SUPPLEMENTAL MATERIALS

Circular permutation profiling by deep sequencing libraries created using transposon mutagenesis

Joshua T. Atkinson[1, ‡], Alicia M. Jones[2, ‡], Quan Zhou,[3]
and Jonathan J. Silberg[2,4,*]

Author affiliations:

1. Systems, Synthetic, and Physical Biology Graduate Program, Rice University, 6100 Main MS-180, Houston, Texas 77005
2. Department of BioSciences, Rice University, MS-140, 6100 Main Street, Houston, TX, 77005
3. Department of Statistics, Rice University, 6100 Main Street, Houston, TX, 77005
4. Department of Bioengineering, Rice University, 6100 Main Street, Houston, TX, 77005

*To whom correspondence should be addressed: Jonathan J. Silberg, Biosciences Department, 6100 Main Street, Houston TX 77005; Tel: 713-348-3849; Email: joff@rice.edu

‡These authors contributed equally to this manuscript.

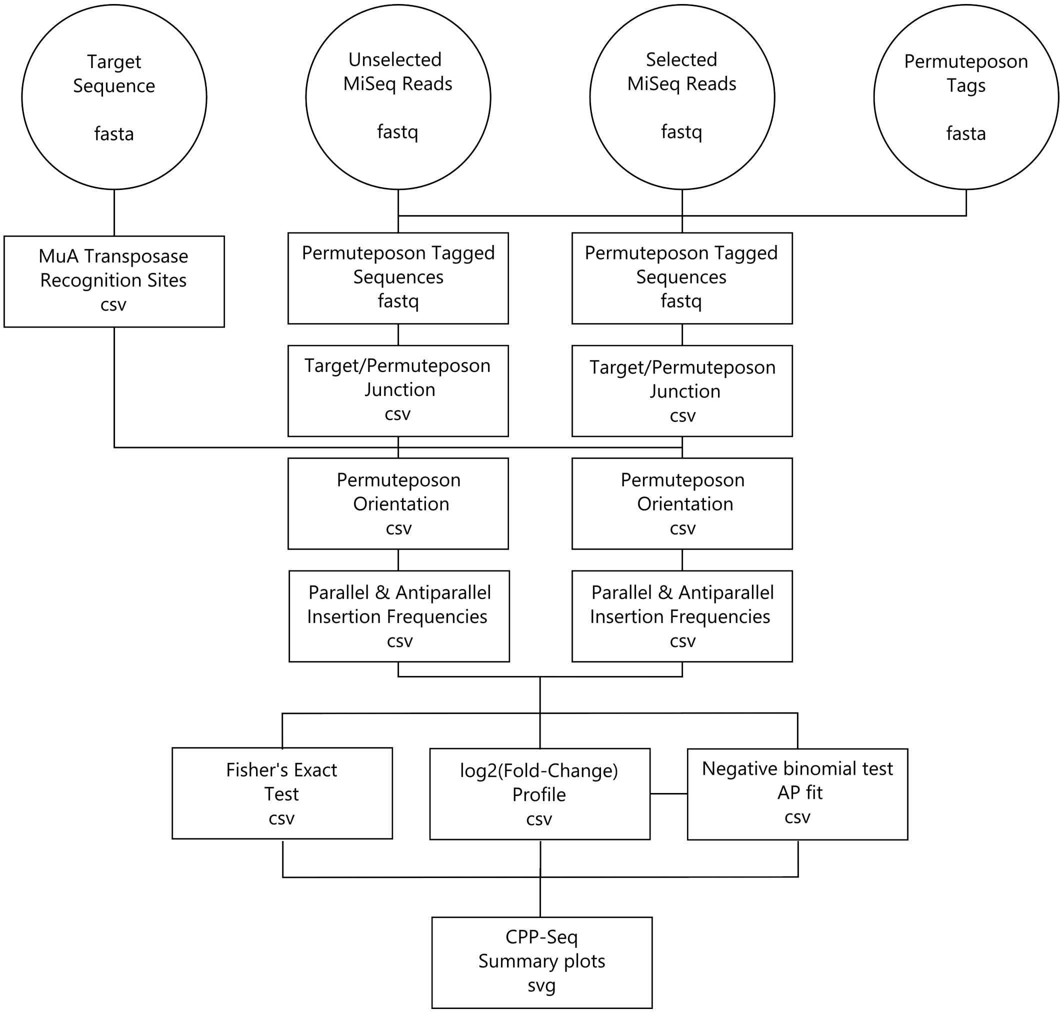

Figure S1. CPP-seq sequence processing pipeline. (1) Four inputs are used with the python script, including the start and stop motifs (Permuteposon tags) shown in Figure 2, the target gene sequence including the linker used to circularize the gene, and deep sequencing data from the unselected and selected libraries; (2) sequences are filtered to identify reads with Permuteposon tags; (3) the 11 bp of AK gene sequence adjacent to the junction is then extracted and searched against a list of all possible 11 bp sequences in the target gene to identify (4) the orientation, *i.e.*, P or AP, and (5) sequence of each read; and (6) the counts from the unselected and selected libraries are compared to asses which present significant changes between the frequency of P and AP in the selected and unselected libraries using Fisher's exact test and a negative binomial model. A profile is generated in this latter step that relates the significance of sequence enrichment to primary structure. (7) Summary plots are generated of the data.

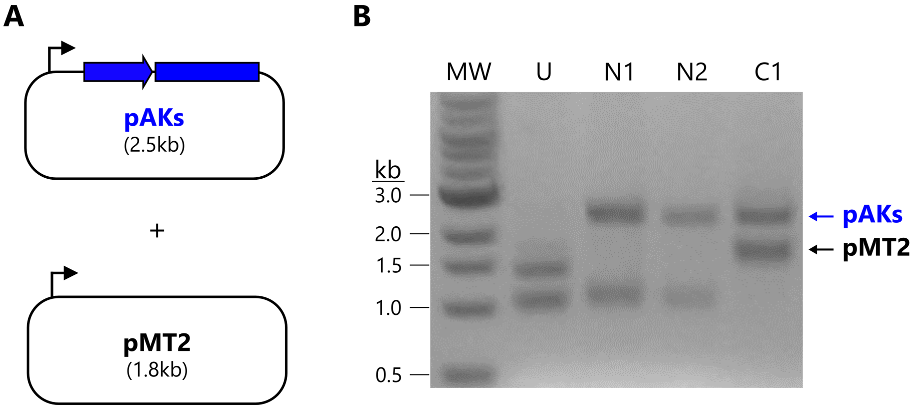

Figure S2. The MuA insertion reaction yields a mixed plasmid population. (**A**) Two plasmids were observed following purification of the naïve library from cells that had been transformed with the library. These include pMT2, which arises from circularlization of the linear P1 permuteposon used for library construction, and vectors encoding different circularly permuted AK genes, designated pAKs. (**B**) The size distribution of the vectors were analyzed before (U) and after incubating with NotI (N1 and N2) and ClaI (C1). Digestion with NotI is expected to only linearize pAKs (not pMT2) because it cuts at a single site within each permuted AK gene. The linear DNA arising from this digestion has a calculated molecular weight of 2.5 kb. ClaI cuts at a single site within both pMT2 and pAKs to yield linear DNA with distinct calculated molecular weights of 1.8 and 2.5 kb, respectively.

