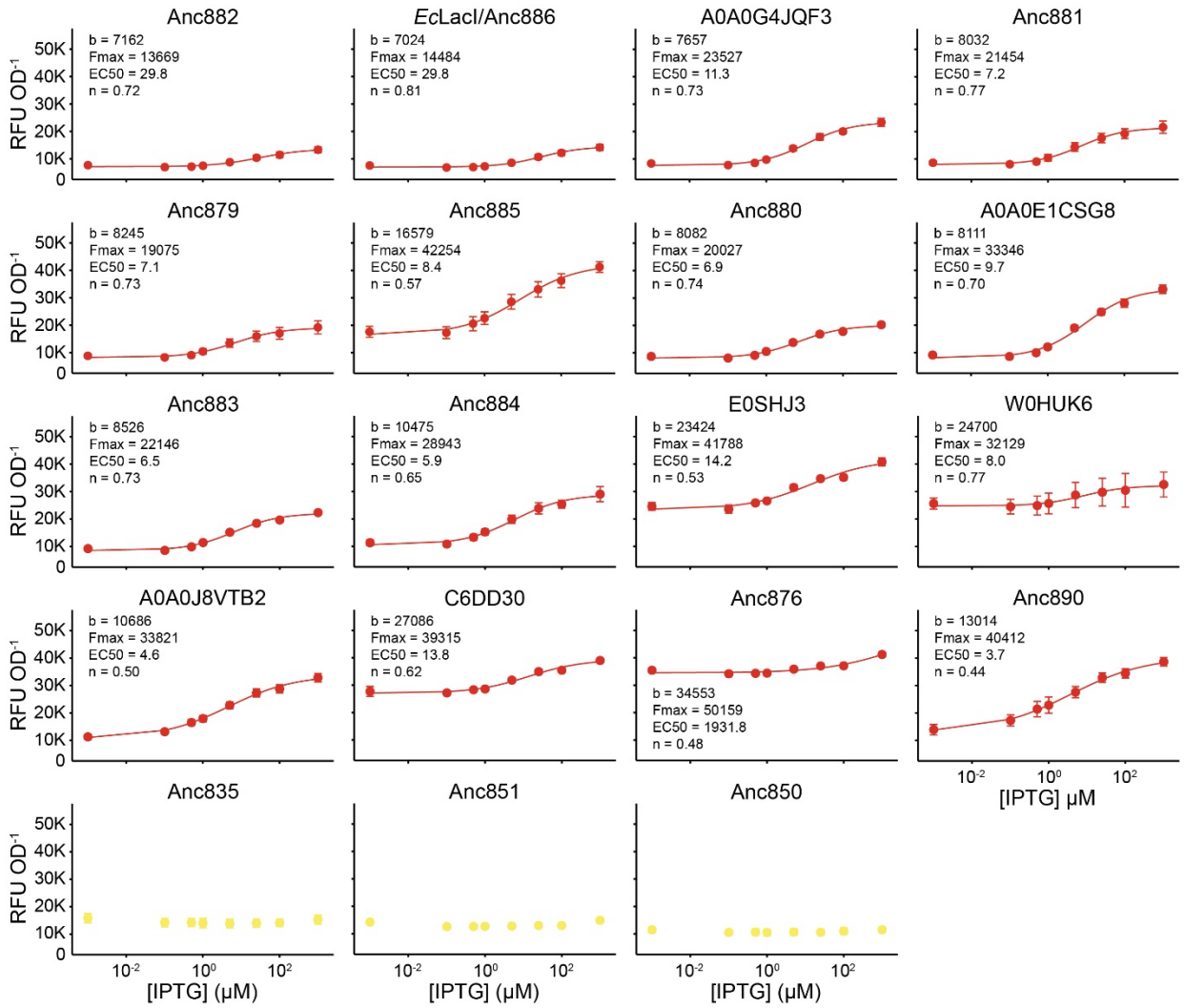
934
935
936 **Supplementary Fig. 5. Post-selection enrichment values of extant and ancestral DBD**
937 **variants.** Log transformed enrichment ratios of all 1158 DBD variants as determined by
938 comparing replicate NGS distributions (*n*=2) of the presorted and sorted populations. Higher
939 enrichment values indicate tighter repression of GFP and increased affinity to LacO. Histograms
940 indicate that few (22 of 1158) DBD variants were enriched and functionally repress LacO (green
941 box).
942



