Adaptation in protein fitness landscapes is facilitated by indirect paths

Nicholas C. Wu[1,2,4,†], Lei Dai[1,3,†], C. Anders Olson[1], James O. Lloyd-Smith[3], Ren Sun[1,2,*]

[1]Department of Molecular and Medical Pharmacology,

[2]Molecular Biology Institute,

[3]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095, USA

[4]Present address: Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

[†]These authors contributed equally to this work

[*]To whom correspondence may be addressed. Email: RSun@mednet.ucla.edu

Abstract

The structure of fitness landscapes is critical for understanding adaptive protein evolution (e.g. antimicrobial resistance, affinity maturation, etc.). Due to limited throughput in fitness measurements, previous empirical studies on fitness landscapes were confined to either the neighborhood around the wild type sequence, involving mostly single and double mutants, or a combinatorially complete subgraph involving only two amino acids at each site. In reality, however, the dimensionality of protein sequence space is higher (20^L, L being the length of the relevant sequence) and there may be higher-order interactions among more than two sites. To study how these features impact the course of protein evolution, we experimentally characterized the fitness landscape of four sites in the IgG-binding domain of protein G, containing 20^4 = 160,000 variants. We found that the fitness landscape was rugged and direct paths of adaptation were often constrained by pairwise epistasis. However, while direct paths were blocked by reciprocal sign epistasis, we found systematic evidence that such evolutionary traps could be circumvented by "extra-dimensional bypass". Extra dimensions in sequence space – with a different amino acid at the site of interest or an additional interacting site – open up indirect paths of adaptation via gain and subsequent loss of mutations. These indirect paths alleviate the constraint on reaching high fitness genotypes via selectively accessible trajectories, suggesting that the heretofore neglected dimensions of sequence space may completely change our views on how proteins evolve.

The fitness landscape is a fundamental concept in evolutionary biology [1–6]. Large-scale datasets combined with quantitative analysis have successfully unraveled important features of empirical fitness landscapes [7–9]. Nevertheless, there is a huge gap between the limited throughput of fitness measurements (usually on the order of 10^2 variants) and the vast size of sequence space. Recently, the bottleneck in experimental throughput has been improved substantially by coupling saturation mutagenesis with deep sequencing [10–16], which opens up unprecedented opportunities to understand the structure of high-dimensional fitness landscapes [17–19].

Previous empirical studies on combinatorially complete fitness landscapes have been limited to subgraphs of the sequence space consisting of only two amino acids at each site (2^L genotypes) [20–25]. Adaptive walks in these subgraphs can only follow "direct paths", where each mutational step reduces the Hamming distance from the starting point to the destination. In sequence space with higher dimensionality (20^L, for a protein sequence with L amino acid residues), however, the extra dimensions may provide additional routes for adaptation. For example, some evolutionary dead ends (i.e. local maxima) may become saddle points and allow for further increase in fitness [26]. In this case, adaptation may proceed via "indirect paths" in sequence space, which involve extra mutations and reversions. The existence of indirect paths has been implied in different contexts [27, 28], but has not been studied systematically so its influence on protein adaptation remains unclear. Another underappreciated property of fitness landscapes is the influence of higher-order interactions. Empirical evidence suggests that pairwise epistasis is prevalent in fitness landscapes [7, 22, 23, 29]. Specifically, sign epistasis between two loci is known to constrain adaptation by limiting the number of selectively accessible paths [20]. Higher-order epistasis (i.e. interactions among more than two loci) has received much less attention and its role in adaptation is yet to be elucidated [28,30].

In this study, we investigated the fitness landscape of all variants (20^4 = 160,000) at four amino acid sites (V39, D40, G41 and V54) in an epistatic region of protein G domain B1 (GB1, 56 amino acids in total) (Supplementary Fig. 1), an immunoglobulin-binding protein expressed in Streptococcal bacteria [31, 32]. The four chosen sites contain 12 of the top 20 positively epistatic interactions among all pairwise interactions in protein GB1, as we previously characterized [33] (Supplementary Fig. 2). Thus the sequence space is expected to cover highly beneficial variants, which presents an ideal scenario for studying adaptive evolution. Briefly, a mutant library containing all amino acid combinations at these four sites was generated by

codon randomization. The "fitness" of protein GB1 variants, as determined by both stability (i.e. the fraction of folded proteins) and function (i.e. binding affinity to IgG-Fc), was measured in a high-throughput manner by coupling mRNA display with Illumina sequencing (Methods, Supplementary Fig. 3A) [34, 35]. The relative frequency of mutant sequences before and after selection allowed us to compute the fitness of each variant relative to the wild type protein (WT).

To understand the impact of epistasis on protein adaptation, we first analyzed subgraphs of sequence space including only two amino acids at each site (Fig. 1A). Each subgraph represented a classical adaptive landscape connecting WT to a beneficial quadruple mutant, analogous to previously studied protein fitness landscapes [9, 20]. Each variant is denoted by the single letter code of amino acids across sites 39, 40, 41 and 54 (for example, WT sequence is VDGV). Each subgraph is combinatorially complete with 2^4 = 16 variants, including WT, the quadruple mutant, and all intermediate variants. We identified a total of 29 subgraphs in which the quadruple mutant was the only fitness peak. By focusing on these subgraphs, we essentially limited the analysis to direct paths of adaptation, where each step would reduce the Hamming distance from the starting point (WT) to the destination (quadruple mutant). Out of 24 possible direct paths, the number of selectively accessible paths (i.e. with monotonically increasing fitness) varied from 12 to 1 among the 29 subgraphs (Fig. 1B). In the most extreme case, only one path was accessible from WT to the quadruple mutant WLFA (Fig. 1A). We also observed a substantial skew in the computed probability of realization among accessible direct paths (Supplementary Fig. 4), suggesting that most of the realizations in adaptation were captured by a small fraction of possible trajectories [20]. These results indicated the existence of sign epistasis and reciprocal sign epistasis, both of which may constrain the accessibility of direct paths [20, 36]. Indeed, we found that these two types of epistasis were prevalent in our fitness landscape (Fig. 1C). Furthermore, we classified the types of all 24 pairwise epistasis in each subgraph and computed the level of ruggedness as $f_{sign} + 2f_{reciprocal}$, where f_{type} was the fraction of each type of pairwise epistasis. As expected, the number of selectively inaccessible direct paths, i.e. paths that involve fitness declines, was found to be positively correlated with the ruggedness induced by pairwise epistasis (Fig. 1D, Pearson correlation = 0.66, p=1.0×10^{-4}) [2].

Our findings support the view that direct paths of protein adaptation are often constrained by pairwise epistasis on a rugged fitness landscape [5, 37]. In particular, adaptation can be trapped when direct paths are

blocked by reciprocal sign epistasis. However, crucially, this analysis was limited to mutational trajectories within a subgraph of the sequence space. In reality, the dimensionality of protein sequence space is higher. Intuitively, when an extra dimension is introduced, a local maximum may become a saddle point and allow for further adaptation – a phenomenon recently proposed under the name "extra-dimensional bypass" [38]. We discovered two distinct mechanisms of bypass, either using an extra amino acid at the same site or using an additional site, that allow proteins to continue adaptation when no direct paths were accessible due to reciprocal sign epistasis (Fig. 2). The first mechanism of bypass, which we termed "conversion bypass", works by converting to an extra amino acid at one of the interacting sites [28]. Consider a simple scenario with only two interacting sites. If the sequence space is limited to 2 amino acids at each site, as in past analyses of adaptive trajectories, the number of neighbors is 2; however, if all 20 possible amino acids were considered, the total number of neighbors would be 38. Some of these 36 extra neighbors may lead to potential routes that circumvent the reciprocal sign epistasis (Fig. 2A). In this case, a successful bypass would require a conversion step that substitutes one of the two interacting sites with an extra amino acid ($00 \rightarrow 20$), followed by the loss of this mutation ($21 \rightarrow 11$). This bypass is feasible only if the original reciprocal sign epistasis is changed to sign epistasis after the conversion. To test whether such bypasses were present in our system, we randomly sampled 10^5 pairwise interactions from the sequence space and analyzed the ∼20,000 reciprocal sign epistasis among them (Methods). More than 40% of the time there was at least one successful conversion bypass and in many cases multiple bypasses were available (Fig. 2B).

The second mechanism of bypass, which we termed "detour bypass", involves an additional site (Fig. 2C). In this case, adaptation can proceed by taking a detour step to gain a mutation at the third site ($000 \rightarrow 100$), followed by the later loss of this mutation ($111 \rightarrow 011$) [27,28]. Detour bypass was observed in our system (Fig. 2D), but was not as prevalent and had a lower probability of success than conversion bypass. Out of 38 possible detour bypasses for a chosen reciprocal sign epistasis, we found that there were on average 1.2 conversion bypasses and 0.27 detour bypasses available. We note, however, that the lower prevalence of detour bypass in our fitness landscape (L=4) does not necessarily mean that it should be expected to be less frequent than conversion bypass in other systems. While the maximum number of possible conversion bypasses is always fixed ($19 \times 2 - 2 = 36$), the maximum number of possible detour bypasses ($19 \times (L-2)$) is proportional to the sequence length L of the entire protein (whereas our study uses a subset $L = 4$). The pervasiveness of extra-dimensional bypasses in our system contrasts with the prevailing view that adaptive

evolution is often blocked by reciprocal sign epistasis, when only direct paths of adaptation are considered. The two distinct mechanisms of bypass both require the use of indirect paths, where the Hamming distance to the destination is either unchanged (conversion) or increased (detour).

In order to circumvent the inaccessible direct paths via extra dimensions, reciprocal sign epistasis must be changed into other types of pairwise epistasis. For detour bypass, this means that the original reciprocal sign epistasis is changed to either magnitude epistasis or sign epistasis in the presence of a third mutation (Supplementary Fig. 5A). There are three possible scenarios where detour bypass can occur (Supplementary Fig. 5B-D). We proved that higher-order epistasis is necessary for the scenario that reciprocal sign epistasis is changed to magnitude epistasis, as well as for one of the two scenarios that reciprocal sign epistasis is changed to sign epistasis (Supplementary Text). This suggests a critical role of higher-order epistasis in mediating detour bypass.

To confirm the presence of higher-order epistasis, we decomposed the fitness landscape by Fourier analysis (Fig. 3A, Methods) [9, 30]. The Fourier coefficients can be interpreted as epistatic interactions of different orders [6, 30], including the main effects of single mutations (the 1^{st} order), pairwise epistasis (the 2^{nd} order), and higher-order epistasis (the 3^{rd} and the 4^{th} order). The fitness of variants can be reconstructed by expansion of Fourier coefficients up to a certain order (Supplementary Fig. 6). In our system with four sites, the 4^{th} order Fourier expansion will always reproduce the measured fitness (i.e. Pearson correlation equals 1). When the 2^{nd} order Fourier expansion does not reproduce the measured fitness (i.e. Pearson correlation less than 1), it indicates the presence of higher-order epistasis. In this way, we identified the 0.1% of subgraphs with greatest fitness contribution from higher-order epistasis (Fig. 3A, red lines) and visualized the corresponding quadruple mutants by the sequence logo plot (Fig. 3B). The skewed composition of amino acids in these subgraphs indicates that higher-order interactions are enriched among specific amino acid combinations of site 39, 41 and 54. This interaction among 3 sites is consistent with our knowledge of the protein structure, where the side chains of sites 39, 41, and 54 can physically interact with each other at the core (Supplementary Fig. 1A) and destabilize the protein due to steric effects (Supplementary Fig. 7).

In the presence of higher-order epistasis, epistasis between any two sites would vary across different genetic backgrounds. We computed the magnitude of pairwise epistasis (ε) between each pair of amino acid

147 substitutions (Methods)[39], and observed numerous instances where the sign of pairwise epistasis depended
148 on genetic background. For example, G41L and V54H were positively epistatic when site 39 was isoleucine
149 [I], but the interaction changed to negative epistasis when site 39 carried a tyrosine [Y] or a tryptophan [W]
150 (Fig. 3C-D). Similar patterns were observed in other pairwise interactions among site 39, 41 and 54, such
151 as G41F/V54A and V39W/V54H (Supplementary Fig. 8). The observed pattern of higher-order epistasis
152 was consistent with the results of the Fourier analysis (Fig. 3B). For example, site 40 was mostly excluded
153 from higher-order epistasis; tyrosine [Y] or tryptophan [W] at site 39 were involved in the most significant
154 higher-order interactions, as they often changed the sign of pairwise epistasis. Higher-order epistasis can
155 also switch the type of pairwise epistasis, such as shifting from reciprocal sign epistasis to magnitude or
156 sign epistasis (Supplementary Fig. 9), which in turn is important for the existence of detour bypass.
157
158 Our analysis on circumventing reciprocal sign epistasis revealed how indirect paths could open up new
159 avenues of adaptation. To study the impact of indirect paths at a global scale, we performed simulated
160 adaptation in the entire sequence space of 160,000 variants. The fitness landscape was completed by im-
161 puting fitness values of the 10,639 missing variants (i.e. 6.6% of the sequence space) that had fewer than
162 10 sequencing read counts in the input library. Our model of protein fitness incorporated main effects of
163 single mutations, pairwise interactions, and three-way interactions among site 39, 41 and 54 (Methods, Sup-
164 plementary Fig. 10). We used predictor selection based on biological knowledge, followed by regularized
165 regression, which has been demonstrated to ameliorate possible bias in the inferred fitness landscape [40].
166 In the complete sequence space, we identified a total of 30 fitness peaks (i.e. local maxima); among them
167 15 peaks had fitness larger than WT and their combined basins of attraction covered 99% of the sequence
168 space (Fig. 4A).
169
170 We then simulated adaptation on the fitness landscape using three different models of adaptive walks (Meth-
171 ods), namely the Greedy Model [6], Correlated Fixation Model [41], and Equal Fixation Model [20]. In the
172 Greedy Model, adaptation proceeds by sequential fixation of mutations that render the largest fitness gain
173 at each step. The other two models assign a nonzero fixation probability to all beneficial mutations, either
174 weighted by (Correlated Fixation Model) or independent of (Equal Fixation Model) the relative fitness gain.
175 Among all the possible adaptive paths to fitness peaks, many of them involved indirect paths, i.e. they em-
176 ployed mechanisms of extra-dimensional bypass (Fig. 4B, Supplementary Fig. 11). We classified each step

on the adaptive paths into three categories based on the change of Hamming distance to the destination (a fitness peak, in this case): "towards (-1)", "conversion (0)", and "detour (+1)" (Fig. 4C). Conversion was found to be pervasive during adaptation in our fitness landscape (17% of mutational steps for Greedy Model, 41% for Correlated Fixation Model, 59% for Equal Fixation Model). The use of detour was less frequent (0.1% of mutational steps for Greedy Model, 1.3% for Correlated Fixation Model, 3.7% for Equal Fixation Model), in accordance with the previous observation that detour bypass was less available than conversion bypass in our fitness landscape with $L = 4$. A conversion step would increase the length of an adaptive path by 1, while a detour step would increase the length by 2. As a result, an indirect path can be substantially longer than a direct path consisting of only "towards" steps. We found that many of the adaptive paths required more than 4 steps, which was the maximal length of a direct path between any variants in this landscape (Fig. 4D). Interestingly, because indirect adaptive paths involved more variants of intermediate fitness, the use of conversion and detour steps depended on the strength of selection. When mutations conferring larger fitness gains were more likely to fix (e.g. Greedy Model and Correlated Fixation Model), adaptation favored direct moves toward the destination, thus leading to a shorter adaptive paths (Fig. 4C-D). This suggests that the strength of selection interacts with the topological structure of fitness landscapes to determine the length and directness of evolutionary trajectories.

