Structure

Identifying allosteric hotspots with dynamics: application to inter- and intra-species conservation

--Manuscript Draft--

Manuscript Number:	STRUCTURE-D-15-00408R
Full Title:	Identifying allosteric hotspots with dynamics: application to inter- and intra-species conservation
Article Type:	Theory
Keywords:	allostery; networks; mathematical models
Corresponding Author:	Mark Gerstein New Haven, CT UNITED STATES
First Author:	Declan Clarke
Order of Authors:	Declan Clarke
	Anurag Sethi
	Shantao Li
	Sushant Kumar
	Richard W.F. Chang
	Jieming Chen
	Mark Gerstein
Abstract:	The rapidly growing volume of data being produced by next-generation sequencing initiatives is enabling more in-depth analyses of conservation than previously possible. Deep sequencing is uncovering disease loci and regions under selective constraint, despite the fact that intuitive biophysical reasons for such constraint are sometimes absent. Allostery may often provide the missing explanatory link. We use models of protein conformational change to identify allosteric residues by finding essential surface cavities and information flow bottlenecks, and we develop a software tool (stress.molmovdb.org) that enables users to perform this analysis on their own proteins of interest. Though fundamentally 3D-structural in nature, our analysis is computationally fast, thereby allowing us to run it across the PDB and to evaluate general properties of predicted allosteric residues. We find that these tend to be conserved over diverse evolutionary time scales. Finally, we highlight examples of allosteric residues that help explain poorly understood disease-associated variants.

Yale University

MB&B
260/266 Whitney Avenue
PO Box 208114
New Haven, CT 06520-8114

Telephone:
203 432 6105
360 838 7861 (fax)
mark@gersteinlab.org
www.gersteinlab.org

Feb. 12 2016

Dear Editors of *Structure*,

Thank you for accepting our work for publication in *Structure*, and we apologize for the delay in submitting our finalized files. We have addressed the last remaining point raised by reviewer #2, and this is detailed in our response. Also, I should point out that Declan Clarke and Anurag Sethi should be considered co-first authors of this work, and this is clearly indicated in the main text file (it appears as if the online submission system for the journal does not enable me to specify co-first authors). We are delighted that our work will be published in your journal, and we hope that this paper (along with our newly-introduced web server) will be of great value to a broad spectrum of your readers. Please feel free to let us know if there is anything else that you need from us, and we look forward to working with you on the proofs for this manuscript.

Yours sincerely,

Mark Gerstein
Albert L. Williams Professor
of Biomedical Informatics

RESPONSE LETTER

Reviewer #2

-- Ref 2.0 – Turning the MC ensemble into a list of sites --

Reviewer Comment	The revisions look good, and address all but one of my queries. Specifically, how the MC ensemble is turned into a list of sites - there is still no indication of how the authors go from the output of their MC simulation, which is the ensemble of structures from the MC trajectory, to a list of ligand binding regions. The text simply states how the MC is performed, followed by "After all candidate sites are identified by these MC simulations...", leaving the reader with no details of how this transformation is performed. I would like to see this clarified in the final manuscript. Otherwise, I believe the protocols and results are now much clearer.
Author Response	We would again like to thank the reviewer for their careful-read through of our work, as well as valuable feedback regarding the clarification. We have added additional details to this formalism in the appropriate section of the supplementary materials.
Excerpt From Revised Manuscript	What form does the resultant MC ensemble take, and how exactly is this MC ensemble turned into a list of candidate sites? Prior to thresholding the list of ranked sites (see below), we generally follow the same formalism detailed in (Mitternacht and Berezovsky, 2011), and a discussion of this is also provided here. We first detail the output provided by a single MC simulation. This MC simulation involves a ligand probing the protein surface through a large number (typically 1,500,000 – 2,000,000, depending on the protein's size) of MC steps in which the ligand explores translational, rotational, and angular degrees of freedom. It is only the ligand that explores these degrees of freedom - the protein remains static throughout this MC simulation. The potential function (see above) usually "pushes" the ligand to favorably occupy a pocket on the protein surface after all steps of the MC simulation are completed. The ligand is thus in contact with a number of residues (typically 10-20) at the end of the simulation. As with the approach taken by Mitternacht and Berezovsky, this list of residues is ordered by local closeness (LC). LC is a geometric quantity that provides a measure of the degree of a residue in the residue-residue contact network; see (Mitternacht and Berezovsky, 2011b) for further discussion of LC. The 10 residues with greatest LC are taken as the final "site" occupied by the ligand at the end of this MC simulation (the remaining residues are not considered to be part of the site). Thus, the output of this single MC simulation is a list of 10 residues on the protein surface such that these residues form a geometrically favorable site for the ligand. Now consider a very large number (typically 5,000 – 10,000, depending on the protein's size) of the MC simulations detailed above. These ~10,000 MC simulations result in ~10,000 sites, where each of these sites is the list of residues in contact with the ligand by the end of the MC simulation.

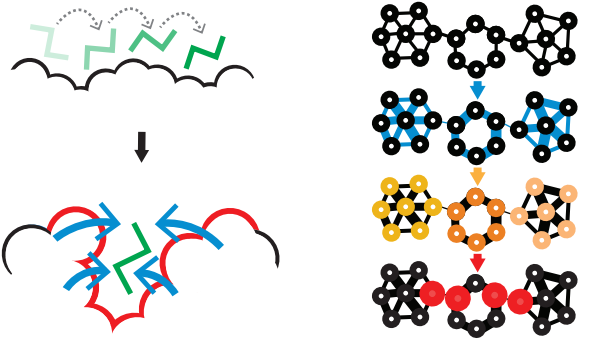
	This long list of sites generally contains many sites with a strong degree of overlap. Thus, to remove redundancy, pairs of sites with extremely high overlap are merged by combining any pair of sites that have a Jaccard similarity of at least 0.7, where the Jaccard similarity between sites i and j is defined as $\|i \cap j\|/\|i \cup j\|$. After merging sites in this way, the residues of a given site are listed by their LC, and no more than 10 residues for a site are used. This entire process results in a list of sites on which binding leverage calculations can be performed.

Graphical Abstract

predicting **allosteric residues** with **dynamics**

surface

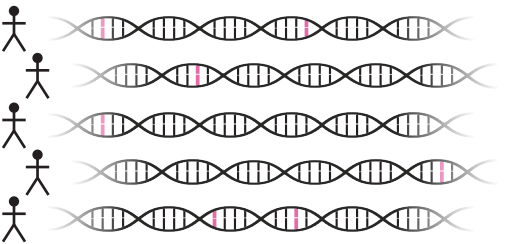
interior



sequence-based analyses

across species

amongst humans



Title:
Identifying allosteric hotspots with dynamics: application to inter- and intra-species conservation

Authors & associated information:
Declan Clarke[a,1], Anurag Sethi[b,c,1], Shantao Li[b,d], Sushant Kumar[b,c], Richard W.F. Chang[e], Jieming Chen[b,f], and Mark Gerstein[b,c,d,2]

[a] Department of Chemistry, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520 USA
[b] Program in Computational Biology and Bioinformatics, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[c] Department of Molecular Biophysics and Biochemistry, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[d] Department of Computer Science, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[e] Yale College, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
[f] Integrated Graduate Program in Physical and Engineering Biology, Yale University, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

[1] D.C. and A.S. contributed equally to this work.
[2] Correspondence should be addressed to M.G. (pi@gersteinlab.org)

Highlights:

- Allostery often provides a biophysical rationale for signatures of conservation
- Models of protein conformational change are used to predict key allosteric residues
- These predicted allosteric residues are conserved across species and amongst humans
- A web tool makes this analysis publically available to the scientific community

ABSTRACT

The rapidly growing volume of data being produced by next-generation sequencing initiatives is enabling more in-depth analyses of conservation than previously possible. Deep sequencing is uncovering disease loci and regions under selective constraint, despite the fact that intuitive biophysical reasons for such constraint are sometimes absent. Allostery may often provide the missing explanatory link. We use models of protein conformational change to identify allosteric residues by finding essential surface cavities and information flow bottlenecks, and we develop a software tool (stress.molmovdb.org) that enables users to perform this analysis on their own proteins of interest. Though fundamentally 3D-structural in nature, our analysis is computationally fast, thereby allowing us to run it across the PDB and to evaluate general properties of predicted allosteric residues. We find that these tend to be conserved over diverse evolutionary time scales. Finally, we highlight examples of allosteric residues that help explain poorly understood disease-associated variants.

INTRODUCTION

The ability to sequence large numbers of human genomes is providing a much deeper view into protein evolution than previously possible. When trying to understand the evolutionary pressures on a given protein, structural biologists now have at their disposal an unprecedented breadth of data regarding patterns of conservation, both across species and amongst humans. As such, there are greater opportunities to take an integrated view of the context in which a protein and its residues function. This view necessarily includes structural constraints such as residue packing, protein-protein interactions, and stability. However, deep sequencing is unearthing a class of conserved residues on which no obvious structural constraints appear to be acting. The missing link in understanding these regions may be provided by studying the protein's dynamic behavior through the lens of the distinct functional and conformational states within an ensemble.

The underlying energetic landscape responsible for the relative distributions of alternative conformations is dynamic in nature: allosteric signals or other external changes may reconfigure and reshape the landscape, thereby shifting the relative populations of states within an ensemble (Tsai *et al.*, 1999). Landscape theory thus provides the conceptual underpinnings necessary to describe how proteins change behavior and shape under changing conditions. A primary driving force behind the evolution of these landscapes is the need to efficiently regulate activity in response to changing cellular contexts, thereby making allostery and conformational change essential components of protein evolution.

Given the importance of allosteric regulation, as well as its role in imparting efficient functionality, several methods have been devised for the identification of likely allosteric residues. Conservation itself has been used, either in the context of conserved residues (Panjkovich and Daura, 2012), networks of co-evolving residues (Halabi *et al*., 2009; Lee *et al*., 2008; Lockless *et al*., 1999; Reynolds *et al*., 2011; Shulman *et al*., 2004; Süel *et al*., 2003), or local conservation in structure (Panjkovich and Daura, 2010). In related studies, both conservation and geometric-based searches for allosteric sites have been successfully applied to several systems (Capra *et al*., 2009).

The concept of 'protein quakes' has been introduced to explain local conformational changes that are essential for global conformation transitions of functional importance (Ansari *et al*., 1985; Miyashita *et al*., 2003). These local changes cause strain within the protein that is relieved by subsequent relaxations (which are also termed functionally important motions) that terminate when the protein reaches the second equilibrium state. Such local perturbations often end with large conformational changes at the focal points of allosteric regulation, and these motions may be identified in a number of ways, including modified normal modes analysis (Miyashita *et al*., 2003) or time-resolved X-ray scattering (Arnlund *et al*., 2014).

In addition to conservation and geometry, protein dynamics have also been used to predict allosteric residues. Normal modes analysis has been used to examine the extent to which bound ligands interfere with low-frequency motions, thereby identifying potentially important residues at the surface (Ming and Wall, 2005; Mitternacht and Berezovsky, 2011; Panjkovich and Daura, 2012). Normal modes have also been used by the Bahar group to identify important subunits that act in a coherent manner for specific

proteins (Chennubhotla and Bahar, 2006; Yang and Bahar, 2005). Rodgers *et al.* have applied normal modes to identify key residues in CRP/FNR transcription factors (Rodgers *et al*., 2013).

With the objective of identifying allosteric residues within the interior, molecular dynamics (MD) simulations and network analyses have been used to identify residues that may function as internal allosteric bottlenecks (Csermely *et al*., 2013; Gasper *et al*., 2012; Rousseau and Schymkowitz, 2005; Sethi *et al*., 2009; Vanwart *et al*., 2012). Ghosh *et al*. (2008) have taken a novel approach of combining MD and network principles to characterize allosterically important communication between domains in methionyl tRNA synthetase. In conjunction with NMR, Rivalta *et al.* have use MD and network analysis to identify important regions in imidazole glycerol phosphate synthase (Rivalta *et al*., 2012).

Though having provided valuable insights, many of these approaches have been limited in terms of scale (the numbers of proteins which may feasibly be investigated), computational demands, or the class of residues to which the method is tailored (surface or interior). Here, we use models of protein conformational change to identify both surface and interior residues that may act as essential allosteric hotspots in a computationally tractable manner, thereby enabling high-throughput analysis. This framework directly incorporates information regarding 3D protein structure and dynamics, and it can be applied on a PDB-wide scale to proteins that exhibit conformational change. Throughout the PDB (Berman *et al*., 2000), the residues identified tend to be conserved both across species and amongst humans, and they may help to elucidate many of the otherwise poorly understood regions in proteins. In a

similar vein, several of our identified sites correspond to human disease loci for which no clear mechanism for pathogenesis had previously been proposed. Finally, we make the software associated with this framework (termed STRESS, for STRucturally-identified ESSential residues) publically available through a tool to enable users to submit their own structures for analysis.

RESULTS

Identifying Potential Allosteric Residues

Allosteric residues at the surface generally play a regulatory role that is fundamentally distinct from that of allosteric residues within the protein interior. While surface residues often constitute the sources or sinks of allosteric signals, interior residues act to transmit such signals. We use models of protein conformational change to identify both classes of residues (Figure 1). Throughout, we term these potential allosteric residues at the surface and interior "surface-critical" and "interior-critical" residues, respectively.

In order to gauge the effectiveness of our approach, we identified and analyzed critical residues within a set of 12 well-studied canonical systems (see Figure S1, as well as Table S1 for rationale regarding the set selection). We then apply this protocol on a large scale across hundreds of proteins for which crystal structures of alternative conformations are available.

Identifying Surface-Critical Residues

Allosteric ligands often act by binding to surface cavities and modulating protein conformational dynamics. The surface-critical residues, some of which may act as latent ligand binding sites and active sites, are first identified by finding cavities using Monte Carlo simulations to probe the surface with a flexible ligand (Figure 1A, top-left). The degree to which cavity occlusion by the ligand disrupts large-scale conformational change is used to assign a score to each cavity – sites at which ligand occlusion strongly interferes with conformational change earn high scores (Figure 1A, top-right), whereas shallow pockets (Figure 1A, bottom-left) or sites at which large-scale motions are largely unaffected (Figure 1A, bottom-right) earn lower scores. Further details are provided in SI Methods section 3.1-a.

This approach is a modified version of the binding leverage framework introduced by Mitternacht and Berezovsky (Mitternacht and Berezovsky, 2011). The main modifications implemented here 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 scores. These modifications were implemented to provide a more selective set of sites; without them, a very large fraction of the protein surface would be occupied by critical sites (Figure S2A). Within our dataset of proteins exhibiting alternative conformations, we find that this modified approach results in an average of ~2 distinct sites per domain (Figure S2A; see Figure S2B for the distribution for distinct sites within entire complexes).

Within the canonical set of 12 proteins, we positively identify an average of 55.6% of the sites known to be directly involved in ligand or substrate binding (see Table 1, Figure S1, and SI Methods section 3.1-a-iv). Some of the sites identified do not

directly overlap with known binding regions, but we often find that these "false positives" nevertheless exhibit some degree of overlap with binding sites (Table S2). In addition, those surface-critical sites that do not match known binding sites may nevertheless correspond to latent allosteric regions: even if no known biological function is assigned to such regions, their occlusion may nevertheless disrupt hitherto unfound large-scale motions.