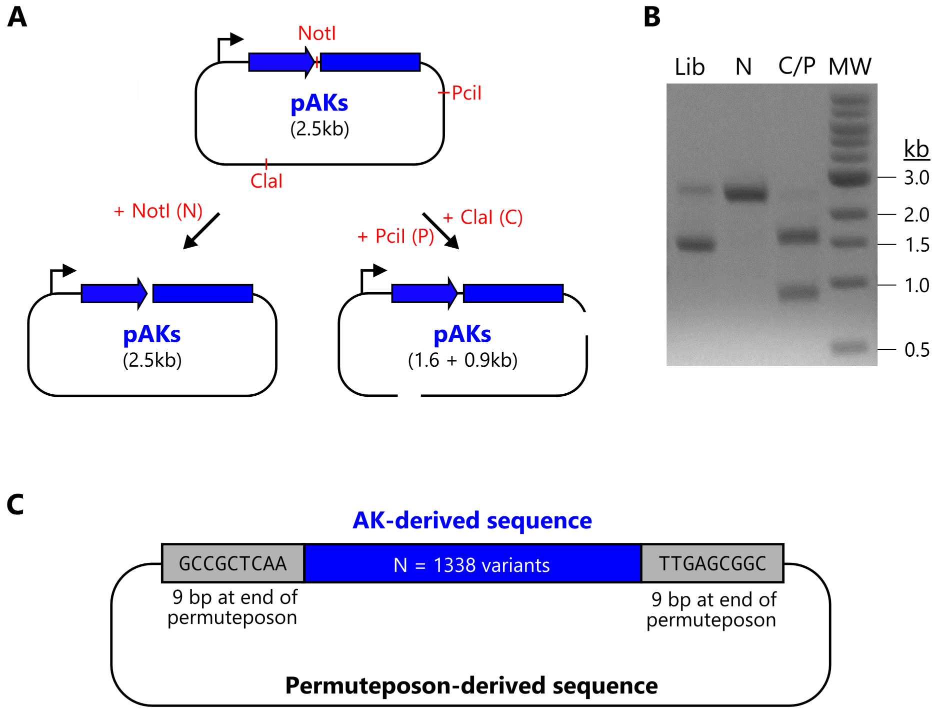

Figure S3. The naïve library appears homogeneous following purification. (**A**) Restriction digest of pAKs with NotI is expected to yield a single 2.5 kb fragment while a ClaI and PciI double digest is expected to yield 1.6 and 0.9 kb fragments. (**B**) Agarose gel electrophoresis analysis of the library before (Lib) and after incubation with NotI (N) or a ClaI/PciI (C/P) mixture reveals the expected bands. The smaller of the two fragments generated by the ClaI/PciI double digest was used for deep mutational scanning. (**C**) To determine the abundance of sequence reads containing a permuteposon, we analyzed the prevalence of a 9 base pair inverted repeat sequence, which is encoded at both ends of the permuteposon.

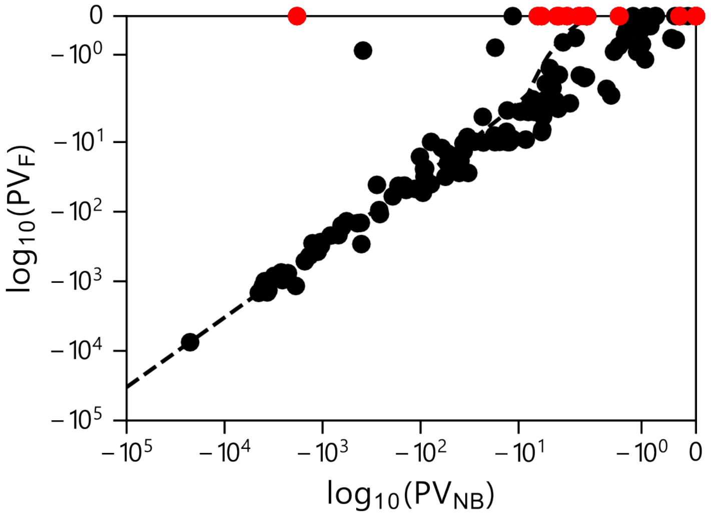

Figure S4. Comparing the p values from Fisher's Test (P_F) and the negative binomial model (P_{NB}). A linear correlation is observed between the two statistical measures (y = 0.332x + 0.631; R^2 = 0.99). The negative binomial model allows for the statistical assessment of 42 additional variants (red circle) beyond the 129 that can be evaluated using Fisher's exact test (black circles). Among the 42 variants, ten presented p values < 0.01