Given that extra-dimensional bypasses can help proteins avoid evolutionary traps, we expect that their existence would facilitate adaptation in rugged fitness landscapes. Indeed, we found that indirect paths increased the number of genotypes with access to each fitness peak (Fig. 4E). In addition, the fraction of genotypes with accessible paths to all 15 fitness peaks increased from from 34% to 93% when indirect adaptive paths were allowed (Supplementary Fig. 11C). We also found that a substantial fraction of beneficial variants (fitness > 1) in the sequence space were accessible from WT only if indirect paths were used (Fig. 4F). Taken together, these results suggest that indirect paths promote evolutionary accessibility in rugged fitness landscapes. This enhanced accessibility would allow proteins to explore more sequence space and lead to delayed commitment to evolutionary fates (i.e. fitness peaks) [28]. Consistent with this expectation, our simulations showed that many mutational trajectories involving extra-dimensional bypass did not fully commit to a fitness peak until the last two steps (Supplementary Fig. 12).

In our analysis, we have limited adaptation to the regime where fitness is monotonically increasing via

sequential fixation of one-step beneficial mutants. When this assumption is relaxed, adaptation can sometimes proceed by crossing fitness valleys [2, 6, 42, 43]. Another simplification in our analysis is to treat all sequences in a "protein space" [44], where two sequences are considered as neighbors if they differ by a single amino-acid substitution. In practice, amino acid substitutions occurring via a single nucelotide mutation are limited by the genetic code, so the total number of one-step neighbors would be reduced from $19L$ to approximately $6L$. We also expect fitness landscapes of different systems to have different topological structure. Even in our system (with >93% coverage of the genotype space), the global structure of the fitness landscape is influenced by the imputed fitness values of missing variants, which can vary when different fitness models or fitting methods are used. Our analysis also ignored measurement errors, but the measurement errors are expected to be very small due to the high reproducibility in the data (Supplementary Fig. 3B). Both imputation of missing variants and measurement errors can lead to slight mis-specification of the topological structure of the fitness landscape. Nevertheless, specific details of a certain fitness landscape do not undermine the generality of our findings on extra-dimensional bypass, higher-order epistasis, and their roles in protein evolution.

Higher-order epistasis has been reported in a few biological systems [28, 45, 46], and is likely to be common in nature [30]. In this study, we uncovered the presence of higher-order epistasis and systematically quantified its contribution to protein fitness. We also revealed the importance of higher-order epistasis in mediating detour bypass, which could promote evolutionary accessibility in rugged fitness landscapes. As we pointed out, the possible number of detour bypasses scales up with sequence length, so it will be interesting to study how extra-dimensional bypass influences adaptation in sequence space of even higher dimensionality. For example, it is plausible that the sequence of a large protein may never be trapped in adaptation [47], so that adaptive accessibility becomes a quantitative rather than qualitative problem. Given the continuing development of sequencing technology, we anticipate that the scale of experimentally determined fitness landscapes will further increase, yet the full protein sequence space is too huge to be mapped exhaustively. Does this mean that we will never be able to understand the full complexity of fitness landscapes? Or perhaps big data from high-throughput measurements will guide us to find general rules? By coupling state-of-the-art experimental techniques with novel quantitative analysis of fitness landscapes, this work takes the optimistic view that we can push the boundary further and discover new mechanisms underlying evolution [9, 48, 49].

Figure Legends

Figure 1. Direct paths of adaptation are constrained by pairwise epistasis. (A) An example of subgraph that contains VDGV (wild type, WT), the quadruple mutant WLFA and all intermediates between them. Each variant in the subgraph is represented by a node. Edges are drawn between nearest neighbors. The arrows in bold represent the only accessible direct path of adaptation from VDGV to WLFA. HD: Hamming Distance. **(B)** We identified a total of 29 subgraphs in which the quadruple mutant was the only fitness peak. The number of accessible direct paths from WT to the quadruple mutant is shown for each subgraph. The maximum number of direct paths is 24. **(C)** The fraction of three types of pairwise epistasis around WT (2091 out of 2166) or randomly sampled from the entire sequence space (10^5 in total). Sign epistasis and reciprocal sign epistasis, both of which can block adaptive paths, are prevalent in the fitness landscape. Classification scheme of epistasis is shown at the top. Each node represents a genotype, which is within a sequence space of two loci and two alleles. Green arrows represent the accessible paths from genotype "00" to a beneficial double mutant "11" (colored in red). **(D)** The number of inaccessible direct paths are positively correlated (Pearson correlation = 0.66, p=1.0×10^{-4}) with the ruggedness induced by sign and reciprocal sign epistasis. The level of ruggedness is quantified as $f_{sign} + 2f_{reciprocal}$, where f_{type} denotes the fraction of each type of pairwise epistasis. The number inside a symbol indicates the number of subgraphs with identical properties.

Figure 2. Two distinct mechanisms of extra-dimensional bypass. (A) Extra amino acids at one of the two interacting sites may open up potential paths that circumvent the reciprocal sign epistasis. The starting point is 00 and the destination is 11 (in red). Green arrows indicate the accessible path. A successful bypass would require a "conversion" step that substitutes one of the two interacting sites with an extra amino acid ($00 \rightarrow 20$), followed by the loss of this mutation later ($21 \rightarrow 11$). The original reciprocal sign epistasis is changed to sign epistasis on the new genetic background after conversion. **(B)** Among ~20,000 randomly sampled reciprocal sign epistasis, >40% of them can be circumvented by at least one conversion bypass (i.e. success, inset). The number of available bypass for the success cases is shown as histogram. **(C)** The second mechanism of bypass involves an additional site. In this case, adaptation involves a "detour" step to gain mutation at the third site ($000 \rightarrow 100$), followed by the loss of this mutation ($111 \rightarrow 011$). The original reciprocal sign epistasis is changed to either magnitude epistasis or sign epistasis on the new genetic

background after detour (Supplementary Fig. 5). **(D)** In comparison to conversion bypass, detour bypass has a lower probability of success (<20%, inset) and is less prevalent.

Figure 3. Evidence of higher-order epistasis. **(A)** The fitness decomposition was performed on all subgraphs without missing variants. The fitness of variants can be reconstructed using Fourier coefficients truncated to a certain order. The Pearson correlation between the measured fitness and the fitness reconstructed by expansion of Fourier coefficients truncated to different orders (from 1^{st} to 4^{th}) is shown for each subgraph. The blue line corresponds to the median Pearson correlation. The top 0.1% subgraphs with fitness contributions from higher-order epistasis (the bottom 0.1% subgraphs ranked by Pearson correlation at 2^{nd} order expansion) are shown in red lines. **(B)** A sequence logo was generated for the quadruple mutants corresponding to the top 0.1% subgraphs with higher-order epistasis. The skewed composition of amino acids indicates that higher-order interactions are enriched among specific amino acid combinations of site 39, 41 and 54. **(C)** The magnitude of pairwise epistasis between G41L and V54H across different genetic backgrounds (i.e. all combinations of amino acids at site 39 and 40) is shown as a heat map. The amino acids of WT are boxed. Epistasis that cannot be determined due to missing variant is colored in grey. **(D)** Altering the genetic background at site 39 changed the positive epistasis ($\varepsilon > 0$) between G41L and V54H to negative epistasis ($\varepsilon < 0$). The fitness of each variant is indicated in the parentheses.

Figure 4. Indirect paths promote evolutionary accessibility. **(A)** 15 peaks had fitness larger than WT and their combined basins of attraction accounted for 99% of the entire sequence space. The size of each basin of attraction is identified by the Greedy Model (Methods). The area of each node is in proportion to the size of the basin of attraction of the corresponding fitness peak. An edge is drawn between fitness peaks that are separated by a Hamming distance of 2. **(B)** A possible adaptive path starting from WT (VDGV) to the fitness peak LYGV. **(C)** The frequency of different types of mutational step are shown. Three models, including the Greedy Model (green), Correlated Fixation Model (blue) and Equal Fixation Model (red), are used to simulate 1,000 adaptive paths starting from each variant in the sequence space. All the adaptive paths end at a fitness peak. **(D)** The distribution of the length of the adaptive path initiated at different starting points. For Correlated Fixation Model and Equal Fixation Model, the length was computed by averaging over 1,000 simulated paths for each starting point. The scale on the left is for Greedy Model. The scale on the right is for Correlated Fixation Model and Equal Fixation Model. **(E)** Indirect paths increased the

number of genotypes accessible to each fitness peak. The 15 peaks are ordered by increasing fitness (from left to right). **(F)** A large fraction of beneficial variants in the sequence space (fitness > 1) were accessible from WT only via indirect paths.

Supplementary Figure Legends

Supplementary Figure 1. The four-site sequence space of protein G. (A) The locations of sites 39, 40, 41, and 54 of protein GB1 are shown on the protein structure. PDB: 1PGA [50]. **(B)** The WT sequence of the nucleotide template [33]. T7 promoter is highlighted in magenta. Randomized sites (39, 40, 41, and 54) are highlighted in red. Poly-GS linkers are highlighted in green. FLAG-tag is highlighted in blue.

Supplementary Figure 2. Positive epistasis is enriched in the four-site sequence space. The distribution of pairwise epistasis measured by Olson et al. [33]. The pairwise epistatic values among sites 39, 40, 41, and 54 are ranked and represented by the red line. The pairwise epistatic values among other sites (all but 39, 40, 41, and 54) are ranked and represented by the black line.

Supplementary Figure 3. Workflow of mRNA display and data validation. (A) The workflow of mRNA display is shown. This is adapted from [33]. **(B)** The fitness values for all single substitution variants and double substitution variants in this study and in our previous study (based on an independently constructed library) [33] are compared. The high correlation (Pearson correlation=0.97) validates the fitness data obtained in this study.

Supplementary Figure 4. Subgraph analysis. We calculated the relative probabilities to realize each accessible path [20] (see Methods). In all the subgraphs analyzed, we found that most of the realizations were captured by a few accessible paths, as demonstrated by the skew in the cumulative probability of realization among different paths. **(A)** Cumulative probability of realization for mutational trajectories from WT (VDGV) to beneficial variants that have a Hamming distance (HD) of 4 from WT. This analysis only included those subgraphs with a reachable quadruple mutation variant (HD = 4 from WT) as the only fitness peak. Correlated Fixation Model is used. The diagonal line indicates the cumulative probability of a subgraph with equal probability of realization for all 24 possible trajectories. The bias of probability of realization in each subgraph was quantified using the Gini index (see Methods) and is shown as a histogram in the inset. **(B)** Same as panel A, except Equal Fixation Model is used instead. **(C and D)** Cumulative probability for mutational trajectories from a deleterious variants that have a Hamming distance (HD) of 4 from WT (VDGV) to WT. This analysis only included those subgraphs with WT being the only fitness

peak and the quadruple variant has a fitness between 0.01 to 1. The diagonal line indicates the cumulative probability of a subgraph with equal probability of realization for all 24 possible trajectories. The bias of probability of realization in each subgraph was quantified using the Gini index and is shown as a histogram in the inset. A total of 526 subgraphs were analyzed. **(C)** Correlated Fixation Model is used. **(D)** Equal Fixation Model is used. **(E)** Number of accessible trajectory from a deleterious variants that have a Hamming distance (HD) of 4 from WT (VDGV) to WT in each subgraph is shown as a barplot. The maximum possible number of accessible trajectory is 24. **(F)** The distribution of number of accessible trajectory is shown as a box plot. "Adaptation from WT" indicates those subgraphs based on the adaptation from WT to a beneficial variant that has a Hamming distance of 4 from WT. "Adaptation to WT" indicates those subgraphs based on the adaptation from a deleterious variant that has a Hamming distance of 4 from WT to WT.

Supplementary Figure 5. Three scenarios of extra-dimensional bypass via an extra site. **(A)** Reciprocal sign epistasis may be bypassed via the involvement of a third site. **(B-D)** There are three possible scenarios. It can be proven that higher-order epistasis is required for the scenarios in **(B)** and **(C)** (Supplementary Text).

Supplementary Figure 6. Fourier analysis decomposes the fitness landscape into epistatic interactions of different orders. Here we show two examples where the fitness contribution from higher-order epistasis is small (bottom 0.1%) in **(A)** and large (top 0.1%) in **(B)**. The fitness of variants can be reconstructed using Fourier coefficients truncated to a certain order. The Fourier coefficients can be interpreted as epistatic interactions of different orders, including the main effects of single mutations (the 1^{st} order), pairwise epistasis (the 2^{nd} order), and higher-order epistasis (the 3^{rd} and the 4^{th} order). **(C)** In our system with four sites, the reconstructed fitness by expansion to the 4^{th} order Fourier coefficients will always reproduce the measured fitness (i.e. Pearson correlation equals 1). If expansion to the 2^{nd} order Fourier coefficients did not reproduce the measured fitness (i.e. Pearson correlation less than 1), it would indicate the presence of higher-order epistasis.

Supplementary Figure 7. Relationship between fitness, size of the protein core, and predicted $\Delta\Delta$G. **(A and B)** The relationship between fitness and the total volume of residue 39, 41, and 54 is shown as a scatter plot for **(A)** all variants, and **(B)** variants with HD of 4 from WT. The blue line indicates the volume of WT (VDGV). The red line indicates the fraction of beneficial variants within a sliding window of ± 20

359 $Å^3$. We postulated that the observed higher-order epistasis was, at least partially, due to the steric effect
360 among site 39, 41, and 54. This was evidenced by the enrichment of beneficial variants when the total
361 volume of these three interacting residues was between $\sim$200 $Å^3$ and $\sim$300 $Å^3$. As the total volume further
362 increased, proportion of beneficial variants dropped. **(C and D)** The relationship between the predicted
363 $\Delta\Delta$G and the total size of residue 39, 41, and 54 for all variants is shown as a scatter plot for **(C)** all vari-
364 ants (Pearson's correlation = 0.41), and **(D)** variants with HD of 4 from WT (Pearson's correlation = 0.34).
365 The purple line represents the linear regression. The predicted $\Delta\Delta$G increased as the total volume of core
366 residues increased, indicating that the protein would be destabilized (i.e. decrease in fitness) when the core
367 was overpacked. Therefore, the higher-order epistasis observed in this study could be partially attributed to
368 the steric effect. Nonetheless, we acknowledged that entropic effect and conformational effect in IgG-FC
369 binding may also contribute to higher-order epistasis.
370
371 **Supplementary Figure 8. Alteration of pairwise epistatic effect under different genetic backgrounds.**
372 **(A)** Pairwise epistatic effect of each substitution pair under each genetic background (total possible genetic
373 backgrounds for each substitution pair = 20 x 20 = 400) was quantified. For each substitution pair, the range
374 of epistasis across different genetic backgrounds is shown in the top panel (brown). For each substitution
375 pair, the standard deviation of epistasis across different genetic backgrounds is shown in the middle panel
376 (green). For each substitution pair, the maximum epistatic value across different genetic backgrounds (ma-
377 genta), the minimum epistatic value across different genetic backgrounds (cyan), and the epistatic value
378 under WT background (grey) are shown. Substitution pair is ranked by the range of epistasis. **(B)** Epistasis
379 between G41F and V54A across different genetic backgrounds (different combination of amino acids in
380 sites 39 and 40) is shown. The epistasis value is color coded. Amino acids of WT are boxed. Epistasis that
381 cannot be determined due to missing variant is colored in grey. **(C)** Epistasis between V39W and V54H
382 across different genetic backgrounds (different combination of amino acids in sites 40 and 41) is shown.
383 The epistasis value is color coded. Amino acids of WT are boxed. Epistasis that cannot be determined due
384 to missing variant is colored in grey.
385
386 **Supplementary Figure 9. Higher-order epistasis can change the type of pairwise epistasis.** The type of
387 pairwise interaction could be changed in the presence of higher-order epistasis. **(A)** Reciprocal sign epistasis
388 between G41L-V54G is changed to magnitude epistasis given the mutation at site 39 (K39W). **(B)** Recipro-

cal sign epistasis between G41L-V54G is changed to sign epistasis given the mutation at site 39 (Q39Y).