Dynamical Network Analysis to Identify Interior-Critical Residues

The binding leverage framework described above is intended to capture hotspot regions at the protein surface, but the Monte Carlo search employed is *a priori* excluded from the protein interior. Allosteric residues often act within the protein interior by functioning as essential information flow 'bottlenecks' within the communication pathways between distant regions.

To identify such bottleneck residues, the protein is first modeled as a network, wherein residues represent nodes and edges represent contacts between residues (in much the same way that the protein is modeled as a network in constructing anisotropic network models, see below). In this regard, the problem of identifying interior-critical residues is reduced to a problem of identifying nodes that participate in network bottlenecks (see Figure 1B and SI Methods section 3.1-b for details). Briefly, the network edges are first weighted by the degree of strength in the correlated motions of contacting residues: a strong correlation in the motion between contacting residues implies that knowing how one residue moves better enables one to predict the motion of the other, thereby suggesting a strong information flow between the two residues. The weights are

used to assign ‘effective distances’ between connecting nodes, with strong correlations resulting in shorter effective node-node distances.

Using the motion-weighted network, “communities” of nodes are identified using the Girvan-Newman formalism (Girvan *et al*., 2002). This formalism entails calculating the betweenness of each edge, where the betweenness of a given edge is defined as the number of shortest paths between all pairs of residues that pass through that edge. Each path length is the sum of that path’s effective node-node distances assigned in the weighting scheme above. Each community identified is a group of nodes such that each node within the community is highly inter-connected (in terms of betweenness), but loosely connected to other nodes outside the community. Communities are thus densely inter-connected regions within proteins. The community partitions and the resultant critical residues for the canonical set are given in Figure 2.

Those residues that are involved in the highest-betweenness edges between pairs of interacting communities are identified as the interior-critical residues. These residues are essential for information flow between communities, as their removal would result in substantially longer paths between the residues of one community to those of another.

Software Tool: STRESS (STRucturally-identified ESSential residues)

We have made the implementations for finding surface- and interior-critical residues available through a new software tool, STRESS, which may be accessed at stress.molmovdb.org (Figure 3A). Users may submit a PDB file or a PDB ID corresponding to a structure to be analyzed, and the output provided constitutes the set of identified critical residues.

Running times are minimized by using a scalable server architecture that runs on the Amazon cloud (Figure 3). A light front-end server handles incoming user requests, and more powerful back-end servers, which perform the calculations, are automatically and dynamically scalable, thereby ensuring that they can handle varying levels of demand both efficiently and economically. In addition, the algorithmic implementation of our software is highly efficient, thereby obviating the need for long wait times. Relative to a naïve global Monte Carlo search implementation, local searches supported with hashing and additional algorithmic optimizations for computational efficiency reduce running times considerably (Figures 3B and 3C). A typical protein of ~500 residues takes only about 30 minutes on a 2.6GHz CPU.

High-Throughput Identification of Alternative Conformations

We use a generalized approach to systematically identify instances of alternative conformations throughout the PDB. We first perform multiple structure alignments (MSAs) across sequence-identical structures that are pre-filtered to ensure structural quality. We then use the resultant pairwise RMSD values to infer distinct conformational states (Figure S3; see also SI Methods section 3.2).

The distributions of the resultant numbers of conformations for domains and chains are given in Figures S3D and S3E, respectively, and an overview is given in Figure S3F. We note that the alternative conformations identified arise in an extremely diverse set of biological contexts, including conformational transitions that accompany

ligand binding, protein-protein or protein-nucleic acid interactions, post-translational modifications, changes in oxidation or oligomerization states, etc. The dataset of alternative conformations identified is provided as a resource in File S1 (see also Figure S3G).

Evaluating Conservation of Critical Residues Using Various Metrics and Sources of Data

The large dataset of dynamic proteins culled throughout the PDB, coupled with the high algorithmic efficiency of our critical residue search implementation, provide a means of identifying and evaluating general properties of a large pool of critical residues. In particular, we use a variety of conservation metrics and data sources to measure the inter- and intra-species conservation of the residues within this pool. As discussed below, we find that both surface- (Figures 4A-D) and interior-critical residues (Figures 4E-H) are consistently more conserved than non-critical residues. We emphasize that the signatures of conservation identified not only provide a means of rationalizing many of the otherwise poorly understood regions of proteins, but they also reinforce the functional importance of the residues predicted to be allosteric.

Conservation Across Species

When evaluating conservation across species, we find that both surface- and interior-critical residues tend to be significantly more conserved than non-critical residues with the same degree of burial (Figures 4B and 4F, respectively; note that negative conservation scores designate stronger conservation – see SI Methods section 3.3-a).

Leveraging Next-Generation Sequencing to Measure Conservation Amongst Humans

In addition to measuring inter-species conservation, we have also used fully sequenced human genomes and exomes to investigate conservation among human populations, as many constraints may be species-specific and active in more recent evolutionary history. Commonly used metrics for quantifying intra-species conservation include minor allele frequency (MAF) and derived allele frequency (DAF). Low MAF or DAF values are interpreted as signatures of deleteriousness, as purifying selection is prone to reduce the frequencies of harmful variants (see SI Methods section 3.3-b).

Non-synonymous single-nucleotide variants (SNVs) from the 1000 Genomes dataset (McVean *et al*., 2012) that intersect surface-critical residues tend to occur at lower DAF values than do SNVs that intersect non-critical residues (Figure 4C). Though this difference is not observed to be significant, the significance improves when examining the shift in DAF distributions, as evaluated with a KS test ($p= 0.159$, Figure S4A), and we point out only a limited number of proteins (thirty-two) for which these 1000 Genomes SNVs intersect with surface-critical sites. Furthermore, the long tail extending to lower DAF values for surface-critical residues may suggest that only a subset of the residues in our prioritized binding sites is essential. In contrast to surface-critical residues, however, interior-critical residues intersect 1000 Genomes SNVs with significantly lower DAF values than do non-critical residues (Figure 4G; see also Figure S4B).

When analyzing human polymorphism data, a variety of statistical measures relating SNVs to selective constraint may be calculated, and the results obtained (along with their associated significance levels) are highly dependent on sample size. 1000

Genomes datasets are attractive partially because of their status as a well-established "blue chip" set of variants in human populations. However, given the relatively limited number of proteins that intersect with 1000 Genomes SNVs, we also analyzed the larger dataset provided by the Exome Aggregation Consortium (ExAC) (Exome Aggregation Consortium, 2015). Though this dataset has been released much more recently (and is consequently not yet as well established as 1000 Genomes), ExAC provides sequence data from more than 60,000 individuals, and samples are sequenced at much higher coverage, thereby ensuring better data quality. This larger dataset enables us to more easily examine trends in the data as they relate to critical and non-critical residues.

Using MAF as a conservation metric, we performed a similar analysis using this data. MAF distributions for surface- and non-critical residues in the same set of proteins are given in Figure 4D. Although the mean value of the MAF distribution for surface-critical residues is slightly higher than that of non-critical residues, the median for surface-critical residues is substantially lower than that for non-critical residues, demonstrating that the majority of proteins are such that MAF values are lower in surface- than in non-critical residues. In addition, the overall shifts of these distributions also point to a trend of lower MAF values in surface-critical residues (Figure S4C, KS test p=9.49e-2).

Interior-critical residues exhibit significantly lower MAF values than do non-critical residues in the same set of proteins. MAF distributions for interior- and non-critical residues are given in Figure 4H (see also Figure S4D).

In addition to analyzing overall allele frequency distributions, we also evaluate the *fraction* of rare alleles as a metric for measuring selective pressure. This fraction is

defined as the ratio of the number of rare (i.e., low-DAF or low-MAF) non-synonymous SNVs to the number of all non-synonymous SNVs in a given protein annotation (such as all surface-critical residues of the protein, for example; see SI Methods section 3.3-b). A higher fraction is interpreted as a proxy for greater conservation (Khurana *et al*., 2013; Sethi *et al*., 2015). Using variable DAF (MAF) cutoffs to define rarity for 1000 Genomes (ExAC) SNVs, both surface- and interior-critical residues are shown to harbor a higher fraction of rare alleles than do non-critical residues, further suggesting a greater degree of evolutionary constraint on critical residues (Figure 5).

Comparisons Between Different Models of Protein Motions

The identification of surface- and interior-critical residues entails using sets of vectors (on each protein residue) to describe conformational change. Notably, our framework enables one to determine these vectors in multiple ways. Conformational changes may be modeled using vectors connecting residues in crystal structures from alternative conformations. We term this approach "ACT", for "absolute conformational transitions" (see SI Methods section 3.2-c). The crystal structures of such paired conformations may be obtained using the framework discussed above. The protein motions may also be inferred from anisotropic network models (ANMs) (Atilgan *et al*., 2001). ANMs entail modeling interacting residues as nodes linked by flexible springs, in a manner similar to elastic network models (Fuglebakk *et al*., 2015; Tirion, 1996) or normal modes analysis (Figure 1B). ANMs are not only simple and straightforward to apply on a database scale, but unlike using alternative crystal structures, the motion vectors inferred may be generated using a single structure.

We find that modeling conformational change using vectors from either ACTs or ANMs gives the same general trends in terms of the disparities in conservation between critical and non-critical residues. Our framework is thus general with respect to how the motion vectors are obtained (see Figure 6 and SI Methods section 3.2-c for further details).

Critical Residues in the Context of Human Disease Variants

Directly related to conservation is confidence with which an SNV is believed to be disease-associated. SIFT (Ng and Henikoff, 2001) and PolyPhen (Adzhubei *et al*., 2010) are two tools for predicting SNV deleteriousness. ExAC SNVs that intersect critical residues exhibit significantly higher PolyPhen scores relative to non-critical residues, suggesting the potentially higher disease susceptibility at critical residues (Figure S5). Significant disparities were not observed in SIFT scores (Figure S6).

Using HGMD (Stenson *et al*., 2014) and ClinVar (Landrum *et al*., 2014), we identify proteins with critical residues that coincide with disease-associated SNVs (File S2). Several critical residues coincide with known disease loci for which the mechanism of pathogenicity is otherwise unclear (File S3). The fibroblast growth factor receptor (FGFR) is a case-in-point (Figure 7A). SNVs in FGFR have been linked to craniofacial defects. Dotted lines in Figure 7B highlight poorly understood disease SNVs that coincide with critical residues. In addition, we identify Y328 as a surface-critical residue, which coincides with a disease-associated SNV from HGMD, despite the lack of confident predictions of deleteriousness by several widely used tools for predicting disease-associated SNVs, including PolyPhen (Adzhubei *et al*., 2010), SIFT (Ng and Henikoff, 2001), and SNPs&GO (Calabrese *et al*., 2009). Together, these results suggest

that the incorporation of surface- and interior-critical residues introduces a valuable layer of annotation to the protein sequence, and may help to explain otherwise poorly understood disease-associated SNVs.

DISCUSSION & CONCLUSIONS

The same principles of energy landscape theory that dictate protein folding are integral to how proteins explore different conformations once they adopt their fully folded states. These landscapes are shaped not only by the protein sequence itself, but also by extrinsic conditions. Such external factors often regulate protein activity by introducing allosteric-induced changes, which ultimately reflect changes in the shapes and population distributions of the energetic landscape. In this regard, allostery provides an ideal platform from which to study protein behavior in the context of their energetic landscapes. To investigate allosteric regulation, and to simultaneously add an extra layer of annotation to conservation patterns, an integrated framework to identify potential allosteric residues is essential. We introduce a framework to select such residues, using knowledge of conformational change.

When applied to many proteins with distinct conformational changes in the PDB, we investigate the conservation of potential allosteric residues in both inter-species and intra-human genomes contexts, and find that these residues tend to exhibit greater conservation in both cases. In addition, we identify several disease-associated variants for which plausible mechanisms had been unknown, but for which allosteric mechanisms provide a reasonable rationale.

Unlike the characterization of many other structural features, such as secondary structure assignment, residue burial, protein-protein interaction interfaces, disorder, and even stability, allostery inherently manifests through dynamic behavior. It is only by considering protein motions and changes in these motions can a fuller understanding of allosteric regulation be realized. As such, MD and NMR are some of the most common means of studying allostery and dynamic behavior (Kornev and Taylor, 2015). However, these methods have limitations when studying large and diverse protein datasets. MD is computationally expensive and impractical when studying large numbers of proteins. NMR structure determination is extremely labor-intensive and better suited to certain classes of structures or dynamics. In addition, NMR structures constitute a relatively small fraction of structures currently available.

Despite these limitations in MD and NMR, allosteric mechanisms and signaling pathways may be conserved across many different but related proteins within the same family, suggesting that such computationally- or labor-intensive approaches for all proteins may not be entirely essential. Flock *et al*. have carefully demonstrated that the allosteric mechanisms responsible for regulating G proteins through GPCRs tend to be conserved (Flock *et al*., 2015). Investigations into representative families have also been enlightening in other contexts. In one of the early studies employing network analysis, del Sol *et al*. conduct a detailed study of several allosteric protein families (including GPCRs) to demonstrate that residues important for maintaining the integrity of short paths within residue contact networks are essential to enabling signal transmission between distant sites (del Sol *et al.*, 2006). Another notable result in the same work is that these key residues (which match experimental results) may become redistributed when

the protein undergoes conformational change, thereby changing optimal communication routes as a means of conferring different regulatory properties.

There are several notable implications of our dynamics-based analysis across a database of proteins. Relative to sequence data, allostery and dynamic behavior are far more difficult to evaluate on a large scale. The framework described here enables one to evaluate dynamic behavior in a systemized and efficient way across many proteins, while simultaneously capturing residues on both the surface and within the interior. That this pipeline can be applied in a high-throughput manner enables the investigation of system-wide phenomena, such as the roles of potential allosteric hotspots in protein-protein interaction networks.

It is only by analyzing a large dataset of proteins can one investigate general trends in predicted allosteric residues. In addition, the implementation detailed here enables one to match structural features with the high-throughput data generated through deep sequencing initiatives, which are providing an unprecedented window into conservation patterns, many of which may be human-specific.

We anticipate that, within the next decade, deep sequencing will enable structural biologists to study evolutionary conservation using sequenced human exomes just as routinely as cross-species alignments. Furthermore, intra-species metrics for conservation provide added value in that the confounding factors of cross-species comparisons are removed: different species evolve in various evolutionary contexts and at different rates, and it can be difficult to decouple these different effects from one another. Cross-species metrics of protein conservation entail comparisons between proteins that may be very different in structure and function. Sequence-variable regions across species may not be

conserved, but nevertheless impart essential functionality. Intra-species comparisons, however, can often provide a more direct and sensitive evaluation of constraint.

In particular, selective constraints within human populations are particularly relevant to understanding human disease. Formalisms for analyzing large structural and sequence datasets will become increasingly important in the context of human health. We anticipate that the framework and formalisms detailed here, along with the accompanying web tool we have introduced, will help to further motivate future studies along these directions.

METHODS

An overview of the framework for finding surface- and interior-critical residues is given in Figure 1. Figure S3A provides a schematic of our pipeline for identifying alternative conformations throughout the PDB. Cross-species conservation scores were analyzed in those PDBs for which full ConSurf files are available through the ConSurf server. 1000 Genomes SNVs were taken from the Phase 3 release, and ExAC SNVs were downloaded in May 2015. Further details on all protocols are provided in SI Methods.