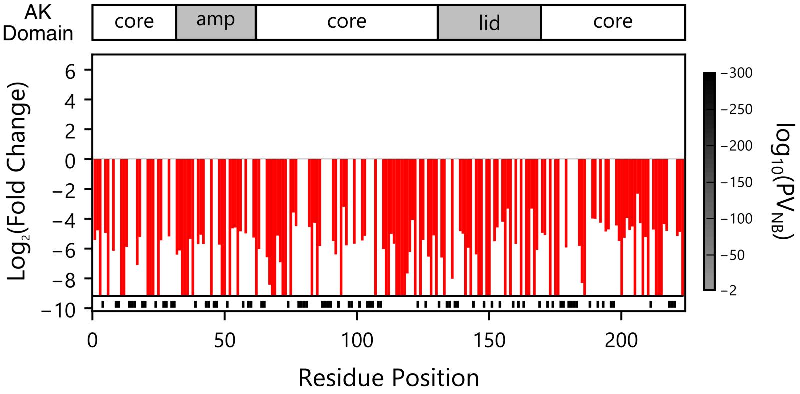

Figure S5. Relationship between AK domain structure and the abundance of AP variants. For each AP variant, the $\log_2$(fold change) is mapped onto the domain location using the AK residue encoded at the beginning of each circularly permuted gene. Each variant is colored based on the p value obtained for the enrichment of P variants as in Figure 5. Variants no longer observed in the selected library (infinitely diluted) are shown as bars that reach the line at the bottom of the graph. Those cognate P and AP variant pairs absent from both the unselected and selected data sets are indicated as black bars below the x axis.

