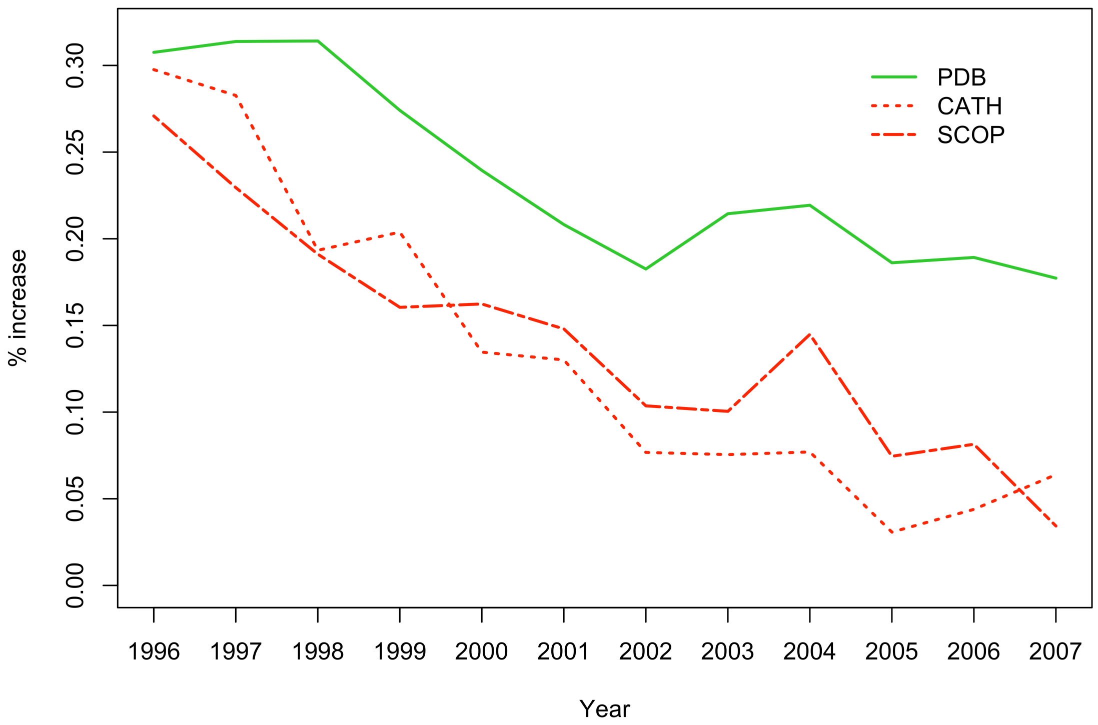
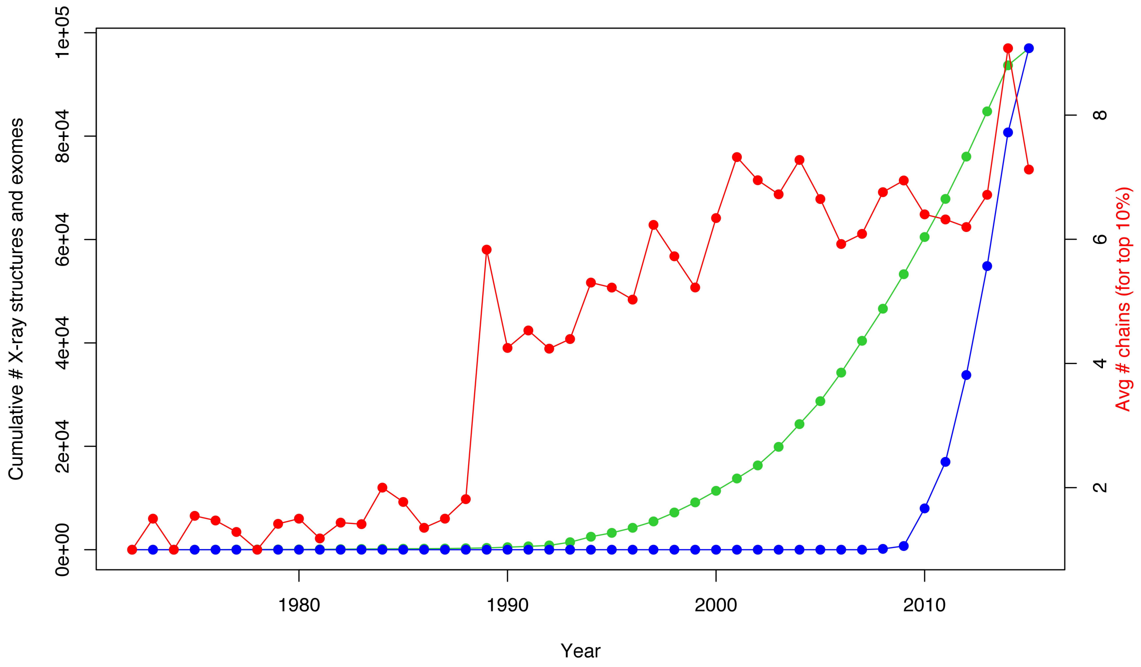
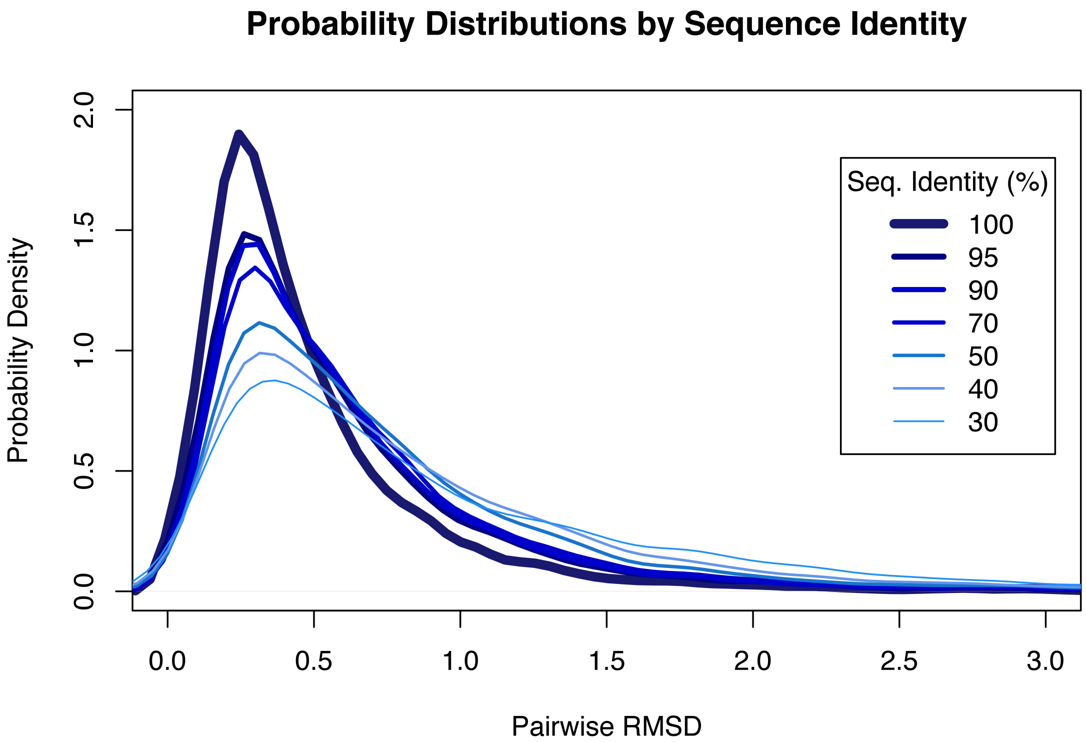


Supp. Fig. 1a: The growth rate of deposited PDB structures from 1996 to 2007, and the concomitant growth rate in the number of folds (as defined by CATH and SCOP). The growing appreciation for dynamic behavior and the importance of conformational heterogeneity is being facilitated by a growing redundancy within the PDB. Such redundancy is represented, for instance, when the same protein is structurally resolved under different conditions, potentially resulting in alternative conformations.

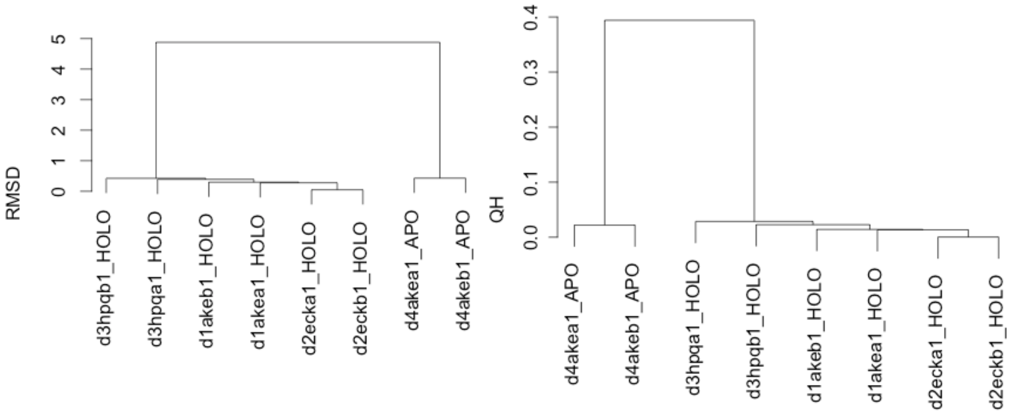


Supp. Fig. 1b: Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa): The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature. Red: Average number of chains per PDB (considering the biological assembly PDB files for the top 10% of PDBs for a each year, as ordered by the number of chains for each structure). Green: Cumulative number of X-Ray structures deposited in the PDB. Blue: Cumulative number of exomes stored in the NCBI Sequence Read Archive (SRA). All data were downloaded in May 2015.

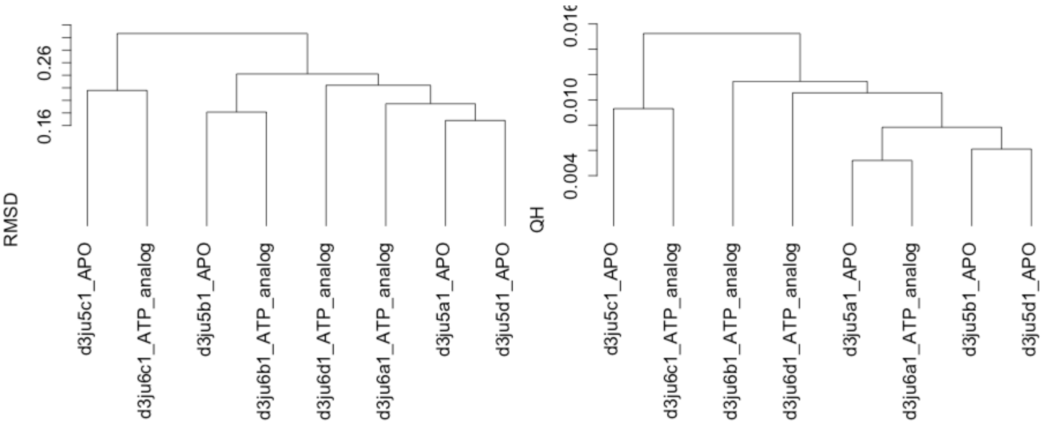


Supp. Fig. 2: Distributions for average pairwise RMSD values across domains within all multiple structure alignments at varying levels of sequence identity.

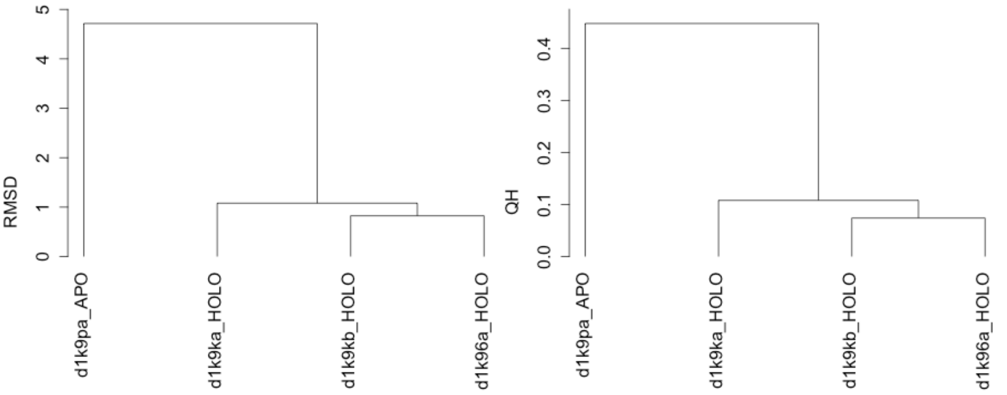
Supp. Fig. 3: Clustering based on RMSD generally matches that used when clustering based on Q_H.