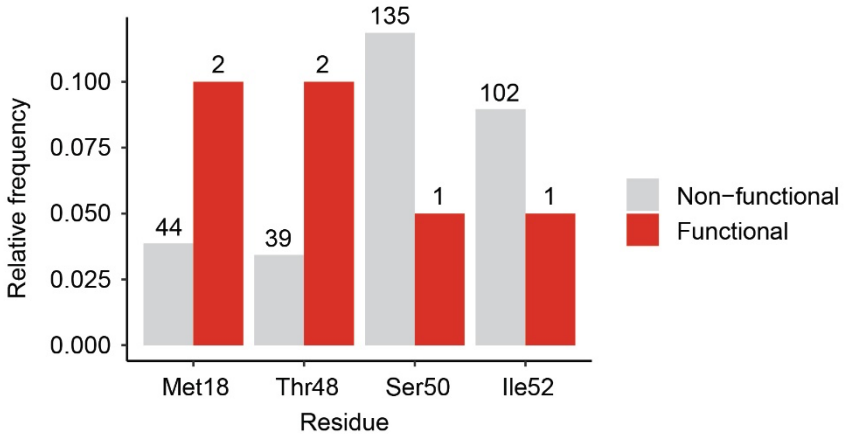
Supplementary Fig. 6. Clonal validation of high-throughput screen. (**a**) Enrichment values of a random subset of depleted ancestral and extant DBD variants derived from the pooled selection and next-generation sequencing experiments. Error bars represent the standard deviation of replicate experiments (*n*=2). (**b**) Clonally assayed mean fluorescence values normalized to OD_{600}. Clones were randomly sampled from the pre-selected population and assayed for GFP repression. All depleted clones exhibit high GFP fluorescence (inability to repress LacO) comparable to our GFP+ reporter cells (green). RFU, relative fluorescence units. OD, optical density.



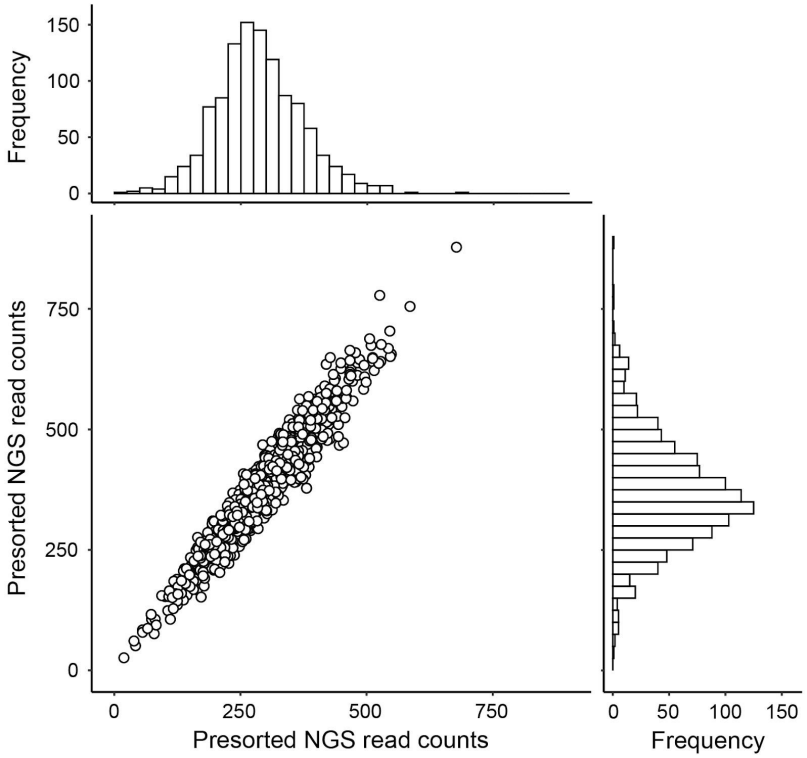
Supplementary Fig. 7. Structural regions of the DBD essential for function. (**a**) Structure (PDB ID: 1EFA) of dimeric extant *E. coli* lac repressor (cartoon) bound to LacO (surface representation). Sequences of functional ancestral and extant DBDs are highly conserved in both the hinge (orange) and recognition helices (blue) (**Fig. 2f**). The hinge helix interacts with the minor groove of LacO, facilitates homodimerization and is involved in allosteric signaling between the LBD and DBD. Functional repressors Anc835, Anc850 and Anc851 are heavily mutated in the hinge helix, but do not relieve GFP repression (**Supplementary Fig. 8**) in the presence of inducer (1mM IPTG), suggesting these mutations may disrupt the allosteric signal transduction pathway. (**b**) Residues of the recognition helix (Y17, Q18, and R22) that facilitate LacO recognition. Y17H and Q18M are the only substitutions tolerated at these critical positions among the functional ancestral and extant DBDs.