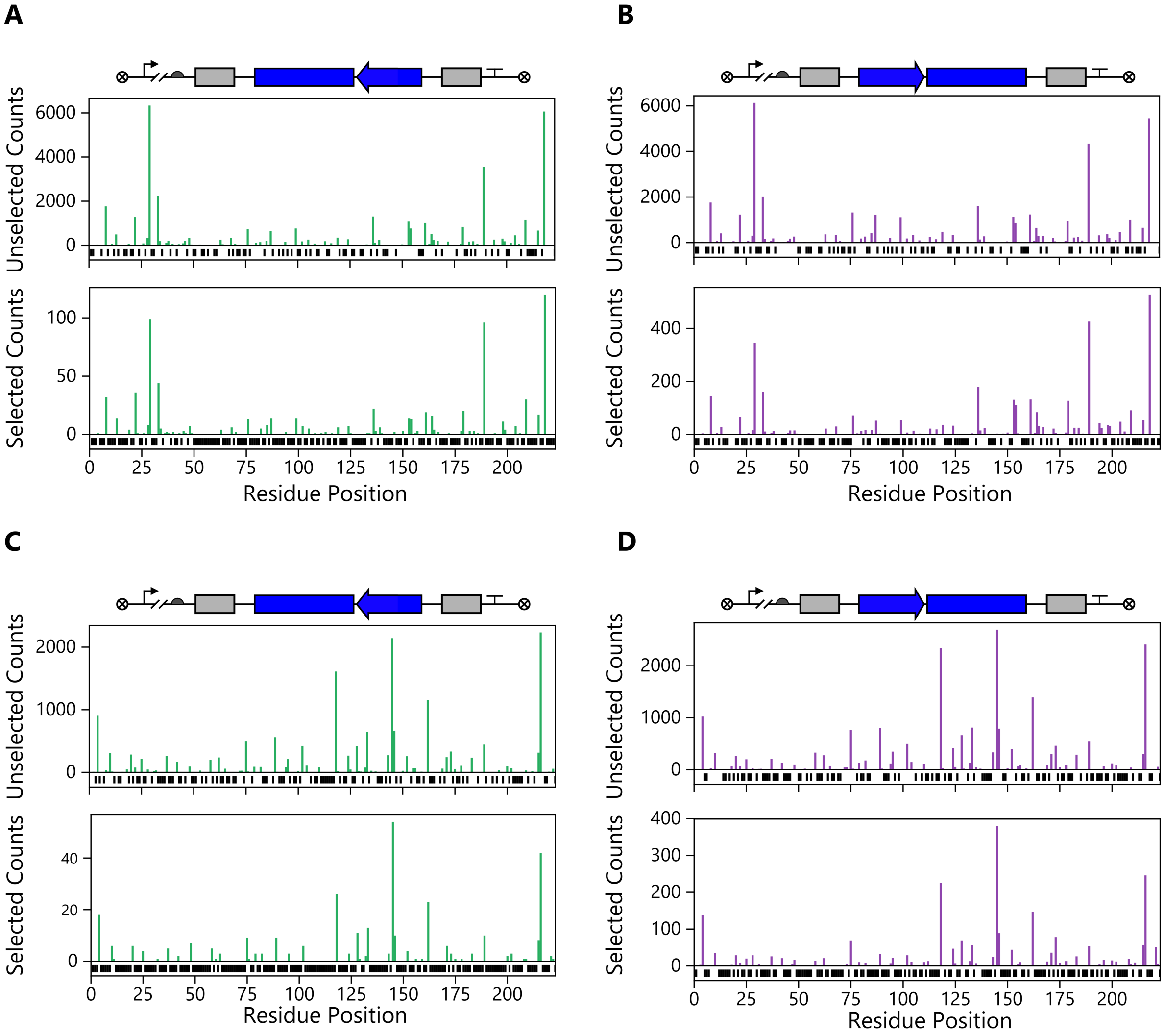

Figure S6. Relationship between sequence abundance of the +1 and -1 frame variants and the AK codon found at the beginning of each permuted gene. A comparison of the number of (**A**) AP and (**B**) P sequences in the +1 frame before (top) and after selecting (bottom) for biological activity. (**C**) A comparison of the number of AP and (**D**) P sequences in the -1 frame before (top) and after selecting (bottom) for biological activity. The residue position represents the AK residue found at the beginning of each ORF. In cases where a P or AP variant was absent, black bars are shown below the x axis.

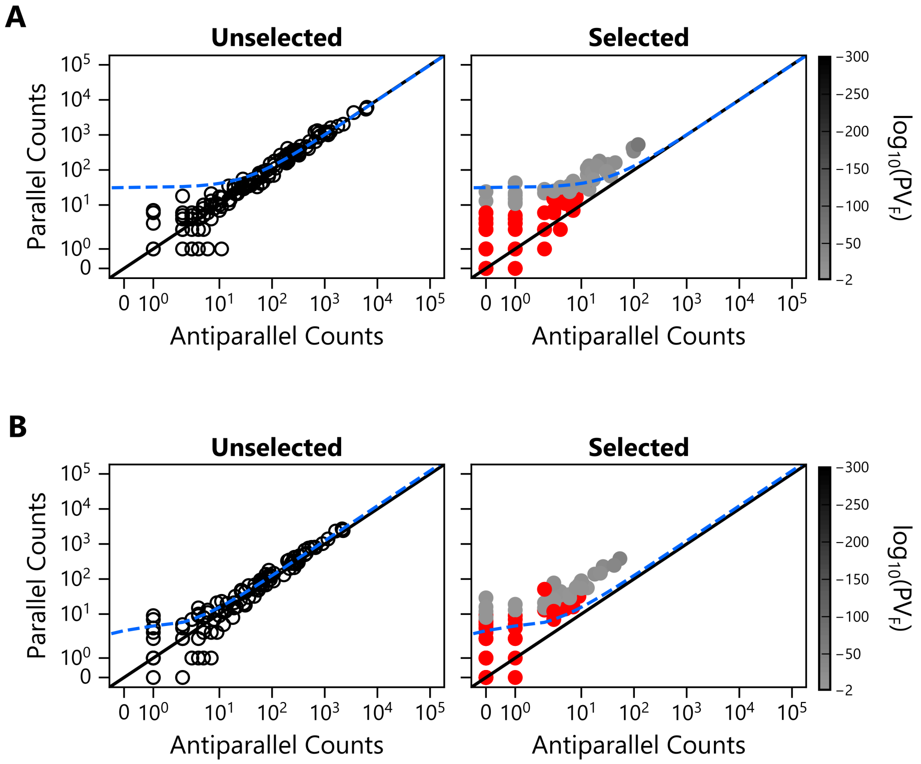

Figure S7. Comparing abundances of identical P and AP in the +1 and -1 frames. In the (**A**) +1 and (**B**) -1 frames, cognate P and AP sequences in the unselected library (open symbols) display a linear correlation shown in purple (y = 0.982x + 31.517; R^2 > 0.98 for +1 frame and y = 1.226x + 3.357; R^2 > 0.98 for -1 frame). Following the selection (closed symbols), some variants have higher AP to P ratio than that observed before the selection. However they still display a strong linear correlation (y = 4.105x + 2.643; R^2 > 0.94 for +1 frame and y = 6.647x + 1.895; R^2 > 0.94 for -1 frame). Selected variants are colored as a function of the p value obtained from Fisher's Exact Test, with variants presenting p-values >0.01 in red, variants displaying p-values ≤10^{-300} in black and those displaying intermediate values shaded as indicated in the bar.