AUTHOR CONTRIBUTIONS

D.C., A.N.S., and M.G. conceived and designed the study. S.K. mapped all variants and participated in their analysis. S.T.L. and R.W.F.C. optimized the software and developed the web server. J.C. aided in ExAC variant analysis. All other authors wrote or edited the manuscript and provided valuable feedback. A.N.S. and M.G. oversaw the project.

ACKNOWLEDGMENTS

DC acknowledges the support of the NIH Predoctoral Program in Biophysics (T32 GM008283-24). We thank Simon Mitternacht for sharing the original source code for binding leverage calculations, as well as Koon-Kiu Yan for helpful discussions and feedback. The authors would like to thank the Exome Aggregation Consortium and the groups that provided exome variant data for comparison. A full list of contributing groups can be found at http://exac.broadinstitute.org/about

REFERENCES

Adzhubei, I. Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods. *7*, 248–249

Ansari, A., Berendzen, J., Bowne, S., Frauenfelder, H., Iben, I.E.T., Sauke, T.B., Shyamsunder, E., and Young, R.D. (1985). Protein states and protein quakes. Proc. Natl. Acad. Sci. U.S.A. 82, 5000–5004.

Arnlund, D., Johansson, L.C., Wickstrand, C., Barty, A., Williams, G.J., Malmerberg, E., Davidsson, J., Milathianaki, D., DePonte, D.P., Shoeman, R.L., *et al*. (2014). Visualizing a protein quake with time-resolved X-ray scattering at a free-electron laser. Nat. Methods. *11*, 923–6.

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I. (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. Biophys. J. *80*, 505–515.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L. and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum. Mutat. *30*, 1237–1244.

Exome Aggregation Consortium, Lek, M., Karczewski, K., Minikel, E., Samocha, K., Banks, E., Fennell, T., O'Donnell-Luria, A., Ware, J., Hill, A., et al. (2015). Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 030338

Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M. and Funkhouser, T.A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput. Biol. *5*, e1000585.

Chennubhotla, C. and Bahar, I. (2006). Markov propagation of allosteric effects in biomolecular systems: application to GroEL–GroES. Mol. Syst. Biol. *2*.

del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol. Syst. Biol. 2(1).

Csermely, P., Korcsmáros, T., Kiss, H.J.M., London, G., and Nussinov, R. (2013). Structure and dynamics of molecular networks: A novel paradigm of drug discovery. Pharmacol. Ther. *138*, 333–408.

Flock, T., Ravarani, C.N.J., Sun, D., Venkatakrishnan, A.J., Kayikci, M., Tate, C.G., Veprintsev, D.B. and Babu, M.M. (2015). Universal allosteric mechanism for Gα activation by GPCRs. Nature *524*, 173–179.

Fuglebakk, E., Tiwari, S.P., and Reuter, N. (2015). Comparing the intrinsic dynamics of multiple protein structures using elastic network models. Biochim. Biophys. Acta - Gen. Subj. *1850*, 911–922.

Gasper, P.M., Fuglestad, B., Komives, E.A., Markwick, P.R.L., and McCammon, J.A. (2012). Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. Proc. Natl. Acad. Sci. U. S. A. *109*, 21216–22.

Ghosh, A., and Vishveshwara, S. (2008). Variations in Clique and Community Patterns in Protein Structures during Allosteric Communication: Investigation of Dynamically Equilibrated Structures of Methionyl tRNA Synthetase Complexes. Biochemistry.

47, 11398-11407.

Girvan, M., Girvan, M., Newman, M.E.J., and Newman, M.E.J. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. U. S. A. *99*, 7821–7826.

Halabi, N., Rivoire, O., Leibler, S., and Ranganathan, R. (2009). Protein Sectors: Evolutionary Units of Three-Dimensional Structure. Cell *138*, 774–786.

Hubbard, S.J. and Thornton, J.M. (1993). 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London.

Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., *et al*. (2013). Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. Science. *342*, 1235587–1235587.

Kornev, A.P. and Taylor, S.S. (2015). Dynamics-Driven Allostery in Protein Kinases. Trends Biochem. Sci. *xx*, 1–20.

Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–5.

Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J., and Ranganathan, R. (2008). Surface Sites for Engineering Allosteric Control in Proteins. Science *322*, 438-442.

Lockless, S.W., Ranganathan, R., Kukic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri, P., Pollastri, G., Socolich, M., Lockless, S.W., *et al*. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. BMC

Bioinformatics *15*, 295–299.

McVean, G.A., Altshuler (Co-Chair), D.M., Durbin (Co-Chair), R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., Flicek, P., *et al*. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Ming, D. and Wall, M.E. (2005). Quantifying allosteric effects in proteins. Proteins *59*, 697–707.

Mitternacht, S. and Berezovsky, I.N. (2011). Binding leverage as a molecular basis for allosteric regulation. PLoS Comput. Biol. *7*, e1002148.

Miyashita, O., Onuchic, J.N., and Wolynes, P.G. (2003). Nonlinear elasticity, protein quakes, and the energy landscapes of functional transitions in proteins. Proc. Natl. Acad. Sci. *100*, 12570–12575.

Ng, P.C. and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. Genome Res. *11*, 863–874.

Panjkovich, A. and Daura, X. (2012). Exploiting protein flexibility to predict the location of allosteric sites. BMC Bioinformatics *13*, 273.

Panjkovich, A. and Daura, X. (2010). Assessing the structural conservation of protein pockets to study functional and allosteric sites: implications for drug discovery. BMC Struct. Biol. *10*, 9.

Reynolds, K.A., McLaughlin, R.N., and Ranganathan, R. (2011). Hot Spots for Allosteric Regulation on Protein Surfaces. Cell *147*, 1564–1575.

Rivalta, I., Sultan, M.M., Lee, N.-S., Manley, G. a., Loria, J.P., and Batista, V.S. (2012). PNAS Plus: Allosteric pathways in imidazole glycerol phosphate synthase. Proc.

Natl. Acad. Sci. *109*, E1428–E1436.

Rodgers, T.L., Townsend, P.D., Burnell, D., Jones, M.L., Richards, S.A., McLeish, T.C.B., Pohl, E., Wilson, M.R., and Cann, M.J. (2013). Modulation of Global Low-Frequency Motions Underlies Allosteric Regulation: Demonstration in CRP/FNR Family Transcription Factors. PLoS Biol. *11*, e1001651.

Rousseau, F. and Schymkowitz, J. (2005). A systems biology perspective on protein structural dynamics and signal transduction. Curr. Opin. Struct. Biol. *15*, 23–30.

Sethi, A., Eargle, J., Black, A.A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA:protein complexes. Proc. Natl. Acad. Sci. U. S. A. *106*, 6620–5.

Sethi, A., Clarke, D., Chen, J., Kumar, S., Galeev, T.R., Regan, L., and Gerstein, M. (2015). Reads meet rotamers: structural biology in the age of deep sequencing. Curr. Opin. Struct. Biol. *35*, 125-134.

Shulman, A.I., Larson, C., Mangelsdorf, D.J., and Ranganathan, R. (2004). Structural determinants of allosteric ligand activation in RXR heterodimers. Cell *116*, 417–429.

Stenson, P.D., Mort, M., Ball, E. V., Shaw, K., Phillips, A.D., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1–9.

Süel, G.M., Lockless, S.W., Wall, M.A., and Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat. Struct. Biol. *10*, 59–69.

Tirion, M.M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-

Parameter, Atomic Analysis. Phys. Rev. Lett. *77*, 1905–1908.

Tsai, C., Ma, B. and Nussinov, R. (1999). Folding and binding cascades: Shifts in energy landscapes. Proc. Natl. Acad. Sci. U. S. A. *96*, 9970–9972.

Vanwart, A.T., Eargle, J., Luthey-Schulten, Z., and Amaro, R.E. (2012). Exploring residue component contributions to dynamical network models of allostery. J. Chem. Theory Comput. *8*, 2949–2961.

Yang, L.W. and Bahar, I. (2005). Coupling between catalytic site and collective dynamics: A requirement for mechanochemical activity of enzymes. Structure *13*, 893–904.

CAPTIONS

Figure 1. Schematic overviews of methods for finding surface- and interior-critical residues. (*A*) A simulated ligand probes the protein surface in a series of Monte Carlo simulations (top-left). The cavities identified may be such that occlusion by the ligand strongly interferes with conformational change (top-right; such a site is likely to be identified as surface-critical, in red), or they may have little effect on conformational change, as in the case of shallow pockets (bottom-left) or pockets in which large-scale motions do not drastically affect pocket volume (bottom-right). (*B*) Interior-critical residues are identified by weighting residue-residue contacts (edges) on the basis of correlated motions, and then identifying communities within the weighted network. Residues involved in the highest-betweenness interactions between communities (in red) are selected as interior-critical residues.

Figure 2. Community partitioning for canonical systems. Different network communities are colored differently, and communities were identified using the dynamical network-based analysis with the GN formalism discussed in the main text and in SI Methods section 3.1-b. Residues shown as spheres are interior-critical residues, and they are colored based on community membership, and black lines connecting pairs of critical residues represent the highest-betweenness edges between the corresponding communities. See also Table S3.

Figure 3. STRESS web server front page, running times, and server architecture. (A) The server enables users to either provide PDB IDs or to upload their own PDB files for proteins of interest. Users may opt to identify surface-critical residues, interior-critical residues, or both. A thin front-end server handles incoming user requests, and more powerful back-end servers perform the heavier algorithmic calculations. The back-end servers are dynamically scalable, making them capable of handling wide fluctuations in user demand. Amazon's Simple Queue Service is used to coordinate between user requests at the front end and the back-end compute nodes: when the front-end server receives a request, it adds the job to the queue, and back-end servers pull that job from the queue when ready. Source code is available through Github (github.com/gersteinlab/STRESS). (B) Running times are shown for systems of various sizes. Shown in red are the running times without optimizing for speed, and green shows running times with algorithmic optimization. (C) The same data is represented as a log-log plot. The slopes of these two approaches demonstrate that our algorithm reduces the computational complexity by an order of magnitude. Our speed-optimized algorithm scales at $O(n^{1.3})$, where n is the number of residues.

Figure 4. Multiple metrics and datasets reveal that critical residues tend to be conserved. Surface- and interior-critical residues (red) in phosphofructokinase (PDB 3PFK) are given in (*A*) and (*E*), respectively. Distributions of cross-species conservation scores, 1000 Genomes SNV DAF averages, and ExAC SNV MAF averages for surface- and non-critical residue sets are given in (*B*), (*C*), and (*D*), respectively. The same distributions corresponding to interior- and non-critical residue sets are given in (F), (G),

and (H), respectively. In (B), mean inter-species conservation scores for surface-critical sets are -0.131, whereas non-critical residue sets with the same degree of burial have a mean score of +0.059 ($p < 2.2e-16$). In (F), mean inter-species conservation scores for interior-critical sets are -0.179, whereas non-critical residue sets with the same degree of burial have a mean score of -0.102 ($p=3.67e-11$). In (C), means for surface- and non-critical sets are 9.10e-4 and 8.34e-4, respectively ($p=0.309$); corresponding means in (*D*) are 4.09e-04 and 2.26e-04, respectively ($p=1.49e-3$). In (*G*), means for interior- and non-critical sets are 2.82e-4 and 3.12e-3, respectively ($p=1.80e-05$); corresponding means in (*H*) are 3.08e-05 and 3.27e-04, respectively ($p=7.98e-09$). N = 421, 32, 84, 517, 31, and 90 structures for panels B, C, D, F, G, and H, respectively. P-values are based on Wilcoxon-rank sum tests. See SI Methods for further details. See also Figures S2 and S4.

Figure 5. Critical residues are shown to be more conserved, as measured by the fraction of rare alleles. Protein regions with high fractions of *rare* variants are believed to be more sensitive to sequence variants than other regions, thereby explaining why such variants occur infrequently in the population. Panels *(A)* and *(C)* show distributions for rare (low DAF) non-synonymous SNVs (taken from the 1000 Genomes dataset) in which the critical residues are defined to be the surface-critical *(A)* and interior-critical *(C)* residues. Panels *(B)* and *(D)* show distributions for rare (low MAF) non-synonymous SNVs (taken from the ExAC dataset) in which the critical residues are defined to be the surface-critical *(B)* and interior-critical *(D)* residues. For varying thresholds to define rarity, there are more structures in which the fraction of rare variants is higher in critical residues than in non-critical residues. Cases in which the fraction is equal in both

categories are not shown. We consider all structures such that at least one critical and at least one non-critical residue intersect a non-synonymous SNV. Panels *(A)*, *(B)*, *(C)*, and *(D)* represent data from 31, 90, 32, and 84 structures, respectively.

Figure 6. Modeling protein conformational change through a direct use of crystal structures from alternative conformations using absolute conformational transitions (ACT). *(A)* Distributions (155 structures) of the mean conservation scores on surface-critical (red) and non-critical residues with the same degree of burial (blue). *(B)* Distributions (159 structures) of the mean conservation scores for interior-critical (red) and non-critical residues with the same degree of burial (blue). Mean values are given in parentheses. Results for single-chain proteins are shown, and p-values were calculated using a Wilcoxon rank sum test. See also Figure S3.

Figure 7. Potential allosteric residues add a layer of annotation to structures in the context of disease-associated SNVs. The structure shown (*A*) is that of the fibroblast growth-factor receptor (FGFR) in VMD Surf rendering, with HGMD SNVs shown in orange, bound to FGF2, in ribbon rendering (PDB 1IIL). (*B*) A linear representation of structural annotation for FGFR. Dotted lines highlight loci which 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 P21802). See also Figures S5 and S6.

Table 1. Statistics on the surfaces of *apo* structures within the canonical set of proteins

For each *apo* structure within the canonical set of proteins, statistics relating surface-critical sites to known ligand-binding sites are reported. The surface of a given structure is defined to be the set of all residues that have a relative solvent accessibility of at least 50%, where relative solvent accessibility is evaluated using all heavy atoms in both the main-chain and side-chain of a given residue. Mean values are given in the bottom row. NACCESS is used to calculate relative solvent accessibility (Hubbard and Thornton, 1993). *Column 1*: Protein name and PDB IDs for each structure; *Column 2*: among these surface residues, the fraction that constitute surface-critical (SC) residues; *Column 3*: among surface residues, the fraction that constitute known ligand-binding (LB) residues (known ligand-binding residues are taken to be those within 4.5 Angstroms of the ligand in the *holo* structure; Table S1); *Column 4*: the Jaccard similarity between the sets of residues represented in columns 2 and 3 (i.e., surface-critical and known-ligand binding residues), where values given in parentheses represent the expected Jaccard similarity, given a null model in which surface-critical and ligand-binding residues are randomly distributed throughout the surface (for each structure, 10,000 simulations are performed to produce random distributions, and the expected values reported here constitute the mean Jaccard similarity among the 10,000 simulations for each structure); *Column 5*: the number of distinct surface-critical sites identified in each structure; *Column 6*: the

number of known ligand-binding sites in each structure; *Column 7*: the number of known ligand-binding sites which are positively identified within the set of surface-critical sites, where a positive match occurs if a majority of the residues in a surface-critical site coincide with the known ligand-binding site; *Column 8*: The fraction of ligand-binding sites captured is simply the ratio of the values in column 7 to those in column 6. See also Figure S1, Table S1, and Table S2.

