The spatial architecture of protein function and adaptation

Richard N. McLaughlin Jr1,2, Frank J. Poelwijk1,2, Arjun Raman1,2, Walraj S. Gosal1,2 & Rama Ranganathan1,2

Statistical analysis of protein evolution suggests a design for natural proteins in which sparse networks of coevolving amino acids (termed sectors) comprise the essence of three-dimensional structure and function1–3. However, proteins are also subject to pressures deriving from the dynamics of the evolutionary process itself—the ability to tolerate mutation and to be adaptive to changing selection pressures4–6. To understand the relationship of the sector architecture to these properties, we developed a high-throughput quantitative method for a comprehensive single-mutation study in which every position is substituted individually to every other amino acid. Using a PDZ domain (PSD95pdz1) model system, we show that sector positions are functionally sensitive to mutation, whereas non-sector positions are more tolerant to substitution. In addition, we find that adaptation to a new binding specificity initiates exclusively through variation within sector residues. A combination of just two sector mutations located near and away from the ligand-binding site suffices to switch the binding specificity of PSD95pdz quantitatively towards a class-switching ligand. The localization of functional constraint and adaptive variation within the sector has important implications for understanding and engineering proteins.

A basic tenet of biology is that the amino acid sequence of proteins specifies their three-dimensional structure and biochemical function, at least in a physiological setting7,8. Statistical coupling analysis (SCA)9–12 is a quantitative approach for understanding the information content of protein sequences through a generalization of the principle of evolutionary conservation. The underlying premise is that the pattern of energetic couplings between residues in a protein—the functional constraints between amino acids—might be exposed through a statistical analysis of coevolution of those residue positions in a family of homologous sequences. A main conclusion of SCA is that most residues in proteins evolve nearly independently, without much influence from even their immediate structural environment, whereas about 20% of amino acids (Fig. 1a and Supplementary Fig. 1) are organized into physically contiguous networks of coevolving amino acids, termed protein sectors1,2. Sectors are typically built around protein active sites, but connect to distant functional surfaces through pathways of residue interactions in the protein core7. For example, in the PDZ family of protein interaction modules, the sector connects the ligand-binding pocket with an allosteric site (Fig. 1b, asterisk) on the opposite surface13. Sectors are found in every protein family studied so far and are related to conserved functional activities, suggesting that this structural feature is a general property of natural proteins14–16.

An important next step is to understand why the design of natural proteins should look like the sector architecture, with distributed sparse networks of cooperatively acting residues embedded within an environment of weakly coupled amino acids. One reason might be necessity for native folding and function, but this seems unlikely. The marked recent advances in physics-based protein design are largely based on homogeneous optimization of local interactions in protein structures17,18, and alternative explanations for long-range communication within proteins have been suggested19. Our proposal is that the sector is the natural consequence of evolutionary constraints not typically considered in protein engineering or biophysical models, primarily of the need for adaptive variation in response to fluctuating conditions of fitness. By placing the constraints on native folding and function on sector positions, this architecture might provide the capacity for rapid adaptive variation through mutation of a few cooperatively acting residues. If so, the plurality of non-sector positions, regardless of structural location, should display much more mutation tolerance and less adaptive potential.

Previous studies have tested the role of sectors in protein function using targeted mutagenesis of a few amino acid positions20–23. Although useful, these studies cannot convincingly test the sector hypothesis posed here, mainly due to the limited scale of experimentation. To address this, we developed a quantitative high-throughput method based on next-generation sequencing suitable for a large-scale mutational analysis of proteins in a cellular context (Fig. 2 and Supplementary Fig. 2). The method is implemented here for comprehensive single mutagenesis in one representative member of the PDZ family of protein interaction modules (PSD95pdz1) as a model system, but could be used for the study of many other proteins and for higher-order mutational studies (F.J.P. and R.R., unpublished observations).

The method involves three components: (1) a bacterial two-hybrid (B2H) system modified from a previous study21 in which the ability of PSD95pdz1 to bind its cognate ligand (-TKNYKTSV-COOH, derived from the cysteine-rich interactor of PDZ (CRIP1)25) is quantitatively linked to the expression of enhanced green fluorescent protein (eGFP) (Supplementary Fig. 2a); (2) a fluorescence-activated cell sorting (FACS) step, in which bacterial populations carrying large libraries of mutations in the protein are selected for those cells displaying

Figure 1 | Sector architecture in the PDZ domain family. a, b. The PDZ sector (blue spheres) shown in a cartoon (a) or space filling (b) representation of the structure of rat PSD95pdz1 (Protein Data Bank (PDB) accession 1BE9). Yellow stick bonds represent the co-crystallized peptide ligand, with ligand positions numbered (0, -1, -2). The sector comprises a sparse network of residues built around the ligand-binding pocket and connecting to a distant surface site (marked with asterisk) through a subset of amino acid interactions within the protein core.