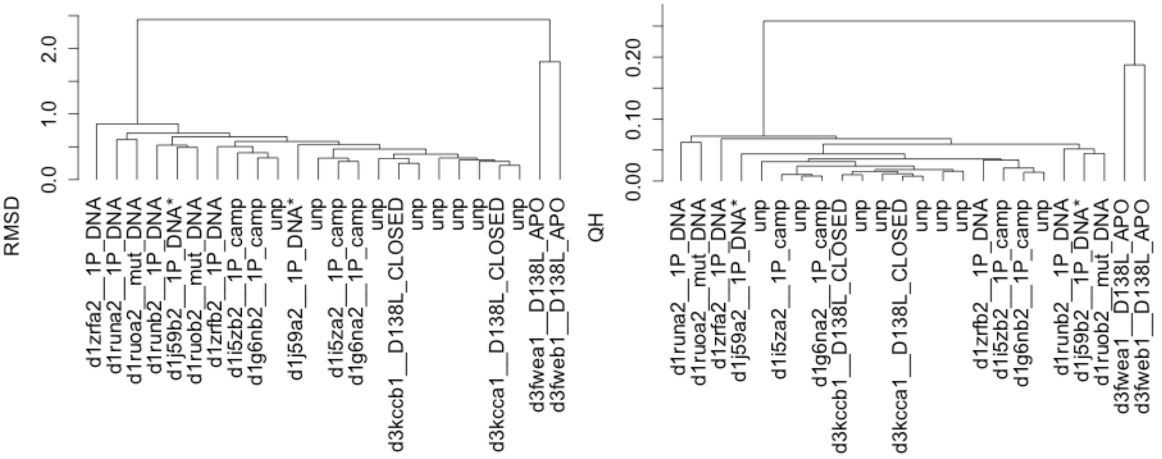
3a) Adenylate kinase



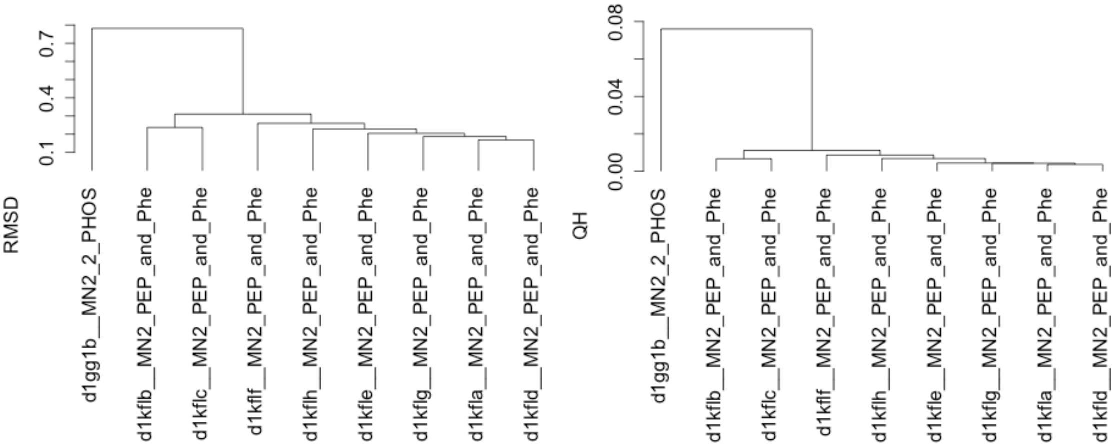
3b) Arginine kinase



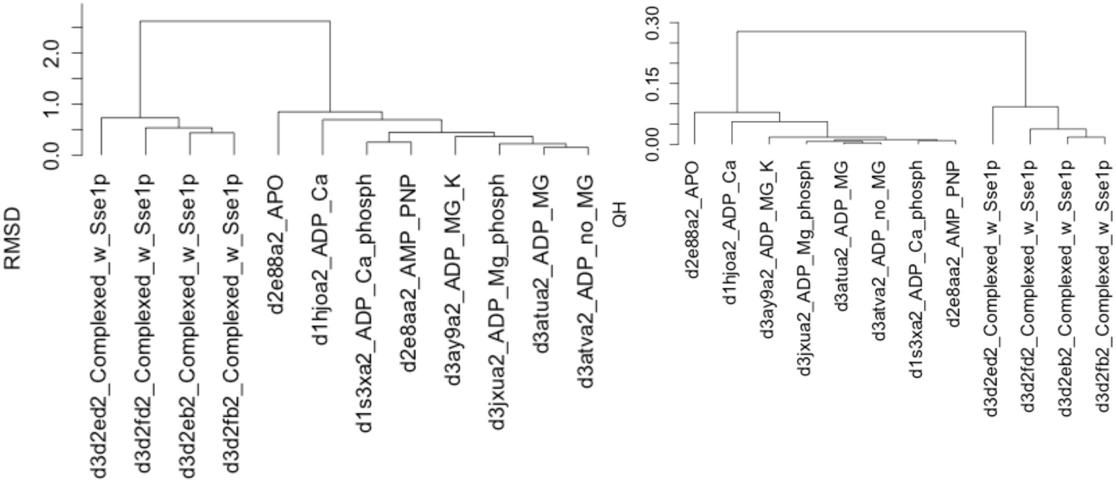
3c) Calcyclin



3d) Catabolite activator protein (CAP)

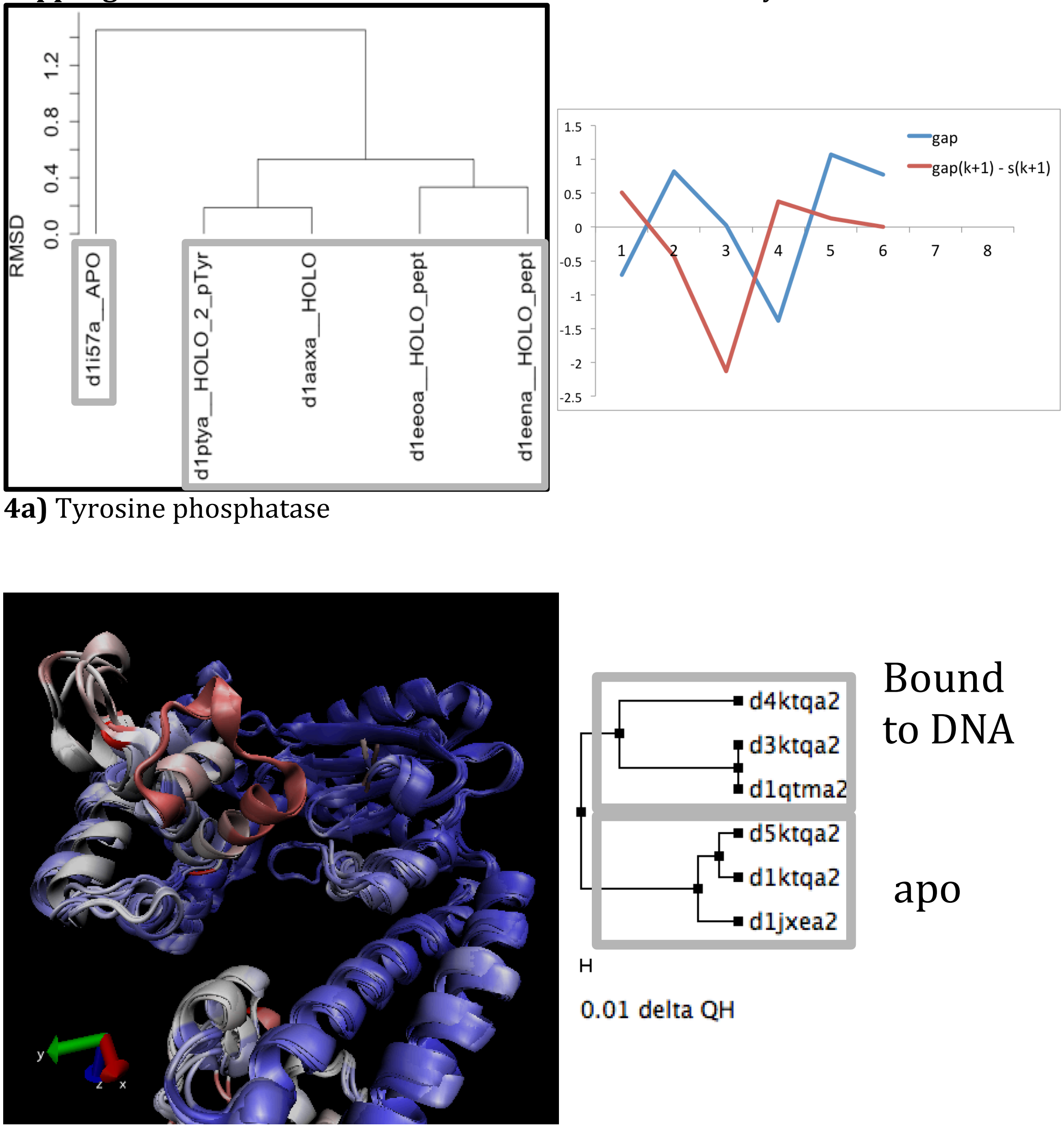


3e) DAHPS



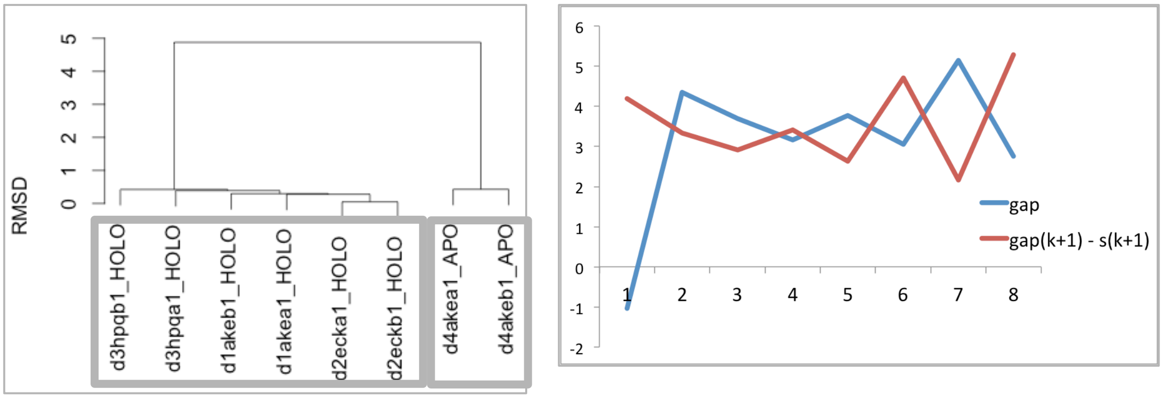
3f) hsp ATPase

Supp Fig. 4: K-values and annotations for several canonical systems.