Supplementary Figure 10. Lasso regression. Coefficients of the statistical model were fit by lasso regression on the measured fitness values of 119,884 non-lethal variants (see Methods). **(A)** 10-fold CV (cross-validation) MSE (mean squared errors) of lasso regression with varying penalty parameter λ. The black line indicates the 10-fold CV MSE of ordinary least squares regression (i.e. penalty parameter is zero). The red lines indicate the standard deviation. $\lambda = 10^{-4}$ is chosen for imputing the fitness values of missing variants. **(B)** The number of nonzero coefficients in the model with varying penalty parameter λ. **(C)** Comparison between the predicted fitness values and the measured fitness values (Pearson correlation=0.93).

Supplementary Figure 11. Indirect paths in adaptation. **(A)** A mutational trajectory initiated from PIWI under Greedy Model, which ended at the fitness peak, FWLG. **(B)** One of the shortest mutational trajectories from WT (VDGV) to a beneficial mutation (VHGL). **(C)** Histogram of the number of fitness accessible from a given genotype. The fraction of genotypes accessible to 15 fitness peaks increased substantially when indirect paths are allowed in adaptation.

Supplementary Figure 12. Delay of commitment in mutational trajectories involving extra-dimensional bypass. An entropy of evolutionary outcome was calculated for each of the 160,000 variants. Given a variant v with n accessible fitness peaks, the entropy of evolutionary outcome was then computed as follow:

$$\mathrm{E}ntropy_v = \sum_{i=1}^{n} -P_i \times \ln(P_i) \tag{1}$$

where P_i represented the frequency of reaching the fitness peak i among 1,000 simulated mutational trajectories from variant v following Correlated Fixation Model.

The entropy of evolutionary fates at each step along an adaptive path is shown. Adaptive paths with the same number of steps are grouped together. We observed that many mutational trajectories that involved extra-dimensional bypass did not fully commit to a fitness peak (entropy = 0) until the last two steps. Each grey line represents a mutational trajectory in each category. Only 100 randomly sampled trajectories are

416 shown due to the difficulty in visualizing a large number of lines on the graph. The median entropy at each

417 step in each category is represented by the red line.

Supplementary Text

Here we prove that higher-order epistasis is required for two possible scenarios of extra-dimensional bypass via an additional site (Supplementary Fig. 5). For a fitness landscape defined on a Boolean hypercube, we can expand the fitness as Taylor series [51].

$$\begin{aligned}
f_{000} &= \alpha_0 \\
f_{001} &= \alpha_0 + \alpha_1 \\
f_{010} &= \alpha_0 + \alpha_2 \\
f_{100} &= \alpha_0 + \alpha_3 \\
f_{011} &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_{12} \\
f_{101} &= \alpha_0 + \alpha_1 + \alpha_3 + \alpha_{13} \\
f_{110} &= \alpha_0 + \alpha_2 + \alpha_3 + \alpha_{23} \\
f_{111} &= \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3 + \alpha_{12} + \alpha_{13} + \alpha_{23} + \alpha_{123}
\end{aligned} \tag{2}$$

To prove that higher-order epistasis is present is equivalent to prove that $\alpha_{123} \neq 0$. The fitness difference between neighbors is visualized by the directed edges that go from low-fitness variant to high-fitness variant, thus each edge represents an inequality. No cyclic paths are allowed in this directed graph.

The reciprocal sign epistasis (Supplementary Fig. 5A) gives,

$$000 \leftarrow 001 : \alpha_1 < 0 \tag{3}$$

$$000 \leftarrow 010 : \alpha_2 < 0 \tag{4}$$

$$001 \rightarrow 011 : \alpha_2 + \alpha_{12} > 0 \tag{5}$$

$$010 \rightarrow 011 : \alpha_1 + \alpha_{12} > 0 \tag{6}$$

The detour step ($000 \rightarrow 100$) and the loss step ($111 \rightarrow 011$) are required for extra-dimensional bypass,

$$000 \rightarrow 100 : \alpha_3 > 0 \tag{7}$$

$$011 \leftarrow 111 : \alpha_3 + \alpha_{13} + \alpha_{23} + \alpha_{123} < 0 \tag{8}$$

For the remaining 6 edges, there are 3 possible configurations (Supplementary Fig. 5B-D). For the scenario illustrated in **(B)**, we have

$$100 \rightarrow 101 : \alpha_1 + \alpha_{13} > 0 \tag{9}$$

$$100 \rightarrow 110 : \alpha_2 + \alpha_{23} > 0 \tag{10}$$

Combining inequality (3) and (9) gives

$$\alpha_{13} > 0 \tag{11}$$

Combining inequality (4) and (10) gives

$$\alpha_{23} > 0 \tag{12}$$

Combining the above two inequalities with (7) and (8), we arrive at

$$\alpha_{123} < 0 \tag{13}$$

For the scenario in **(C)**, the proof of higher-order epistasis is similar. We have (the yellow edge)

$$001 \rightarrow 101 : \alpha_3 + \alpha_{13} > 0 \tag{14}$$

Combining the above inequality with (4), (8) and (10), we arrive at

$$\alpha_{123} < 0 \tag{15}$$

For the scenario in **(D)**, when $\alpha_3 + \alpha_{13} < 0$, all the inequalities can be satisfied with $\alpha_{123} = 0$. So higher-order epistasis is not necessary in this case.

Methods

Mutant library construction

Two oligonucleotides (Integrated DNA Technologies, Coralville, IA), 5'-AGT CTA GTA TCC AAC GGC NNS NNS NNK GAA TGG ACC TAC GAC GAC GCT ACC AAA ACC TT-3' and 5'-TTG TAA TCG GAT CCT CCG GAT TCG GTM NNC GTG AAG GTT TTG GTA GCG TCG TCG T-3' were annealed by heating to 95^{o}C for 5 minutes and cooling to room temperature over 1 hour. The annealed nucleotide was extended in a reaction containing 0.5 uM of each oligonucleotide, 50 mM NaCl, 10 mM Tris-HCl pH 7.9, 10 mM $MgCl_2$, 1 mM DTT, 250 uM each dNTP, and 50 units Klenow exo- (New England Biolabs, Ipswich, MA) for 30 mins at 37^{o}C. The product (cassette I) was purified by PureLink PCR Purification Kit (Life Technologies, Carlsbad, CA) according to manufacturer's instructions.

A constant region was generated by PCR amplification using KOD DNA polymerase (EMD Millipore, Billerica, MA) with 1.5 mM $MgSO_4$, 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dTTP), 0.05 ng protein GB1 wild type (WT) template, and 0.5 uM each of 5'-TTC TAA TAC GAC TCA CTA TAG GGA CAA TTA CTA TTT ACA TAT CCA CCA TG-3' and 5'-AGT CTA GTA TCC TCG ACG CCG TTG TCG TTA GCG TAC TGC-3'. The sequence of the WT template consisted of a T7 promoter, 5' UTR, the coding sequence of Protein GB1, 3' poly-GS linkers, and a FLAG-tag (Supplementary Fig. 1B) [33]. The thermocycler was set as follows: 2 minutes at 95^{o}C, then 18 three-step cycles of 20 seconds at 95^{o}C, 15 seconds at 58^{o}C, and 20 seconds at 68^{o}C, and 1 minute final extension at 68^{o}C. The product (constant region) was purified by PureLink PCR Purification Kit (Life Technologies) according to manufacturer's instructions. Both the purified constant region and cassette I were digested with BciVI (New England Biolabs) and purified by PureLink PCR Purification Kit (Life Technologies) according to manufacturer's instructions.

Ligation between the constant region and cassette I (molar ratio of 1:1) was performed using T4 DNA ligase (New England Biolabs). Agarose gel electrophoresis was performed to separate the ligated product from the reactants. The ligated product was purified from the agarose gel using Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA) according to manufacturer's instructions. PCR amplification was then performed using KOD DNA polymerase (EMD Millipore) with 1.5 mM $MgSO_4$, 0.2 mM of each dNTP

(dATP, dCTP, dGTP, and dTTP), 4 ng of the ligated product, and 0.5 uM each of 5'-TTC TAA TAC GAC TCA CTA TAG GGA CAA TTA CTA TTT ACA TAT CCA CCA TG-3' and 5'-GGA GCC GCT ACC CTT ATC GTC GTC ATC CTT GTA ATC GGA TCC TCC GGA TTC-3'. The thermocycler was set as follows: 2 minutes at 95^{o}C, then 10 three-step cycles of 20 seconds at 95^{o}C, 15 seconds at 56^{o}C, and 20 seconds at 68^{o}C, and 1 minute final extension at 68^{o}C. The product, which is referred as "DNA library", was purified by PureLink PCR Purification Kit (Life Technologies) according to manufacturer's instructions.

Affinity selection by mRNA display

Affinity selection by mRNA display [34, 35] was performed as described (Supplementary Fig. 3A) [33]. Briefly, The DNA library was transcribed by T7 RNA polymerase (Life Technologies) according to manufacturer's instructions. Ligation was performed using 1 nmol of mRNA, 1.1 nmol of 5'-TTT TTT TTT TTT GGA GCC GCT ACC-3', and 1.2 nmol of 5-/5Phos/-d(A)21-(9)3-ACC-Puromycin by T4 DNA ligase (New England Biolabs) in a 100 uL reaction. The ligated product was purified by urea PAGE and translated in a 100 uL reaction volume using Retic Lysate IVT Kit (Life Technologies) according to manufacturer's instructions followed by incubation with 500 mM final concentration of KCl and 60 mM final concentration of $MgCl_2$ for at least 30 minutes at room temperature to increase the efficiency for fusion formation [52]. The mRNA-protein fusion was then purified using ANTI-FLAG M2 Affinity Gel (Sigma-Aldrich, St. Louis, MO). Elution was performed using 3X FLAG peptide (Sigma-Aldrich). The purified mRNA-protein fusion was reverse transcribed using SuperScript III Reverse Transcriptase (Life Technologies). This reverse transcribed product, which was referred as "input library", was incubated with Pierce streptavidin agarose (SA) beads (Life Technologies) that were conjugated with biotinylated human IgG-FC (Rockland Immunochemicals, Limerick, PA). After washing, the immobilized mRNA-protein fusion was eluted by heating to 95^{o}C. The eluted sample was referred as "selected library".

Sequencing library preparation

PCR amplification was performed using KOD DNA polymerase (EMD Millipore) with 1.5 mM $MgSO_4$, 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dTTP), the selected library, and 0.5 uM each of 5'-CTA CAC GAC GCT CTT CCG ATC TNN NAG CAG TAC GCT AAC GAC AAC G-3' and 5'-TGC TGA ACC GCT

CTT CCG ATC TNN NTA ATC GGA TCC TCC GGA TTC G-3'. The underlined "NNN" indicated the position of the multiplex identifier, GTG for input library and TGT for post-selection library. The thermocycler was set as follows: 2 minutes at 95^{o}C, then 10 to 12 three-step cycles of 20 seconds at 95^{o}C, 15 seconds at 56^{o}C, and 20 seconds at 68^{o}C, and 1 minute final extension at 68^{o}C. The product was then PCR amplified again using KOD DNA polymerase (EMD Millipore) with 1.5 mM $MgSO_4$, 0.2 mM of each dNTP (dATP, dCTP, dGTP, and dTTP), the eluted product from mRNA display, and 0.5 uM each of 5'-AAT GAT ACG GCG ACC ACC GAG ATC TA CAC TCT TTC CCT ACA CGA CGC TCT TCC G-3' and 5'-CAA GCA GAA GAC GGC ATA CGA GAT CGG TCT CGG CAT TCC TGC TGA ACC GCT CTT CCG-3'. The thermocycler was set as follows: 2 minutes at 95^{o}C, then 10 to 12 three-step cycles of 20 seconds at 95^{o}C, 15 seconds at 56^{o}C, and 20 seconds at 68^{o}C, and 1 minute final extension at 68^{o}C. The PCR product was then subjected to 2 x 100 bp paired-end sequencing using Illumina HiSeq 2500 platform. Raw sequencing data have been submitted to the NIH Short Read Archive under accession number: BioProject PRJNA278685.

We were able to compute the fitness for 93.4% of all variants from the sequencing data. The fitness measurements in this study were highly consistent with our previous study on fitness of single and double mutants in protein GB1 (Pearson correlation = 0.97, Supplementary Fig. 3B) [33].

Sequencing data analysis

The first three nucleotides of both forward read and reverse read were used for demultiplexing. If the first three nucleotides of the forward read were different from that of the reverse read, the given paired-end read would be discarded. For both forward read and reverse read, the nucleotides that were corresponding to the codons of protein GB1 sites 39, 40, 41, and 54 were extracted. If coding sequence of sites 39, 40, 41, and 54 in the forward read and that in the reverse read did not reverse-complement each other, the paired-end read would be discarded. Subsequently, the occurrence of individual variants at the amino acid level for site 39, 40, 41, and 54 in both input library and selected library were counted, with each paired-end read represented 1 count. Custom python scripts and bash scripts were used for sequencing data processing. All scripts are available upon request.

Calculation of fitness

The fitness (w) for a given variant *i* was computed as:

$$w_i = \frac{count_{i,selected}/count_{i,input}}{count_{WT,selected}/count_{WT,input}} \tag{16}$$

where $\text{count}_{i,selected}$ represented the count of variant *i* in the selected library, $\text{count}_{i,input}$ represented the count of variant *i* in the input library, $\text{count}_{WT,selected}$ represented the count of WT (VDGV) in the selected library, and $\text{count}_{WT,input}$ represented the count of WT (VDGV) in the input library.

Therefore, the fitness of each variant, w_i, could be viewed as the fitness relative to WT (VDGV), such that $w_{WT} = 1$. Variants with $\text{count}_{input} < 10$ were filtered to reduce noise. The fraction of all possible variants that passed this filter was 93.4% (149,361 out of 160,000 all possible variants).

The fitness of each single substitution variant was referenced to our previous study [33], because the sequencing coverage of single substitution variants in our previous study was much higher than in this study ($\sim$100 fold higher). Hence, our confidence in computing fitness for a single substitution variant should also be much higher in our previous study than this study. Subsequently, the fitness of each single substitution in this study was calculated by multiplying a factor of 1.159 by the fitness of that single substitution computed from our previous study [33]. This is based on the linear regression analysis between the single substitution fitness as measured in our previous study and in this study, which had a slope of 1.159 and a y-intercept of $\sim$0.

Magnitude and type of pairwise epistasis

The three types of pairwise epistasis (magnitude, sign and reciprocal sign) were classified by ranking the fitness of the four variants involved [53].

To quantify the magnitude of epistasis (ε) between substitutions *a* and *b* on a given background variant *BG*, the relative epistasis model [39] was employed as follows:

$$\varepsilon_{ab,BG} = \ln(\frac{w_{ab}}{w_{BG}}) - \ln(\frac{w_a}{w_{BG}}) - \ln(\frac{w_b}{w_{BG}}) \quad (17)$$

where w_{ab} represents the fitness of the double substitution, $\ln(w_a)$ and $\ln(w_b)$ represents the fitness of each of the single substitution respectively, and w_{BG} represents the fitness of the background variant.

As described previously [33], there exists a limitation in determining the exact fitness for very low-fitness variants in this system. To account for this limitation, several rules were adapted from our previous study to minimize potential artifacts in determining ε [33]. We previously determined that the detection limit of fitness (w) in this system is $\sim$0.01 [33].