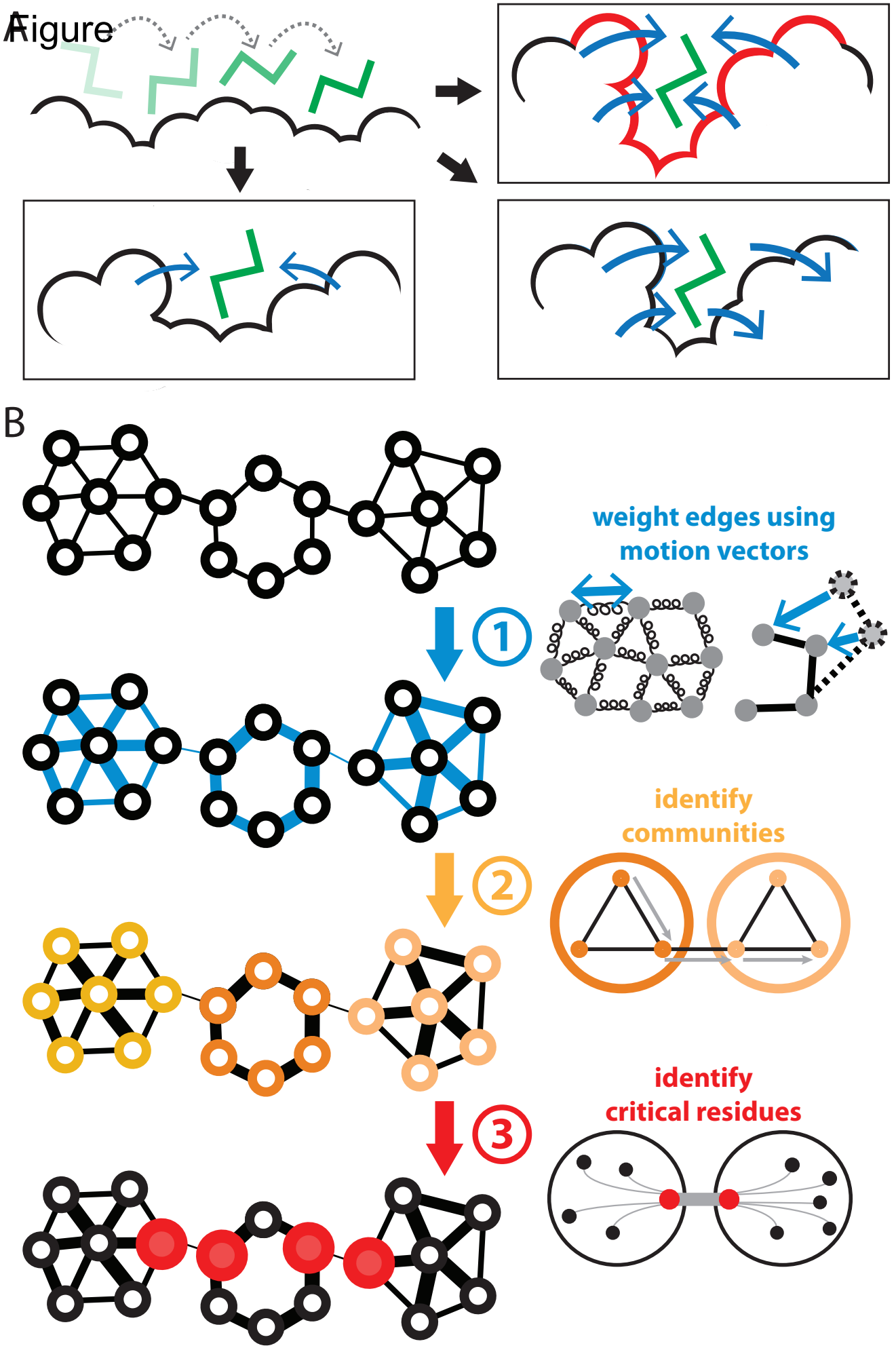
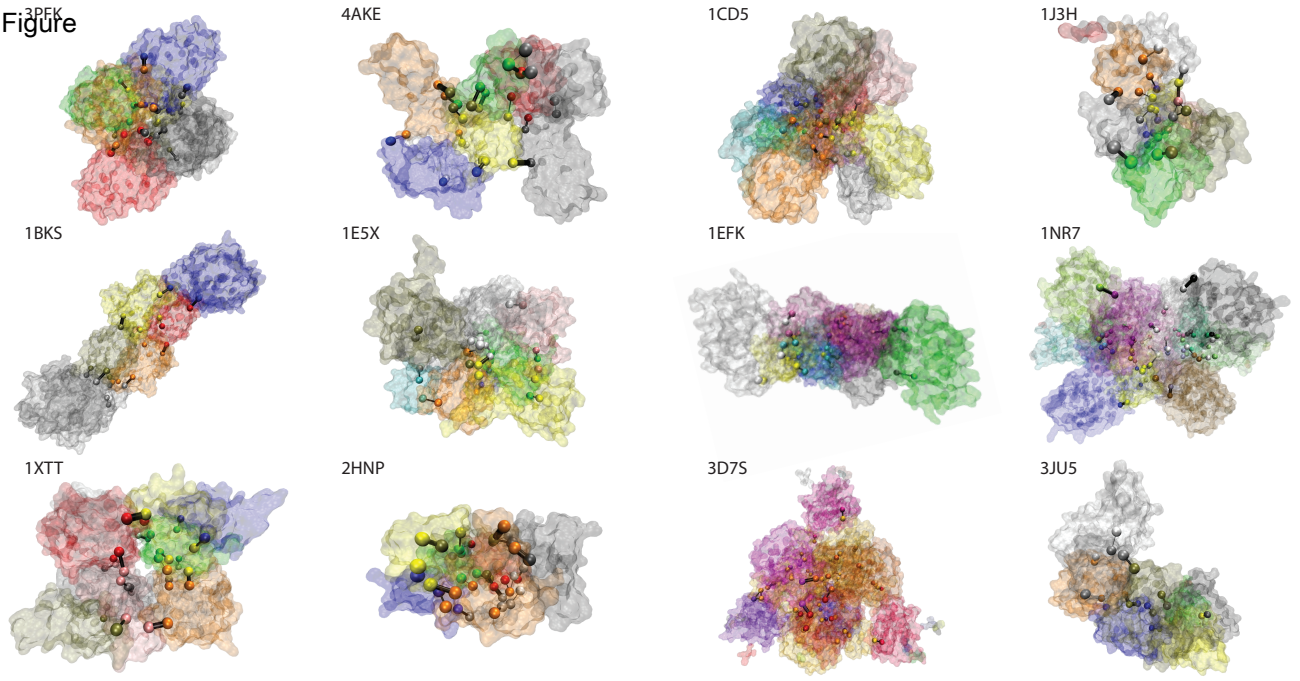
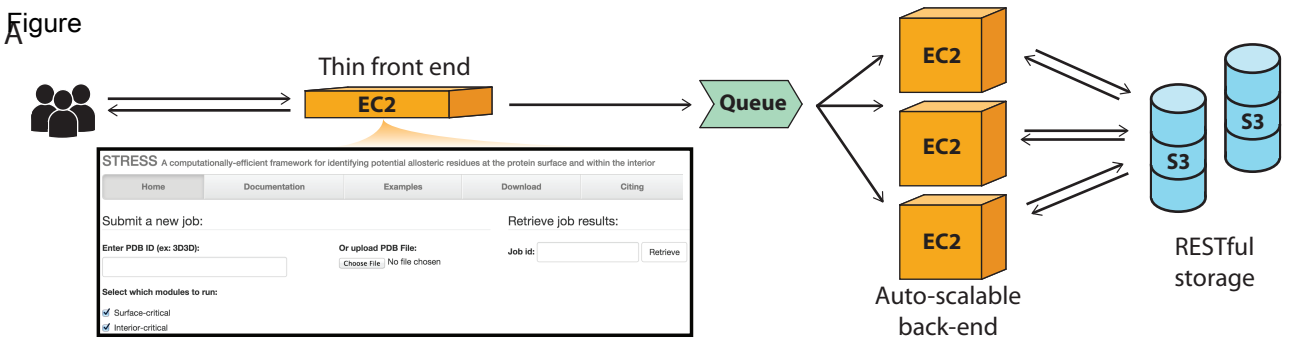


Figure
B
weight edges using
motion vectors
1
identify
communities
2
identify
critical residues
3

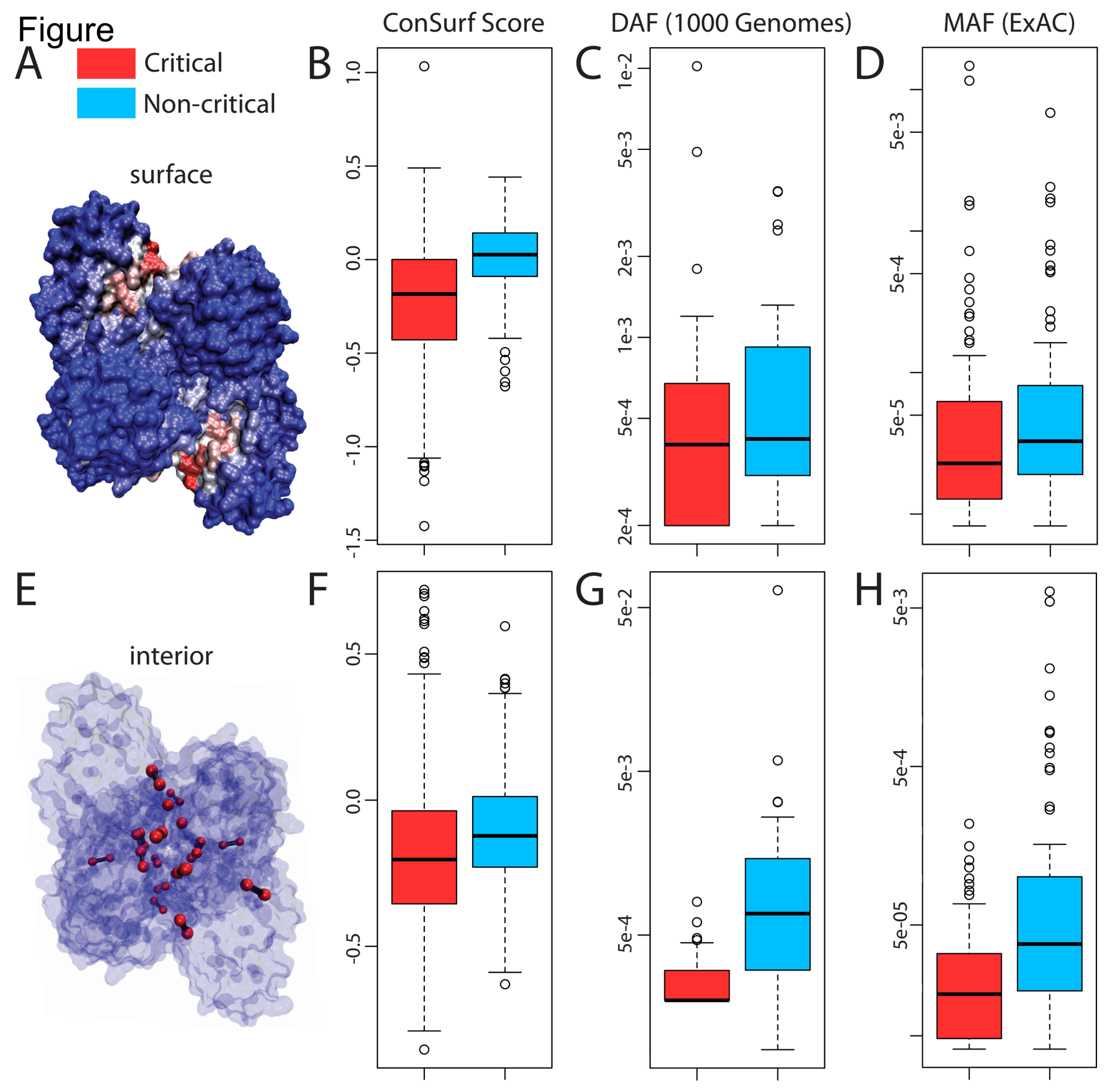
Figure



Figure


A
Thin front end
EC2
Queue
EC2
EC2
EC2
S3
S3
RESTful storage
Auto-scalable back-end
STRESS A computationally-efficient framework for identifying potential allosteric residues at the protein surface and within the interior
Home
Documentation
Examples
Download
Citing
Submit a new job:
Retrieve job results:
Enter PDB ID (ex: 3D3D):
Or upload PDB File:
Choose File
No file chosen
Job id:
Retrieve
Select which modules to run:
Surface-critical
Interior-critical
B
Running time (hours)
Num. Residues
Naïve implementation
Optimized implementation
C
Running time (sec, log scale)
Num. Residues (log scale)
slope: 2.344
slope: 1.307

Figure


A
Critical
Non-critical
surface
B
ConSurf Score
1.0
0.5
0.0
-0.5
-1.0
-1.5
C
DAF (1000 Genomes)
1e-2
5e-3
2e-3
1e-3
5e-4
2e-4
D
MAF (ExAC)
5e-3
5e-4
5e-5
E
interior
F
0.5
0.0
-0.5
G
5e-2
5e-3
5e-4
H
5e-3
5e-4
5e-05

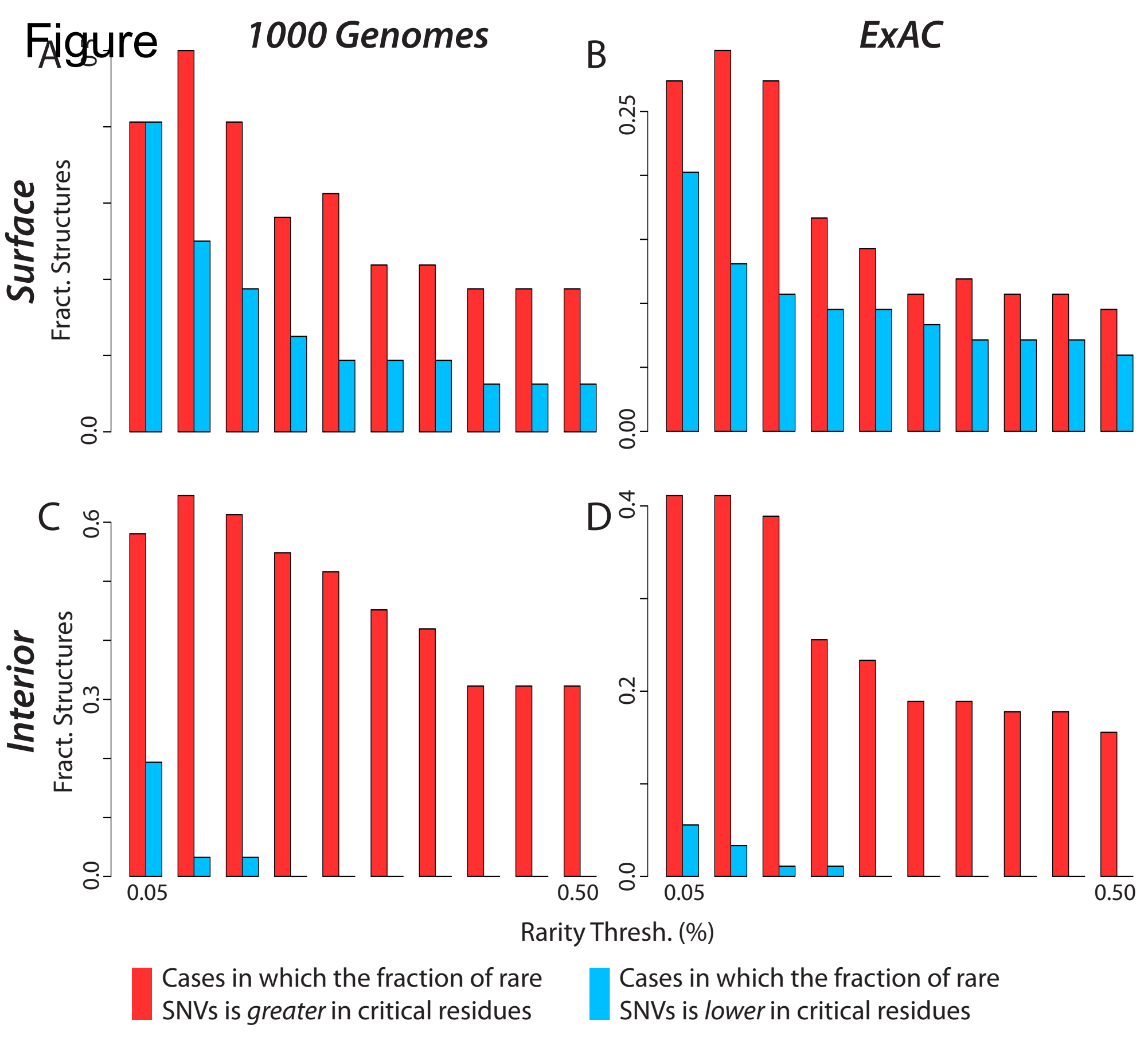


Figure
1000 Genomes
ExAC
A
B
C
D
Surface
Interior
Fract. Structures
0.0
0.3
0.6
0.00
0.25
0.2
0.4
0.05
0.50
Rarity Thresh. (%)
Cases in which the fraction of rare SNVs is greater in critical residues
Cases in which the fraction of rare SNVs is lower in critical residues