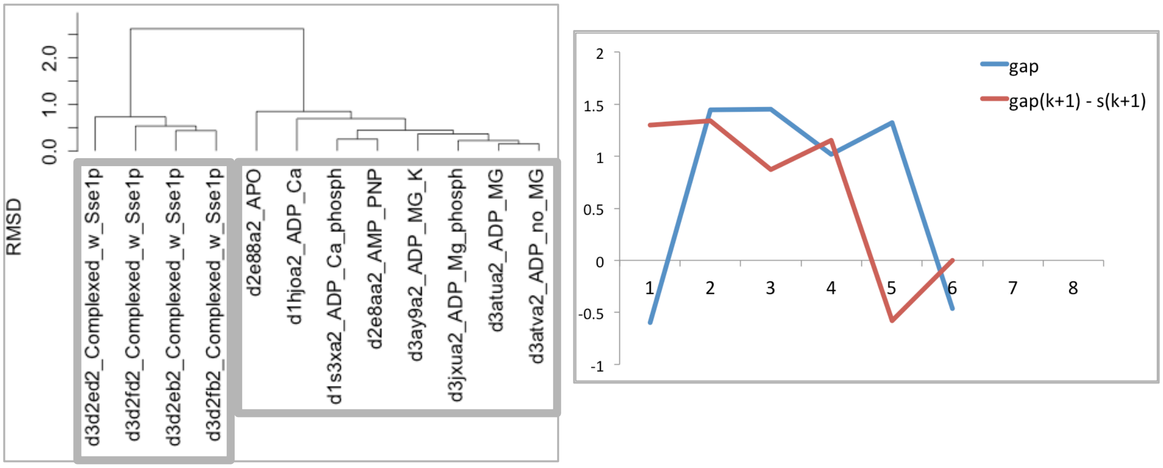


4a) Tyrosine phosphatase

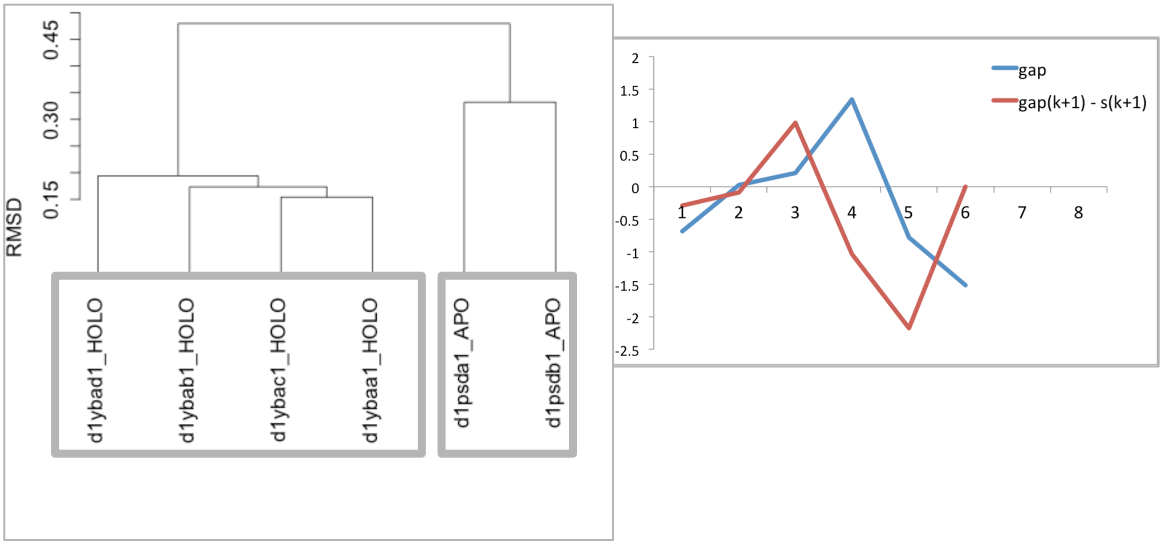
4b) DNA Pol I



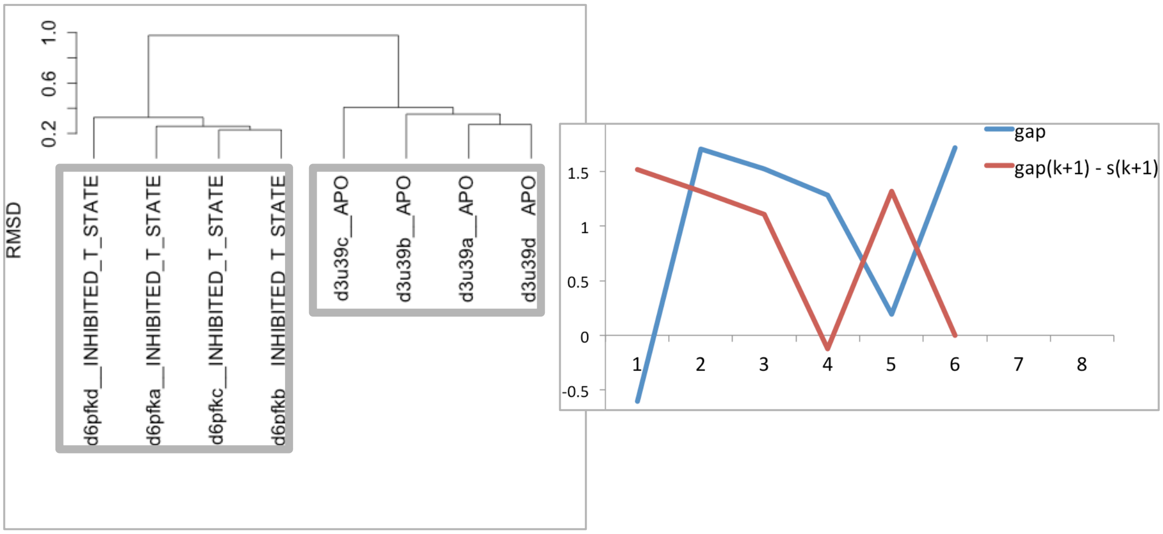
4c) Adenylate kinase



4d) hsp ATPase

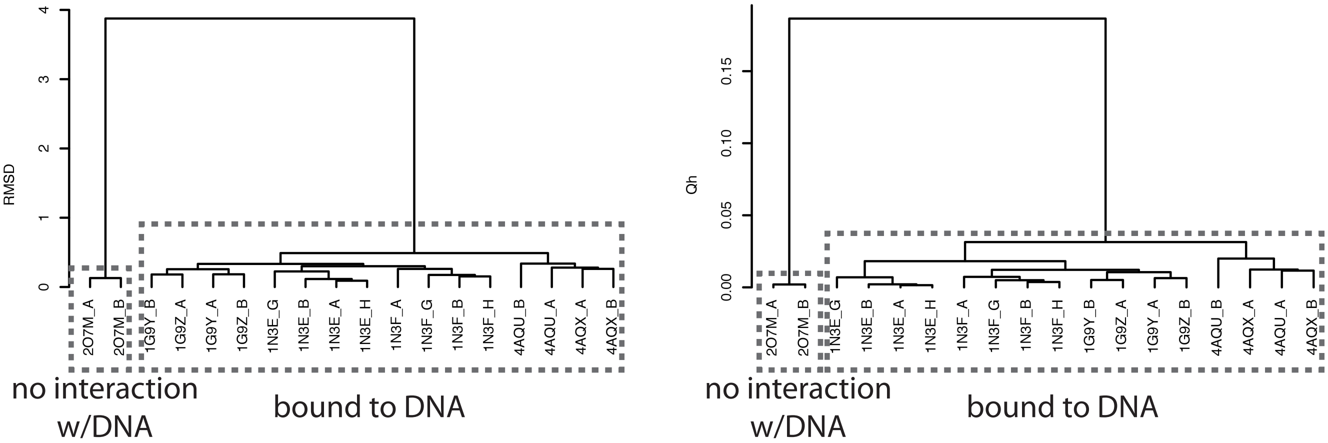


4e) Phosphoglycerate dehydrogenase

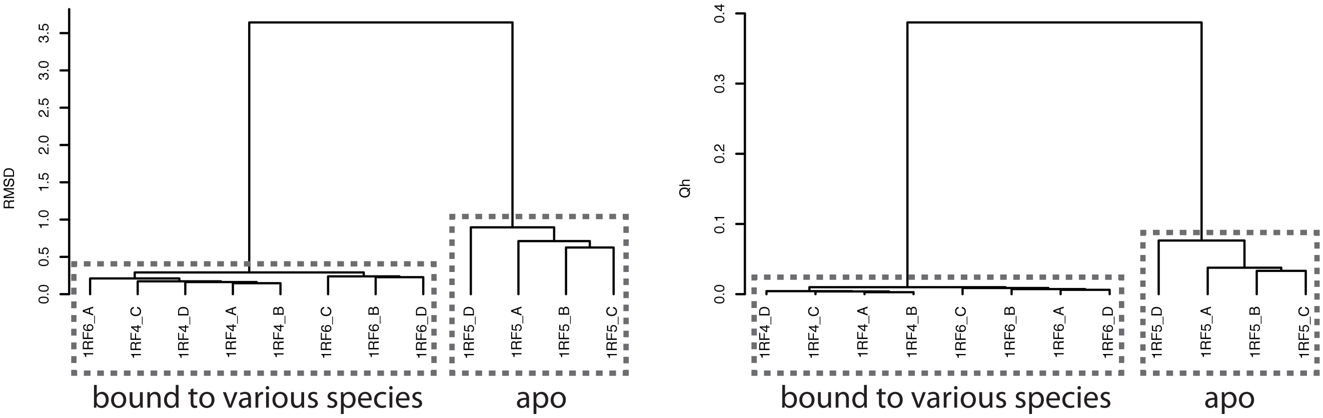


4f) Phosphfructokinase

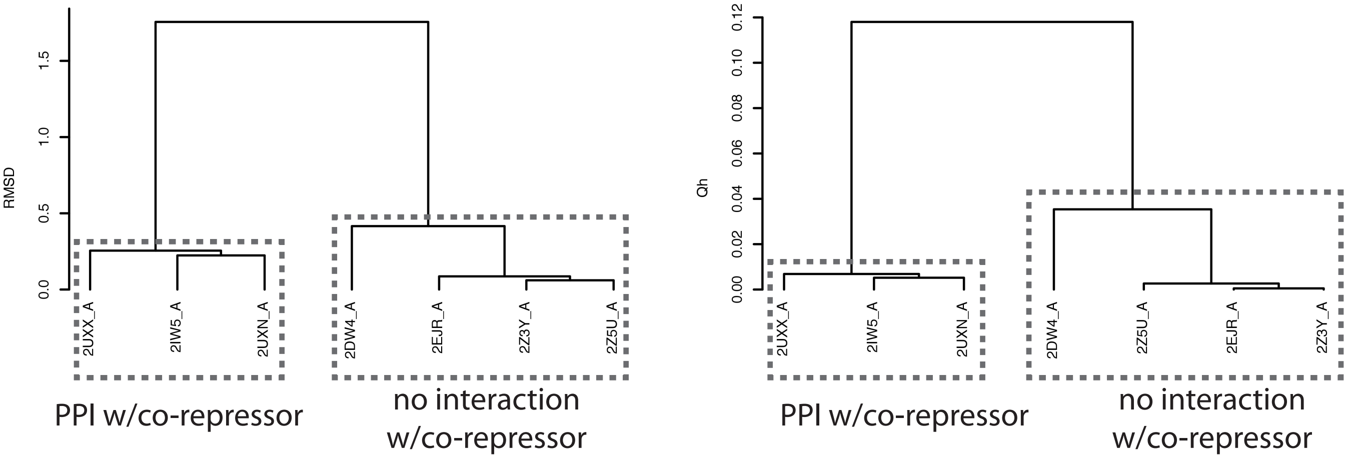
4g) I-CreI



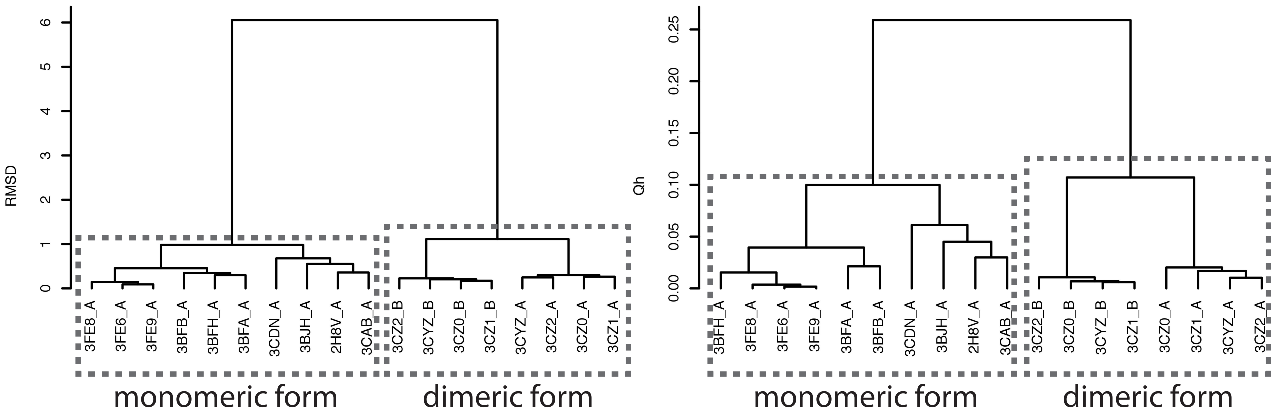
4h) 5-enolpyruvylshikimate-3-phosphate synthase



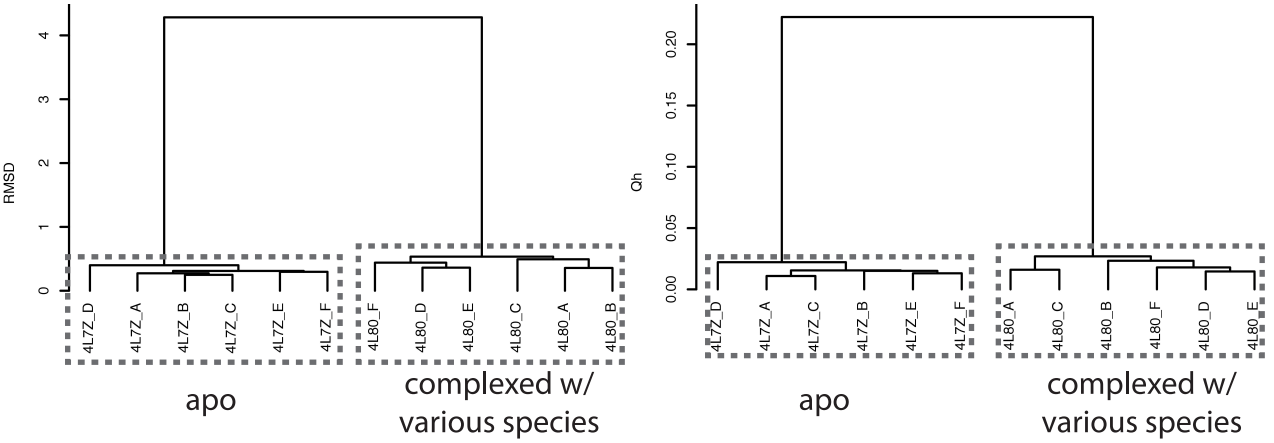
4i) Lysine-specific histone demethylase 1