Rule 1) if $\max(\frac{w_{ab}}{w_{BG}}, \frac{w_a}{w_{BG}}, \frac{w_b}{w_{BG}}) < 0.01$, $\varepsilon_{ab,BG,adjusted} = 0$

Rule 2) if $\min(w_a, w_b, \frac{w_a}{w_{BG}}, \frac{w_b}{w_{BG}}) < 0.01$, $\varepsilon_{ab,BG,adjusted} = \max(0, \varepsilon_{ab,BG})$

Rule 3) if $\min(w_{ab}, \frac{w_{ab}}{w_{BG}}) < 0.01$, $\varepsilon_{ab,BG,adjusted} = \min(0, \varepsilon_{ab,BG})$

Rule 1 prevents epistasis being artifically estimated from low-fitness variants. Rule 2 prevents overestimation of epistasis due to low fitness of one of the two single substitutions. Rule 3 prevents underestimation of epistasis due to low fitness of the double substitution. To compute the epistasis between two substitutions, *a* and *b*, on a given background variant *BG*, $\varepsilon_{ab,BG,adjusted}$ would be used if one of the above three rules was satisfied. Otherwise, $\varepsilon_{ab,BG}$ would be used.

Fourier analysis

Fitness decomposition was performed on all subgraphs without missing variants (109,235 subgraphs in total). We decomposed the fitness landscape into epistatic interactions of different orders by Fourier analysis [9, 54]. The Fourier coefficients given by the transform can be interpreted as epistasis of different orders [6, 30].

For a binary sequence $\vec{z}$ with dimension L (z_i equals 1 if mutation is present at position i, or 0 otherwise), the Fourier decomposition theorem states that the fitness function $f(\vec{z})$ can be expressed as [51]:

$$f(\vec{z}) = \sum_{\vec{k}} \hat{f}_{\vec{k}} (-1)^{\vec{z}\cdot\vec{k}} \tag{18}$$

The formula for the Fourier coefficients $\hat{f}_{\vec{k}}$ is then:

$$\hat{f}_{\vec{k}} = \frac{1}{2^L} \sum_{\vec{z}} f(\vec{z}) (-1)^{\vec{z}\cdot\vec{k}} \tag{19}$$

For example, we can expand the fitness landscape up to the second order, i.e. with linear and quadratic terms

$$f(\vec{\sigma}) = \hat{f}_0 + \sum_i \hat{f}_{\vec{e_i}} \sigma_i + \sum_{i<j} \hat{f}_{\vec{e_i}+\vec{e_j}} \sigma_i \sigma_j + \cdots \tag{20}$$

where $\sigma_i \equiv (-1)^{z_i} \in \{+1, -1\}$, and $\vec{e_i}$ is a unit vector along the i^{th} direction. In our analysis of subgraphs, there are a total of $2^4 = 16$ terms in the Fourier decomposition, with $\binom{4}{i}$ terms for the i^{th} order ($i = 0, 1, 2, 3, 4$). We can expand the fitness landscape up to a given order by ignoring all higher-order terms in Equation 18. In this paper, we refer to higher-order epistasis as non-zero contribution to fitness from the 3^{rd} order terms and beyond.

Imputing the fitness of missing variants

The fitness values for 10,639 variants (6.6% of the entire sequence space) were not directly measured (read count in the input pool = 0) or were filtered out because of low read counts in the input pool (see section "Calculation of fitness"). To impute the fitness of these missing variants, we performed regularized regression on fitness values of observed variants using the following model [40, 55]:

$$log(f) = \alpha_0 + \sum_{i=1}^{N_M} \beta_i M_i + \sum_{j=1}^{N_P} \gamma_j P_j + \sum_{k=1}^{N_T} \delta_k T_k \tag{21}$$

Here, f is the protein fitness. α_0 is the intercept that represents the log fitness of WT; β_i represents the main effect of a single mutation, i; M_i is a dummy variable that equals 1 if the single mutation i is present in the sequence, or 0 if the single mutation is absent; and $N_M = 19 \times \binom{4}{1} = 76$ is the total number

of single mutations. Similarly, γ_j represents the effect of interaction between a pair of mutations; P_j is the dummy variable that equals either 1 or 0 depending on the presence of that those two mutations; and $N_P = 19^2 \times \binom{4}{2} = 2166$ is the total number of possible pairwise interactions. In addition to the main effects of single mutations and pairwise interactions, the three-way interactions among sites 39, 41 and 54 are included in the model, based on our knowledge of higher-order epistasis (Fig. 3). δ_k represents the effect of three-way interactions among sites 39, 41 and 54; T_k is the dummy variable that equals either 1 or 0 depending on the presence of that three-way interaction; and $N_T = 19^3 = 6859$ is the total number of three-way interactions. Thus, the total number of coefficients in this model is 9,102, including main effects of each site (i.e. additive effects), interactions between pairs of sites (i.e. pairwise epistasis), and a subset of three-way interactions (i.e. higher-order epistasis).

Out of the 149,361 variants with experimentally measured fitness values, 119,884 variants were non-lethal ($f > 0$) and were used to fit the model coefficients using lasso regression (Matlab R2014b). Lasso regression adds a penalty term $\lambda \sum |\theta|$ (θ stands for any coefficient in the model) when minimizing the least squares, thus it favors sparse solutions of coefficients (Supplementary Fig. 10B). We calculated the 10-fold cross-validation MSE (mean squared errors) of the lasso regression for a wide range of penalty parameter λ (Supplementary Fig. 10A). $\lambda = 10^{-4}$ is chosen. For measured variants, the model-predicted fitness values were highly correlated with the actual fitness values (Pearson correlation=0.93, Supplementary Fig. 10C). We then used the fitted model to impute the fitness of the 10,639 missing variants and complete the entire fitness landscape.

Simulating adaptation using three models for fixation

Python package "networkx" was employed to construct a directed graph that represented the entire fitness landscape for sites 39, 40, 41, and 54. A total of 4^{20} = 160,000 nodes were present in the directed graph, where each node represented a 4-site variant. For all pairs of variants separated by a Hamming distance of 1, a directed edge was generated from the variant with a lower fitness to the variant with a higher fitness. Therefore, all successors of a given node had a higher fitness than the given node. A fitness peak was defined as a node that had 0 out-degree. Three models, namely the Greedy Model [6], the Correlated Fixation Model [41], and the Equal Fixation Model [20], were employed in this study to simulate the mutational steps in

adaptive trajectories. The Greedy Model represents adaptive evolution of a large population with pervasive clonal interference [6]. The Correlated Fixation Model represents adaptive evolution of a population under the scheme of strong-selection/weak-mutation (SSWM), which assumes that the time to fixation is much shorter than the time between mutations, and the fixation probability of a given mutation is proportional to the improvement in fitness. The Equal Fixation Model represents a simplified scenario of adaptation where all beneficial mutations fix with equal probability [20]. Under all three models, the probability of fixation of a deleterious or neutral mutation is 0. Considering a mutational trajectory initiated at a node, n_i with a fitness value of w_i, where n_i has M successors, (n_1, n_2, ... n_M) with fitness values of (w_1, w_2, ... w_M). Then the probability that the next mutational step is from n_i to n_k, where $k \in$ (1, 2, ... M), is denoted $P_{i \to k}$ and called the probability of fixation, and can be computed for each model as follows.

For the Greedy Model (deterministic model),

$$\text{if } w_k = max(w_1, w_2, ...w_M),\ P_{i \to k} = 1 \tag{22}$$

$$\text{otherwise, } P_{i \to k} = 0 \tag{23}$$

For the Correlated Fixation Model (non-deterministic model),

$$P_{i \to k} = \frac{w_k - w_i}{\sum_{n=1}^{M} (w_n - w_i)} \tag{24}$$

For the Equal Fixation Model (non-deterministic model),

$$P_{i \to k} = \frac{1}{M} \tag{25}$$

To compute the shortest path from a given variant to all reachable variants, the function "single_source_shortest_path" in "networkx" was used. If the shortest path between a low-fitness variant and a high-fitness variant does not exist, it means that the high-fitness variant is inaccessible. If the shortest path is longer than the Hamming

Distance between two variants, it means that adaptation requires indirect paths.

Analysis of direct paths within a subgraph

In the subgraph analysis shown in Supplementary Fig. 4, the fitness landscape was restricted to 2 amino acids at each of the 4 sites (the WT and adapted alleles). There was a total of 2^4 variants, hence nodes, in a given subgraph. Only those subgraphs where the fitness of all variants was measured directly were used (i.e. any subgraph with missing variants was excluded from this analysis). Mutational trajectories were generated in the same manner as in the analysis of the entire fitness landscape (see subsection "Simulating adaptation using three models for fixation"). In a subgraph with only one fitness peak, the probability of a mutational trajectory from node *i* to node *j* via intermediate *a*, *b*, and *c* was as follows:

$$P_{i\to a\to b\to c\to j} = P_{i\to a} \times P_{a\to b} \times P_{b\to c} \times P_{c\to j} \tag{26}$$

To compute the Gini index for a given set of mutational trajectories from node i to node j, the probabilities of all possible mutational trajectories were sorted from large to small. Inaccessible trajectories were also included in this sorted list with a probability of 0. This sorted list with t trajectories was denoted as ($P_{i\to j,1}$, $P_{i\to j,2}$, ... $P_{i\to j,t}$), where $P_{i\to j,1}$ was the largest and $P_{i\to j,t}$ was the smallest. This sorted list was converted into a list of cumulative probabilities, which is denoted as ($A_{i\to j,1}$, $A_{i\to j,2}$, ... $A_{i\to j,t}$), where $A_{i\to j,t} = \sum_{n=1}^{t} P_{i\to j,t}$.

The Gini index for the given subgraph was then computed as follows:

$$\text{Gini index} = \frac{2 \times \sum_{n=1}^{t-1} (A_{i\to j,n}) + A_{i\to j,t} - t}{t-1} \tag{27}$$

Visualization

Sequence logo was generated by WebLogo (http://weblogo.berkeley.edu/logo.cgi) [56].

The visualization of basins of attraction (Fig. 4A) was generated using Graphviz with "fdp" as the option for layout.

$\Delta\Delta$G prediction

The $\Delta\Delta$G prediction was performed by the ddg_monomer application in Rosetta software [57] with the parameters from row 16 of Table I in Kellogg et al. were used [58].

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

We would like to thank Jesse Bloom and Joshua Plotkin for helpful comments on early versions of the manuscript. N.C.W. was supported by Philip Whitcome Pre-Doctoral Fellowship, Audree Fowler Fellowship in Protein Science, and UCLA Dissertation Year Fellowship. L.D. was supported by HHMI Postdoctoral Fellowship from Jane Coffin Childs Memorial Fund for Medical Research. R.S. was supported by NIH R01 DE023591. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Contributions

N.C.W., C.A.O., and R.S. designed the experiment, N.C.W. and C.A.O. performed the experiments, N.C.W. processed the sequencing data, L.D. and N.C.W. analyzed the fitness landscape, J.O.L.S. provided important intellectual inputs, L.D. and N.C.W. wrote the manuscript, J.O.L.S. and R.S. revised the manuscript.

References

[1] Kauffman, S. and Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* **128**(1), 11–45, Sep (1987).

[2] Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M., and Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**(7126), 383–386 (2007).

[3] Romero, P. A. and Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* **10**(12), 866–876, Dec (2009).

[4] Hartl, D. L. What can we learn from fitness landscapes? *Current opinion in microbiology* **21**, 51–57 (2014).

[5] Kondrashov, D. A. and Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends in Genetics* **31**(1), 24–33 (2015).

[6] de Visser, J. A. G. M. and Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* **15**(7), 480–490, Jul (2014).

[7] Kouyos, R. D., Leventhal, G. E., Hinkley, T., Haddad, M., Whitcomb, J. M., Petropoulos, C. J., and Bonhoeffer, S. Exploring the complexity of the hiv-1 fitness landscape. *PLoS Genet* **8**(3), e1002551 (2012).

[8] Barton, J. P., Kardar, M., and Chakraborty, A. K. Scaling laws describe memories of host-pathogen riposte in the hiv population. *Proc Natl Acad Sci U S A* **112**(7), 1965–1970, Feb (2015).

[9] Szendro, I. G., Schenk, M. F., Franke, J., Krug, J., and De Visser, J. A. G. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment* **2013**(01), P01005 (2013).

[10] Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., and Fields, S. High-resolution mapping of protein sequence-function relationships. *Nat Methods* **7**(9), 741–746, Sep (2010).

[11] Hietpas, R. T., Jensen, J. D., and Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc Natl Acad Sci U S A* **108**(19), 7896–7901, May (2011).

[12] Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., Gros, P.-A., and Tenaillon, O. Capturing the mutational landscape of the beta-lactamase tem-1. *Proc Natl Acad Sci U S A* **110**(32), 13067–13072, Aug (2013).

[13] Wu, N. C., Young, A. P., Al-Mawsawi, L. Q., Olson, C. A., Feng, J., Qi, H., Chen, S.-H., Lu, I.-H., Lin, C.-Y., Chin, R. G., Luan, H. H., Nguyen, N., Nelson, S. F., Li, X., Wu, T.-T., and Sun, R. High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Sci Rep* **4**, 4942 (2014).

[14] Thyagarajan, B. and Bloom, J. D. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife* **3** (2014).

[15] Qi, H., Olson, C. A., Wu, N. C., Ke, R., Loverdo, C., Chu, V., Truong, S., Remenyi, R., Chen, Z., Du, Y., et al. A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis c viral fitness and drug sensitivity. *PLoS Pathog* **10**(4), e1004064 (2014).

[16] Stiffler, M. A., Hekstra, D. R., and Ranganathan, R. Evolvability as a function of purifying selection in tem-1 β-lactamase. *Cell* **160**(5), 882–892, Feb (2015).

[17] Jiménez, J. I., Xulvi-Brunet, R., Campbell, G. W., Turk-MacLeod, R., and Chen, I. A. Comprehensive experimental fitness landscape and evolutionary network for small rna. *Proc Natl Acad Sci U S A* **110**(37), 14984–14989, Sep (2013).

[18] Pitt, J. N. and Ferré-D'Amaré, A. R. Rapid construction of empirical rna fitness landscapes. *Science* **330**(6002), 376–379, Oct (2010).

[19] Payne, J. L. and Wagner, A. The robustness and evolvability of transcription factor binding sites. *Science* **343**(6173), 875–877 (2014).

[20] Weinreich, D. M., Delaney, N. F., Depristo, M. A., and Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**(5770), 111–114, Apr (2006).

728 [21] Lunzer, M., Miller, S. P., Felsheim, R., and Dean, A. M. The biochemical architecture of an ancient
729 adaptive landscape. *Science* **310**(5747), 499–501, Oct (2005).

730 [22] O'Maille, P. E., Malone, A., Dellas, N., Andes Hess, Jr, B., Smentek, L., Sheehan, I., Greenhagen,
731 B. T., Chappell, J., Manning, G., and Noel, J. P. Quantitative exploration of the catalytic landscape
732 separating divergent plant sesquiterpene synthases. *Nat Chem Biol* **4**(10), 617–623, Oct (2008).

733 [23] Lozovsky, E. R., Chookajorn, T., Brown, K. M., Imwong, M., Shaw, P. J., Kamchonwongpaisan, S.,
734 Neafsey, D. E., Weinreich, D. M., and Hartl, D. L. Stepwise acquisition of pyrimethamine resistance
735 in the malaria parasite. *Proc Natl Acad Sci U S A* **106**(29), 12025–12030, Jul (2009).

736 [24] Franke, J., Klözer, A., de Visser, J. A. G. M., and Krug, J. Evolutionary accessibility of mutational
737 pathways. *PLoS Comput Biol* **7**(8), e1002134, Aug (2011).

738 [25] Tan, L., Serene, S., Chao, H. X., and Gore, J. Hidden randomness between fitness landscapes limits
739 reverse evolution. *Physical review letters* **106**(19), 198102 (2011).