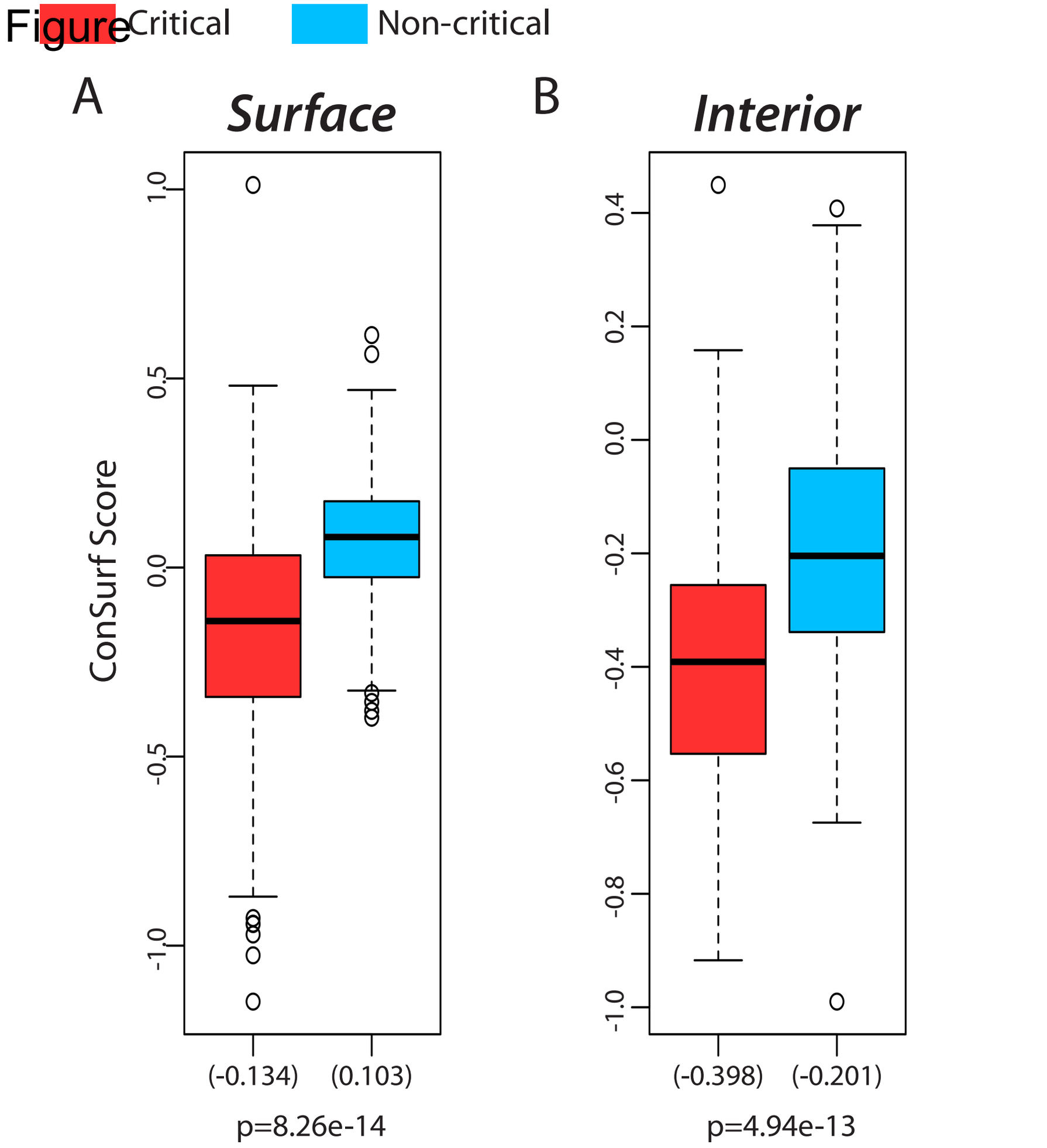
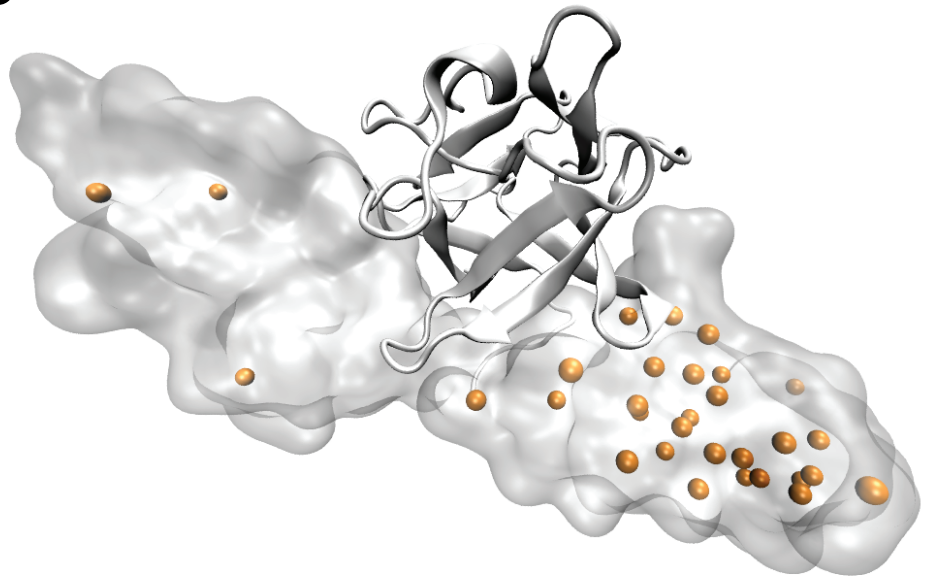


Figure
Critical
Non-critical
A
Surface
B
Interior
ConSurf Score
1.0
0.5
0.0
-0.5
-1.0
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
(-0.134)
(0.103)
p=8.26e-14
(-0.398)
(-0.201)
p=4.94e-13

Figure A



- Predicted allosteric (surface)
- Predicted allosteric (interior)
- Hinge residues
- Buried residues
- Protein-protein interaction site
- Post-translational modifications
- HGMD site (w/o annot. overlap)
- HGMD site (w/annot. overlap)

B

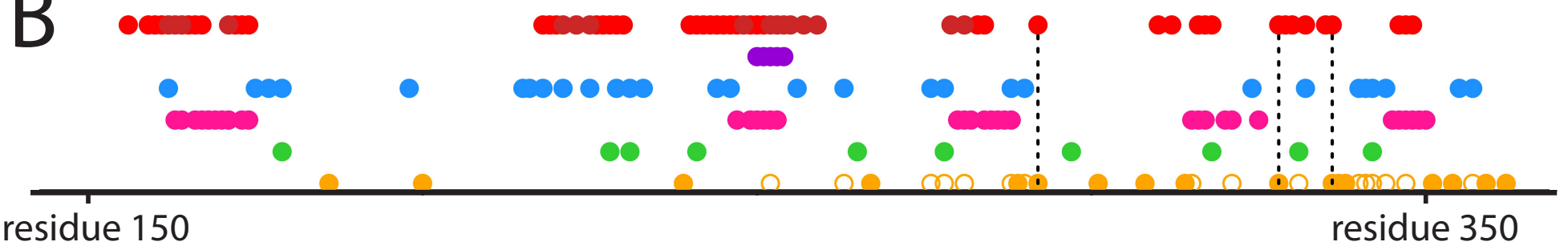


Table 1. Statistics on the surfaces of apo structures within the canonical set of proteins

Protein name (pdb ID)	% Surf (SC res)	% Surf (LB res)	SC-LB overlap	# SC sites	# LB sites	# Overlapping sites	% LB sites identified
Phosphofructokinase (3pfk)	51.0	20.4	0.255 (0.155)	19	3	3	100.0
Adenylate kinase (4ake)	45.4	17.8	0.274 (0.154)	29	2	2	100.0
G-6-P deaminase (1cd5)	58.9	10.0	0.153 (0.096)	24	2	1	50.0
cAMP-dep. prot. kin. (1j3h)	6.6	8.0	0.25 (0.041)	2	1	1	100.0
Trp synthase (1bks)	34.3	9.7	0.079 (0.079)	24	4	1	25.0
Thr synthase (1e5x)	20.7	9.3	0.139 (0.077)	17	3	2	66.7
Hum. malic enzyme (1efk)	5.5	8.6	0.03 (0.036)	10	10	0	0.0
Glu dehydrogenase (1nr7)	14.9	17.5	0.187 (0.102)	45	24	6	25.0
P-ribosyltransferase (1xtt)	29.8	19.6	0.295 (0.154)	31	5	5	100.0
Tyr phosphatase (2hnp)	73.9	13.3	0.16 (0.134)	25	2	2	100.0
Asp transcarbamoylase (3d7s)	26.7	13.7	0.054 (0.064)	26	9	0	0.0
Arg kinase (3ju5)	1.6	3.9	0 (0.013)	1	2	0	0.0
mean	30.8	12.7	0.156 (0.092)	21.083	5.583	1.917	55.6

Supplemental Information

1 – Supplemental Figures

2 – Supplemental Tables

3 – Supplemental Methods

4 – Supplemental References

1 – Supplemental Figures

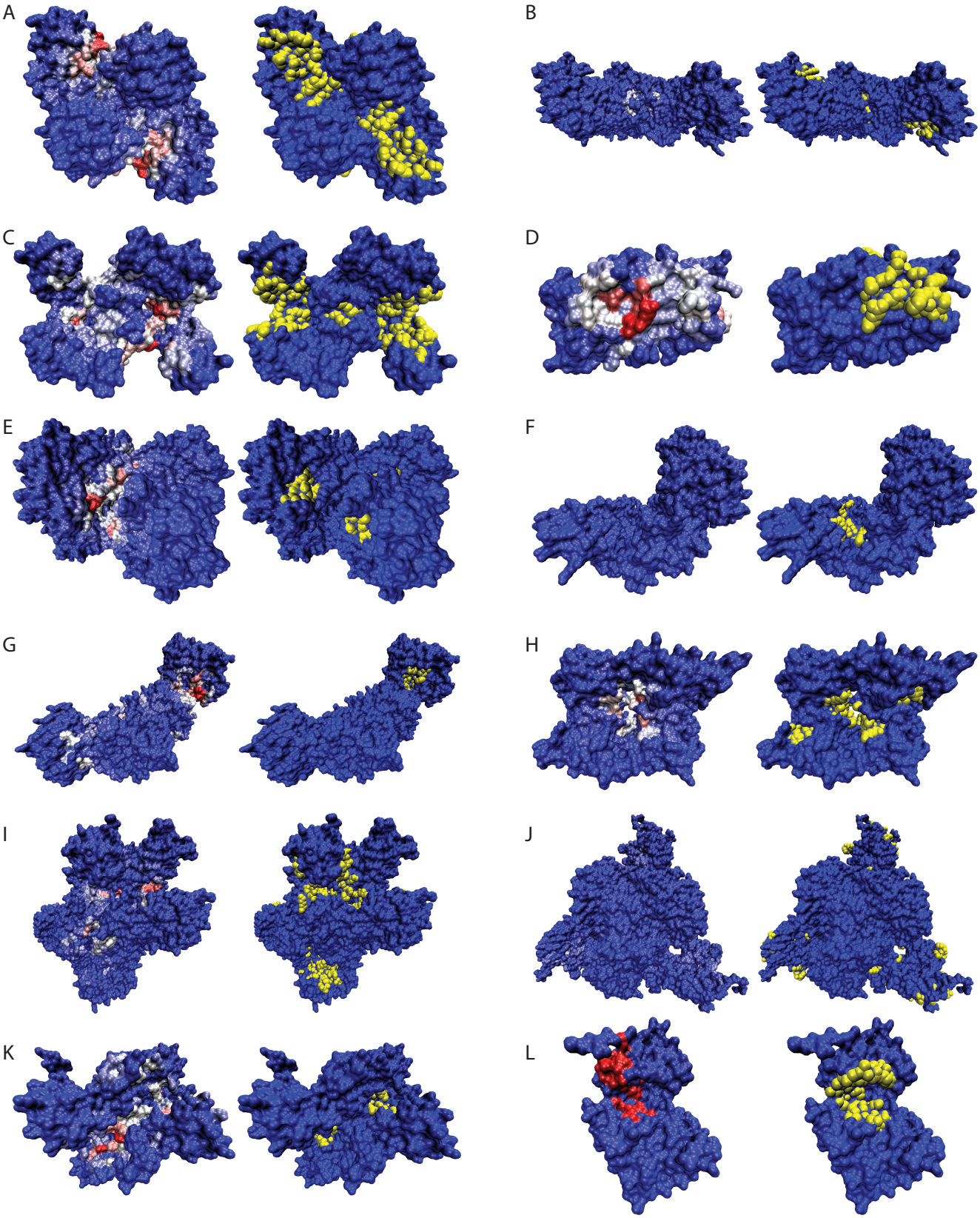


Figure S1, related to Table 1. Canonical proteins with surface-critical and known ligand-binding sites. Left panels show sites that are scored highly (i.e., surface-critical residues, in red). Right panels show residues (yellow) that directly contact ligands, based on the *holo* structure (see Table S1). PDB IDs: *(A)* 3PFK; *(B)* 1EFK; *(C)* 4AKE; *(D)* 2HNP; *(E)* 1CD5; *(F)* 3JU5; *(G)* 1BKS; *(H)* 1XTT; *(I)* 1NR7; *(J)* 3D7S; *(K)* 1E5X; *(L)* 1J3H.