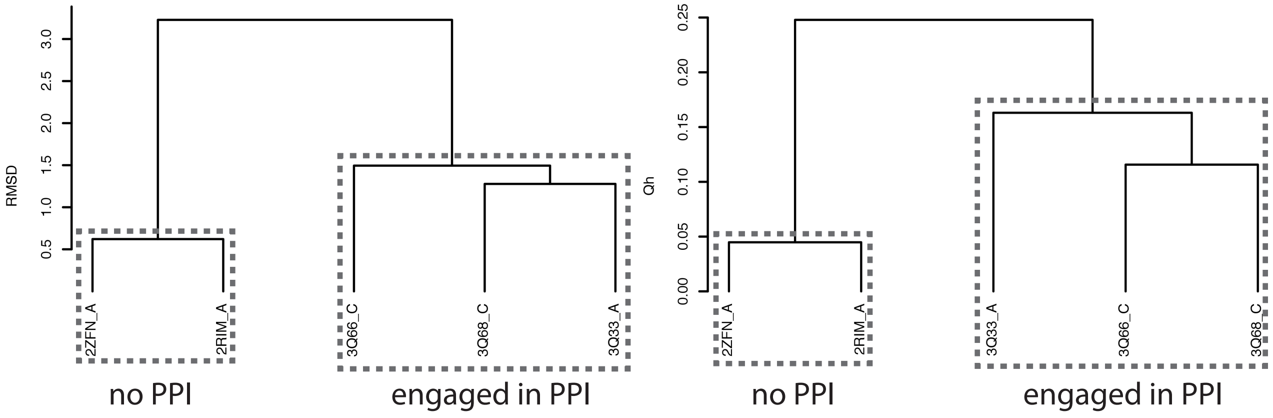
4j) Pheromone binding protein



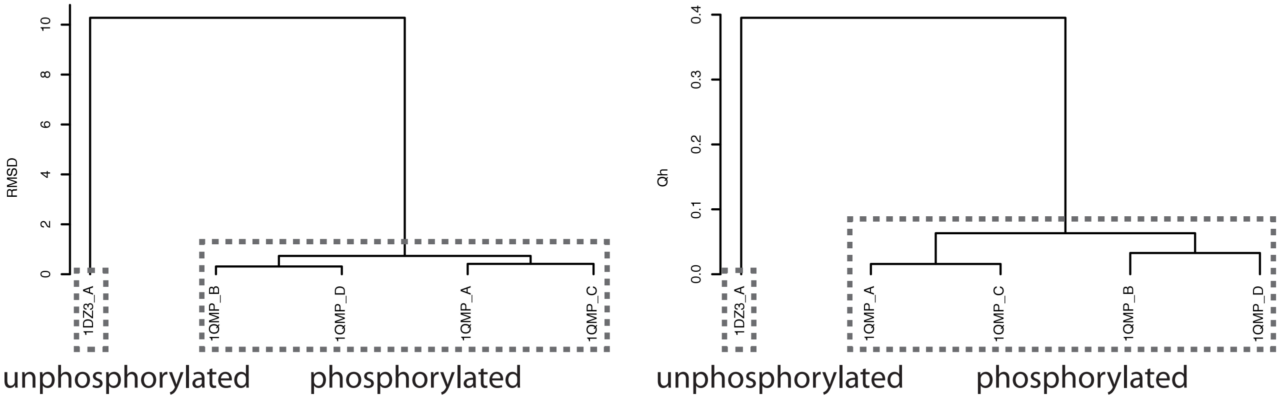
4k) HpaI aldolase



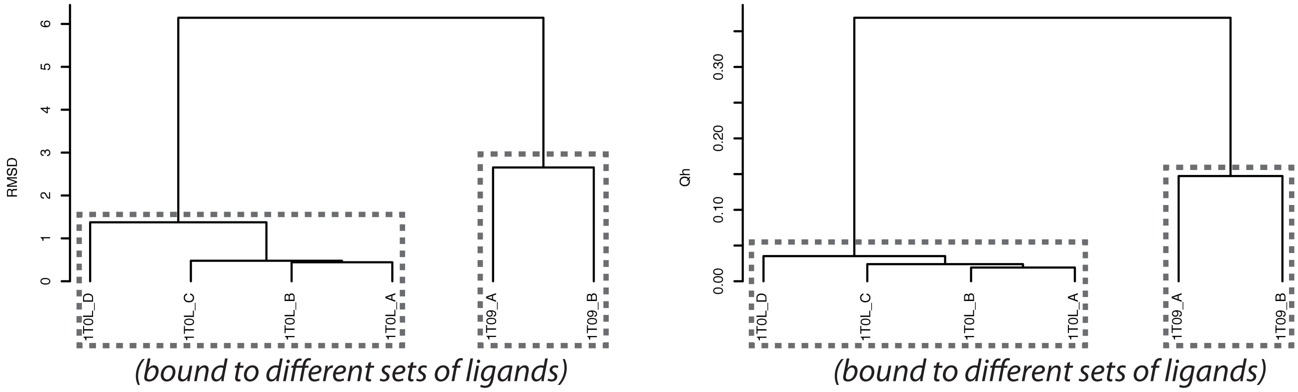
4l) Histone acetyltransferase RTT109



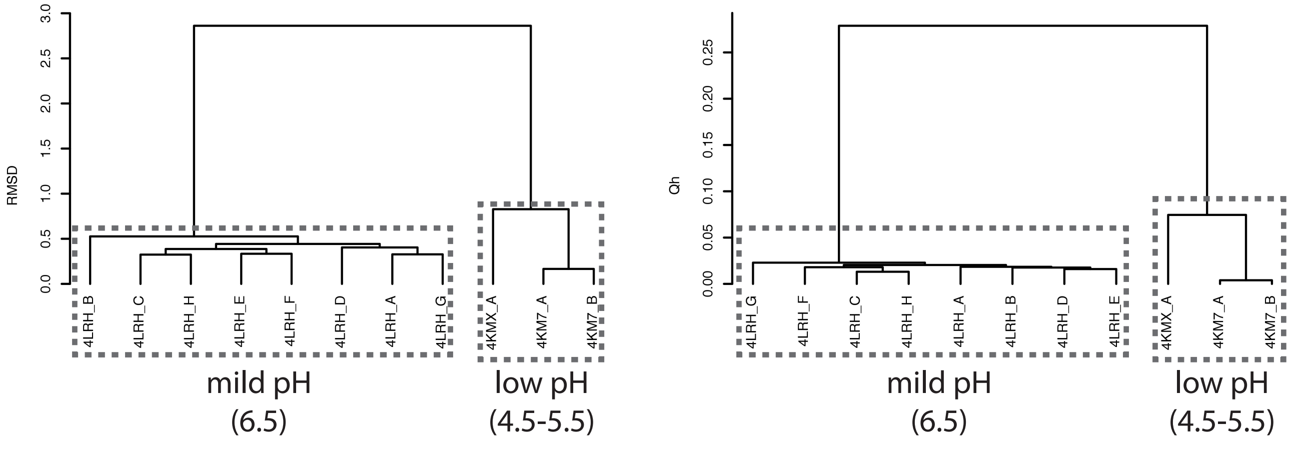
4m) Stage 0 sporulation protein A



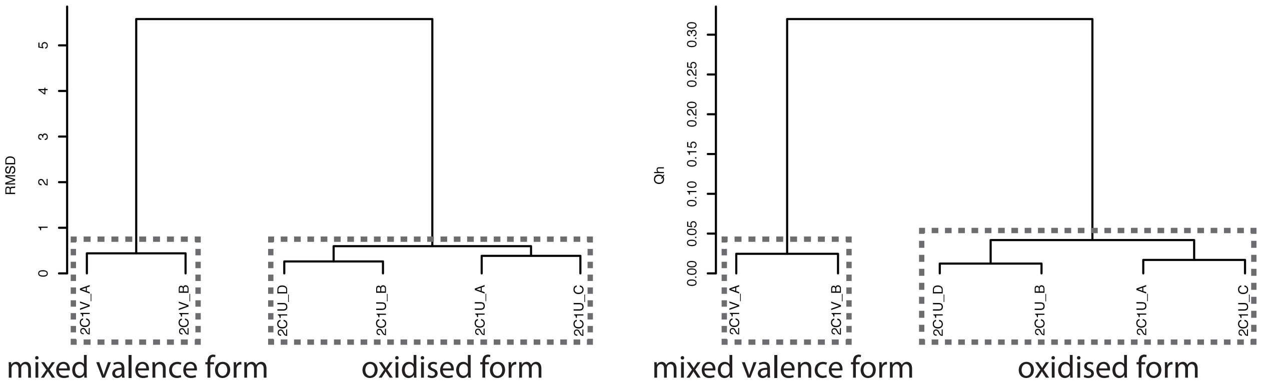
4n) Isocitrate dehydrogenase



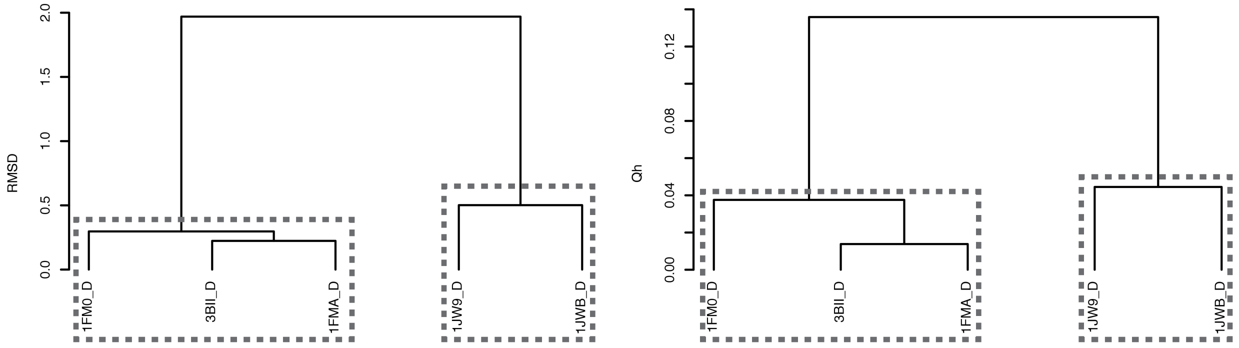
4o) Human folate receptor alpha



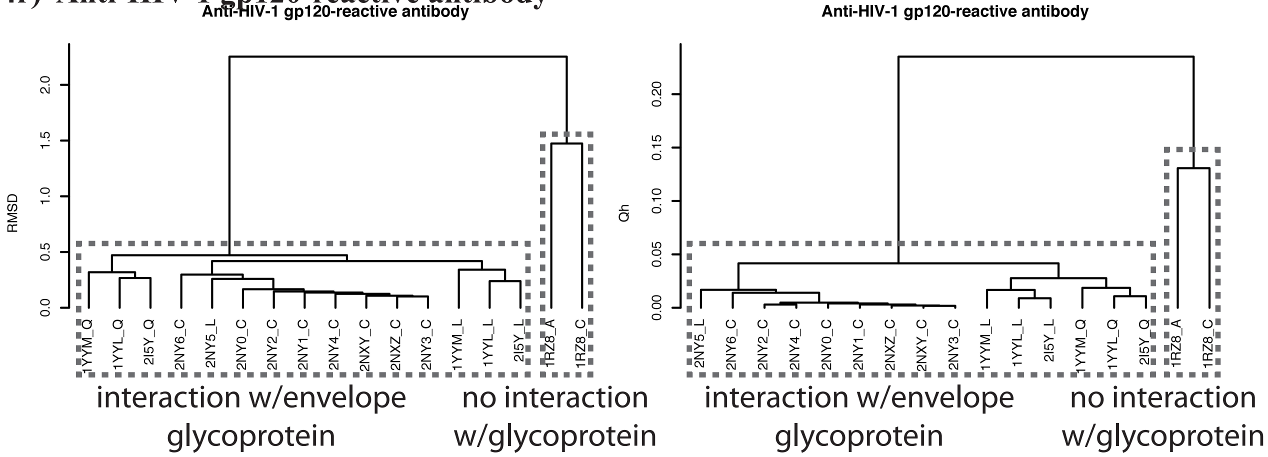
4p) Cytochrome C peroxidase



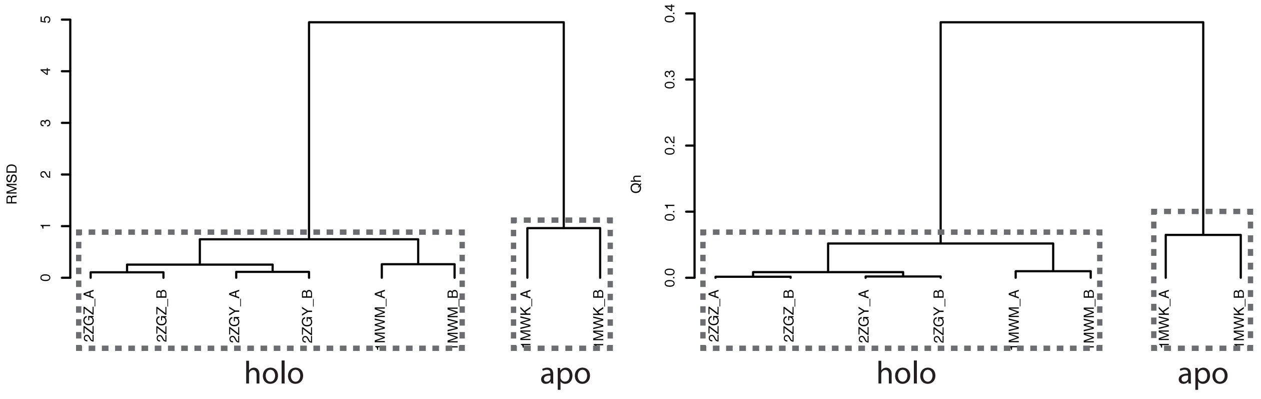
4q) Molybdopterin-converting factor subunit 1