740 [26] Cariani, P. A. Extradimensional bypass. *Biosystems* **64**(1-3), 47–53, Jan (2002).

741 [27] DePristo, M. A., Hartl, D. L., and Weinreich, D. M. Mutational reversions during adaptive protein
742 evolution. *Molecular biology and evolution* **24**(8), 1608–1610 (2007).

743 [28] Palmer, A. C., Toprak, E., Baym, M., Kim, S., Veres, A., Bershtein, S., and Kishony, R. Delayed
744 commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nat Commun* **6**, 7385
745 (2015).

746 [29] Kvitek, D. J. and Sherlock, G. Reciprocal sign epistasis between frequently experimentally evolved
747 adaptive mutations causes a rugged fitness landscape. *PLoS Genet* **7**(4), e1002056, Apr (2011).

748 [30] Weinreich, D. M., Lan, Y., Wylie, C. S., and Heckendorn, R. B. Should evolutionary geneticists worry
749 about higher-order epistasis? *Curr Opin Genet Dev* **23**(6), 700–707, Dec (2013).

750 [31] Sjöbring, U., Björck, L., and Kastern, W. Streptococcal protein g. gene structure and protein binding
751 properties. *Journal of Biological Chemistry* **266**(1), 399–405 (1991).

752 [32] Sauer-Eriksson, A. E., Kleywegt, G. J., Uhlén, M., and Jones, T. A. Crystal structure of the c2 fragment
753 of streptococcal protein g in complex with the fc domain of human igg. *Structure* **3**(3), 265–278 (1995).

754 [33] Olson, C. A., Wu, N. C., and Sun, R. A comprehensive biophysical description of pairwise epistasis
755 throughout an entire protein domain. *Curr Biol* **24**(22), 2643–2651, Nov (2014).

756 [34] Roberts, R. W. and Szostak, J. W. Rna-peptide fusions for the in vitro selection of peptides and proteins.
757 *Proc Natl Acad Sci U S A* **94**(23), 12297–12302, Nov (1997).

758 [35] Olson, C. A., Nie, J., Diep, J., Al-Shyoukh, I., Takahashi, T. T., Al-Mawsawi, L. Q., Bolin, J. M.,
759 Elwell, A. L., Swanson, S., Stewart, R., Thomson, J. A., Soh, H. T., Roberts, R. W., and Sun, R.
760 Single-round, multiplexed antibody mimetic design through mrna display. *Angew Chem Int Ed Engl*
761 **51**(50), 12449–12453, Dec (2012).

762 [36] Tufts, D. M., Natarajan, C., Revsbech, I. G., Projecto-Garcia, J., Hoffmann, F. G., Weber, R. E., Fago,
763 A., Moriyama, H., and Storz, J. F. Epistasis constrains mutational pathways of hemoglobin adaptation
764 in high-altitude pikas. *Mol Biol Evol* **32**(2), 287–298, Feb (2015).

765 [37] Weinreich, D. M., Watson, R. A., and Chao, L. Perspective: Sign epistasis and genetic constraint on
766 evolutionary trajectories. *Evolution* **59**(6), 1165–1174, Jun (2005).

767 [38] Gutiérrez, J. and Maere, S. Modeling the evolution of molecular systems from a mechanistic perspec-
768 tive. *Trends Plant Sci* **19**(5), 292–303, May (2014).

769 [39] Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E., and Cooper, T. F. Negative epistasis between
770 beneficial mutations in an evolving bacterial population. *Science* **332**(6034), 1193–1196, Jun (2011).

771 [40] Otwinowski, J. and Plotkin, J. B. Inferring fitness landscapes by regression produces biased estimates
772 of epistasis. *Proc Natl Acad Sci U S A* **111**(22), E2301–E2309, Jun (2014).

773 [41] Gillespie, J. H. Molecular evolution over the mutational landscape. *Evolution* **38**(5), 1116–1129
774 (1984).

775 [42] Weissman, D. B., Desai, M. M., Fisher, D. S., and Feldman, M. W. The rate at which asexual popula-
776 tions cross fitness valleys. *Theor Popul Biol* **75**(4), 286–300, Jun (2009).

777 [43] Ostman, B., Hintze, A., and Adami, C. Impact of epistasis and pleiotropy on evolutionary adaptation.
778 *Proc Biol Sci* **279**(1727), 247–256, Jan (2012).

779 [44] Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225**, 563–564 (1970).

780 [45] Wang, Y., Arenas, C. D., Stoebel, D. M., and Cooper, T. F. Genetic background affects epistatic
781 interactions between two beneficial mutations. *Biol Lett* **9**(1), 20120328, Feb (2013).

782 [46] Pettersson, M., Besnier, F., Siegel, P. B., and Carlborg, O. Replication and explorations of high-order
783 epistasis using a large advanced intercross line pedigree. *PLoS Genet* **7**(7), e1002180, Jul (2011).

784 [47] Gavrilets, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol Evol* **12**(8), 307–312,
785 Aug (1997).

786 [48] Fisher, D., Lässig, M., and Shraiman, B. Evolutionary dynamics and statistical physics. *Journal of*
787 *Statistical Mechanics: Theory and Experiment* **2013**(01), N01001 (2013).

788 [49] Desai, M. M. Statistical questions in experimental evolution. *Journal of Statistical Mechanics: Theory*
789 *and Experiment* **2013**(01), P01003 (2013).

790 [50] Gallagher, T., Alexander, P., Bryan, P., and Gilliland, G. L. Two crystal structures of the b1
791 immunoglobulin-binding domain of streptococcal protein g and comparison with nmr. *Biochemistry*
792 **33**(15), 4721–4729, Apr (1994).

793 [51] Weinberger, E. D. Fourier and taylor series on fitness landscapes. *Biological Cybernetics* **65**(5), 321–
794 330 (1991).

795 [52] Liu, R., Barrick, J. E., Szostak, J. W., and Roberts, R. W. Optimized synthesis of rna-protein fusions
796 for in vitro protein selection. *Methods Enzymol* **318**, 268–293 (2000).

797 [53] Greene, D. and Crona, K. The changing geometry of a fitness landscape along an adaptive walk. *PLoS*
798 *Comput Biol* **10**(5), e1003520, May (2014).

799 [54] Stadler, P. F. Landscapes and their correlation functions. *Journal of Mathematical chemistry* **20**(1),
800 1–45 (1996).

801 [55] Hinkley, T., Martins, J., Chappey, C., Haddad, M., Stawiski, E., Whitcomb, J. M., Petropoulos, C. J.,
802 and Bonhoeffer, S. A systems analysis of mutational effects in hiv-1 protease and reverse transcriptase.
803 *Nat Genet* **43**(5), 487–489, May (2011).

804 [56] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. Weblogo: a sequence logo generator.
805 *Genome Res* **14**(6), 1188–1190, Jun (2004).

[57] Das, R. and Baker, D. Macromolecular modeling with rosetta. *Annu Rev Biochem* **77**, 363–382 (2008).

[58] Kellogg, E. H., Leaver-Fay, A., and Baker, D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**(3), 830–838, Mar (2011).

Figure 1

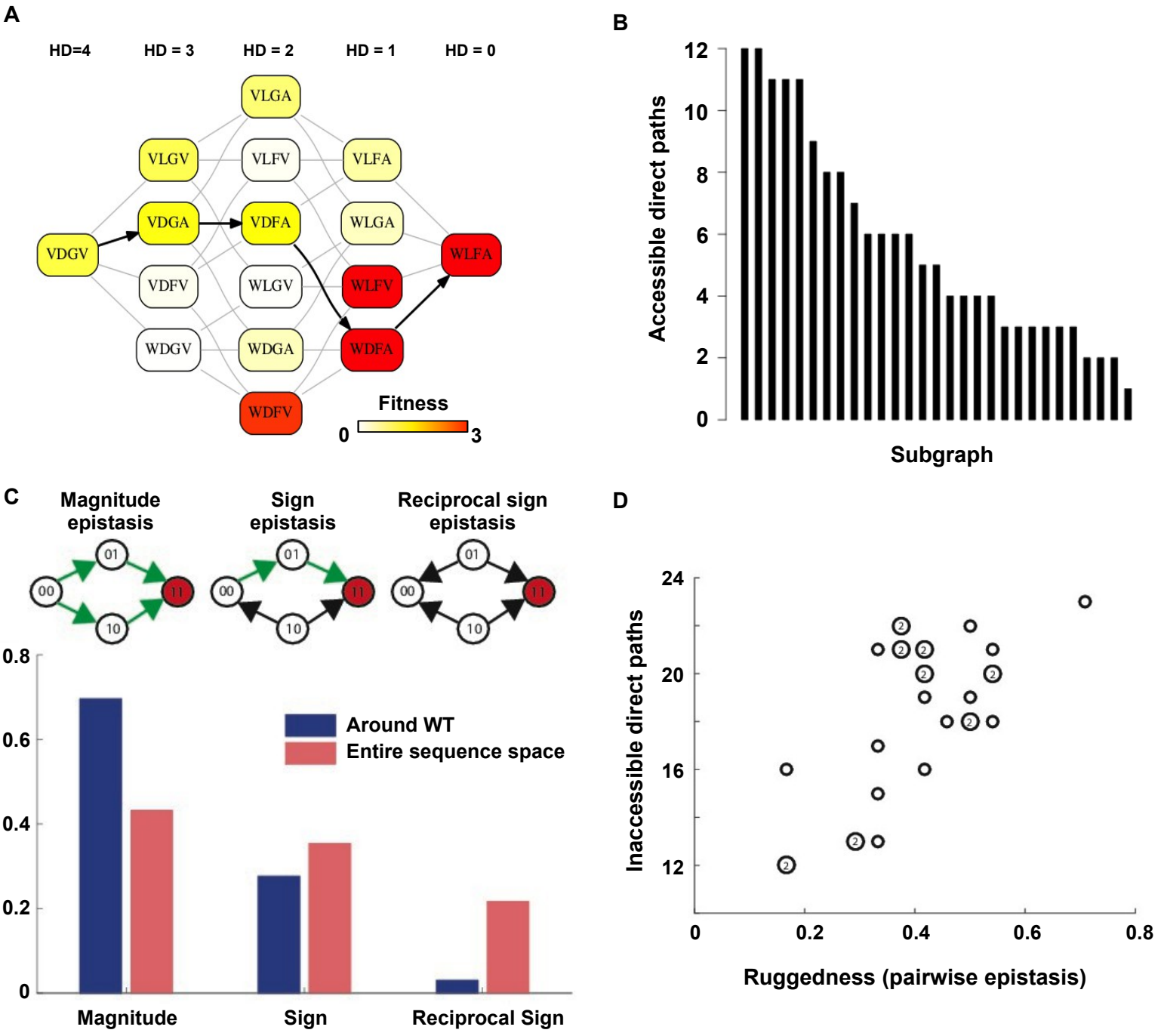

A
HD=4
HD = 3
HD = 2
HD = 1
HD = 0
VLGA
VLGV
VLFV
VLFA
VDGA
VDFA
WLGA
VDGV
WLFA
VDFV
WLGV
WLFV
WDGV
WDGA
WDFA
WDFV
Fitness
0
3
B
Accessible direct paths
12
10
8
6
4
2
0
Subgraph
C
Magnitude epistasis
Sign epistasis
Reciprocal sign epistasis
00
01
10
11
0.8
0.6
0.4
0.2
0
Around WT
Entire sequence space
Magnitude
Sign
Reciprocal Sign
D
Inaccessible direct paths
24
20
16
12
0
0.2
0.4
0.6
0.8
Ruggedness (pairwise epistasis)

Figure 2

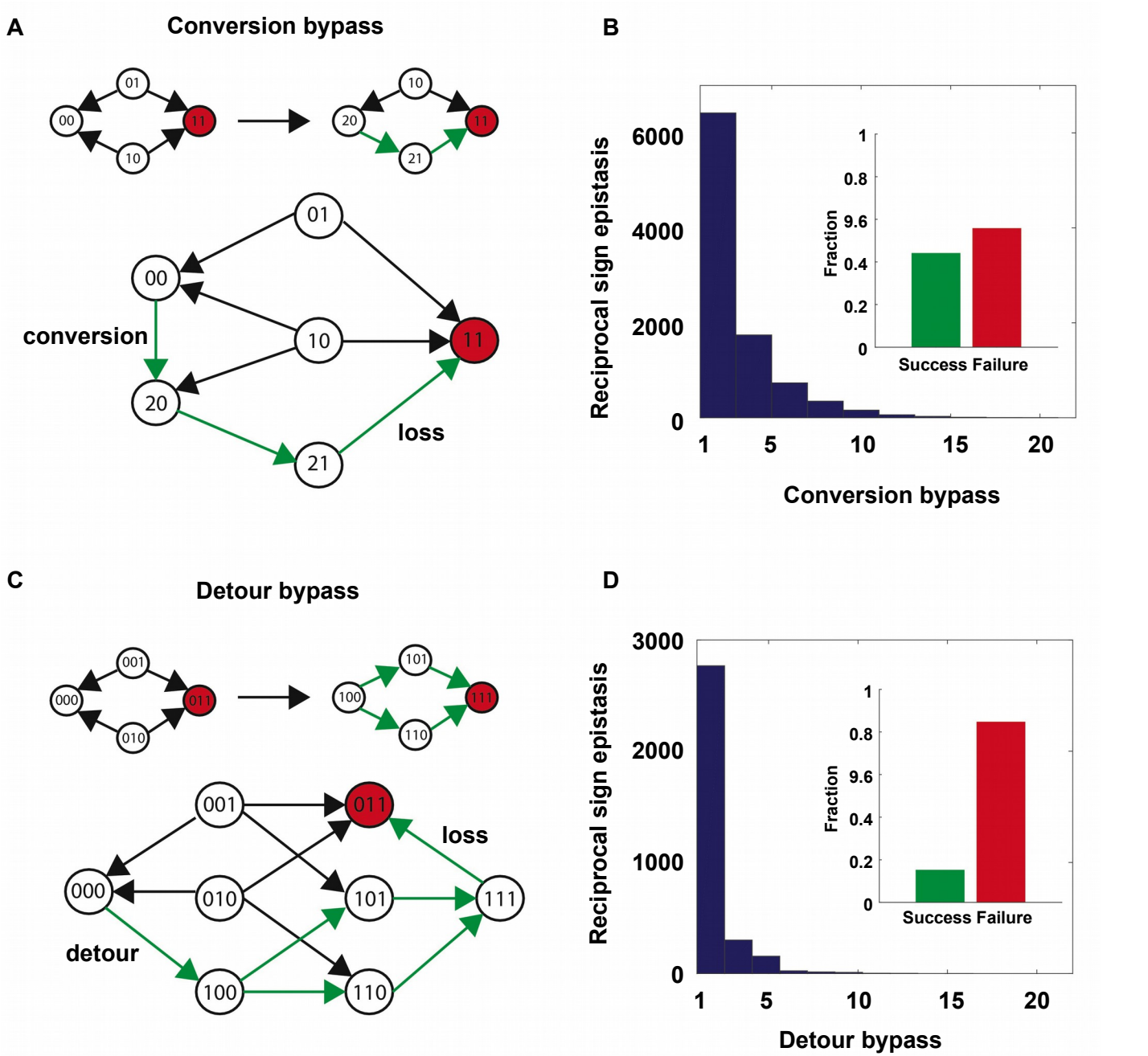

A
Conversion bypass
01
00
11
10
20
11
21
01
00
conversion
10
11
20
loss
21
B
Reciprocal sign epistasis
6000
4000
2000
0
1
5
10
15
20
Conversion bypass
Fraction
1
0.8
9.6
0.4
0.2
0
Success Failure
C
Detour bypass
001
000
011
010
101
100
111
110
001
011
loss
000
010
101
111
detour
100
110
D
Reciprocal sign epistasis
3000
2000
1000
0
1
5
10
15
20
Detour bypass
Fraction
1
0.8
9.6
0.4
0.2
0
Success Failure