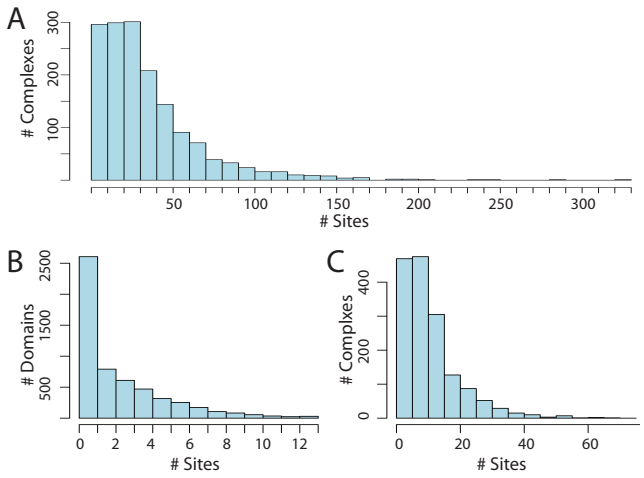


Figure S2, related to Figure 4. Summary statistics for surface-critical sites.
Panel *(A)* shows the distribution of the number of surface-critical sites per complex without applying thresholds, with complexes represented in biological assembly files downloaded from the PDB. Without applying thresholds to the list of ranked surface-critical sites, the protein is often covered with an excess of identified critical sites. Distributions of the numbers of distinct surface-critical sites per domain and per complex are given in panels (*B*) and (*C*), respectively.

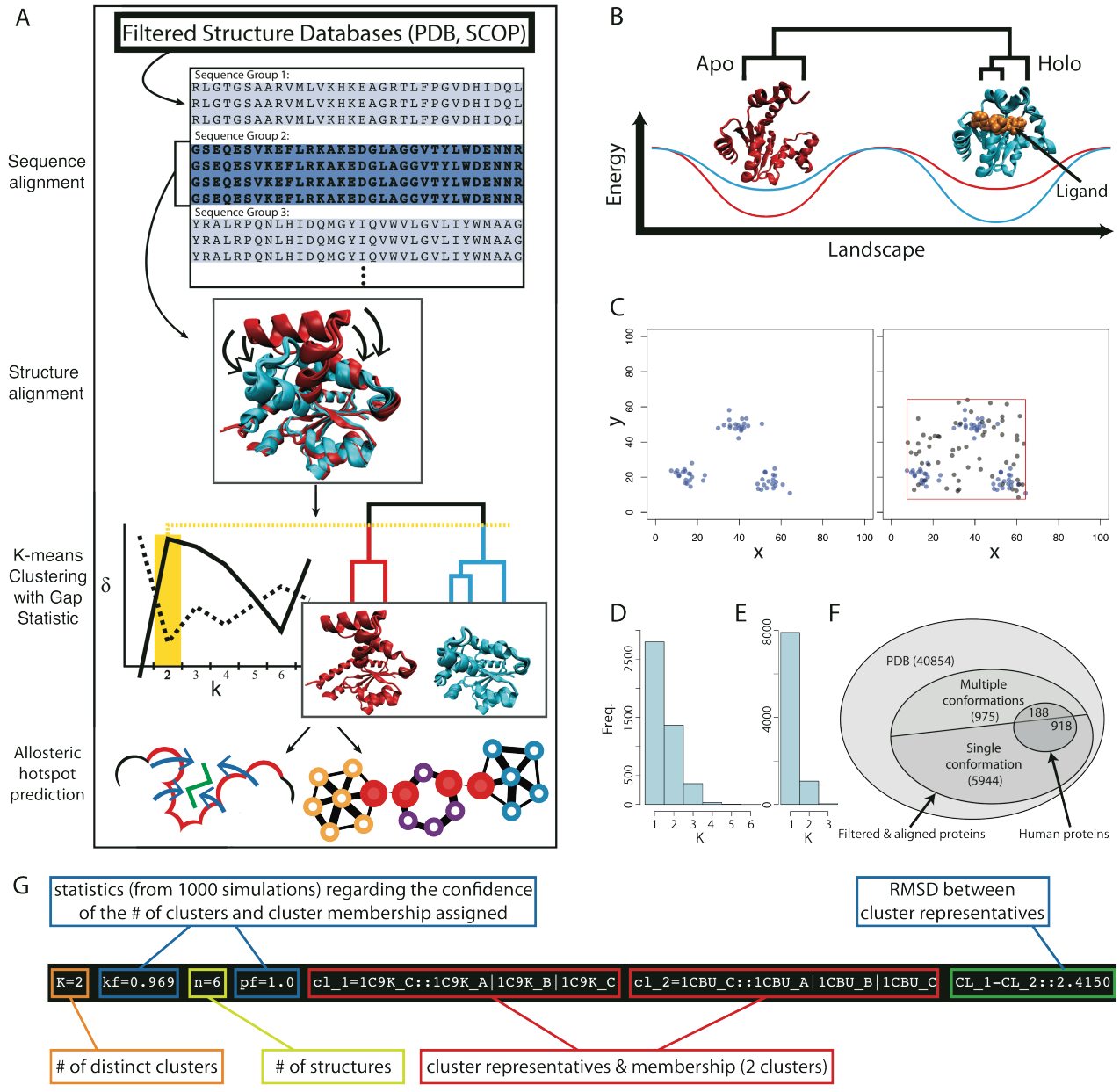

A
Filtered Structure Databases (PDB, SCOP)
Sequence Group 1:
RLGTGSAARVMLVKHKEAGRTLFPGVDHIDQL
RLGTGSAARVMLVKHKEAGRTLFPGVDHIDQL
RLGTGSAARVMLVKHKEAGRTLFPGVDHIDQL
Sequence Group 2:
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
GSEQESVKEFLRKAKEDGLAGGVTYLWDENNR
Sequence Group 3:
YRALRPQNLHIDQMGYIQVWVLGVLIYWMAAG
YRALRPQNLHIDQMGYIQVWVLGVLIYWMAAG
YRALRPQNLHIDQMGYIQVWVLGVLIYWMAAG
Sequence alignment
Structure alignment
K-means Clustering with Gap Statistic
Allosteric hotspot prediction
δ
k
B
Apo
Holo
Ligand
Energy
Landscape
C
y
x
D
Freq.
K
E
K
F
PDB (40854)
Multiple conformations (975)
188
918
Single conformation (5944)
Filtered & aligned proteins
Human proteins
G
statistics (from 1000 simulations) regarding the confidence of the # of clusters and cluster membership assigned
RMSD between cluster representatives
K=2
kf=0.969
n=6
pf=1.0
cl_1=1C9K_C::1C9K_A|1C9K_B|1C9K_C
cl_2=1CBU_C::1CBU_A|1CBU_B|1CBU_C
CL_1-CL_2::2.4150
# of distinct clusters
# of structures
cluster representatives & membership (2 clusters)

Figure S3, related to Figure 6. Pipeline for identifying alternative conformations throughout the PDB.

(A) Pipeline for identifying distinct conformations and critical residues: *Top to bottom*: BLASTClust is applied to the sequences corresponding to a filtered set of structures, thereby providing a large number of sequence-identical sets of proteins (i.e., "sequence groups"). For each sequence-identical group, a multiple structure alignment is performed using STAMP. The example shown here is adenylate kinase. Using the pairwise RMSD values in this structure alignment, the structures are clustered using the UPGMA algorithm, and K-means with the gap statistic (δ) is performed to identify the number of distinct conformations. The plot at left identifies 2 as the optimal value for K: the solid line represents δ(K) values at each value of K, and the dotted line represents $\delta(K+1) - s_{k+1}$ for each value of K (see SI Methods section 3.2-b for details). The structures that exhibit multiple clusters (i.e., those with K > 1) are then taken to exhibit multiple conformations. Finally, surface-critical (bottom-left) and interior-critical (bottom-right) residues are identified on those proteins determined to exist as multiple conformations. *(B)* Energy landscapes to describe distributions of different conformations. Energy landscape theory may be used to describe the relative populations of alternative biological states and conformations (for instance, active/inactive, or *holo*/*apo*). In the *apo* state, the landscape may take the form of the red curve, resulting in most proteins favoring the conformation shown in red. Once binding to ligand, the landscape becomes reconfigured to take the shape in the cyan curve, thereby shifting the distribution of conformations to that shown in cyan. One may use multiple structure alignments for domains or proteins to identify these distinct biological states in a database of structures. The schematized dendrogram represents the partitioning of these structures by a metric such as RMSD. The example shown is a multiple structure alignment of adenylate kinase. SCOP IDs of the *apo* domains: d4akea1 and d4akeb1; those of the *holo* domains: d3hpqb1, d3hpqa1, d2eckb1, d2ecka1, d1akeb1, and d1akea1. *(C)* Intuition behind the k-means algorithm with the gap statistic. The objective is to identify the ideal number of clusters to describe the observed data of 60 points (in blue). This entails defining how well-clustered our observed data appears (given an assigned number of clusters, K) relative to a null model consisting of a randomly distributed set of 60 points (grey) that fall within the same variable ranges as the observed data. Further details are provided by Tibshirani et al, 2001. The distributions of the number conformations (i.e., "K") for domains and chains are given in (*D*) and (*E*), respectively. Only proteins for which K exceeds 1 (for chains) are included in our dataset of multiple conformations. (*F*) Distinct proteins in our dataset within the context of high-quality X-ray structures in the PDB that we structurally aligned. A set of distinct proteins is such that no pair shares more than 90% sequence identity. *(G)* A single annotated entry from our database of alternative conformations. The clustering for the protein adenosylcobinamide kinase is shown. Two distinct conformations are represented in the ensemble of structures. The measure *kf* designates the fraction of times that the optimal value of K (here, K=2) was obtained out of 1000 simulations in which the algorithm (K-means with the gap statistic) obtained this particular value of K. The high *kf* value (0.969) signifies that these structures are very well clustered into two groups. *n* designates the number of distinct structures (PDB chains in this case) in the multiple structure alignment. *pf* designates the fraction of times (out of 1000 simulations of running Lloyd's algorithm, the standard K-means algorithm) that this particular set of structure-group assignments were assigned. In this this example, for all 1000 simulations, 1C9K_C and 1C9K_A were clustered in one group, and 1CBU_A, 1CBU_B, 1CBU_C clustered together. Within each cluster (the two clusters shown as two red boxes), the chain preceding the "::" tag designates the cluster representative (i.e., the structure closest to the Euclidean centroid of the cluster). The last field gives the RMSD values between cluster representatives. See the header information within File S1 for further details.

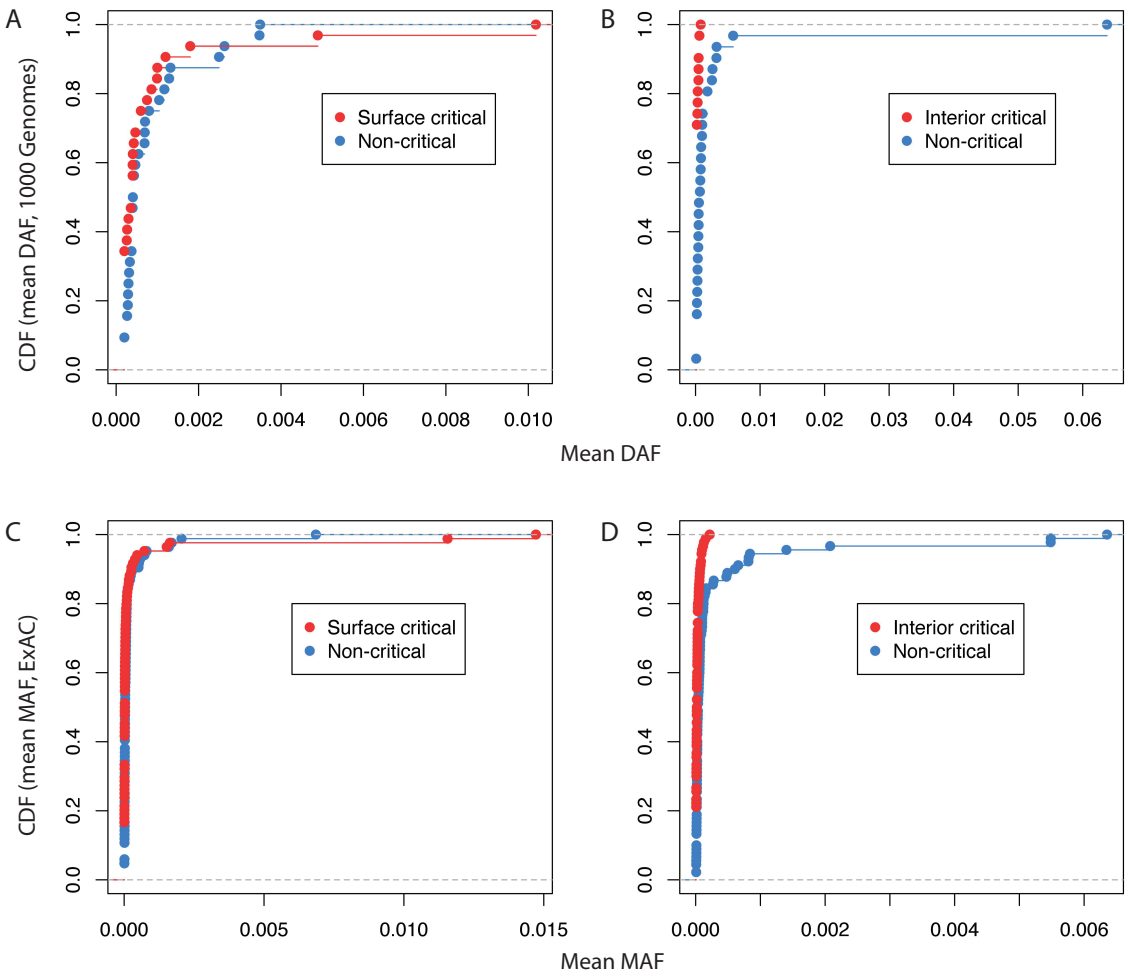


Figure S4, related to Figure 4. Shifts in allele frequency distributions from 1000 Genomes (panels A and B) and ExAC (panels C and D) datasets using two-sample Kolmogorov-Smirnov tests. Cumulative distribution functions for *(A)* mean DAF values of surface-critical and non-critical residues (p-val = 0.159); *(B)* mean DAF values of interior-critical and non-critical residues (p-val = 1.79e-4); *(C)* mean MAF values of surface-critical and non-critical residues (p-val = 9.49e-2); *(D)* mean MAF values of interior-critical and non-critical residues (p-val = 1.75e-4). All p-values are based on tow-sample KS tests.

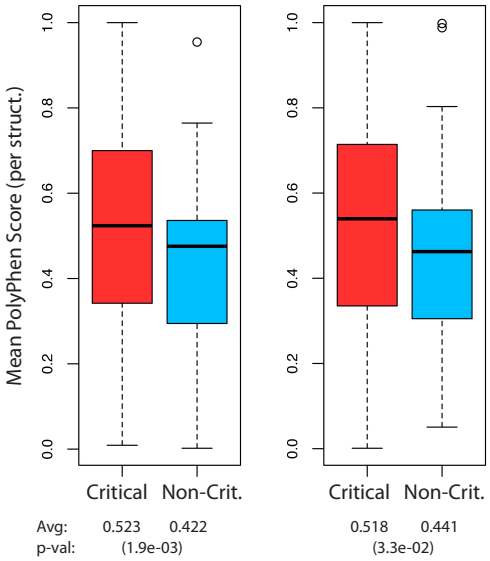


Figure S5, related to Figure 7. Evaluating pathogenicity using PolyPhen scores for critical- and non-critical residues, as identified by ExAC
Left: Distributions (64 structures) of mean PolyPhen values on surface-critical residues (red) and non-critical residues (blue). *Right*: Distributions (70 structures) of mean PolyPhen values on interior-critical 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.