1Green Center for Systems Biology, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA. 2Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050, USA. *Present address: Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109-1024, USA.
eGFP levels above a specified threshold (Supplementary Fig. 2b, c); and (3) Solexa high-throughput sequencing to determine the frequency of each allele in the unselected and selected populations 23–26 (Supplementary Fig. 2c). The effect of each mutation is then expressed as the log frequency of observing each amino acid at position i in the selected (sel) versus the unselected (unsel) population, relative to wild type (WT):

$$\Delta E_i^x = \log \left[ \frac{f_{i,\text{sel}}^{x,\text{WT} \times \text{WT}}}{f_{i,\text{WT},\text{WT}}^{x,\text{WT} \times \text{WT}}} \right] - \log \left[ \frac{f_{i,\text{WT},\text{WT}}^{x,\text{WT} \times \text{WT}}}{f_{i,\text{WT},\text{WT}}^{x,\text{WT} \times \text{WT}}} \right]$$

In this assay, mutations that show no functional effect should show a relative frequency in the selected population that is identical to wild type ($\Delta E_i^x = 0$), and deviations from this expectation should provide a quantitative measure of the functional effect of each mutation.

Tuning of growth and induction parameters and introduction of a volumetrically largest side chain. Substitutions to alanine or cysteine charged at neutral pH (Asp, Glu, Lys and Arg), and by tryptophan, the unfavourable substitution, followed by amino acids that are formally shared with the overall PDZ family (83 total) is individually mutated to every other amino acid (Fig. 2a, 83 positions). To examine the position-specific effects of mutation in PSD95$^{\text{d3}}$, we considered both the full data matrix (Fig. 2a) and the average effect of all mutations per position $\langle \Delta E_i^x \rangle$ (Fig. 2b, c). This analysis indicates a heterogeneous, distributed and physically contiguous network of functional residues in PSD95$^{\text{d3}}$ (Fig. 2b–d). Most positions show little effect on mutation, tolerating nearly every substitution even if radically different in chemical character (Fig. 2a–c and Supplementary Fig. 5). This includes some that are in direct contact with peptide ligand (for example, 326 and 380), and some that are buried in the protein core and largely conserved (for example, 314, 316, 356, 357 and 390). By contrast, a subset of positions (20 out of 83, Fig. 2c) shows significant sensitivity to mutation ($>2\sigma$ from mean). Within the binding pocket, His 372 tolerates essentially no other substitution and residue 323, Phe 325, Ile 375, and Leu 379 show tolerance to only the most chemically conservative mutations. However, outside the direct spatial environment of the ligand, Gly 329, Gly 330, Ile 336, Ala 347, Leu 353, Val 362 and Ala 375 comprise a subset of buried residues that also show significant sensitivity to mutation—the largest average mutational effect in the whole protein comes from position 329.

We used the B2H–sequencing assay to carry out a complete single-mutation scan for the PSD95$^{\text{d3}}$ domain in which every position shared with the overall PDZ family (83 total) is individually mutated to every other amino acid (Fig. 2a, 83 positions x 19 mutations + wild type = 1,578 variants). The data reveal several aspects of the global pattern of mutational sensitivity in the protein. First, the effect of each amino acid substitution averaged over all positions (across rows in the data matrix) is what might be predicted from the chemical properties of the side chains ($\langle \Delta E_i^x \rangle$, Fig. 2a, right). Proline is the most unfavourable substitution, followed by amino acids that are formally charged at neutral pH (Asp, Glu, Lys and Arg), and by tryptophan, the volumetrically largest side chain. Substitutions to alanine or cysteine introduce the least perturbation on average, consistent with their use for scanning mutation or solvent accessibility studies, respectively.