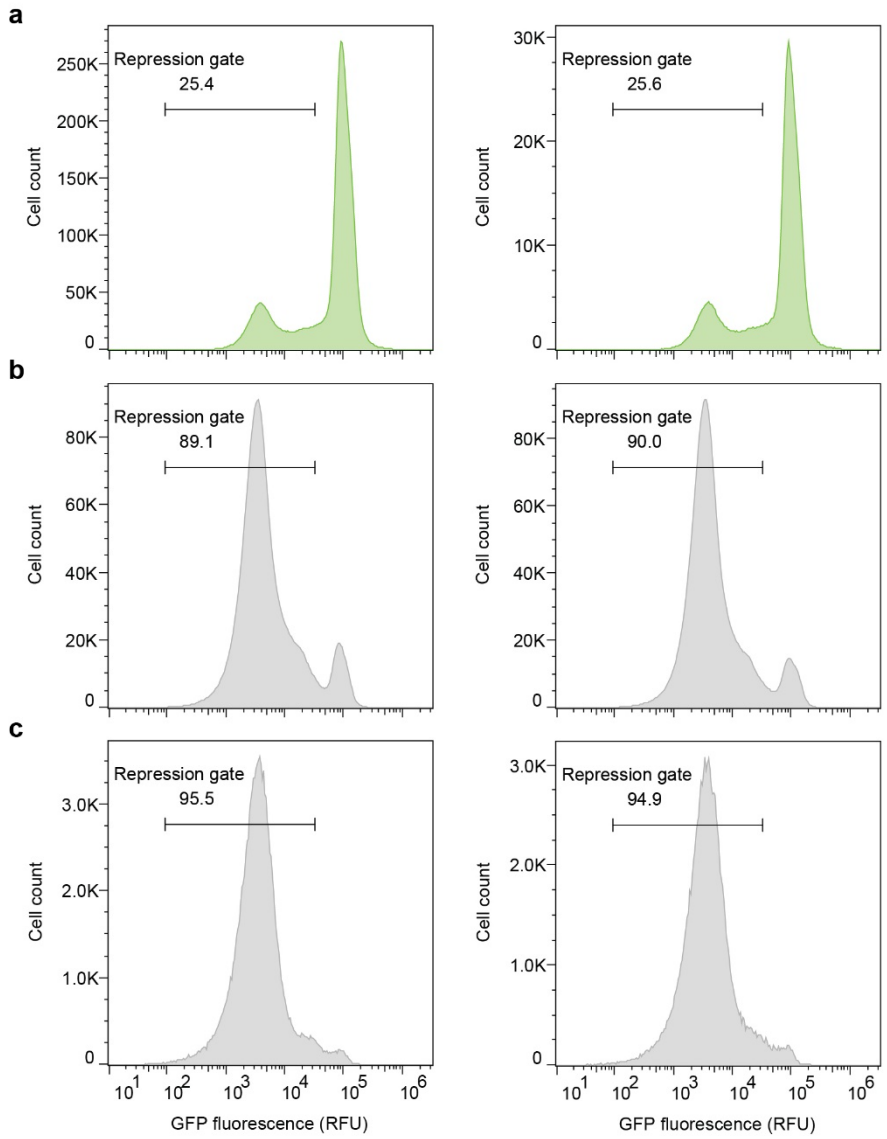
Supplementary Fig. 8. Induction assay of repression competent extant and ancestral DBD variants. The gratuitous inducer IPTG binds to the LBD of *E. coli* LacI to allosterically relieve LacO repression. Dose response curves were fit to the Langmuir-Hill Equation for inducible variants (red). IPTG did not relieve LacO repression for Anc835, Anc851, and Anc850 (yellow). Basal fluorescence, b, is a measure of LacO affinity in the absence of inducer. Fmax is the maximum fluorescence signal achieved at 1mM IPTG. EC50 is the half maximal effective concentration, and n is the Hill coefficient, which is a measure of cooperativity. Error bars represent the standard deviation of replicate experiments ($n \geq 2$).