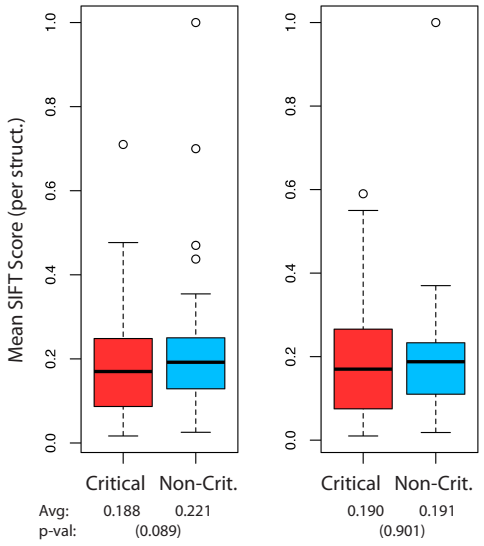


Figure S6, related to Figure 7. Evaluating pathogenicity using mean SIFT scores for critical- and non-critical residues, as identified by ExAC
Left: Distributions (63 structures) of mean SIFT values on surface-critical residues (red) and non-critical residues (blue). *Right*: Distributions (65 structures) of mean SIFT values on interior-critical 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.

2 – Supplemental Tables

HOLO	APO
1ake (**AP5**)	4ake
3cep (**G3P, IDM, PLP**)	1bks (**PLP**)
1hor (**AGP**, *PO4*, [& **16G** in pdb 1HOT])	1cd5
2c2b (**SAM**, [& **LLP** in pdb 2c2g])	1e5x
1gz3 (**ATP, FUM**, *OXL*)	1efk (MAK)
1atp (**ATP**)	1j3h
1hwz (**GLU**, **GTP, NDP** [& **ADP** in PDB 1NQT])	1nr7
1xtu (**CTP, U5P**)	1xtt (*ACY*, **U5P**)
1aax **(BPM** [& **892** in PDB 1T49])	2hnp
7at1 (**ATP**, **MAL**, **PCT** [& **CTP** in PDB 1RAC], [& **PAL** in PDB 1D09])	3d7s
3ju6 (**ANP**, **ARG**)	3ju5
6pfk (PGA [& **F6P + ADP** in PDB 4PFK])	3pfk (*PO4*)

Table S1, related to Table 1. Set of 12 canonical proteins, organized by state (*apo* or *holo*)
These 12 proteins were chosen to constitute the canonical set for several reasons: the allosteric mechanisms of their natural ligands are well understood, and both the *holo* and *apo* states for each system are available and clearly distinguishable; in addition, these proteins have been extensively investigated in the contexts of both binding leverage and allostery in general. Ligands are given in parentheses (those in bold text designate the ligands used to define residues involved in ligand-binding interactions).

n	Mean fract. of ligand-binding sites captured
6	0.56
5	0.59
4	0.65
3	0.69
2	0.79
1	0.84

Table S2, related to Table 1. Capturing known-ligand binding sites at varying thresholds
Here, *n* designates the number of residues within a surface-critical site that overlap with known ligand-binding residues. For the calculations reported above and in the main text, this value is taken to be n=6. Because each surface-critical site typically has 10 residues, and never has more than 10 residues, this criterion enforces that a majority of surface-critical residues within a given site overlap with known ligand-binding residues in order to be counted as a site match. However, as this threshold (n) is relaxed to lower values, the fraction of captured known ligand-binding sites improves rapidly, suggesting that surface-critical sites generally lie close to known ligand binding sites in many cases.

Concordance Btwn Community Detection Methods: GN vs. Infomap

Protein (PDB, # residues)	*Community Detection Method: GN | InfoMap*			
	Modularity	**# Communities**	**# Critical Residues**	**% of GN critical residues matching those in Infomap**
tRNA synthetase (1N78, 542)	0.71 | 0.68	14 | 25	47 | 109	0.28 (0.20)
Adenylate kinase (4AKE, 428)	0.73 | 0.70	11 | 20	39 | 82	0.90 (0.19)
Arginine Kinase (3JU5, 728)	0.72 | 0.69	12 | 28	41 | 142	0.22 (0.19)
Tyrosine Phosphatase (2HNP, 278)	0.59 | 0.59	7 | 15	27 | 70	0.26 (0.25)
Phosphoribosyltransferase (1XTT, 846)	0.72 | 0.68	9 | 32	36 | 174	0.22 (0.21)
cAMP-dep. PK (1J3H, 332)	0.66 | 0.64	11 | 19	36 | 78	0.33 (0.23)
Anthranilate synthase (1I7Q, 1418)	0.75 | 0.69	12 | 46	51 | 288	0.31 (0.20)
Malic enzyme (1EFK, 2212)	0.81 | 0.72	17 | 70	74 | 425	0.18 (0.19)
Threonine synthase (1E5X, 884)	0.73 | 0.69	13 | 36	43 | 192	0.28 (0.22)
G-6-P Deaminase (1CD5, 1596)	0.79 | 0.72	18 | 54	58 | 266	0.16 (0.17)
Phosphofructokinase (3PFK, 1276)	0.76 | 0.68	10 | 51	45 | 307	0.24 (0.24)
Tryptophan synthase (1BKS, 1294)	0.77 | 0.69	10 | 46	41 | 284	0.24 (0.22)
Means	***0.73 | 0.68***	***12.0 | 36.8***	***44.8 | 201.4***	***0.3***

Table S3, related to Figure 2. Comparing the two network module identification algorithms GN & Infomap

Though both GN (values to the left of "|" symbols throughout the table) and Infomap (values to the right) decompose networks to give similar modularity, the number of communities, and hence the number of critical residues connecting communities, is substantially larger when decomposing networks using Infomap than using GN.

3 - Supplemental Methods

3.1 Identifying Potential Allosteric Residues

Allosteric residues are predicted both on the surface and within the protein interior. In this study, these two sets of predicted allosteric residues are termed "surface-critical" and "interior-critical" residues, respectively. Notably, allosteric sites on the surface play mechanistic roles that are generally different from those within the interior: while surface sites often function as the source points or sinks of allosteric signals, the interior acts to transmit such information. Thus, different approaches are needed for selecting these two sets of residues. For both, biological assembly files from the PDB are used as the input to our analysis (Berman et al., 2000).

3.1-a Identifying Surface-Critical Residues

With the objective of identifying potential allosteric residues on the protein surface, we employ a modified version of the binding leverage method for identifying likely ligand binding sites (Figure 1A), as described previously (Mitternacht and Berezovsky, 2011). Allosteric signals may be transmitted over large distances by a mechanism in which the allosteric ligand has a global effect on a protein's functionally important motions. For instance, introducing a bulky ligand into the site of an open pocket may disrupt large-scale motions if those motions normally entail that the pocket become collapsed over the course of a motion (Figure 1A). Such a modulation of the global motions may affect activity within sites that are distant from the allosteric ligand-binding site. We point the reader to work by Mitternacht and Berezovsky for a more detailed discussion regarding the binding leverage method (Mitternacht and Berezovsky, 2011), though a general overview of the approach, along with a detailed discussion of the changes we have implemented, are given below.

3.1-a-i Monte Carlo Simulations & Parameterization to Identify Candidate Allosteric Sites on the Surface

The surface of most proteins is a highly dense patchwork of pockets, ridges, protrusions, and clefts. Throughout this complex topology, there are many potential sites that may confer allosteric regulation upon binding by natural or artificial ligands. Thus, as a first step to identifying surface-critical sites, we aim to identify surface pockets that are capable of accommodating small ligands. These candidate allosteric sites are generated by Monte Carlo (MC) simulations in which a simple flexible ligand (comprising of 4 "atoms" linked by bonds of fixed length 3.8 Angstroms, but variable bond and dihedral angles) explores the protein's surface. The number of MC simulations is set to 10 times the number of residues in the protein structure, and the number of MC steps per simulation in our implementation is set to 10,000 times the size of the simulation box, as measured in Angstroms. The size of this simulation box is set to twice the maximum size of the PDB along any of the x, y or z-axes. *Apo* structures were used when probing protein surfaces for putative ligand binding sites in the canonical set of proteins.

Throughout the MC simulation, a simple square well potential (i.e., modeling hard-sphere interactions) is used to model the attractive and repulsive energy terms associated with the ligand's interaction with the protein surface. In the unmodified implementation of the method, these energy terms depend only on the ligand atom's distance to *alpha carbon atoms* in the protein – other heavy atoms or biophysical properties are not considered.

Our approach and set of applications differ from those previously developed in several key ways. When running MC simulations to probe the protein surface and generate candidate binding sites, we use all heavy atoms in the protein when evaluating a ligand's affinity for each location. By including all heavy atoms (i.e., as oppose to using the protein's alpha carbon atoms exclusively), our hope is to generate a more selective set of candidate sites. Indeed, the use of alpha carbon atoms alone leaves 'holes' in the protein which do not actually exist in the context of the dense topology of side chain atoms. Thus, by including all heavy atoms, we hope to reduce the number of false positive candidate sites, as well as more realistically model ligand binding affinities in general.

In the original binding leverage framework, an interaction between a ligand atom and an alpha carbon atom in the protein contributes -0.75 to the binding energy if the interaction distance is within the

range of 5.5 to 8 Angstroms. Interaction distances greater than 8 Angstroms do not contribute to the binding energy, but distances in the range of 5.0 to 5.5 are repulsive, and those between 4.5 to 5.0 Angstroms are strongly repulsive (distances below 4.5 Angstroms are not permitted). However, given the much higher density of atoms interacting with the ligand in our all-heavy atom model of each protein, it is necessary to accordingly change the energy parameters associated with the ligand's binding affinity.

The determination of how these parameters should be changed in an all-heavy atom model is fundamentally a problem of *optimization*. Thus, how are these parameters optimized in the potential function? We determined which combination of parameters best predicts known ligand binding sites in threonine synthase (1E5X), phosphoribosyltransferase (1XTT), tyrosine phosphatase (2HNP), arginine kinase (3JU5), and adenylate kinase (4AKE). Specifically, the parameters to be optimized include (1) the ranges of favorable and unfavorable interactions (i.e., the *widths* of the potential function) and (2) the attractive and repulsive energies themselves (i.e., the *depths* and *heights* of the potential function).

For well *depths*, we tested models using several attractive potentials, ranging from -0.05 to -0.75, including all intermediate factors of 0.05. For well *widths*, we first tried using the cutoff distances originally used (attractive in the range of 5.5 to 8.0 Angstroms, repulsive in the range of 5.0 to 5.5, and strongly repulsive in the range of 4.5 to 5.0). However, these cutoffs, which were originally devised to model the ligand's affinity to the alpha carbon atom skeleton alone, were observed to be inappropriate when including all heavy atoms. Thus, in addition to sampling various well widths, we also performed the simulations using revised sets of distance cutoffs. The optimized set of parameters were as follows (here, $D_{lig\text{-}prot}$ designates the distance, in Angstroms, between a ligand atom and a protein atom):

widths	*depths & heights*
$\infty > D_{lig\text{-}prot} \geq 4.5$:	Energy = 0
$4.5 > D_{lig\text{-}prot} \geq 3.5$:	Energy = - 0.35 (attractive)
$3.5 > D_{lig\text{-}prot} \geq 3.0$:	Energy = +10 (repulsive)
$3.0 > D_{lig\text{-}prot} \geq 0.0$:	Energy = +10000 (strongly repulsive: effectively prohibited)

In addition to optimizing these parameters within the potential function, we also determined that setting the number of MC steps to 10,000 times the size of the simulation box (see above) provided the best convergence across multiple simulations on the same protein – that is, this number of steps better enabled us to recapture the same set of sites when running the simulations multiple times.

What form does the resultant MC ensemble take, and how exactly is this MC ensemble turned into a list of candidate sites? Prior to thresholding the list of ranked sites (see below), we generally follow the same formalism detailed in (Mitternacht and Berezovsky, 2011), and a discussion of this is also provided here. We first detail the output provided by a single MC simulation. This MC simulation involves a ligand probing the protein surface through a large number (typically 1,500,000 – 2,000,000, depending on the protein's size) of MC steps in which the ligand explores translational, rotational, and angular degrees of freedom. It is only the ligand that explores these degrees of freedom - the protein remains static throughout this MC simulation. The potential function (see above) usually "pushes" the ligand to favorably occupy a pocket on the protein surface after all steps of the MC simulation are completed. The ligand is thus in contact with a number of residues (typically 10-20) at the end of the simulation. As with the approach taken by Mitternacht and Berezovsky, this list of residues is ordered by local closeness (LC). LC is a geometric quantity that provides a measure of the degree of a residue in the residue-residue contact network; see (Mitternacht and Berezovsky, 2011b) for further discussion of LC. The 10 residues with greatest LC are taken as the final "site" occupied by the ligand at the end of this MC simulation (the remaining residues are not considered to be part of the site). Thus, the output of this single MC simulation is a list of 10 residues on the protein surface such that these residues form a geometrically favorable site for the ligand.

Now consider a very large number (typically 5,000 – 10,000, depending on the protein's size) of the MC simulations detailed above. These ~10,000 MC simulations result in ~10,000 sites, where each of these sites is the list of residues in contact with the ligand by the end of the MC simulation. This long list of sites generally contains many sites with a strong degree of overlap. Thus, to remove redundancy, pairs of sites with extremely high overlap are merged by combining any pair of sites that have a Jaccard similarity of at least 0.7, where the Jaccard similarity between sites i and j is defined as $|i \cap j|/|i \cup j|$. After merging sites in this way, the residues of a given site are listed by their LC, and no more than 10 residues for a site are used. This entire process results in a list of sites on which binding leverage calculations can be performed.

3.1-a-ii Binding Leverage Calculations

Once candidate pocket sites are identified using the protocol outlined above, an obvious question is whether these sites can function allosterically by influencing global low-frequency motions of the protein. In order to rank the candidate sites by the degree to which they can impart such allosteric properties, the binding leverage associated with each candidate site is calculated.

First, normal modes analysis is applied to generate a model of the protein's low-frequency motions (alternatively, one may use direct displacement vectors between two structures; see SI Methods section 3.2-c). To generate these modes, we use the alpha carbon atoms in building the protein's elastic networks. Using default parameters, we take the top 10 (lowest-frequency) available non-trivial Fourier normal modes generated using the Molecular Modeling Toolkit (MMTK) (Hinsen, 2000). Specifically, these 10 low-frequency modes are produced using the "representative structures" within each cluster of a multiple structure alignment (for details on representative structures, see SI Methods section 3.2-b). Note that this exact same method for producing the modes was also used in the identification of interior-critical residues (see below).