Figure 2 | Complete single mutagenesis in PSD95$^{\text{d3}}$. a. The data matrix showing $\Delta E_i^x$—the functional cost of every mutation x at each position i relative to wild-type PSD95$^{\text{d3}}$—colorimetrically, with blue representing loss-of-function and red representing gain-of-function mutations. The wild-type amino acid at each position is indicated by bold squares in the grid. The average functional cost of each amino acid substitution over all positions ($\langle \Delta E_i^x \rangle$) is shown at the right. b. The functional cost of all amino acid substitutions at each position shown as the average taken over each column ($\langle \Delta E_i^x \rangle_x$). c. A histogram of the data in panel b indicates positions with a significant effect ($>2\sigma$, 20 out of 83). d. Mapping of the 20 functionally significant positions on the PSD95$^{\text{d3}}$ structure, the peptide ligand shown as yellow stick bonds. These positions comprise a distributed, physically contiguous network built around the binding pocket and extending through the protein structure.
We measured the difference in the functional effect of mutations when binding either the wild-type or T₂₋F ligands, a global analysis of the context dependence (or epistasis) between every mutation of every position in PSD⁹⁵₋₋₋ and the T₂₋₋₋ ligation on the peptide ligand (Fig. 4a, b). The data show that nearly all mutations at positions in PSD⁹⁵₋₋₋—even if they have an absolute effect on ligand binding—show the same effect on binding the T₂₋₋₋ ligand as for the wild-type and ligand, resulting in no significant epistasis with the T₂₋₋₋ mutation (Fig. 4c). These positions are insensitive to the switch in peptide identity and would therefore be non-adaptive with regard to this perturbation. However, mutations at nine positions in PSD⁹⁵₋₋₋ show statistically significant epistasis with the T₂₋₋₋ mutation (Fig. 4b, c); these positions show a mutational response that depends on the identity of the target ligand. These positions uniformly show an effect on mutation that is less deleterious for the T₂₋₋₋ ligand than for the wild-type and ligand (Fig. 4b, red pixels). All of these positions are within the sector (Fig. 4d, see overlap between sector (blue mesh) and epistatic positions (red)), and map out a pattern of epistasis for the T₂₋₋₋ mutation that involves a spatially distributed network in the PDZ protein structure comprising residues both near and far from the minus two ligand position (Figs. 1a and 4d).

A subset of the sector positions (322, 330, 336 and 372) shows such extreme epistasis that the positions flip in the direction of mutational effect between the wild-type and T₂₋₋₋ ligands (Fig. 4a, b, d). Mutations at these positions destabilize or are neutral for binding to the wild-type and ligand, but are favorable for binding to the T₂₋₋₋ ligand and would therefore be selected when challenged to bind this class-switching variant. Notably, only one of the selectable sites (372) directly contacts the T₋ ligand position. The other sites are located either one shell (329, 330) or two shells (336) of residues away from the binding pocket, or act at a distance from an epitope contacting the terminal carboxylate of the ligand (322; Fig. 4d). We also find that for these few sites of selection for T₂₋₋₋, it is not merely one or two mutations, but nearly every substitution that shows the effect of stabilizing binding to the T₂₋₋₋ ligand (Fig. 4b and Supplementary Fig. 7).

These data describe many potential mutational routes for switching the binding specificities PSD⁹⁵₋₋₋ towards the class II T₋ ligand. For example, mutation of position 330, which does not contact ligand (Fig. 4d), to threonine is expected to moderately destabilize binding to wild-type CRIPT ligand, but is the most favourable of all single mutants for the T₋ ligand (Fig. 4e). Also, His372Ala, a mutation at a position directly linking the −2 position of the ligand and position 330 (Fig. 4d), is expected to decrease binding for CRIPT more strongly than the Gly330Thr mutation, but to increase binding for T₋ ligand to a similar degree (Fig. 4e). Binding affinities for Gly330Thr, His372Ala and the double mutant combination were measured using a fluorescence polarization assay and show excellent consistency with the B2H data (Fig. 4f). His372Ala converts PSD⁹⁵₋₋₋ from a protein with an approximately 45-fold preference for the CRIPT ligand to one with about a 14-fold preference for the T₋ ligand, a partial specificity switch from a single mutation. By contrast, Gly330Thr converts PSD⁹⁵₋₋₋ to a domain with an unexpected phenotype: high-affinity but non-specific recognition of both CRIPT and T₋ ligands. Such a phenotype could be evolutionarily important when a mutational path characterized by a promiscuous but biologically functional intermediate is advantageous. Finally, the combination of both mutations (Gly330Thr and His372Ala) complements the specificity switch; this double mutant displays an approximately 45-fold preference for the T₋ ligand. These data demonstrate that short, cooperative pathways of mutation within the sector can suffice to change functional specificity quantitatively.