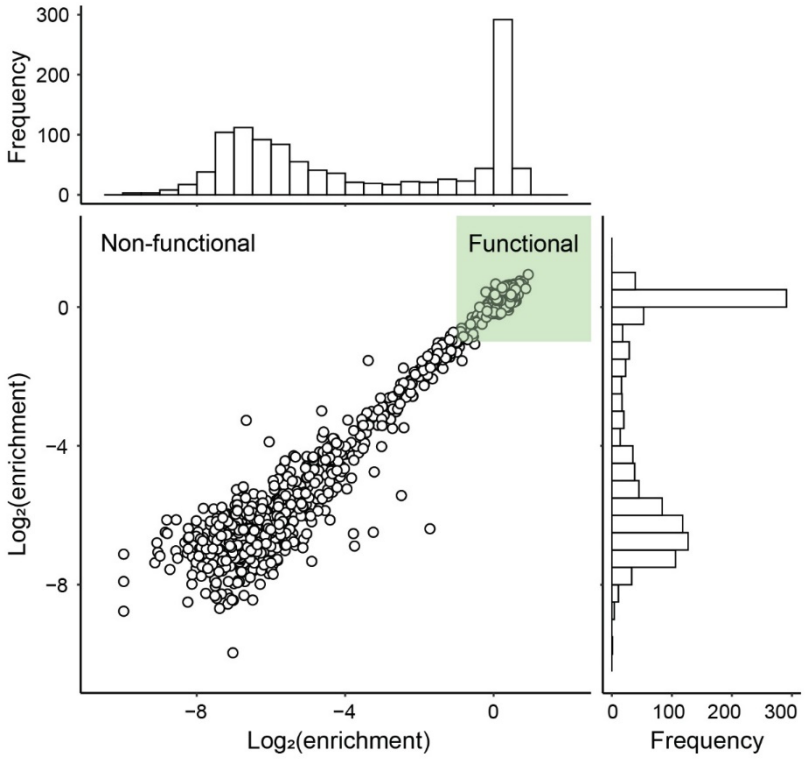
Supplementary Fig. 9. Gln18Met substitution provides additional plasticity for LacO repression. Relative amino acid frequencies of Met18, Thr48, Ser50, and Ile52 for non-functional (gray) and functional (red) phylogenetic DBDs. The Gln18Met substitution in *E. coli* LacI has been previously shown to increase affinity to LacO by 2.5-fold.[39] Both functional repressors Anc849 and Anc850 have the Gln18Met substitution, and this increased LacO affinity may contribute to the loss of IPTG inducibility (**Supplementary Fig. 8**) and enable less favorable substitutions (Ser50 and Ile52 are more prominent among non-functional DBDs) to be tolerated in Anc851.