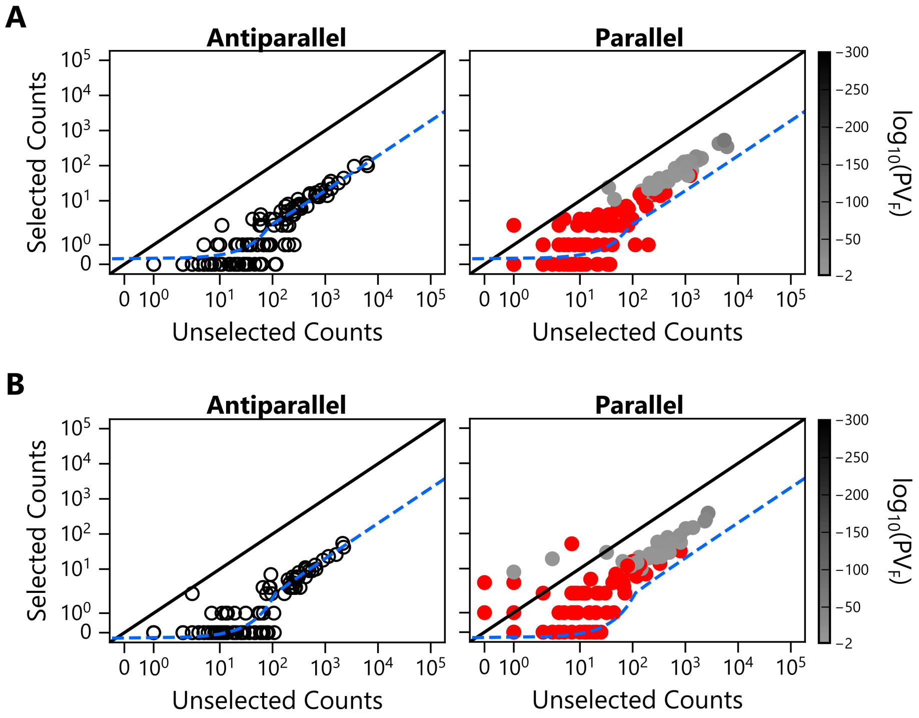

Figure S8. Effect of selection on sequence abundance in the +1 and -1 frames. The abundance of each P and AP sequence in (**A**) the +1 frame and (**B**) -1 frame before and after selection. The black lines represent the expectation if there was no dilution or enrichment following the selection. The AP variants (open circles) display a strong linear correlation shown in blue (y = 0.019x + 0.292; R^2 > 0.95 for +1 frame and y = 0.02x - 0.273; R^2 > 0.95 for -1 frame). Selected variants are colored as a function of the p value obtained from Fisher's Exact Test, with variants presenting p-values >0.01 in red, variants displaying p-values ≤10^{-300} in black and those displaying intermediate values shaded as indicated in the bar. The P variants also display a linear correlation (y = 0.08x + 1.199; R^2 > 0.92 for +1 frame and y = 0.111x - 0.475; R^2 > 0.93 for -1 frame).