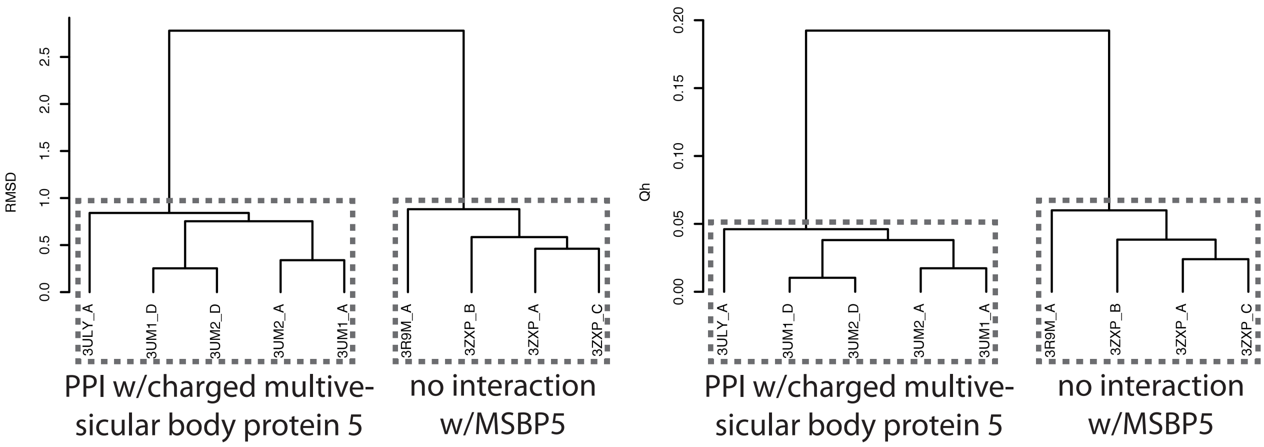
4r) Anti-HIV-1 gp120-reactive antibody



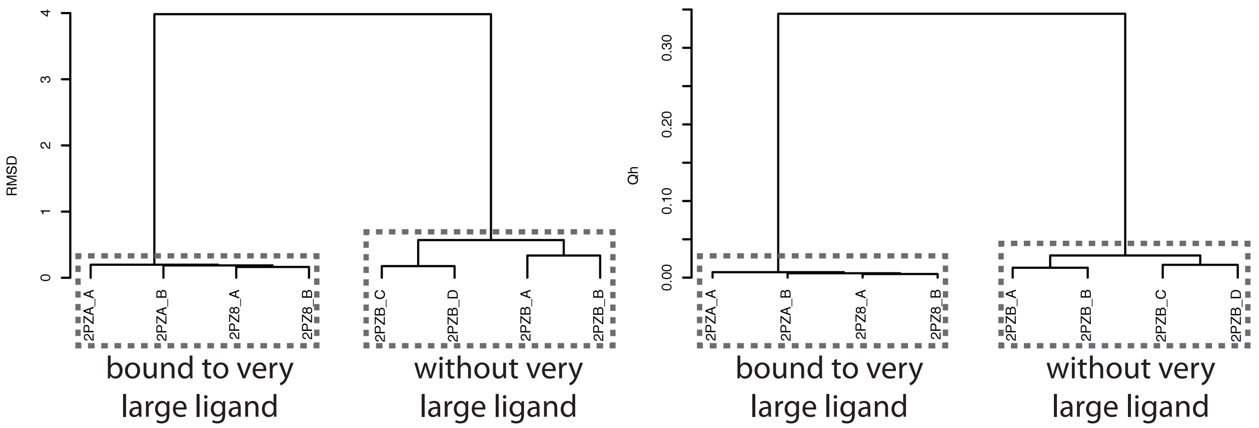
4s) Plasmid segregation protein parM



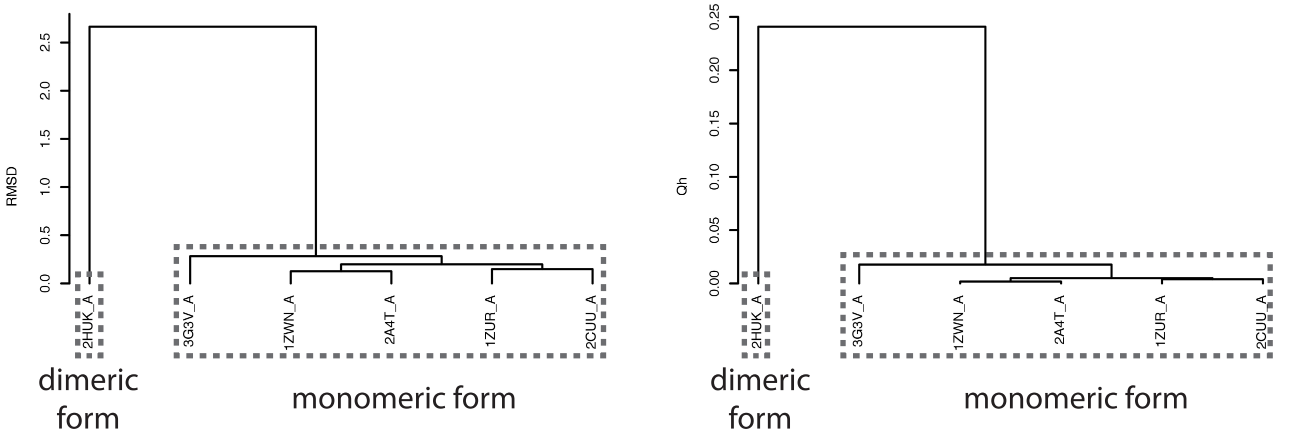
4t) BRO1 domain-containing protein BROX



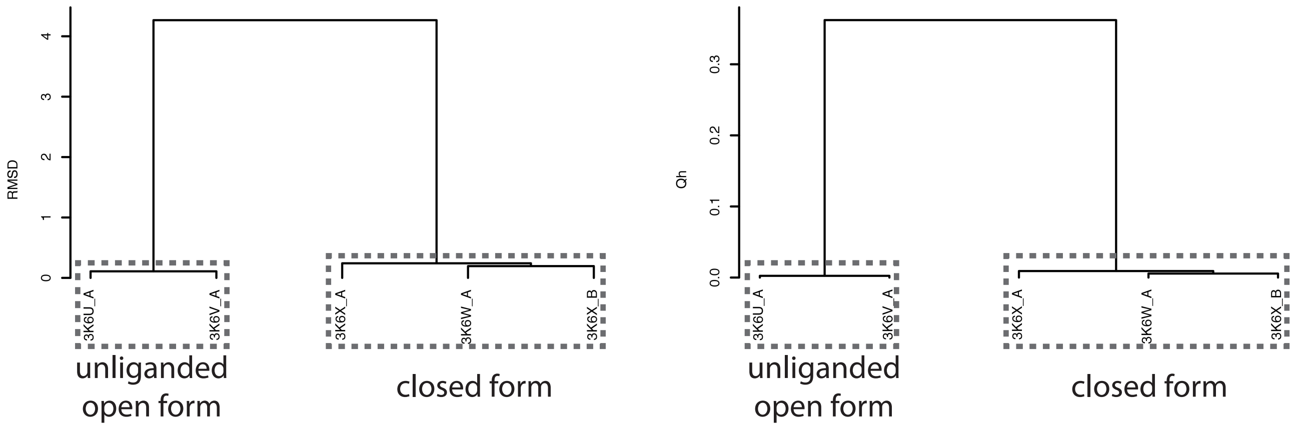
4u) NAD+ synthetase



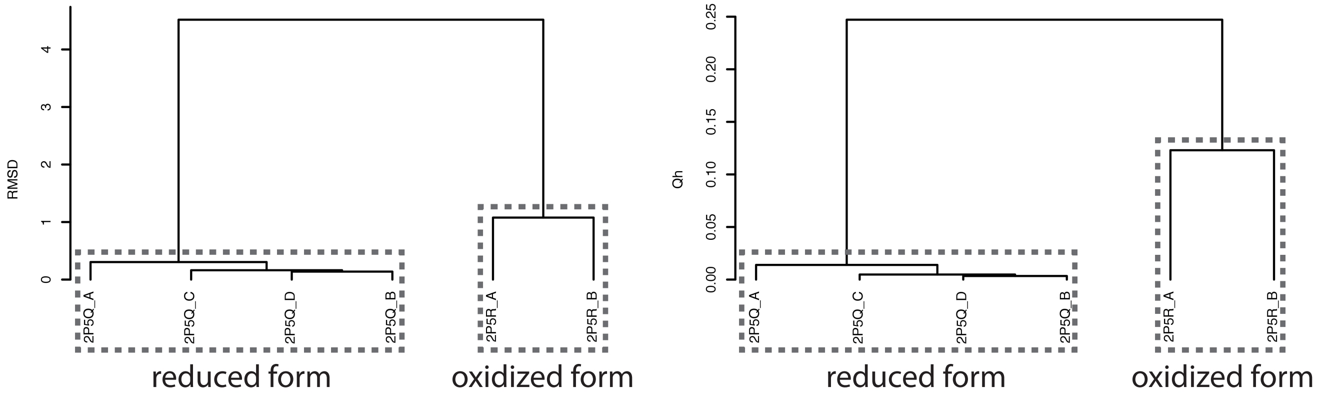
4v) Lysozyme



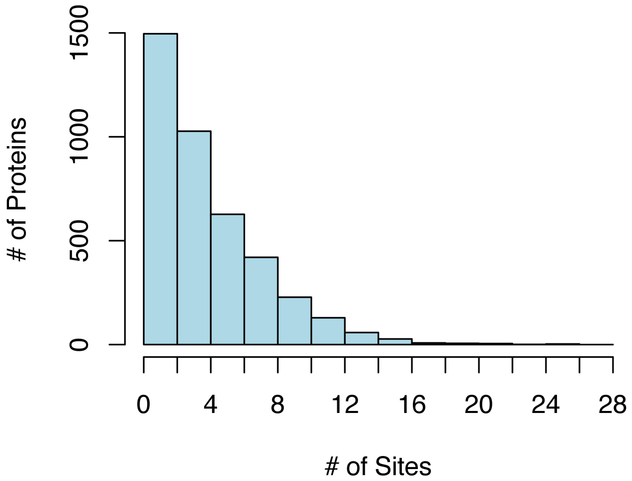
4w) Solute-binding protein MA_0280



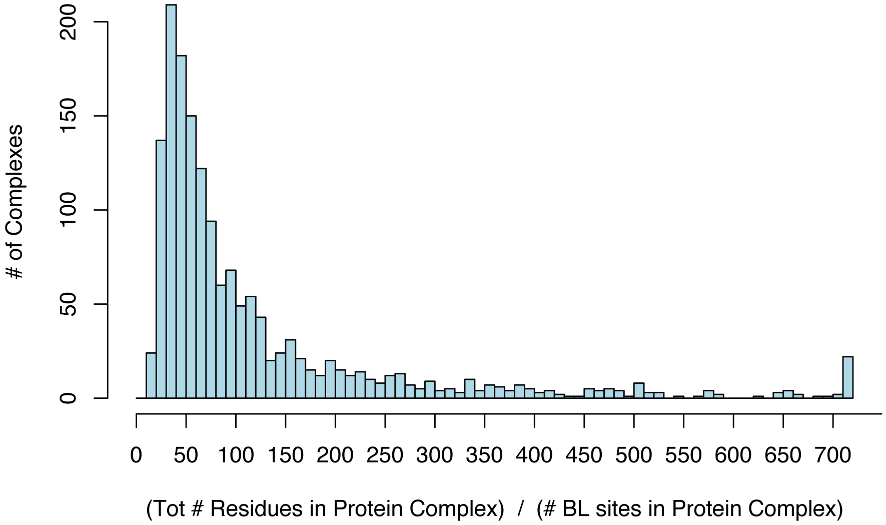
4x) Glutathione peroxidase 5



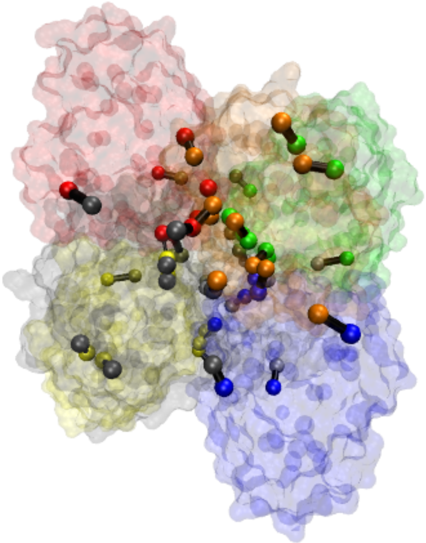
Supp Fig. 5



5a) Number of binding sites per protein (PDB chain)

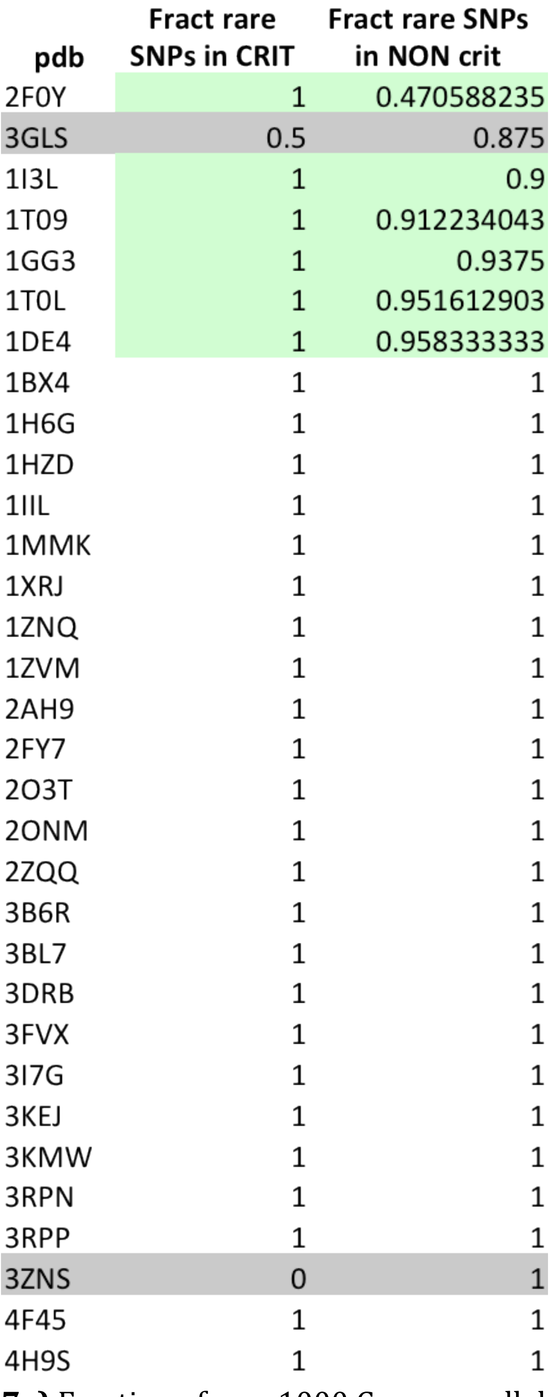


5b) Density of prioritized sites with respect to number of residues in complex



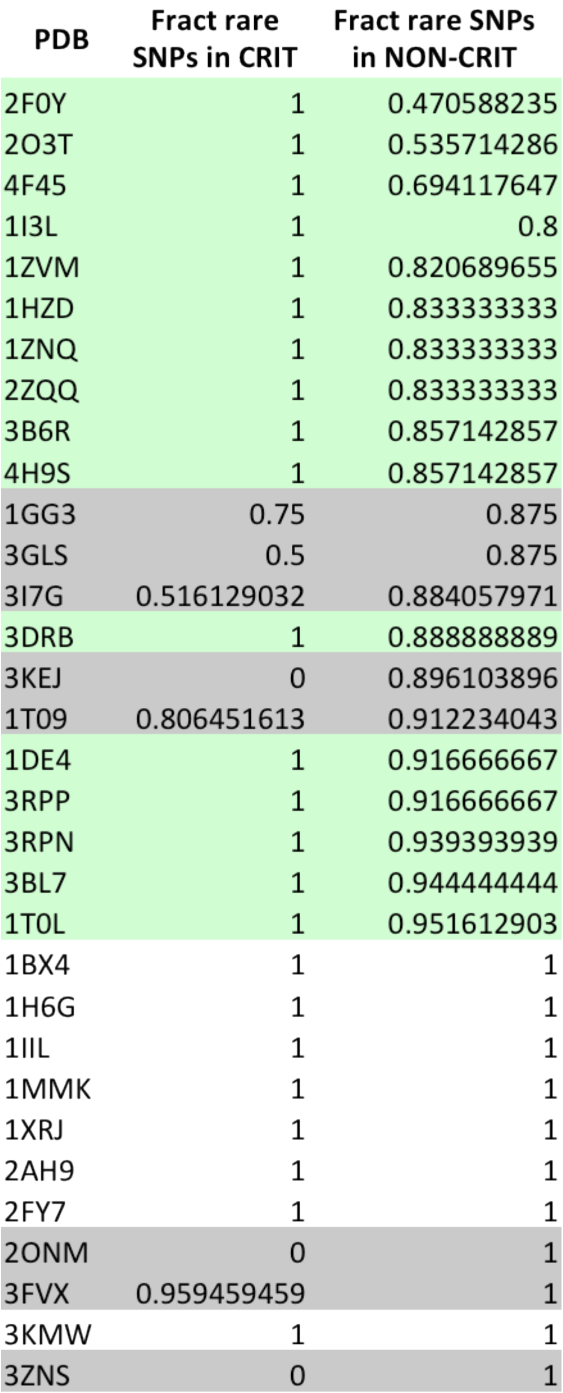
Supp Fig. 6: Communities identified by dynamical network-based analysis. Different communities are colored differently. Residues shown as spheres are critical residues. The thickness of a black links between a pair of residues is proportional to that pair's associated betweenness. The protein shown is phosphfructokinase (PDB ID 3PFK).