In summary, the data presented here indicate that protein robustness and adaptability can be explained through a model in which the main functional constraints are loaded in the sector—a sparse, collectively evolving network within the protein structure. By saturation point mutagenesis, we find that sector positions selectively resist variation when challenged with wild-type ligand, but can flip to promote...
ligand is shown in yellow stick bonds, the $T_{-2}$ position is indicated in red and sector positions are in blue mesh. The positions showing mutational epistasis with $T_{-2}F$ comprise a physically distributed network propagating from the $T_{-2}$ position, and are entirely composed of sector positions. The asterisks are as in b, e. The B2H–sequencing data for sector positions 330 and 372 when binding CRIPIT (Fig. 2c) or $T_{-2}F$ (Supplementary Fig. 7) ligands suggest mutations for altering the specificity of PSD95$^{pdz3}$ towards $T_{-2}F$; the wild-type amino acid is shown in bold. f, Binding affinities for purified PSD95$^{pdz3}$ carrying the Gly330Thr and His372Ala mutations both singly and together. Gly330Thr displays high affinity but non-specific binding for CRIPIT and $T_{-2}F$ ligands, His372Ala shows a partial specificity switch towards $T_{-2}F$ and the double mutant represents a complete specificity switch for these two ligands.

**METHODS SUMMARY**

SCA and sector identification were carried out using version 5.0 of the SCA MATLAB toolbox. The software and a script for carrying out the calculations is available for download from our laboratory website (http://systems.swmed.edu/rr_lab).

Comprehensive single-mutation libraries in PSD95$^{pdz3}$ were constructed by oligonucleotide-directed mutagenesis, randomizing each codon to NNS, in which N represents a mixture of all four bases, and S represents a mixture of G and C. NNS libraries for each codon were separately amplified by PCR, mixed in an equimolar ratio, and cloned into pZS22-PDZ3 (see later) to make the library.

The B2H system consists of three plasmids: (1) pZS22-PDZ3 (kanamycin resistant), providing IPTG-inducible expression of PSD95$^{pdz3}$ variants fused to the DNA-binding domain of bacteriophage $\lambda$-cl; (2) pZA31-RNA2-CRIPT (or $T_{-2}F$) (chloramphenicol resistant), providing anhydrotetracycline-inducible
expression of the amino-terminal domain of *Escherichia coli* RNA polymerase α-subunit fused to the target peptide ligand; and (3) pZE1RM-eGFP (ampicillin resistant), containing the target promoter driving the enhanced green fluorescent protein (eGFP). MC4100-Z1 *E. coli* cells transformed with these three plasmids were grown in ZYM-505 media, diluted with inducers to attendance at 600 nm (D_{600nm}) of 0.4, grown for 2 h, and subjected to FACS sorting (BD FACSAria, gates set at top 10% and 25% of eGFP distribution for wild-type PSD95^{ΔN}). Sorted cells were grown in ZYM-505 for 12 h, miniprepped and amplified by PCR to prepare samples for Solexa paired-end sequencing completed using Solexa v4 PE-flowcell (University of Texas Southwestern genomics core) and analysed through CLC Genomics Workbench and self-coded software.

For biophysical measurements, PSD95^{ΔN} variants were expressed in BL21(DE3) *E. coli* cells as glutathione S-transferase (GST) fusions and purified using affinity chromatography and cleavage of the GST tag. Binding affinities were determined using tetramethylrhodamine (TMR)-labelled target peptides, monitoring fluorescence polarization of TMR on a Victor^V^ plate reader as a function of PSD95^{ΔN} concentration.

**Full Methods** and any associated references are available in the online version of the paper.

Received 18 July 2011; accepted 10 August 2012.

Published online 7 October 2012.