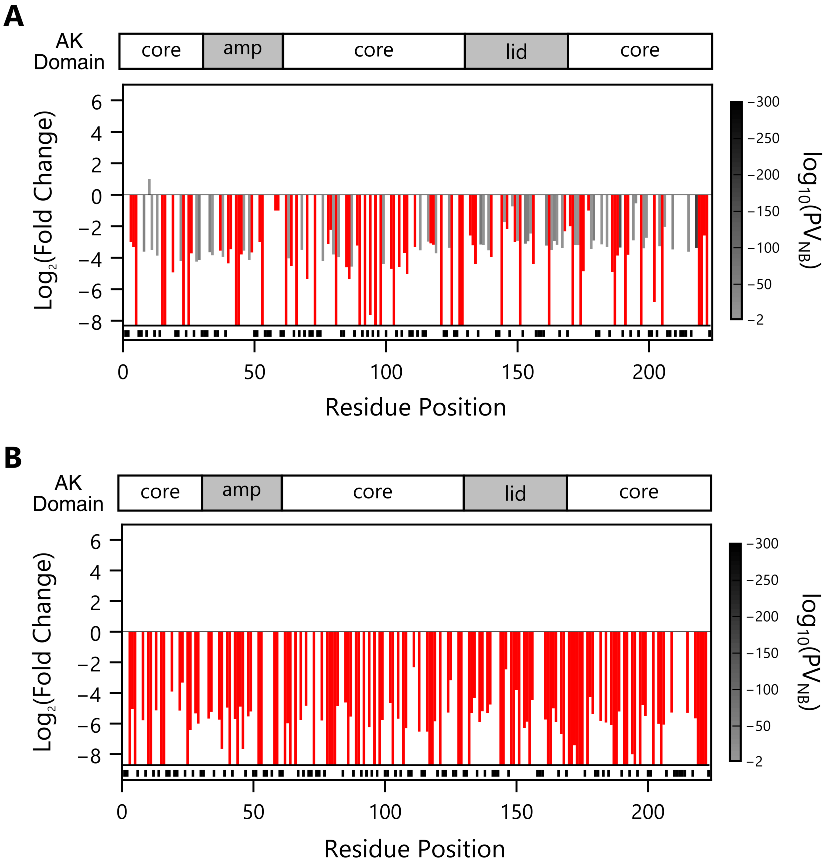

Figure S9. Structure and sequence enrichment for the +1 frame variants. For each P variant (**A**) and AP variant (**B**) in the +1 frame, the $\log_2$(fold change) is mapped onto the AK domain location of the residue encoded by the first codon in the permuted gene and color coded based on the p-values obtained using the approach outlined in Figure 6. Those cognate P and AP variant pairs absent from both the unselected and selected data sets are indicated as black bars shown below the x axis.

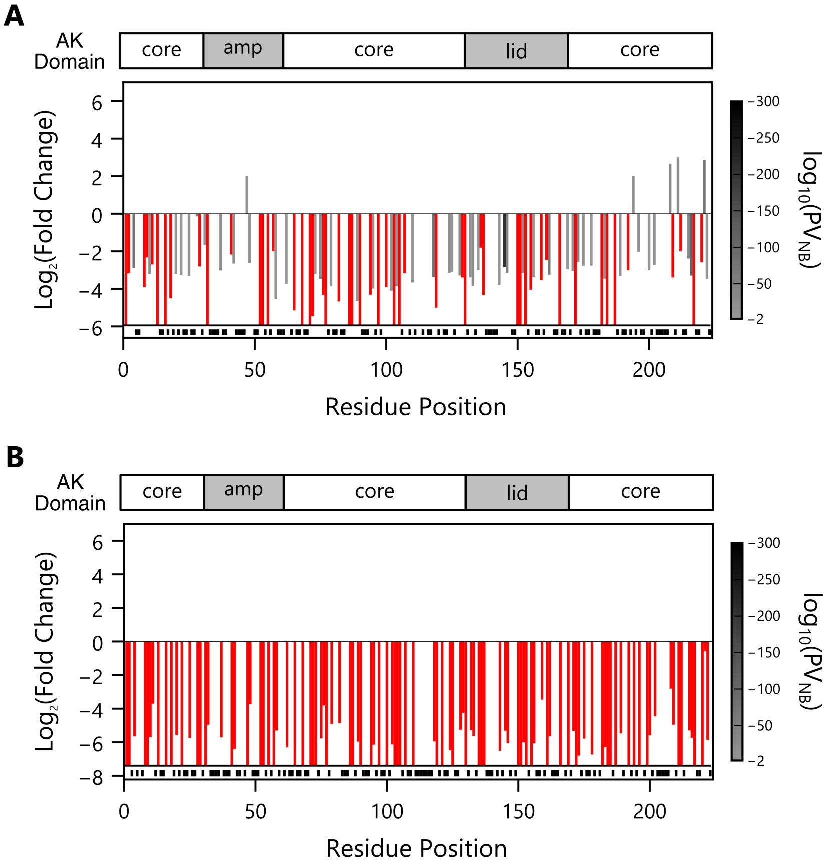

Figure S10. Structure and sequence enrichment for the -1 frame variants. For each P variant (**A**) and AP variant (**B**) in the **-**1 frame, the $\log_2$(fold change) is mapped onto the AK domain location of the residue encoded by the first codon in the permuted gene and color coded based on the p-values obtained using the approach outlined in Figure 6.