Supp Fig. 7



pdb	Fract rare SNPs in CRIT	Fract rare SNPs in NON crit
2F0Y	1	0.470588235
3GLS	0.5	0.875
1I3L	1	0.9
1T09	1	0.912234043
1GG3	1	0.9375
1T0L	1	0.951612903
1DE4	1	0.958333333
1BX4	1	1
1H6G	1	1
1HZD	1	1
1IIL	1	1
1MMK	1	1
1XRJ	1	1
1ZNQ	1	1
1ZVM	1	1
2AH9	1	1
2FY7	1	1
2O3T	1	1
2ONM	1	1
2ZQQ	1	1
3B6R	1	1
3BL7	1	1
3DRB	1	1
3FVX	1	1
3I7G	1	1
3KEJ	1	1
3KMW	1	1
3RPN	1	1
3RPP	1	1
3ZNS	0	1
4F45	1	1
4H9S	1	1

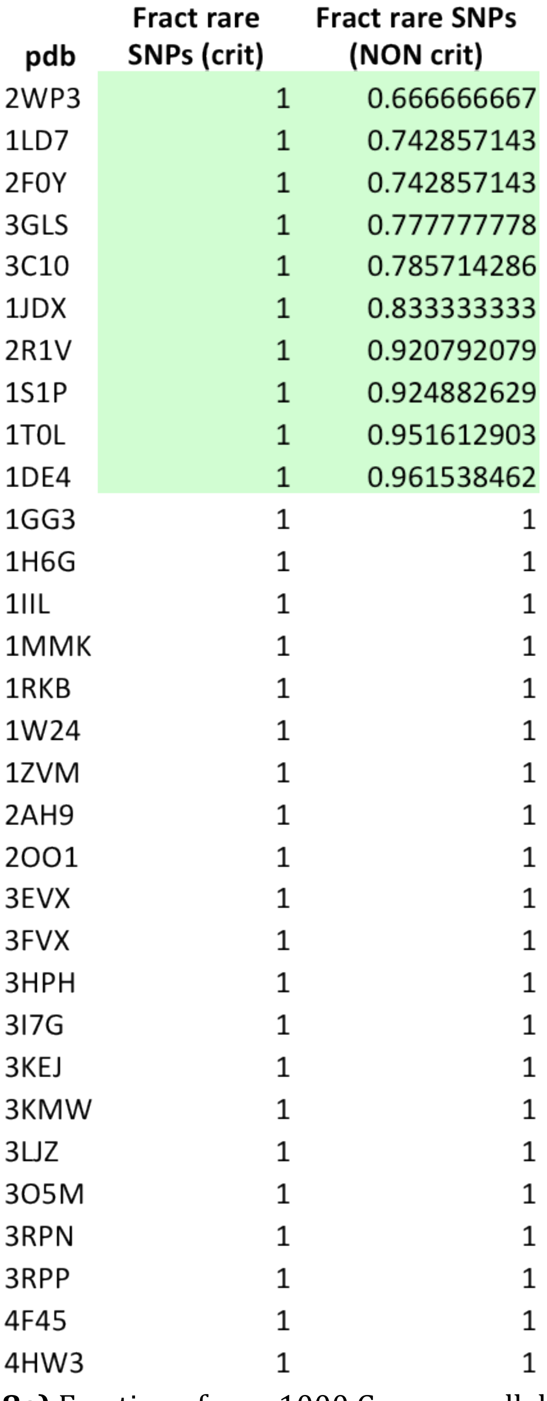
7a) Fraction of rare 1000 Genomes alleles (using a DAF cutoff of 0.05%) for surface critical and non-critical residues. Green is used to highlight cases for which the fraction of rare variants is higher in critical residues than in non-critical residues, and gray designates cases for which the opposite trend is observed.



PDB	Fract rare SNPs in CRIT	Fract rare SNPs in NON-CRIT
2F0Y	1	0.470588235
2O3T	1	0.535714286
4F45	1	0.694117647
1I3L	1	0.8
1ZVM	1	0.820689655
1HZD	1	0.833333333
1ZNQ	1	0.833333333
2ZQQ	1	0.833333333
3B6R	1	0.857142857
4H9S	1	0.857142857
1GG3	0.75	0.875
3GLS	0.5	0.875
3I7G	0.516129032	0.884057971
3DRB	1	0.888888889
3KEJ	0	0.896103896
1T09	0.806451613	0.912234043
1DE4	1	0.916666667
3RPP	1	0.916666667
3RPN	1	0.939393939
3BL7	1	0.944444444
1T0L	1	0.951612903
1BX4	1	1
1H6G	1	1
1IIL	1	1
1MMK	1	1
1XRJ	1	1
2AH9	1	1
2FY7	1	1
2ONM	0	1
3FVX	0.959459459	1
3KMW	1	1
3ZNS	0	1

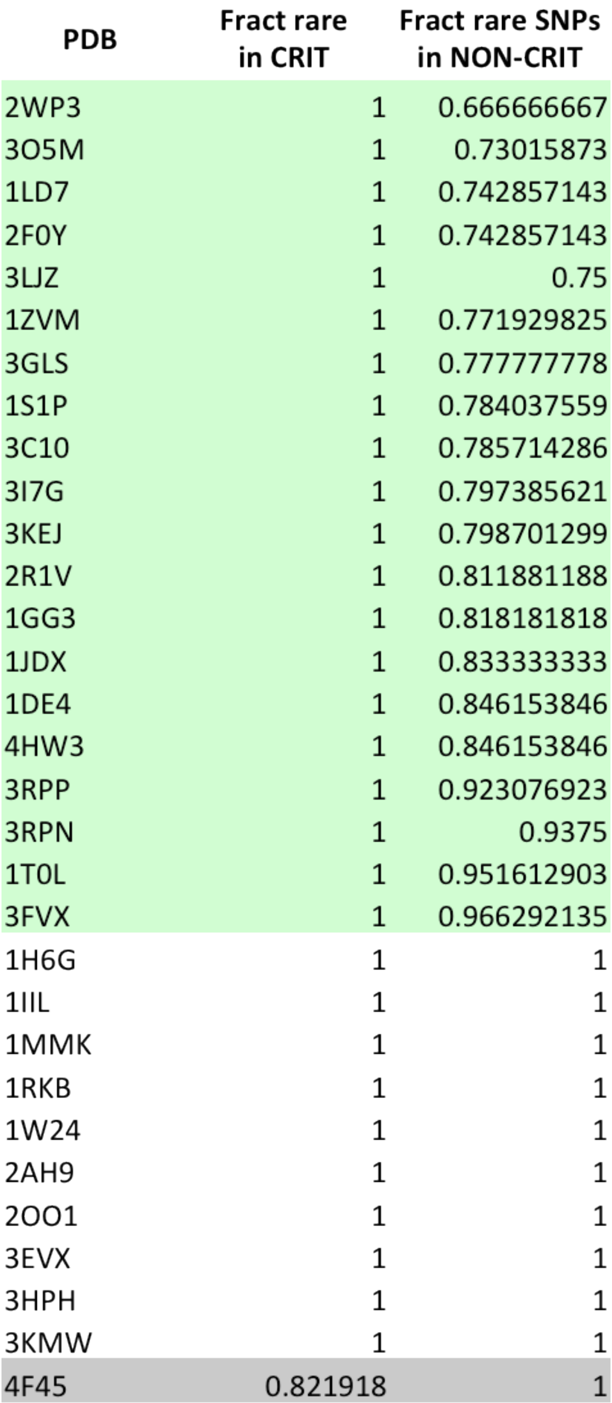
7b) Fraction of rare 1000 Genomes alleles (using a DAF cutoff of 0.01%) for surface critical and non-critical residues. Green is used to highlight cases for which the fraction of rare variants is higher in critical residues than in non-critical residues, and gray designates cases for which the opposite trend is observed.

Supp Fig. 8



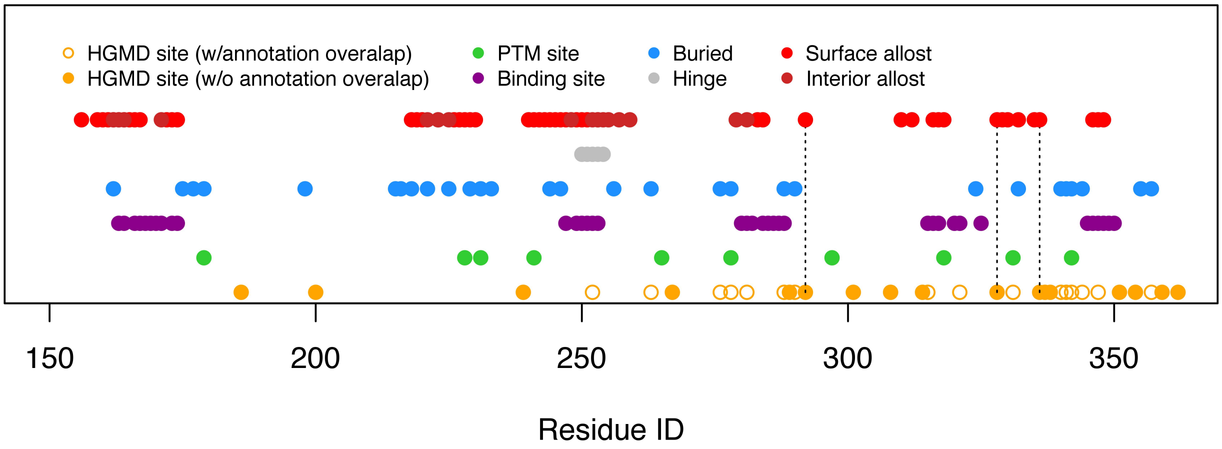
pdb	Fract rare SNPs (crit)	Fract rare SNPs (NON crit)
2WP3	1	0.666666667
1LD7	1	0.742857143
2F0Y	1	0.742857143
3GLS	1	0.777777778
3C10	1	0.785714286
1JDX	1	0.833333333
2R1V	1	0.920792079
1S1P	1	0.924882629
1T0L	1	0.951612903
1DE4	1	0.961538462
1GG3	1	1
1H6G	1	1
1IIL	1	1
1MMK	1	1
1RKB	1	1
1W24	1	1
1ZVM	1	1
2AH9	1	1
2OO1	1	1
3EVX	1	1
3FVX	1	1
3HPH	1	1
3I7G	1	1
3KEJ	1	1
3KMW	1	1
3LJZ	1	1
3O5M	1	1
3RPN	1	1
3RPP	1	1
4F45	1	1
4HW3	1	1

8a) Fraction of rare 1000 Genomes alleles (using a DAF cutoff of 0.05%) for interior critical and non-critical residues. Green is used to highlight cases for which the fraction of rare variants is higher in critical residues than in non-critical residues.

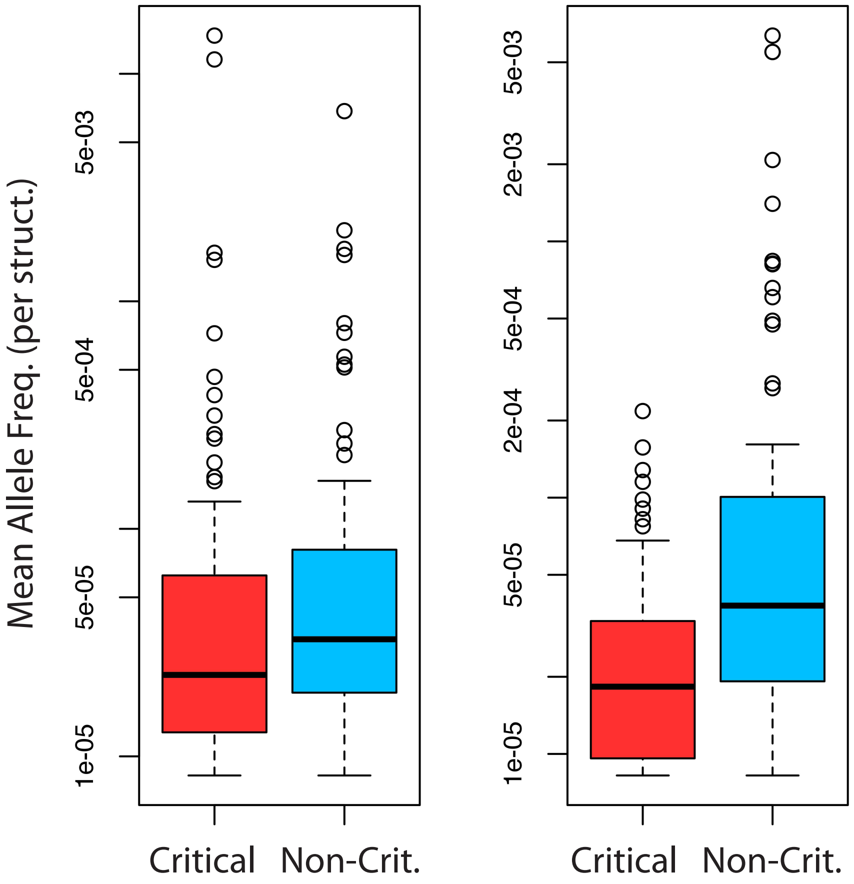


PDB	Fract rare in CRIT	Fract rare SNPs in NON-CRIT
2WP3	1	0.666666667
3O5M	1	0.73015873
1LD7	1	0.742857143
2F0Y	1	0.742857143
3LJZ	1	0.75
1ZVM	1	0.771929825
3GLS	1	0.777777778
1S1P	1	0.784037559
3C10	1	0.785714286
3I7G	1	0.797385621
3KEJ	1	0.798701299
2R1V	1	0.811881188
1GG3	1	0.818181818
1JDX	1	0.833333333
1DE4	1	0.846153846
4HW3	1	0.846153846
3RPP	1	0.923076923
3RPN	1	0.9375
1T0L	1	0.951612903
3FVX	1	0.966292135
1H6G	1	1
1IIL	1	1
1MMK	1	1
1RKB	1	1
1W24	1	1
2AH9	1	1
2OO1	1	1
3EVX	1	1
3HPH	1	1
3KMW	1	1
4F45	0.821918	1

8b) Fraction of rare 1000 Genomes alleles (using a DAF cutoff of 0.01%) for interior critical and non-critical residues. Green is used to highlight cases for which the fraction of rare variants is higher in critical residues than in non-critical residues, and gray designates cases for which the opposite trend is observed.