**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank W. Russ, P. Mishra and other members of the Ranganathan laboratory for contributions to this work, W. Wakeland and C. Liang for assistance with Solexa sequencing, E. Curry and A. Mobley for assistance with flow cytometry, and M. Elowitz for providing the pZ plasmids. We acknowledge support from the University of Texas Southwestern Graduate School and Pharmacology Training Grant (R. N. M.), the Helen Hay Whitney Fellowship program (F.J.P.) and support from the University of Texas Southwestern Graduate School and Pharmacology Training Grant (R. N. M.), the Helen Hay Whitney Fellowship program (F.J.P.) and support from the National Institutes of Health (RO1EY018720-05, R.R.). The Robert A. Welch Foundation (I-1366, R.R.) and the Green Center for Systems Biology (R.R.).

**Author Contributions** R.N.M. and R.R. developed the research plan and experimental strategy. R.N.M. built the B2H assay, collected the data and wrote and executed the code for analysis of the B2H and sequencing data. F.J.P. and W.S.G. improved the dynamic range of the B2H assay. A.R. carried out the mutational analysis in Fig. 4 e, f. R.N.M. and R.R. analysed the data writing the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.R. (rama.ranganathan@utsouthwestern.edu).
NNS library construction. Comprehensive single mutant libraries were constructed using oligonucleotide-directed mutagenesis of PSD95δΔ. To mutate each position in PSD95δΔ (positions 311–393), two mutagenic oligonucleotides (one sense, one antisense) were synthesized (IDT) that contain sequence complementary to 15 base pairs (bp) on either side of the targeted position. For the targeted position, the oligonucleotides contain NNS codons, in which N is a mixture of A, T, C and G, and S is a mixture of G and C. This biased randomization results in 32 possible codons with all 20 amino acids sampled—a significant increase in library complexity with a loss of amino acid coarseness. One round of PCR was carried out with either the sense or antisense oligonucleotide and a flaunting antisense or sense oligonucleotide. A second PCR round using a combination of the first round products and both flanking primers produced the full-length double stranded product. For the 83 positions randomized here, this constituted 83 × 2 first round PCR reactions and 83 second round reactions, for a total of 249 PCR reactions. All reactions yielded a single intense band on an agarose gel. PCR product concentrations were measured using PicoGreen (Invitrogen), pooled in equimolar ratios, purified, digested and ligated into the B2H λ-cl fusion expression vector. Each ligation was purified, eluted into 7 μl dH2O, and measured in the B2H assay as described earlier. Each transformation yielded greater than 10^6 variants.

To permit coverage of the full PSD95δΔ gene (~300 bp) with 75 base pair-ended reads by Solexa sequencing, we split the PSD95δΔ sequence into three subgroups (positions 311–338, 339–365 and 366–393, respectively). The NNS-mutagenesis products for the positions of each subgroup were mixed and ligated as a single library. Each subgroup was independently subject to the B2H assay, FACS, amplification and sample preparation for Solexa sequencing (see later).

The B2H assay. Electrocumponent MC4100-Z1 cells containing pZE1RM-eGFP (amcinolin resistant) and pZ31-RNAx-CRIPT (chloramphenicol resistant) plasmids were transformed with 1 μl of 20 ng μl^-1 pZZS2-PDZ3-WT (kanamycin resistant) plasmid and recovered for 1 h in ZYM-505 media. To quantify library complexity, 1 μl recovered transformation mixture was plated on LB plus kanamycin. The entire 1 ml transformation was then added to 10 ml ZYM-505 (ref. 31 plus kanamycin (30 μg ml^-1) and amcinolin (50 μg ml^-1) and chloramphenicol (25 μg ml^-1) in a 50 ml baffled flask, grown at 6 °C at 37 °C at 225 r.p.m., then diluted 10 μl culture into 10 ml ZYM-505 plus antibiotics, and grown 12 h at 37 °C at 225 r.p.m., and finally measured for attenuation (D) at 600 nm. A 35 μl aliquot of each culture was added to one well of a 48-well plate containing 500 μl LB with antibiotics plus anhydrotracycline (100 ng ml^-1) plus isopropylthigalactoside (IPTG) (100 μM) for a final D_600 nm = 0.6, and incubated at 18 °C. 150 r.p.m. for 2 h for induction. Induced cells were diluted to 30 μl cells in 1 ml filter-sterilized M9 plus 0.4% glucose for flow cytometry. Before analysis or sorting, cells were passed through a 30-gauge needle for disaggregation to single cells.