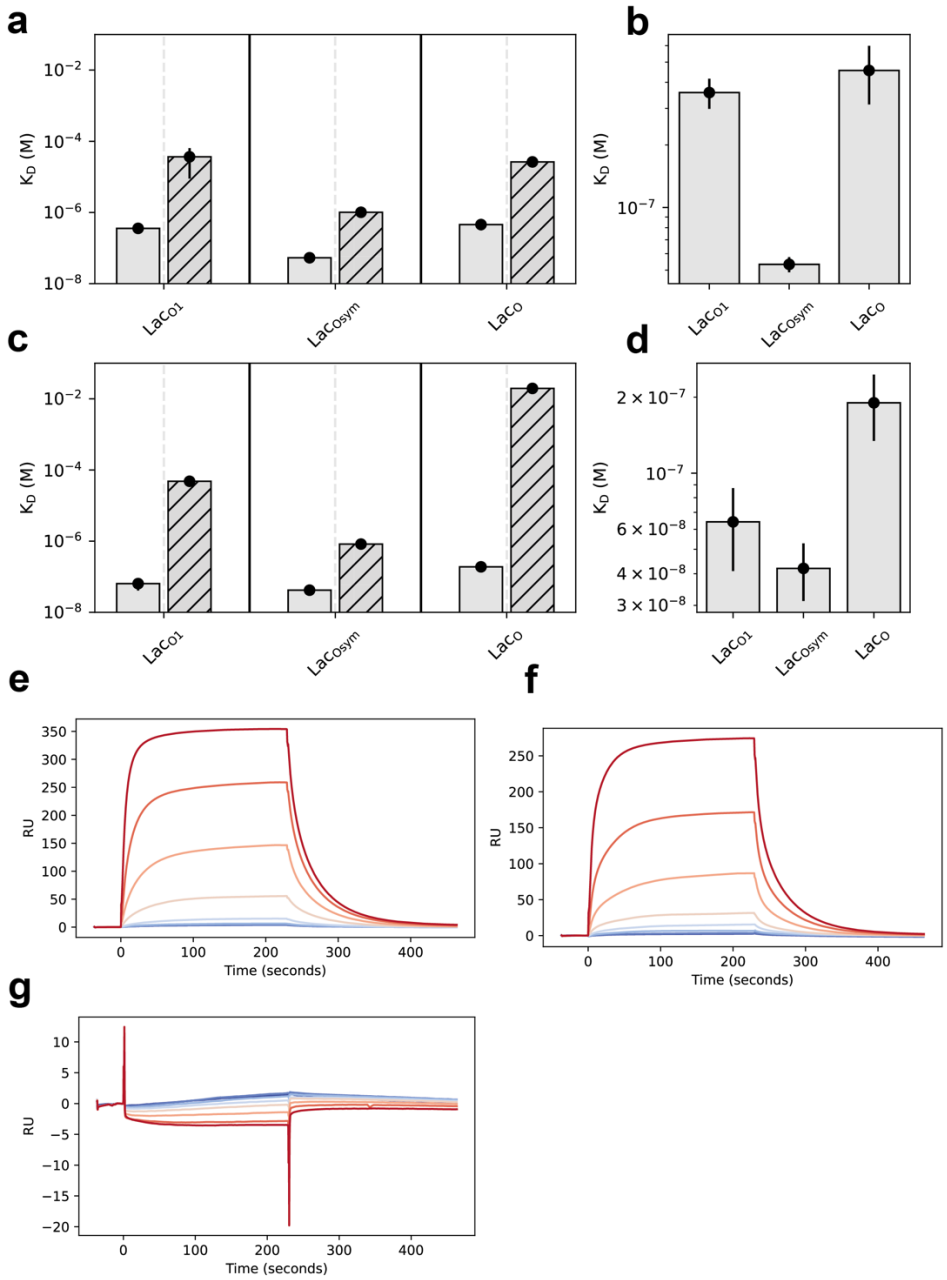
Supplementary Fig. 10 Presorted library distribution of *E. coli* LacI DBD deep mutational scanning variants. Raw read counts of all 1121 DBD single mutant variants from replicate next-generation sequencing runs (*n*=2). All DMS variants were present in the pooled library (median=320±50 reads) prior to selection of functional repressors. Histograms of presorted read counts show that DBD variants were approximately normally distributed with minimal skew (skewness=0.38±0.04).