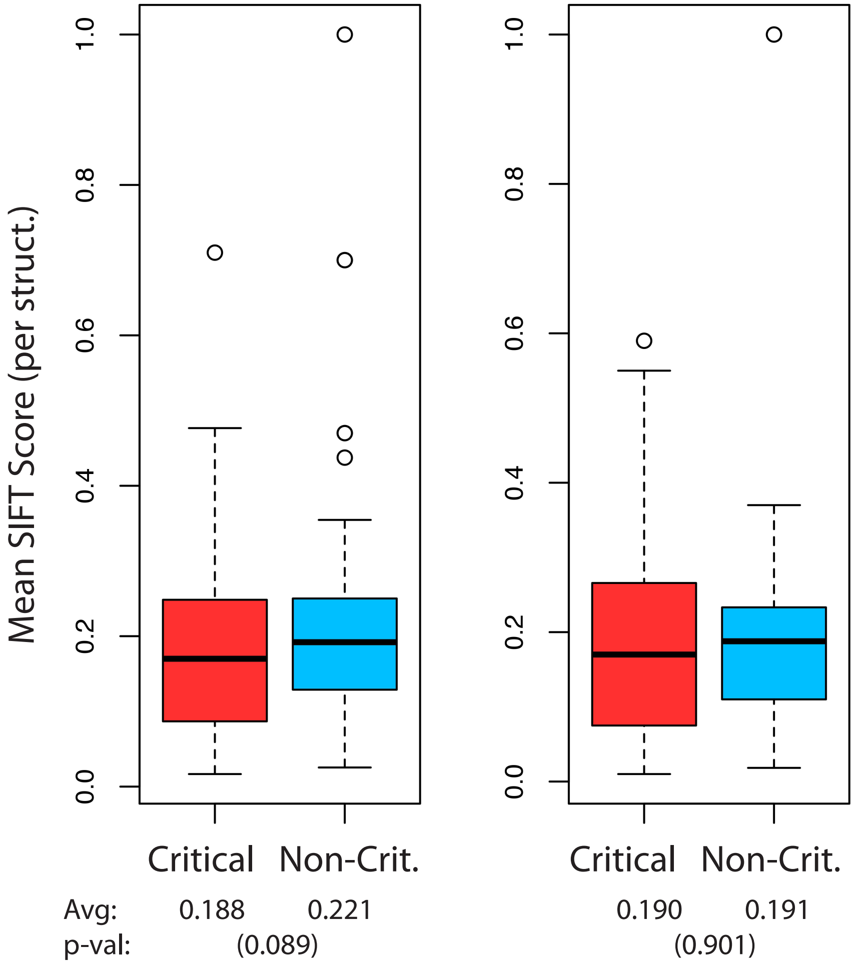


Supp Fig. 9: Disease mutations afflicting FGFR2 in the context of critical residues and biological annotation. Shown below is chain E of the PDB 1IIL, which corresponds to the FGFR2. Dotted lines highlight loci that correspond to HGMD sites that coincide with critical residues, but for which other annotations fail to coincide. Deeply-buried residues are defined to be those that exhibit a relative solvent-exposed surface area of 5% or less, and binding site residues are defined as those for which at least one heavy atom falls within 4.5 Angstroms of any heavy atom in the binding partner (heparin-binding growth factor 2). The loci of PTM sites were taken from UniProt (accession no. P21802).

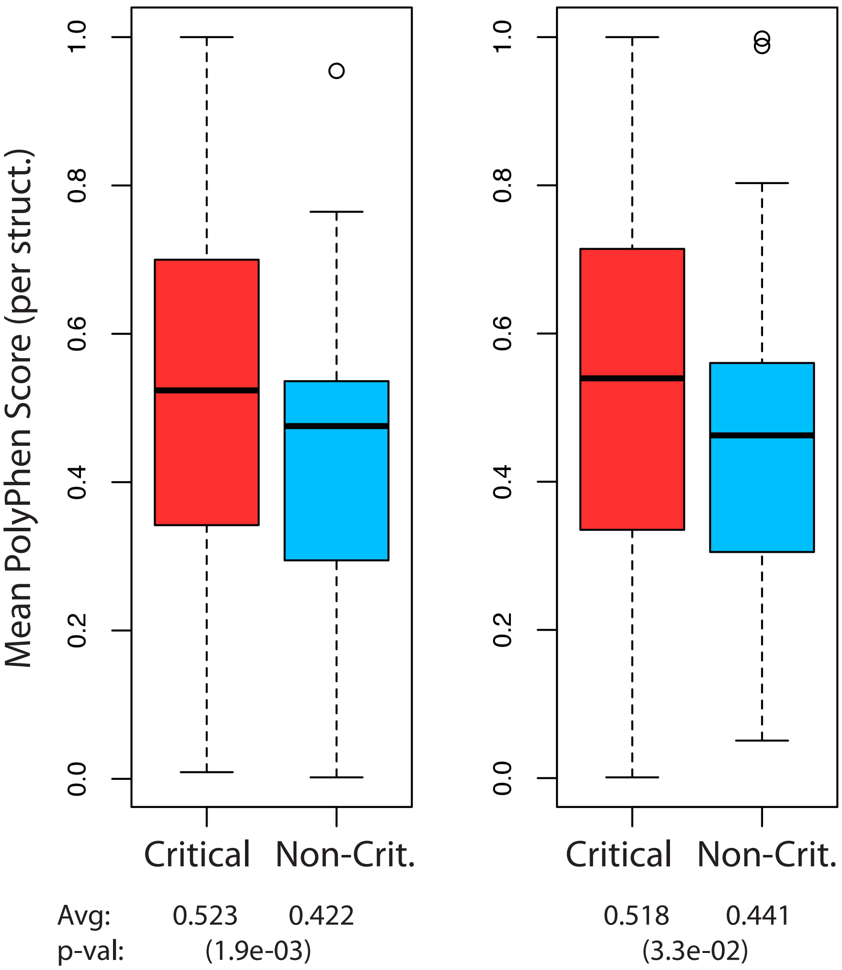


Avg:	4.09E-04	2.26E-04	3.08E-05	3.27E-04
p-val:	(1.49e-03)		(7.98e-09)	

Supp Fig. 10: Mean allele frequencies (AF) for critical- and non-critical residues, as identified by ExAC. *Left*: Distribution of mean minor allele frequencies (MAF) on critical surface residues (red) and non-critical residues (blue). *Right*: Distribution of mean AF values on critical interior residues (red) and non-critical residues (blue). Overall mean values and p-values are given below plots.

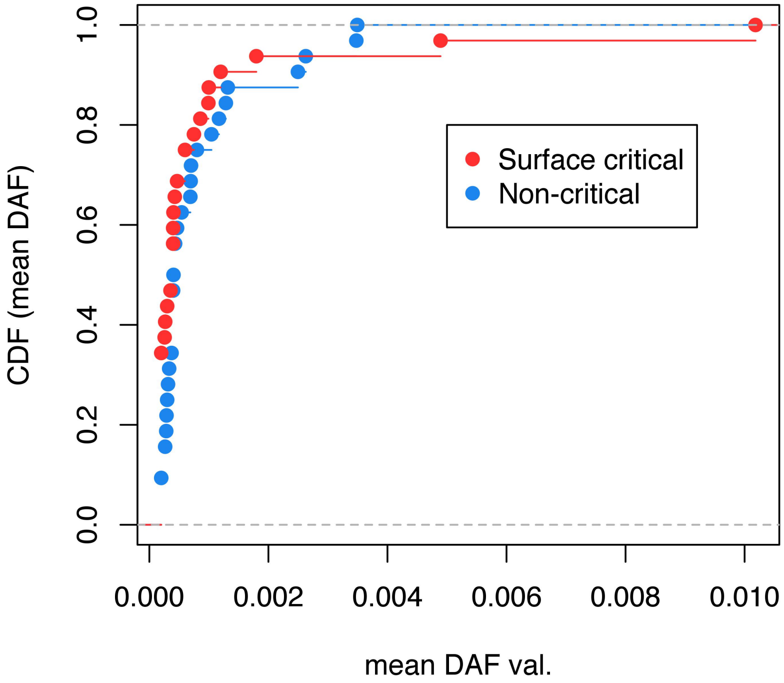


Supp Fig. 11: Mean SIFT scores for critical- and non-critical residues, as identified by ExAC. *Left*: Distribution of mean SIFT values on critical surface residues (red) and non-critical residues (blue). *Right*: Distribution of mean SIFT values on critical interior residues (red) and non-critical residues (blue). Overall mean values and p-values are given below plots. Note that lower SIFT scores denote more damaging variants.

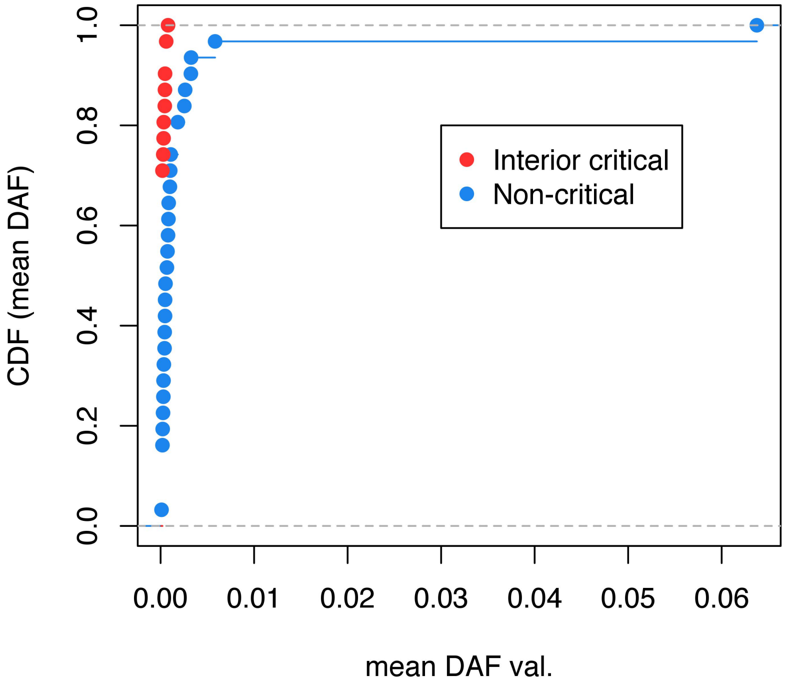


Supp Fig. 12: Mean PolyPhen scores for critical- and non-critical residues, as identified by ExAC. *Left*: Distribution of mean PolyPhen values on critical surface residues (red) and non-critical residues (blue). *Right*: Distribution of mean PolyPhen values on critical interior residues (red) and non-critical residues (blue). Overall mean values and p-values are given below plots. Note that higher PolyPhen scores denote more damaging variants.

Supp Fig. 13: Potential shifts in DAF distributions (in 1000 Genomes) using two-sample Kolmogorov-Smirnov tests

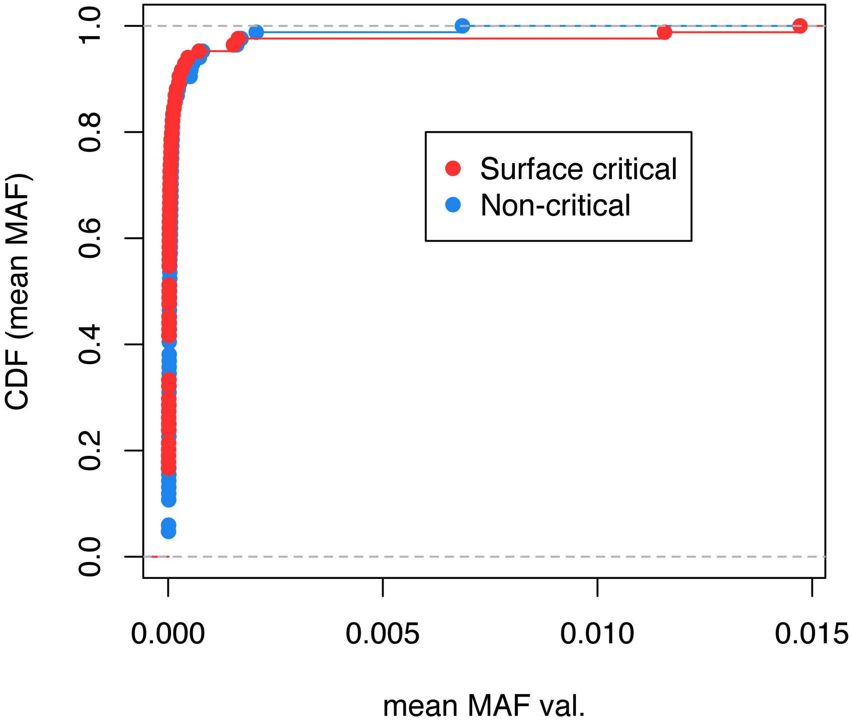


13a) Cumulative distribution functions for mean DAF values of critical surface and non-critical residues (p-val = 0.080).