All flow cytometry was performed with standardized settings on a BD FACsCan, setup to measure GFP fluorescence (FL1). Cell sorting was performed on the BD FACSaria by technicians in the UT Southwestern Medical Center cytometry core. For library selections, flow-cytometry gates were placed relative to the fluorescence distribution of WT-PDZ2 to control for systematic assay-to-assay variability. For each NNS library, two populations were collected at gates set at the top 10% and 20% of the WT-PDZ2 distribution. The B2H assay library of PDZ3 mutants, a positive cell population numbering greater than 1,000 times the complexity of the library was collected. Cells were sorted into chilled rich medium (ZYM-505 without antibiotics, 4 °C), and the collection tube was kept chilled in the cytometer during sorting to maximize cell viability. Typical viability of sorted cell populations were >70% when plated on selective medium.

Identification of λ-cl Glu34Pro to increase B2H dynamic range. An initial problem with dynamic range in the B2H assay was high basal activity, such that eGFP was expressed to some degree even without PDZ domain function. A mechanistic explanation for this high basal activity was suggested by the observation that the λ-cl binding sites of the eGFP but not the PDZ domain closely match to the RNA polymerase σ subunit as to activate transcription through direct (PDZ-independent) contact. Previous studies showed only one λ-cl position (Glu34) underlies this non-specific mode of transactivation[24–25]. To identify λ-cl variants with reduced basal activity, a library of all possible amino acid substitutions at position 34 (Glu34X) was cloned and expressed in the B2H assay as described earlier. Induced B2H Glu34X libraries containing wild-type PSD95δΔ and CRIP peptide (high-affinity interaction) were plated on selective medium, and the GFP intensity of hundreds of colonies was visually assessed using a wide-field fluorescence microscope. Twenty colonies with high intensity were picked, pooled, grown in liquid culture and mini-prepared to isolate plasmid DNA. This library of high intensity clones was then induced in the B2H assay containing a peptide that should not bind wild-type PSD95δΔ (TKNYKQGKK) for negative selection. This library was plated on selective medium and four colonies with low intensity were picked and sequenced. All four colonies contained the Glu34Pro λ-cl variant,
suggesting that this variant should provide an increased dynamic range for the B2H assay. This prediction was confirmed in the standard B2H assay using a set of PSD95<sup>pdz3</sup> mutants as described in Supplementary Fig. 2.

**Solexa sequencing.** Sorted cell populations were diluted into ZYM-505 plus kanamycin and grown 12 h at 37°C with shaking (250 r.p.m.). Overnight cultures were centrifuged and mini-prepped (Promega Wizard Plus SV miniprep kit). Purified DNA was quantified (Nanodrop ND-1000 Spectrophotometer), and 200 ng of plasmid DNA per 50 μl PCR reaction was used as template for the first round of adaptor addition. To preserve the ratio of template alleles, we used a large template concentration and few amplification cycles (16 cycles). This first PCR reaction added the Solexa paired-end sequencing oligonucleotide annealing site as well as a 3-bp barcode that indicates the origin of the sample (input or selected library, selection gate). The second PCR reaction added the remainder of the sequencing oligonucleotide annealing site and the annealing site for the flow cell oligonucleotide. All oligonucleotides were purchased from IDT as 100 nM syntheses with standard purification. Each PCR reaction included 5% dimethylsulphoxide (DMSO) and produced a single intense band on an agarose gel.

The second round PCR products were purified (ZYMO DNA clean and concentrator-5 Kit) and eluted in 20 μl dH<sub>2</sub>O. Purified PCR products were quantified (Invitrogen quant-IT picogreen dsDNA quantification kit) in triplicate using lambda-DNA as a standard. PCR products were diluted to 10 nM and 8 pmol was loaded onto a Solexa v4 PE-flow cell in the University of Texas Southwestern Genome Sequencing Core that yielded 250,000–300,000 clusters per lane.

Owing to the unbalanced nature of the first bases of each PCR product, a PhiX control lane was used for matrix and phasing calculations as per manufacturer recommendation.

Sequences from the Illumina RTA base-caller were imported into CLC Genomics Workbench as `.qseq` files and trimmed for quality using a cut-off of 0.05 for the modified Mott algorithm. Bases that did not pass the trim filter were deleted from each read, and reads shorter than 49 bp were discarded. Reads were sorted into groups according to the 3-bp barcode contained in each PCR product, and barcode groups were exported as FASTQ files for further analysis. Custom software written in MATLAB was used to count the number of occurrences of each allele in each population. The functional effect of each allele was calculated as the average of the value from the two FACS gates (Supplementary Fig. 4).