Supplementary Fig. 11. Isolation of repression competent *E. coli* LacI DBD deep mutational scanning variants using FACS. (**a**), Distribution of GFP fluorescence of the pre-selected library using flow cytometry. A larger fraction of low-fluorescence cells (25.4-25.6%) was observed for the DMS library than the phylogenetic library, which indicates that many more DMS variants repress LacO. (**b-c**), Distribution of GFP fluorescence after one (**b**) or two (**c**) rounds of FACS-based selection to isolate the low fluorescence subpopulation. The indicated repression gate was used to enrich for repression competent variants. RFU, relative fluorescence units.

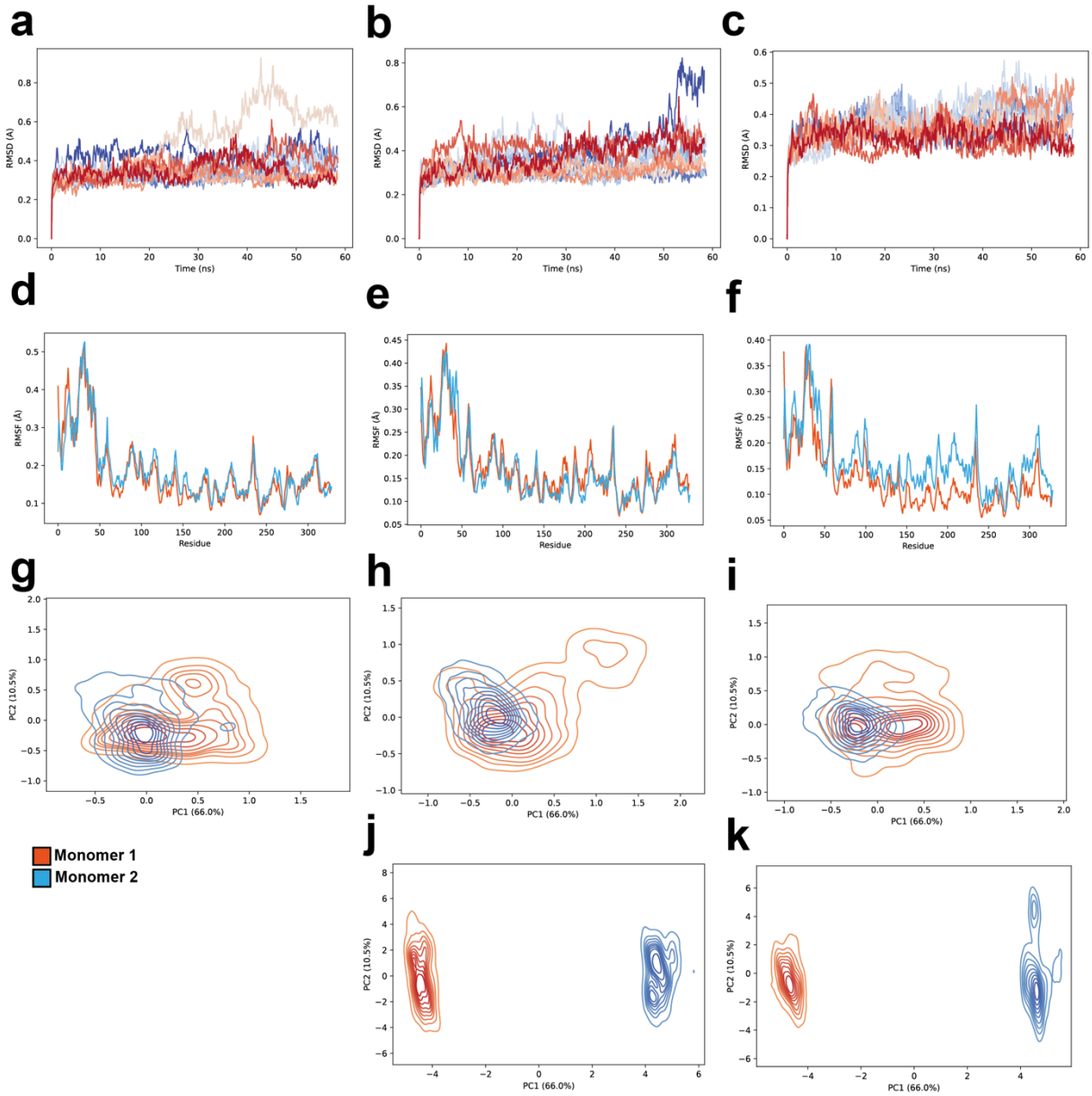


Supplementary Fig. 12. Post-selection enrichment values of *E. coli* LacI DBD deep mutational scanning variants. Log transformed enrichment ratios of all 1158 DBD variants as determined by comparing replicate NGS distributions (*n*=2) of the presorted and sorted populations. Higher enrichment values indicate tighter repression of GFP and increased affinity to LacO. Histograms indicate that 400 (35%) DBD variants were enriched and functionally repress LacO (green box).

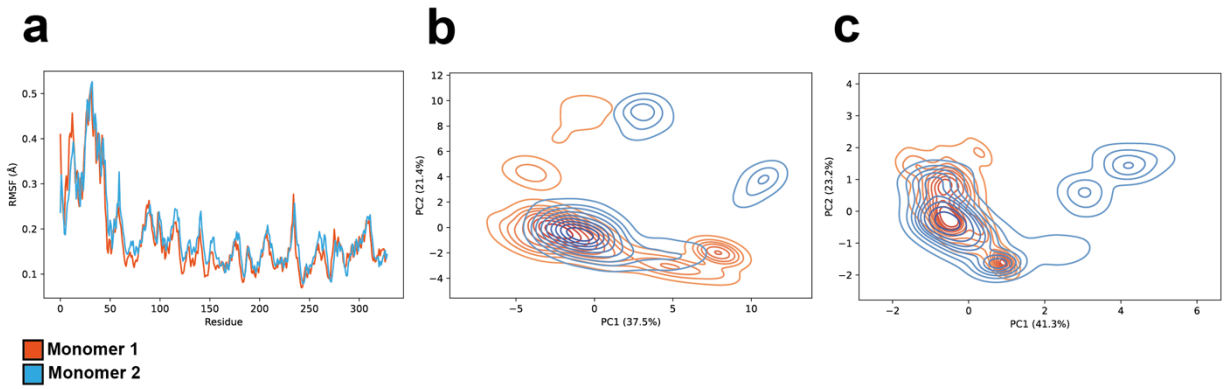


Supplementary Fig. 13 Surface plasmon resonance kinetics and sensograms. (a) K_d constants of Anc880 for $LacO_1$, $LacO_{sym}$ and LacO when repressed (solid) and induced with 70 μM IPTG (hashed). (b) Zoomed in view of