Figure 3

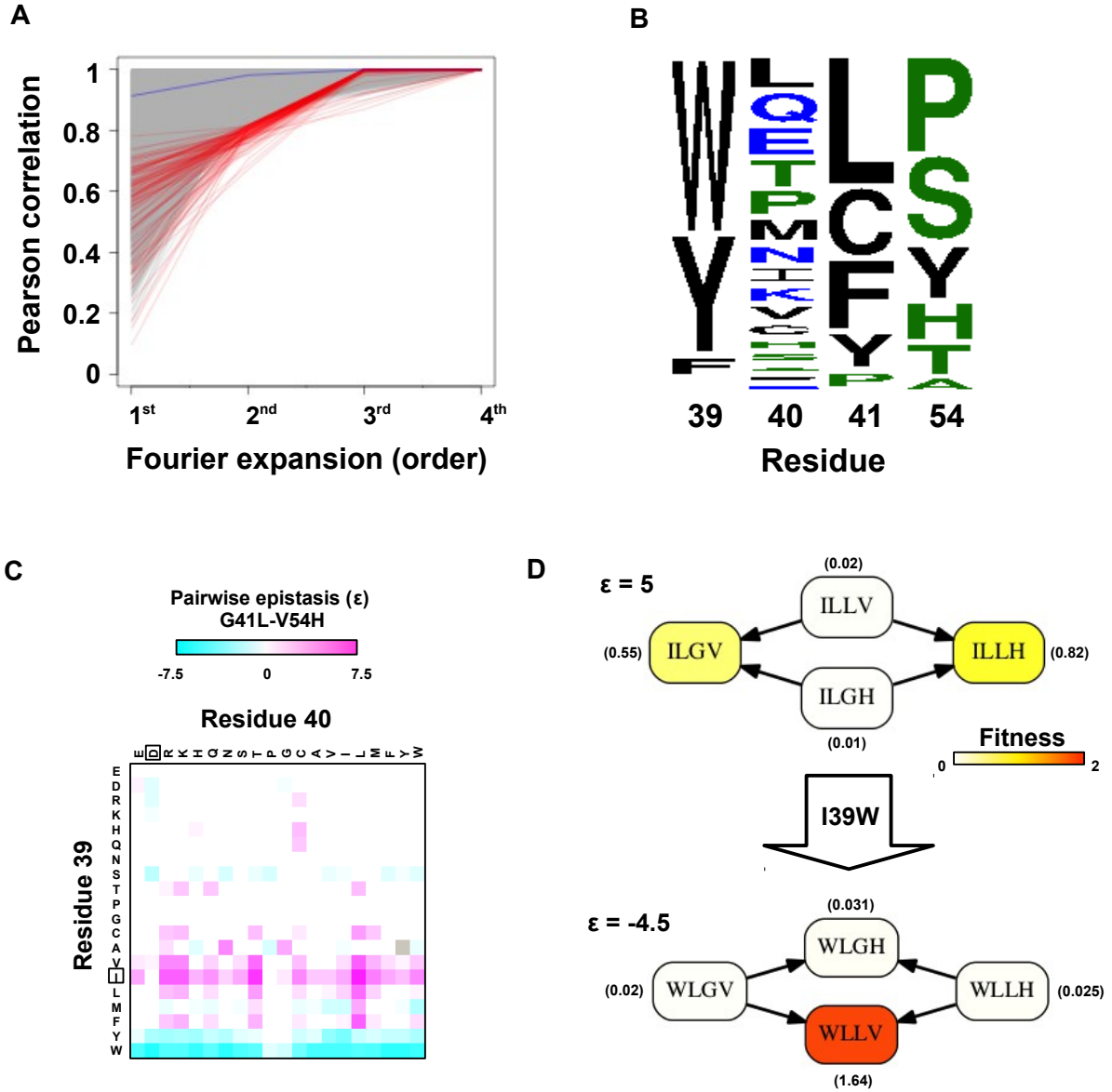
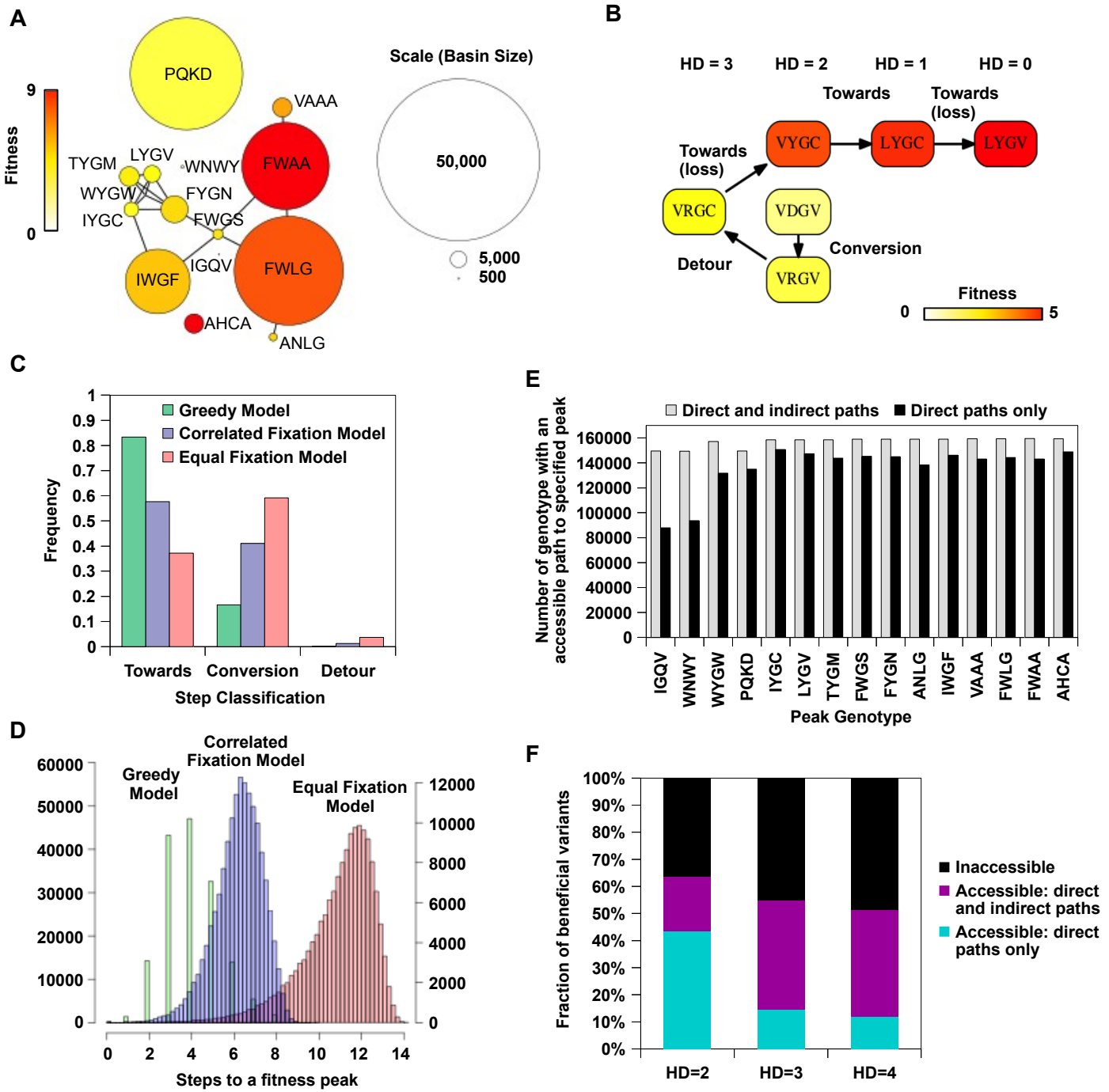
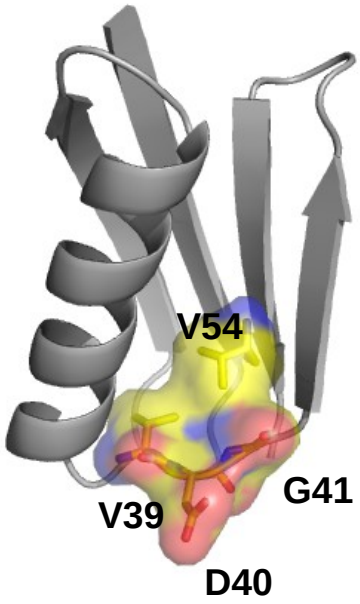

A
Pearson correlation
1
0.8
0.6
0.4
0.2
0
1st
2nd
3rd
4th
Fourier expansion (order)
B
39
40
41
54
Residue
C
Pairwise epistasis (ε)
G41L-V54H
-7.5
0
7.5
Residue 40
Residue 39
D
ε = 5
(0.02)
ILLV
(0.55)
ILGV
ILLH
(0.82)
ILGH
(0.01)
Fitness
0
2
I39W
ε = -4.5
(0.031)
WLGH
(0.02)
WLGV
WLLH
(0.025)
WLLV
(1.64)

Figure 4


A
Fitness
9
0
PQKD
VAAA
FWAA
TYGM
LYGV
WNWY
WYGW
FYGN
IYGC
FWGS
IWGF
IGQV
FWLG
AHCA
ANLG
Scale (Basin Size)
50,000
5,000
500
B
HD = 3
HD = 2
HD = 1
HD = 0
Towards
Towards (loss)
Towards (loss)
VYGC
LYGC
LYGV
VRGC
VDGV
Conversion
Detour
VRGV
Fitness
0
5
C
Frequency
Greedy Model
Correlated Fixation Model
Equal Fixation Model
Towards
Conversion
Detour
Step Classification
D
Correlated Fixation Model
Greedy Model
Equal Fixation Model
Steps to a fitness peak
E
Direct and indirect paths
Direct paths only
Number of genotype with an accessible path to specified peak
IGQV
WNWY
WYGW
PQKD
IYGC
LYGV
TYGM
FWGS
FYGN
ANLG
IWGF
VAAA
FWLG
FWAA
AHCA
Peak Genotype
F
Fraction of beneficial variants
HD=2
HD=3
HD=4
Inaccessible
Accessible: direct and indirect paths
Accessible: direct paths only

Supplementary Figure 1

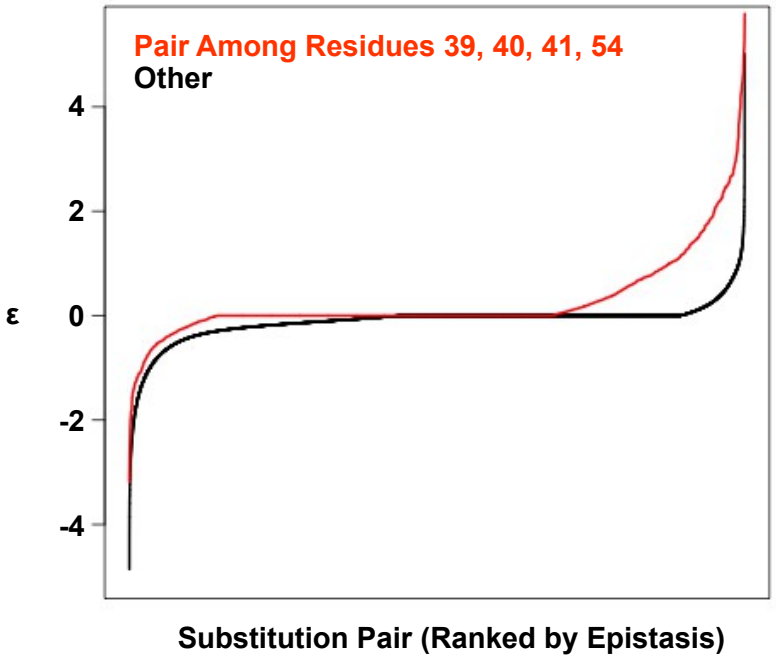
(A)



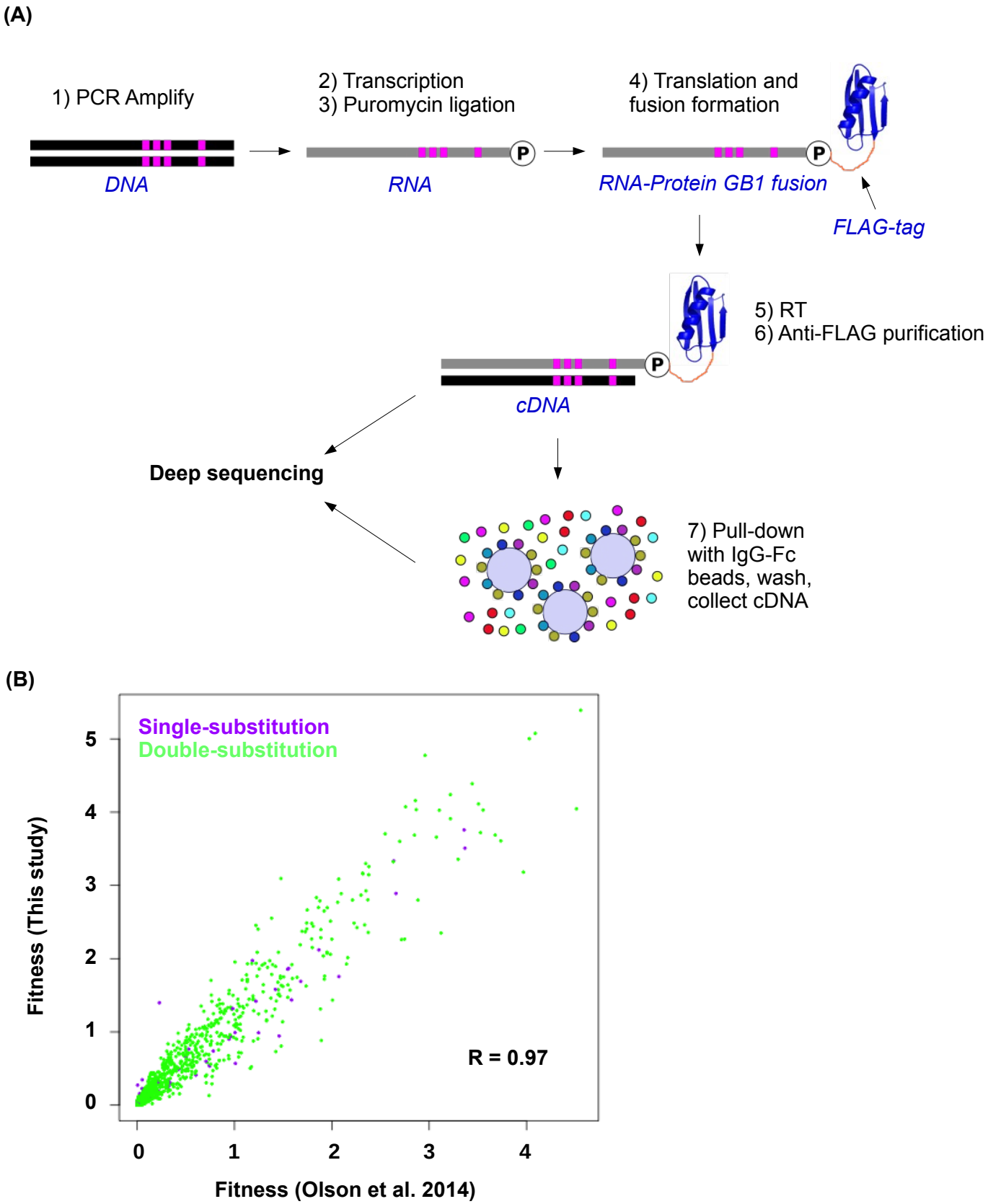
(B)

```
5'-TTCTAATACGACTCACTATAGGGACAATTACTATTTACATATCCACC

   Met GlnTyrLysLeuIle LeuAsnGlyLysThr LeuLysGlyGluThr ThrThrGluAlaVal
   ATG CAGTACAAGCTGATT CTGAACGGTAAGACG CTGAAAGGTGAGACG ACCACCGAAGCTGTA

   AspAlaAlaThrAla GluLysValPheLys GlnTyrAlaAsnAsp AsnGlyValAspGly
   GACGCTGCTACTGCA GAGAAGGTGTTCAAG CAGTACGCTAACGAC AACGGCGTCGACGGT

   GluTrpThrTyrAsp AspAlaThrLysThr PheThrValThrGlu
   GAATGGACCTACGAC GACGCTACCAAAACC TTCACGGTTACCGAA

   SerGlyGlySerAspTyrLysAspAspAspAspLysGlySerGlySer
   TCCGGAGGATCCGATTACAAGGATGACGACGATAAGGGTAGCGGCTCC-3'
```