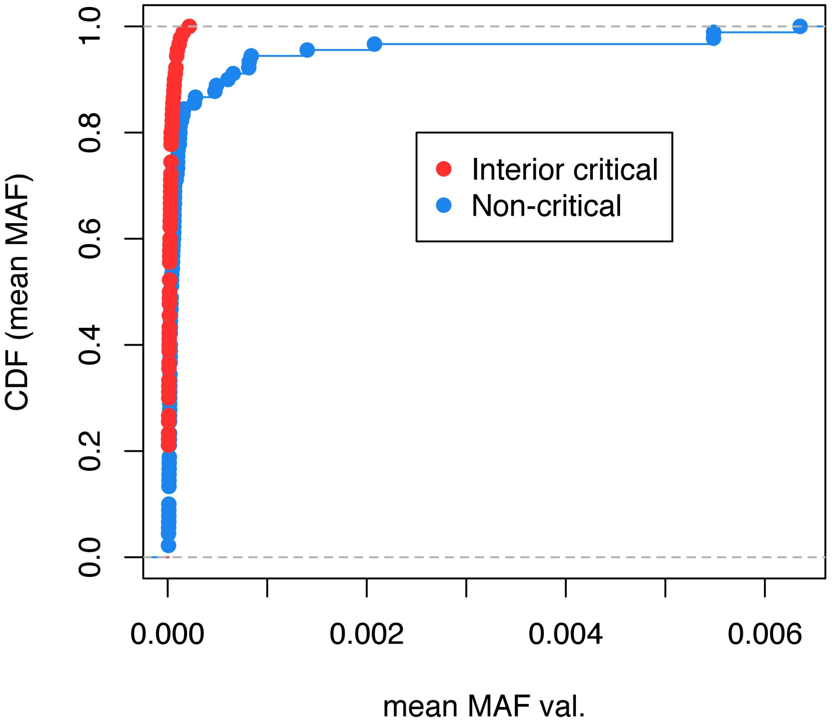


13b) Cumulative distribution functions for mean DAF values of interior critical and non-critical residues (p-val = 8.9E-5).

Supp Fig. 14: Potential shifts in mean minor allele frequency distributions (in ExAC) using two-sample Kolmogorov-Smirnov tests

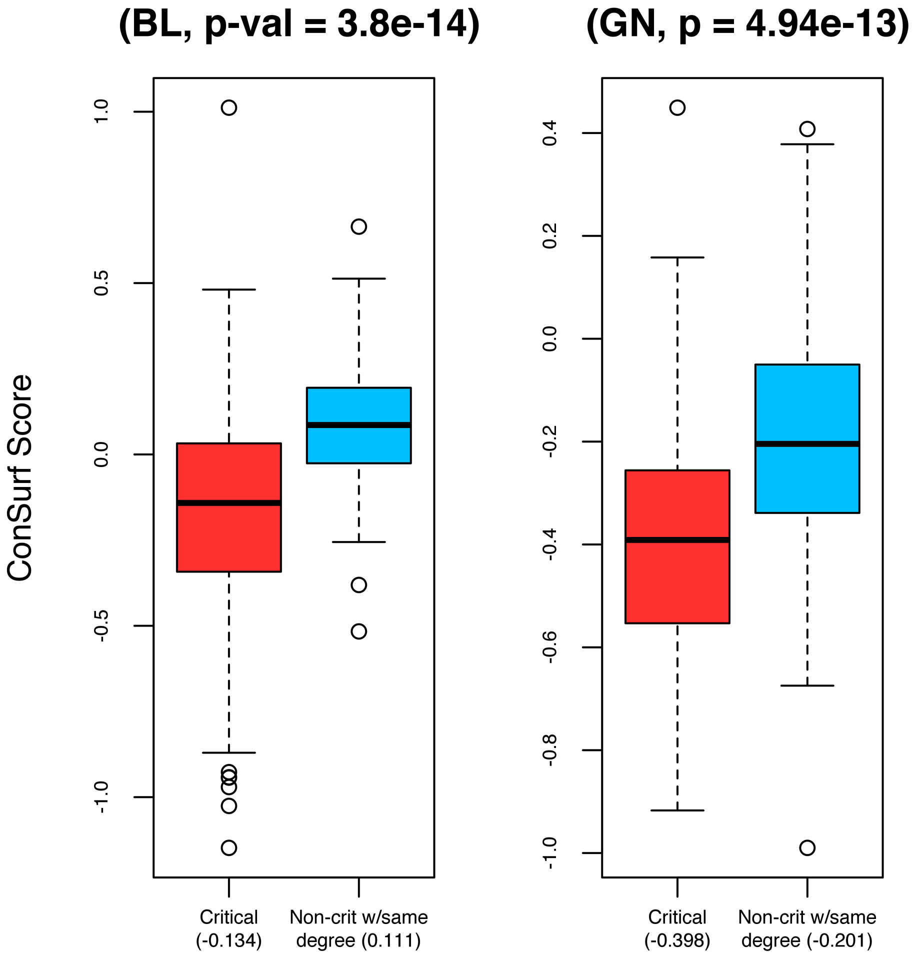


14a) Cumulative distribution functions for mean minor allele frequencies of surface critical and non-critical residues (p-val = 0.0475)



14b) Cumulative distribution functions for mean minor allele frequencies of interior critical and non-critical residues (p-val = 8.7E-5).

Supp Fig. 15: Modeling protein conformational change through a direct use of crystal structures from alternative conformations using absolute conformational transitions (ACT).



Left: Distribution of the mean conservation scores values on critical surface residues (red) and non-critical residues (blue). *Right*: Distribution of the mean conservation scores for critical interior residues (red) and non-critical residues (blue).

Alternative Conformations in Domains

We first worked with domains to probe for intra-domain conformational changes. Better structure alignments are generally possible at the domain level. The filtered dataset of domains contains 79% of all available crystal structures in the PDB (as of December 2013). PDB-wide MSAs across sequence-similar groups reveal that, in agreement with expectation, average pairwise root-mean-square deviation (RMSD) values increase at lower levels of sequence identity, as do Q_H values (Q_H, an alternative metric to RMSD, quantifies the degree to which residue-residue distances differ between two conformations, and is detailed in [[cite]] and Methods) (Supp. Fig. 2).

Modified Binding Leverage Framework for Identifying Known Ligand-Binding Sites

It has previously been shown that it is especially difficult to identify the sites in aspartate transcarbamoylase (Mitternacht and Berezovsky, 2011); excluding aspartate transcarbamoylase from this analysis results in finding an average of 65% of known biological sites. These statistics are achieved by covering an average of 15% of proteins' residues (Supp. Table 2), even though more than 15% of the proteins' residues are involved in ligand- or substrate-binding for most proteins (Supp. Table 3).

Obtaining Models of Protein Motions by Directly Using Displacement Vectors from Alternative Conformations

This more direct model of conformational change, which we term absolute conformational transitions (ACT), may be applied in a straightforward manner to single-chain proteins. When we use ACT to apply the modified binding leverage framework for such single-chain proteins, we observe that our surface critical residues are significantly more conserved than are non-critical residues (Supp. Fig. 15, left). The same trend is observed when ACT is applied in our dynamical network analysis for identifying interior critical residues (Supp. Fig. 15, right).

For the binding leverage framework, each candidate site is scored on the basis of the degree to which occlusion by the ligand conflicts with the large-scale motions of the protein (Fig. 1, bottom left; see Methods). These motions are taken from anisotropic network models (ANMs), but the results do not change drastically if we directly use the alternative conformations as given by the crystal structures (see discussion below). The main modifications to the formalism previously described include the use of heavy atoms in the protein during the Monte Carlo search, in addition to an automated means of thresholding the list of ranked sites to give a more selective set of candidate sites.

Decomposing Proteins into Modules Using Two Different Algorithms

Many algorithms have been devised to extract the community structure of networks. In a comprehensive study comparing different algorithms (Lancichinetti et al, 2009), an information theory-based approach (Rosvall et al, 2007), was shown to be one of the strongest. This method effectively reduces the network community detection problem to a problem in information compression: the prominent features of the network are extracted in this compression process, giving rise to distinct modules (more details are provided in Rosvall et al, 2007).

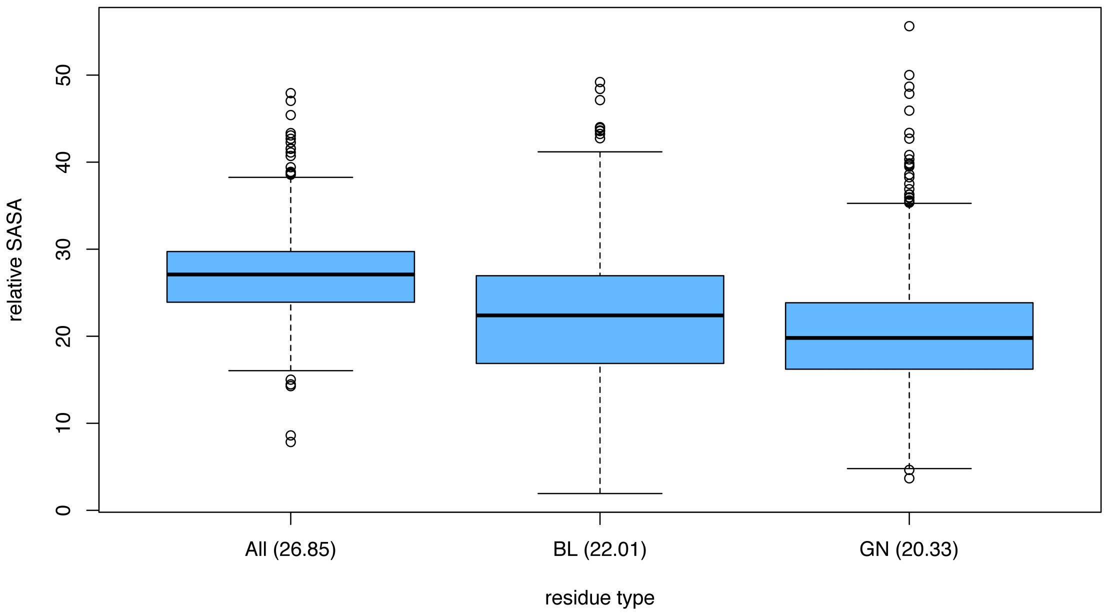
Perhaps surprisingly, even though both methods achieve similar network modularity, we find that Infomap (see Methods and Rosvall et al, 2007) produces at least twice the number of communities relative to that of GN, and it thus generates many more critical residues (Supp. Table 5). For the canonical set of proteins, GN and Infomap generated an average of 12.0 and 36.8 communities, respectively (corresponding to an average of 44.8 and 201.4 critical residues, respectively). Thus, given that GN produces a more selective set of residues for each protein, the focus of our analyses is based on GN (corresponding results for Infomap are available in the in the Supplement).

Although the critical residues identified by GN do not always correspond to those identified by Infomap, the mean fraction of GN-identified critical residues that match Infomap-identified residues is 0.30 (the expected mean is 0.21, p-value=0.058), which further justifies our decision to focus on GN). Furthermore, we observe that obvious structural communities are detected when applying both methods (i.e., a community generated by GN is often the same as that generated by Infomap, and in other cases, a community generated by GN is often composed of sub-communities generated by Infomap).

As noted, the modularity from the network partitions generated by GN and Infomap are very similar (for the 12 canonical systems, the mean modularity for GN and Infomap is 0.73 and 0.68, respectively). Presumably, GN modularity values are consistently at least as high as those in Infomap because GN explicitly optimizes modularity in partitioning the network, whereas Infomap does not.

Comparisons between essential residues identified by binding leverage and dynamical network analysis

To better characterize the residues that we identify as critical by the binding leverage framework ("BL residues") and dynamical network analysis ("GN residues"), we evaluated their solvent accessible surface area (SASA). Relative SASA values (which represent the solvent accessibility of a residue relative to that residue in an extended ALA-x-ALA tripeptide) were obtained using NACCESS (Hubbard et al, 1993). For each protein complex in our large dataset, we calculated the mean relative SASA for each of these two classes of critical residues, as well as the mean relative SASA for all residues in each complex [[see FIGURE below naccess__avgs_box.jpg]]. As discussed, since GN residues are involved in high-betweenness edges connecting communities, many GN residues tend to be interior to the protein, thus explaining why GN residues exhibit the lowest mean SASA. BL residues, in contrast, exhibit mean SASA values (centered at 22.01) that are intermediate between GN residues (20.33) and the protein-wide average (26.85). BL residues are excluded from the deep interior of the protein. Nevertheless, as BL hotspots occur within clefts and pockets, the associated residues should be partially buried.



In addition to sequence conservation, we evaluated the structure conservation of BL residues and GN residues using several different approaches (see Methods for details). In brief, for each set of proteins in our dataset, multiple structure alignments were generated for the associated domains at the Fold, Superfamily, and Family levels within SCOP (Fox et al, 2014), and the structural conservation of critical residues was evaluated by 1) looking for co-occurrence of these critical residues within each multiple structure alignment (i.e., we determine whether a given critical residue in a protein overlaps with a critical residue in another protein within the multiple structure alignment); and 2) evaluating the Qres scores of these critical residues (Qres is a metric that quantifies local structural similarity in an alignment; see Methods for details).