repressed K_d constants for Anc880 binding. (c) K_d constants of Anc881 for $LacO_1$, $LacO_{sym}$ and LacO when repressed (solid) and induced with 70 μM IPTG (hashed). (d)) Zoomed in view of repressed K_d constants for Anc881 binding. (e) Sensogram for Anc880 LacO binding at protein concentrations of 250 nM – 3.91 nM in a log2 dilution series. (f) Sensogram for Anc881 LacO binding at protein concentrations of 250 nM –

1020 3.91 nM in a log2 dilution series. (g) Sensogram for Anc882 binding random DNA at protein
1021 concentrations of 250 nM – 3.91 nM in a log2 dilution series**.** There is no detectable binding for
1022 random DNA.



Supplementary Fig. 14. Molecular dynamics simulations of Anc880, Anc882 and EcLacI. (a – c). RMSD evolution of Anc882, Anc880 and EcLacI, respectively. The 10 independent replicates are distinguished by colour. (d-f) RMSF split by monomer for Anc882, Anc880 and EcLacI, respectively. (g-i) PCA of backbone alpha carbons within only the DBD for Anc882, Anc880 and EcLacI, respectively. (j-k) PCA of backbone alpha carbons over the full regulator structure for Anc880 and EcLacI.



Supplementary Fig. 15. Molecular dynamics simulations of apo-Anc882. (a) RMSF of apo-Anc882 trajectories split by monomer. (b) PCA of backbone alpha carbons over the full apo-Anc882 structure. (c) PCA of backbone alpha carbons over the apo-Anc882 DBD. Motions between monomers in the apo-Anc882 trajectory are highly similar.