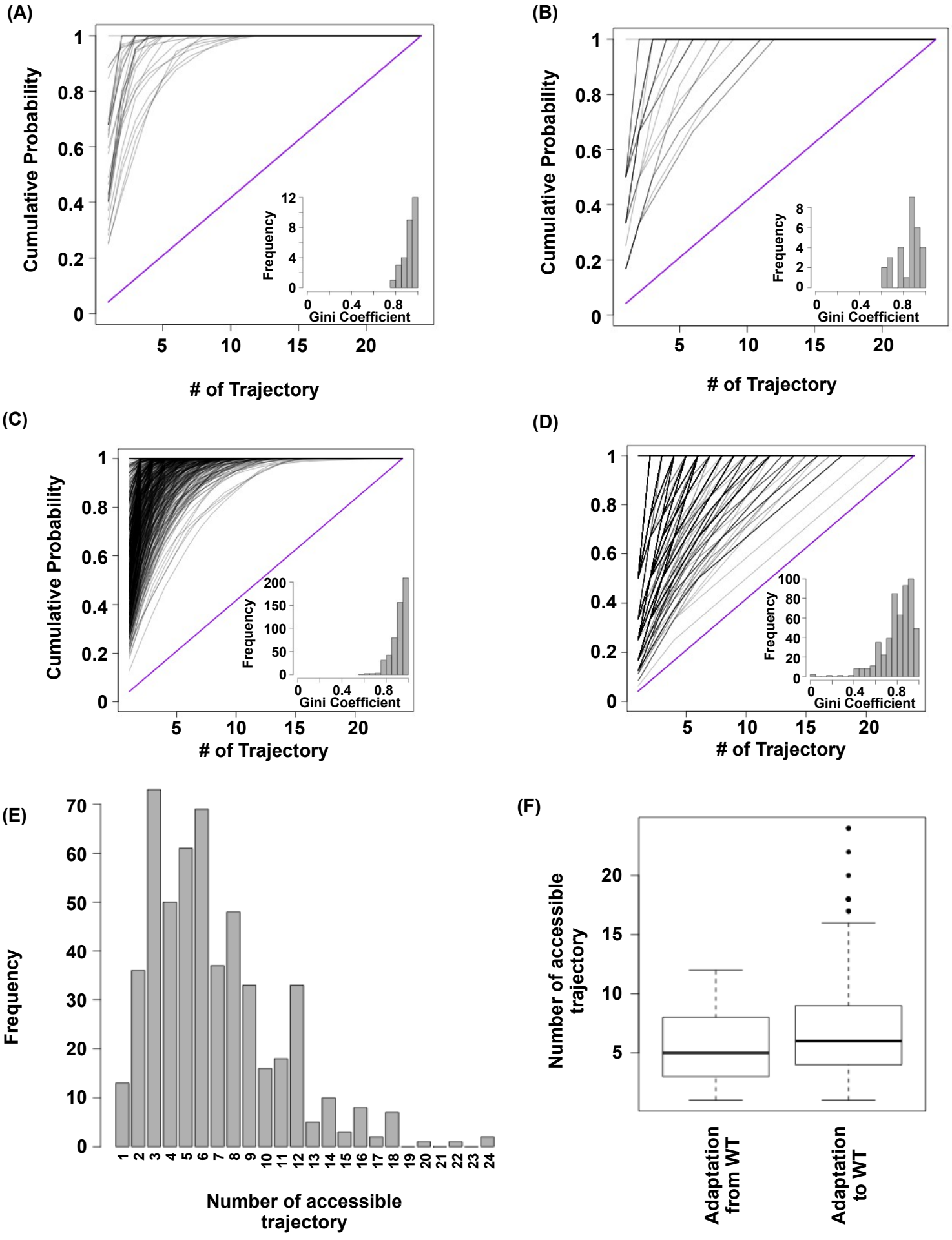
Supplementary Figure 2



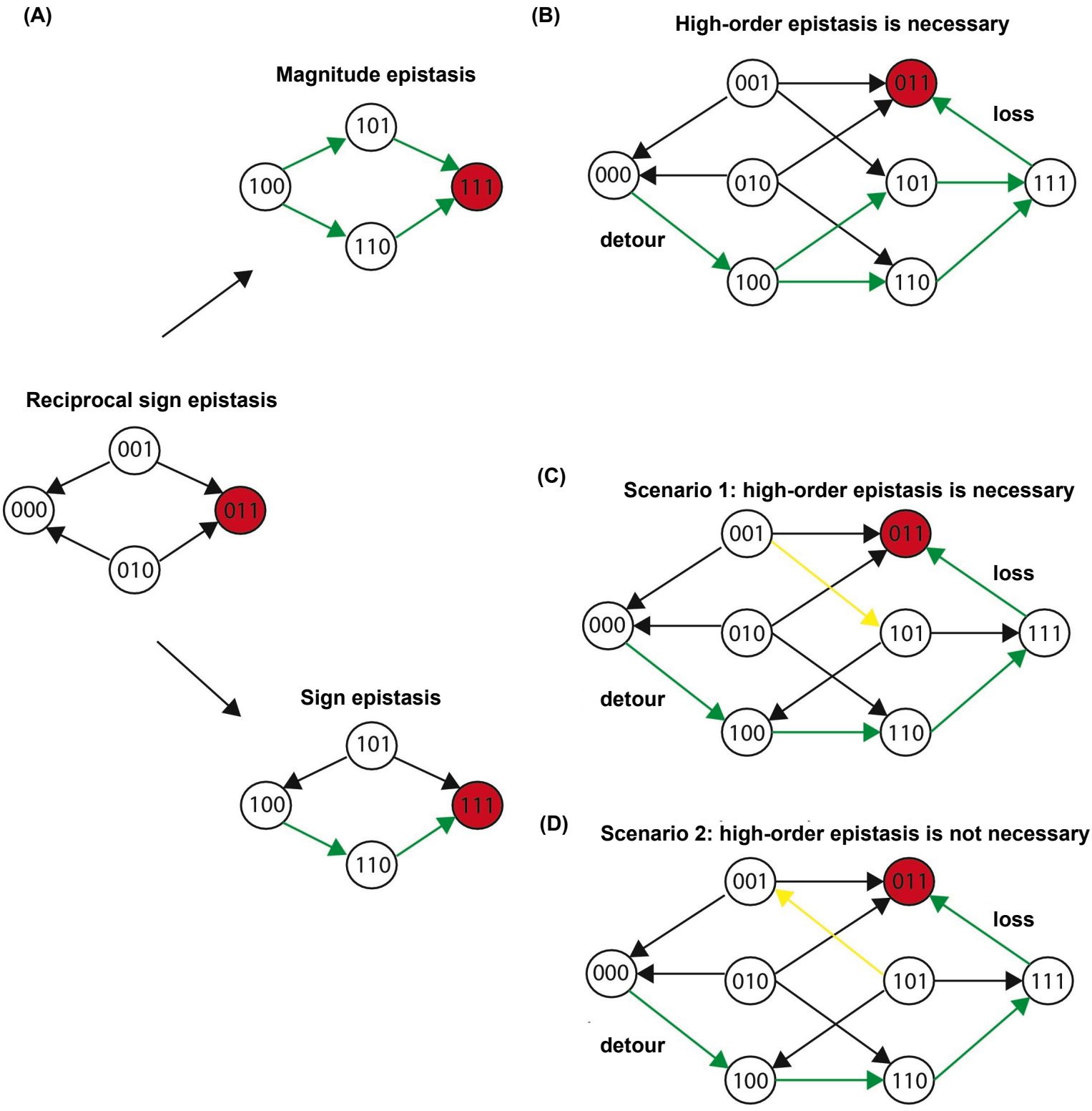
Supplementary Figure 3



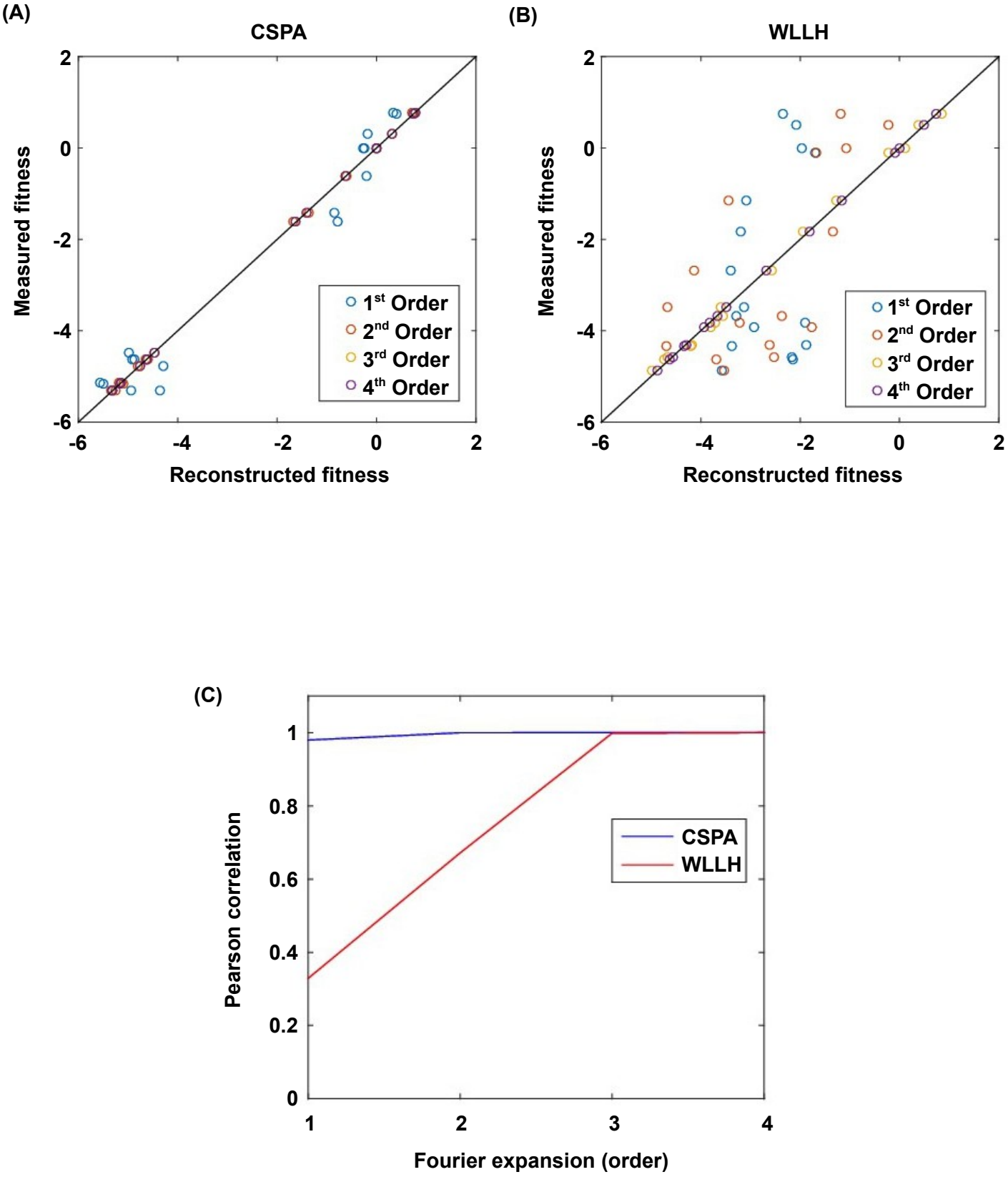
Supplementary Figure 4



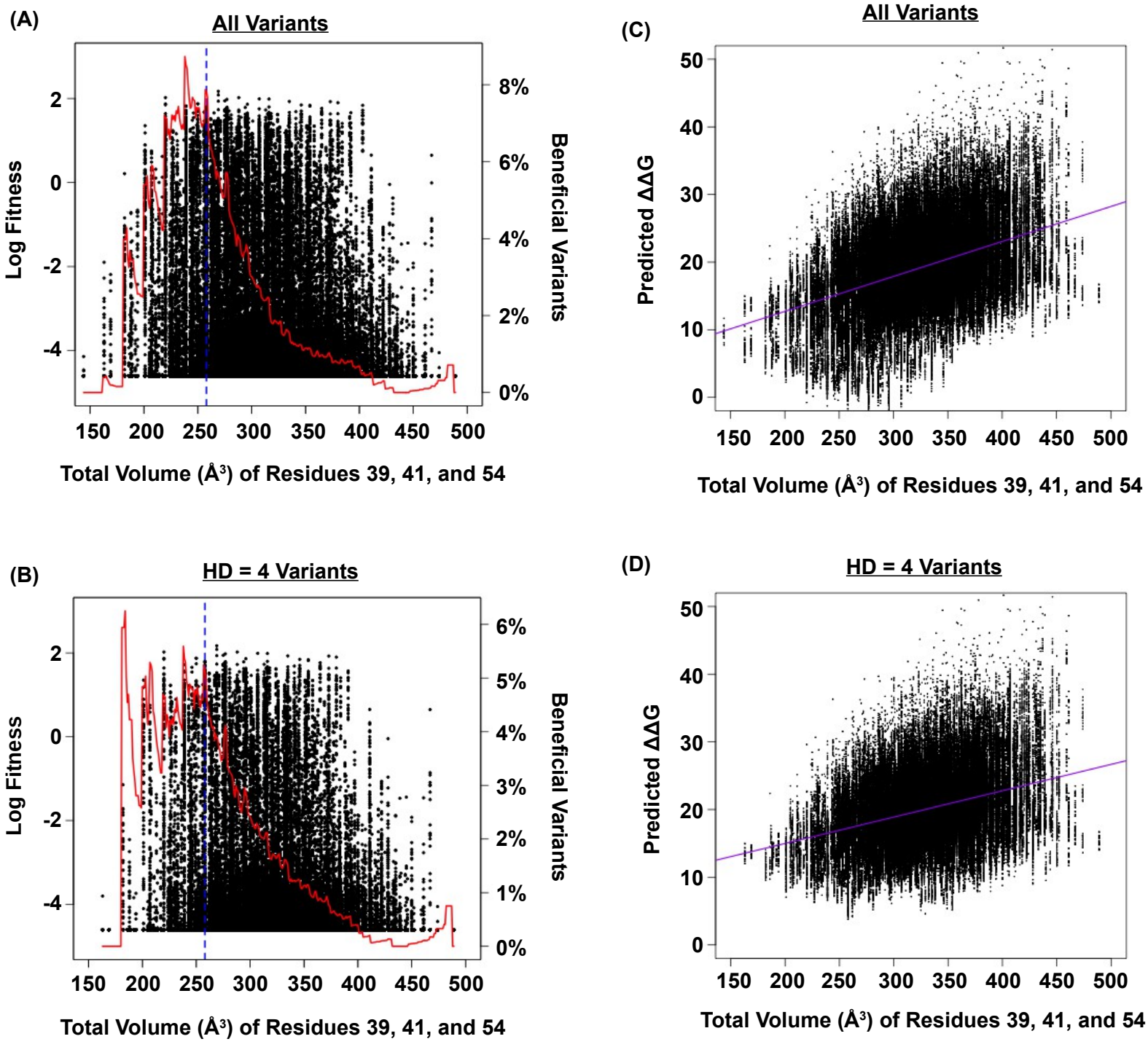
Supplementary Figure 5



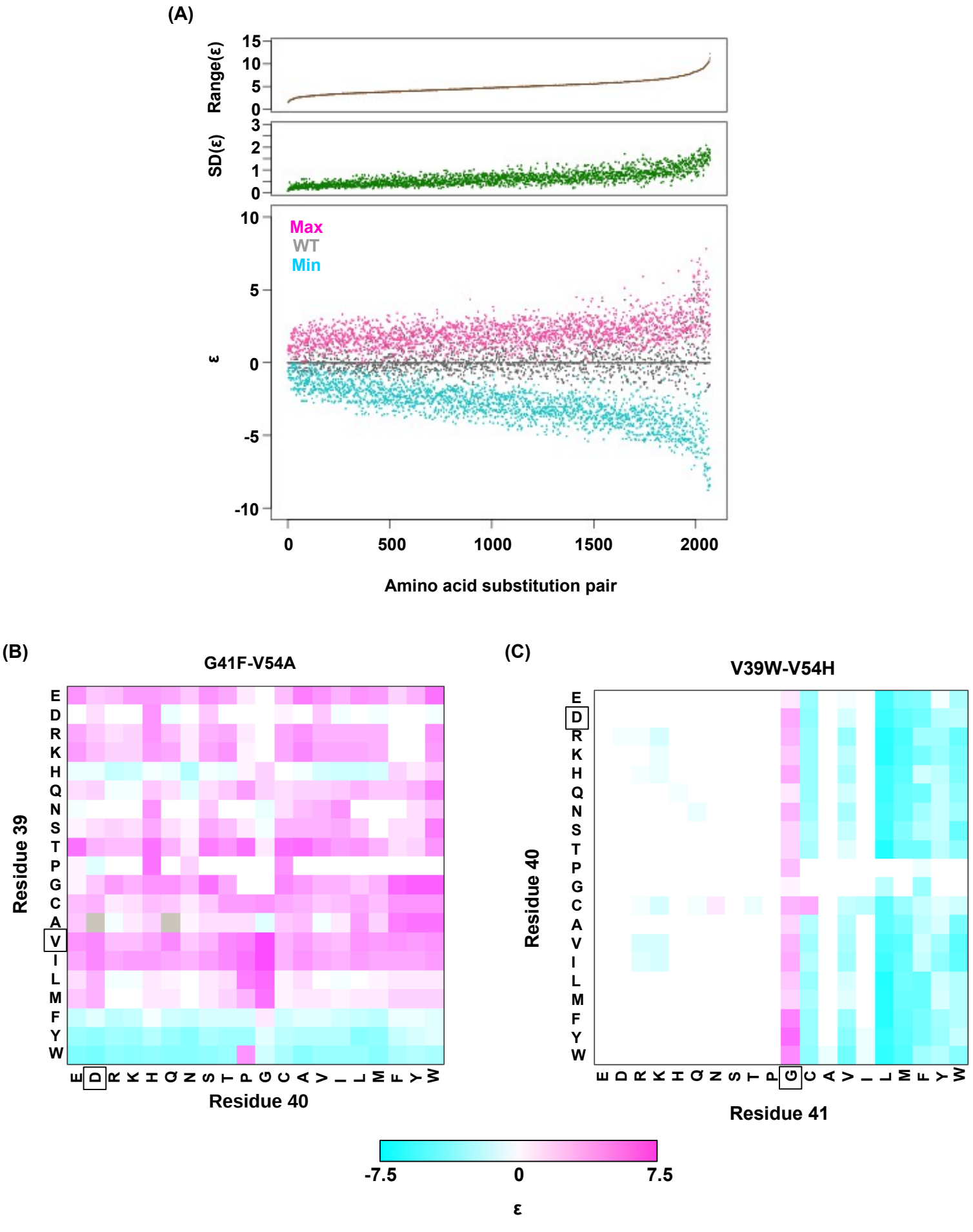
Supplementary Figure 6



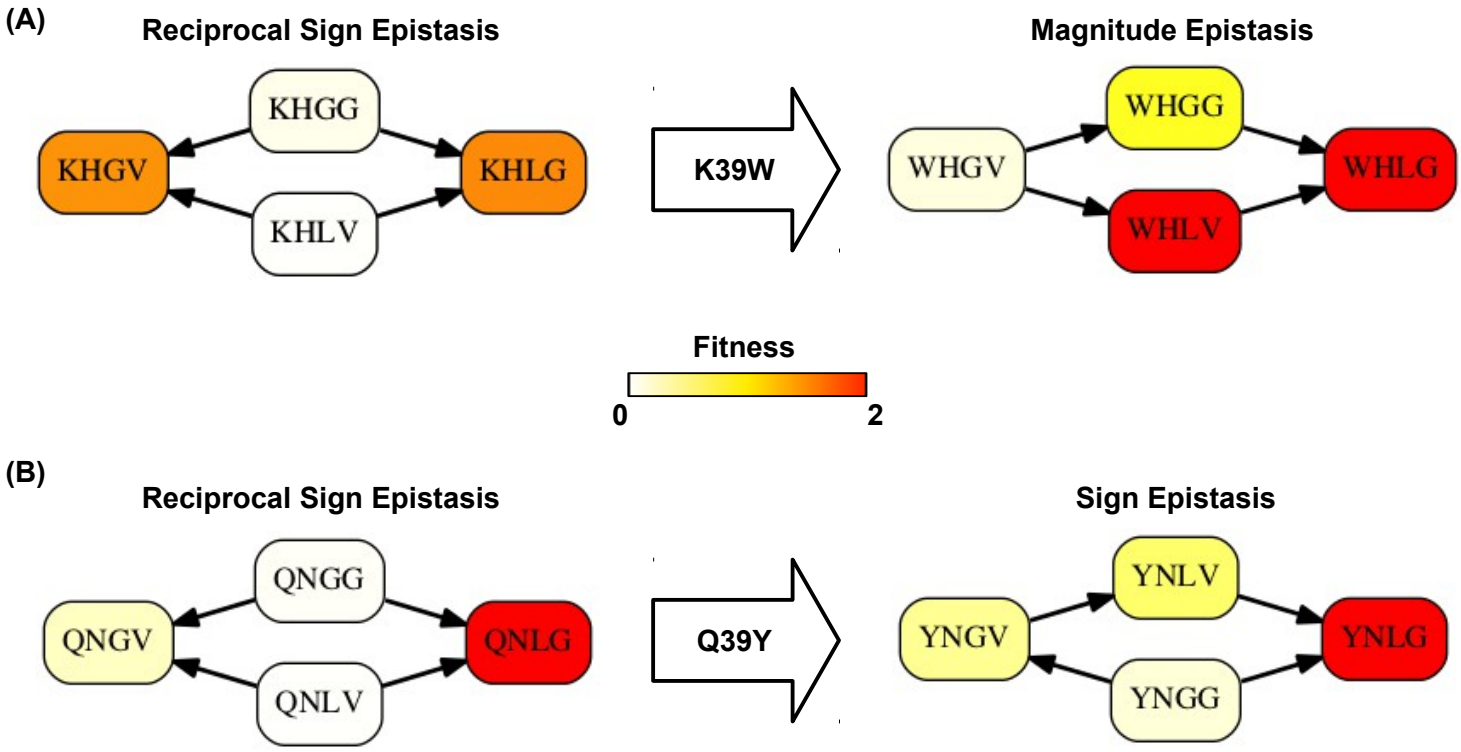
Supplementary Figure 7



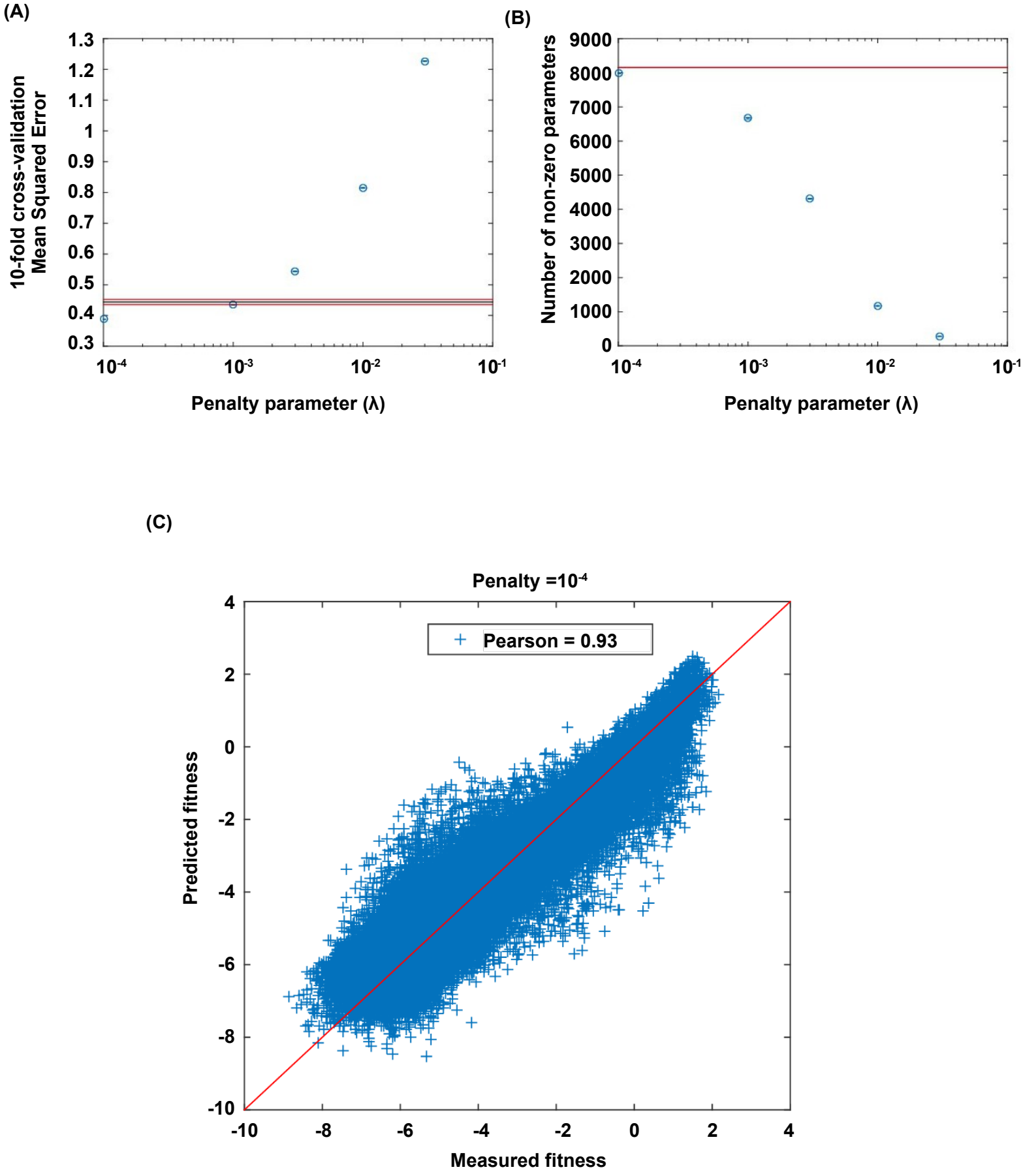
Supplementary Figure 8



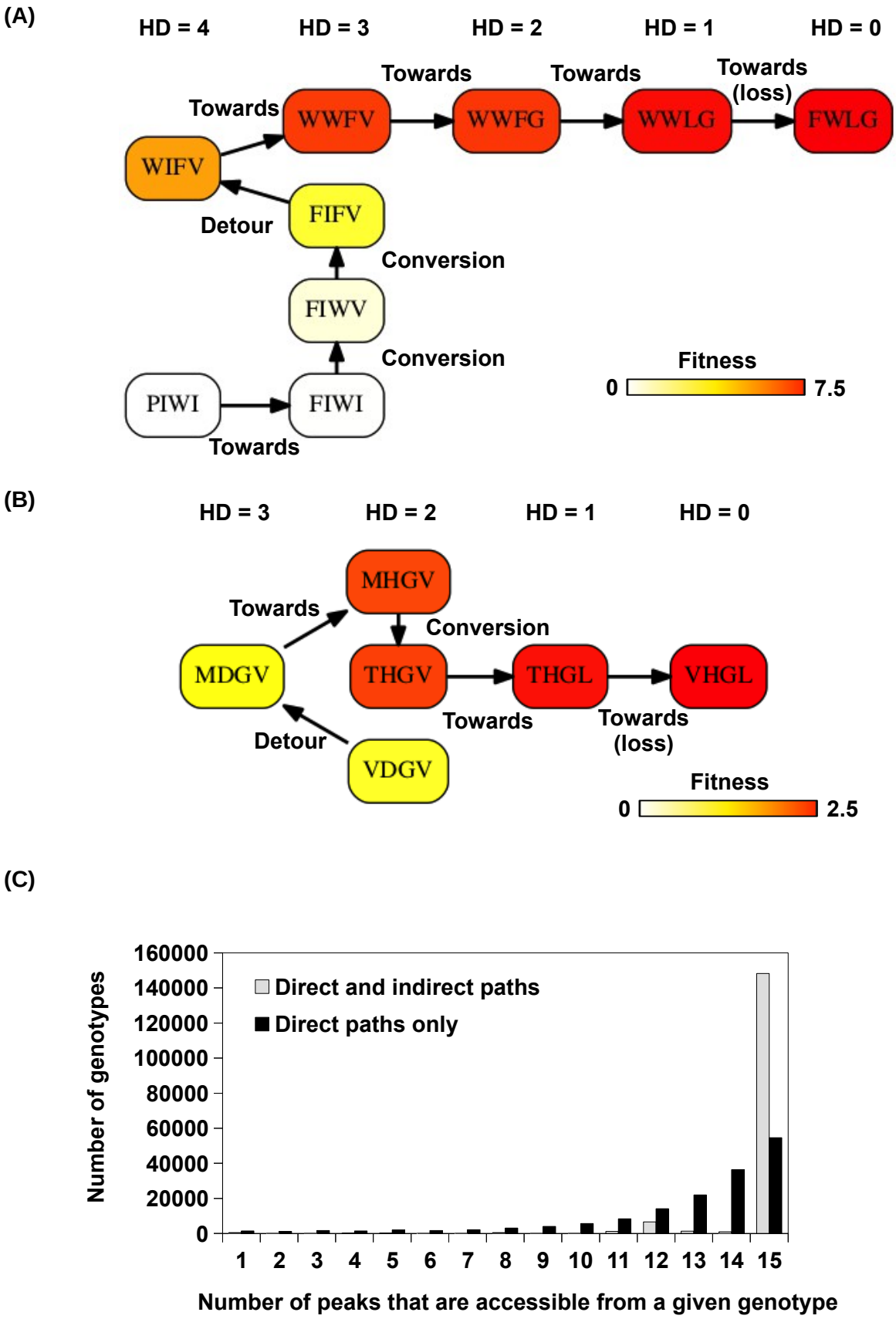
Supplementary Figure 9



Supplementary Figure 10



Supplementary Figure 11



Supplementary Figure 12