Once the 10 modes are produced, each of the candidate sites is then scored based on the degree to which deformations in the site couple to the low-frequency modes; that is, those sites which are heavily deformed as a result of the normal mode fluctuations (Figure 1A, top-right) receive a high score (termed the binding leverage for that site), whereas shallow sites with few interacting residues (Figure 1A, bottom-left) or sites that undergo minimal change over the course of a mode fluctuation (Figure 1A, bottom-right) receive low binding leverage scores. Specifically, the binding leverage score for a given site is calculated as

$$binding\ leverage\ =\ \sum_{m=1}^{10}\left(\sum_{i}\sum_{j}\Delta d_{ij(m)}^{2}\right)$$

Here, the outer sum is taken over the 10 modes, and the pair of inner sums are taken over all pairs of residues (i,j) such that the line connecting the pair lies within 3.0 Angstroms of any atom within the simulated ligand. The value $\Delta d_{ij(m)}$ for each residue pair (i,j) represents the change in the distance between residues i and j when this distance is calculated using mode m. Thus, one may think of binding leverage as qualitatively predicting the extent to which a surface pocket is deformed when the protein undergoes conformational transitions.

3.1-a-iii Defining & Applying Thresholds to Select High-Confidence Surface-Critical Sites

As discussed in the main text, without applying thresholds to the list of ranked surface sites that remain after running the binding leverage calculations, a very large number of sites would occupy the protein surface (Figure S2A). Thus, it is necessary to trim and process this list. To do so, we borrow from principles in energy gap theory (Bryngelson et al., 1995). Details regarding the calculations for establishing a threshold on the number of sites are given here.

For each of the N candidate binding sites in what we call "pre-processed ranked list of sites" (produced by the procedure outlined above), we calculate the value $\partial \mathtt{BL(j)}/\Delta \mathtt{BL}$. Here, $\mathtt{j}$ is the j^{th} site to appear in the pre-processed ranked list of sites, with this list of sites being ranked in descending order of each site's binding leverage score (see above). $\partial \mathtt{BL(j)}$ is defined as the difference in the binding leverage scores of the j^{th} site and the $(j\text{-}1)^{th}$ site in the ranked list. Because the list of sites is organized in descending order of binding leverage scores, $\partial \mathtt{BL(j)} \geq 0$. $\Delta \mathtt{BL}$ is a constant defined as:

$$\Delta \mathtt{BL}\ =\ \max_{binding\_leverage\_score}\ -\ \min_{binding\_leverage\_score}$$

$\Delta \mathtt{BL}$ is thus the difference in the binding scores associated with the very top site and very bottom site in this ranked list. Qualitatively, a large value for $\partial \mathtt{BL(j)}/\Delta \mathtt{BL}$ indicates that there is a large drop in binding leverage scores in going from site j to site $(j\text{-}1)$ within the pre-processed ranked list.

We then consider those sites with the highest $\partial \mathtt{BL}/\Delta \mathtt{BL}$ values – specifically, we consider the top 5.5% of sites in terms of $\partial \mathtt{BL}/\Delta \mathtt{BL}$. Thus, we are considering site j if there is a very large gap in binding leverage scores between sites j and $(j\text{-}1)$. The lowest-occurring site within *this* considered list of high $\partial \mathtt{BL}/\Delta \mathtt{BL}$ values demarcates a threshold beyond which we reject all lower sites within the pre-processed ranked list, leaving only what we call the "processed ranked list of sites".

We then go up from to bottom through the top of this processed ranked list of sites, and for each site, we determine the Jaccard similarity with all sites above. If the Jaccard similarity with any site above

exceeds 0.7, then the lower site is removed from the processed ranked list. The final list obtained after performing these Jaccard similarity filters is taken to represent the set of surface-critical sites on a structure.

In counting the final number of truly *distinct* surface-critical sites for any given structure, we remove redundant sites within the set of surface-critical sites obtained in the process above, as some of the sites within this set may still exhibit considerable mutual overlap. A site *i* within the list of surface-critical sites is said to be redundant if it is assigned a redundancy score that exceeds 0.4, where

$$redundancy\_score(i) = |\{R_{site\_i}\} \cap \{R_{sites>i}\}| / N_{res\_i}$$

Here, $\{R_{site\_i}\}$ is the set of residues in site *i*, $\{R_{sites>i}\}$ is the union of residues in all accepted sites above site *i* in the list of sites, N_{res_i} is the number of residues in site *i*, and the |…| notation in the denominator of this ratio is meant to designate the number of residues in the indicated intersection. If this redundancy score is less than 0.4, then site *i* is considered to be truly distinct from all other sits, and it is included in the list of distinct sites. If the redundancy score exceeds 0.4, then the site overlaps with another site on the surface, and it is thus rejected from the set of accepted distinct sites. Finally, the total number of sites in the accepted set of sites is taken as the number of distinct sites for a structure.

3.1-a-iv Known Ligand-Binding Sites at the Surface

Known ligand-binding residues of an *apo* structure are taken to be those within 4.5 Angstroms of the ligand in the corresponding *holo* structure (Table S1). Within the canonical set of proteins, surface-critical sites overlap with an average of 56% of the known-ligand binding sites (Table 1). It has previously been shown that the sites in aspartate transcarbamoylase (PDB ID 3D7S) are especially difficult to identify (Mitternacht and Berezovsky, 2011); excluding aspartate transcarbamoylase results in finding an average of 61% of known biological ligand binding sites. In addition, we emphasize that many of the "false positives" (sites predicted to be important allosterically, but do not correspond to known ligand binding sites) may nevertheless function as latent allosteric sites. Such sites potentially may impart allosteric properties through previously uncharacterized ligands or through artificial ligands (such as drugs targeted to specific proteins).

3.1-b Dynamical Network Analysis to Identify Interior-Critical Residues

As discussed, allosteric residues within the protein interior often act to transmit signals. The identification of such residues is accomplished by a network formalism (Figure 1B), wherein the objective is to identify network nodes (i.e., residues) that are essential for communication between communities (i.e., groups of highly inter-connected residues of the contact map). This first entails representing a protein structure as a network of interacting residues, and then weighting the connections (edges) between these residues using information about inferred protein motions. Once the edges are weighted, the network is decomposed into distinct modules, and the residues that are identified as being important for inter-module communication are identified as the interior-critical residues. The details of this formalism are provided here.

3.1-b-i Network Formalism and Weighting Scheme

The network representing interacting residues is first constructed. An edge between residues *i* and *j* is drawn if any heavy atom within residue *i* is located within 4.5 Angstroms of any heavy atom within residue *j*, and we exclude the trivial cases of pairs of residues that are adjacent in sequence, which are not considered to be in contact within the network.

Network edges are then weighted on the basis of correlated motions of the interacting residues, with these motions provided by the same ANMs that had been used in the identification of surface-critical residues (as with the identification of surface-critical residues, it is also possible to model conformational changes by using information regarding pairs of distinct conformations; see the SI Methods section 3.2-c). Again, the 10 lowest-frequency non-trivial modes are produced using the "representative structures" (see discussion in SI Methods section 3.2-b) within each cluster of a multiple structure alignment for a given protein. We emphasize that, although ANMs are more coarse-grained than molecular dynamics, our use of ANMs is motivated by their much faster computational efficiency, which is a required feature for our database-scale analysis.

The edge weighting scheme is performed as follows: an "effective distance" D_{ij} for an edge between interacting residues i and j is set to $D_{ij} = -\log(|C_{ij}|)$, where C_{ij} designates the correlated motions between residue i and j:

$$C_{ij} = Cov_{ij} / \sqrt{(\langle \mathbf{r}_i^2 \rangle \langle \mathbf{r}_j^2 \rangle)}$$

where

$$Cov_{ij} = \langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$$

Here, $\mathbf{r}_i$ and $\mathbf{r}_j$ designate the vectors associated with residues i and j (respectively) under a particular mode. The brackets in the term $\langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$ indicate that we are taking the mean value for the dot product $\mathbf{r}_i \bullet \mathbf{r}_j$ over the 10 lowest-frequency non-trivial modes.

An example may help to clarify this. If two interacting residues exhibit a *high* degree of correlated motion, then the motion of one may tell us about the motion of the other, suggesting a strong flow of energy or information between the two residues, resulting in a *low* value for D_{ij}: a strong correlation (or a strong anti-correlation) between nodes i and j result in a value for $|C_{ij}|$ that is close to 1. This gives a low value for D_{ij} ($-\log(|C_{ij}|) \approx 0$). Thus, given a strong correlated motion, this effective distance D_{ij} between residues i and j is very short. This small D_{ij} means that any path involving this pair of residues is likewise shorter as a result, thereby more likely placing this pair of residues within a shortest path, and more likely rendering this pair a bottleneck pair. In sum, this edge-weighting scheme is such that a high correlation in motion results in a short effective distance, thereby more likely rendering this a bottleneck pair of residues.

In the opposite scenario, for interacting residues with poor correlation values ($C_{ij} \approx 0$), a large effective distance D_{ij} results, thereby making it more difficult for the pair of residues to lie within shortest paths or within the same community.

Once all connections between interacting pairs of residues are appropriately weighted and the communities are assigned using the Girvan-Newman (GN) algorithm (Girvan et al., 2002) with these effective distances, a residue is deemed to be critical for allosteric signal transmission (i.e., an interior-critical residue) if it is involved in the highest-betweenness edge connecting two distinct communities. A given edge's *betweenness* is taken to be the number of shortest paths involving that edge, where a path length is the sum of its associated effective edge distances (see above). The shortest distance between each ${}_NC_2$ pair of nodes in the network of N residues is calculated with the Floyd–Warshall algorithm. See Figure 2 for examples of community partitions and associated interior-critical residues.

3.1-b-ii Decomposing Proteins into Modules Using Different Algorithms

We use the GN formalism to identify the community structure of networks as part of our framework to identify interior-critical residues. By identifying the "community structure", we are referring to the problem of finding the optimal partitioning of a network into different "modules" (i.e., communties), such that each node within a module is highly connected to other nodes within the same module, and minimally conntected to other nodes in outside modules. However, although we employ GN, many other algorithms have been devised to identify community structure.

In a study comparing different algorithms (Lancichinetti and Fortunato, 2009), an information theory-based approach (Rosvall and Bergstrom, 2007) was shown to be one of the strongest methods. This approach (termed "Infomap") 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; further details are provided in (Rosvall and Bergstrom, 2007).

Perhaps surprisingly, even though both GN and Infomap achieve similar network modularity (with GN being slightly better), Infomap produces at least twice the number of communities relative to that of GN when applied to protein structures, and it thus generates many more interior-critical residues (Table S3). Within the set of 12 canonical proteins, GN and Infomap generate an average of 12.0 and 36.8 communities, respectively. This corresponds to an average of 44.8 and 201.4 interior-critical residues when using GN and Infomap, respectively. Thus, given that GN produces a more selective set of residues for each protein, we use GN throughout our analyses.

Although the critical residues identified by GN do not always correspond to those identified by Infomap, the mean fraction of GN-identified interior-critical residues that match Infomap-identified residues is 0.30 (the expected mean, based on a uniformly-random distribution of critical residues

throughout the protein, is 0.21, p-value=0.058). Furthermore, we observe that obvious structural communities are detected when applying both methods: 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. In addition, the modularity from the network partitions generated by GN and Infomap are comparable. For the 12 canonical systems, the mean modularity for GN and Infomap is 0.73 and 0.68, respectively. 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.

Together, these results suggest that both GN and Infomap generate similar partitions. Roughly, the set of interior-critical residues identified by GN partially constitute a subset of those identified with Infomap. If these sets of residues were completely different, then the choice between GN and Infomap would be difficult, as the results in our downstream conservation analyses would then be highly sensitive to our community detection method of choice. Given that the two residue sets are not disjoint, our choice of GN over infomap was largely guided by the fact that GN is far more selective in identifying important network elements (i.e., interior-critical residues), as evidenced in Table S3. In contrast, Infomap generates a much less selective set of interior-critical residues.

3.1-c STRESS (STRucturally-identified ESSential residues)

We have developed an easy-to-use web tool in order to enable those in the structural biology community to identify surface- and interior-critical residues within their own proteins of interest. Our server has been designed to be both user-friendly and highly efficient.

We use local searching supported by hashing to perform a local search in each sampling step of the Monte Carlo simulations, which takes constant time. This approach brings down the asymptotic computational complexity by an order of magnitude, relative to a simpler implementation without optimization (Figures 3B and 3C). The time complexity of the core computation, Monte Carlo sampling, is $O(|T||S|)$, where T and S are simulation trials and steps for each trial, respectively. After carefully profiling and optimizing for speed (with optimizations introduced through changes in the workflow, data structures, numerical arithmetic, etc.), a typical case takes ~30 minutes on a E5-2660 v3 (2.60GHz) core.

In terms of operation, our tool utilizes two types of servers: front-end servers that handle incoming HTTP requests and back-end servers that perform algorithmic calculations (Figure 3A). Communication between these two types of servers is handled by Amazon's Simple Queue Service (SQS). When our front-end servers receive a new request, they add the job to the queue and then return to requests immediately. Our back-end servers poll the queue for new jobs and run them when capacity is available. Amazon's Elastic Beanstalk offers several features that enable us to dynamically scale our web application. We use Auto Scaling to automatically adjust the number of back-end servers backing our application based on predefined conditions, such as the number of jobs in the queue and CPU utilization. Elastic Load Balancer automatically distributes incoming network traffic. This system ensures that we are able to handle varying levels of demand in a reliable and cost-effective manner. Since we may have multiple servers backing our tool simultaneously (some handling HTTP requests and some performing calculations, any of which may be terminated at any time by Auto Scaling), it is important that our servers are stateless. We thus store input and output files remotely in an S3 bucket, which is accessible to each server via RESTful conventions. The corresponding source code and README files are made available through Github (github.com/gersteinlab/STRESS).

3.2 High-Throughput Identification of Alternative Conformations

There are many proteins within the PDB for which multiple distinct conformations are available. In many cases, a large number of structures may represent a relatively small number of conformations. We have sought identify such alternative conformations using a structural clustering scheme as part of our framework for identifying critical residues. The purpose of developing this clustering scheme is three-fold:

1) We are interested in those structures that exhibit distinct conformations, as we are focusing on cases for which pronounced conformational change plays an essential role in allostery.

2) The clustering scheme ultimately enables us to perform an important control. Namely, it enables us to address the question: are the results robust to alternative methods of inferring information about conformational change? ANMs provide only one means of defining the vectors for predicted conformational change. However, another approach is to use the direct displacement vectors from the crystal structures of alternative conformations. This alternative constitutes a method that we term "absolute conformational change" (ACT) in the manuscript.
3) ANMs constitute the bulk of our analysis, so we must be confident that the structures analyzed are suitable: if a given protein is not believed to undergo significant conformational change, it may not be appropriate to apply ANMs, as the ANMs can incorrectly predict large-scale conformational change where no such change is believed to occur.

An overview of our pipeline is provided in Figure S3A. Broadly, we perform MSAs for thousands of structures, with each alignment consisting of sequence-identical groups. Within each alignment, we cluster structures using RMSD to determine the distinct conformational states. We then use models of protein conformational transitions to identify surface- and interior-critical residues.

3.2-a Database-Wide Multiple Structure Alignments

FASTA files of all SCOP domains were downloaded from the SCOP website (version 2.03) (Fox et al., 2014; Murzin et al., 1995). We first worked with domains to probe for intra-domain conformational changes, as better alignments are generally possible at the domain level. For all other analyses reported, all results are based on groups of structures that are 100% sequence identical. We removed structures with resolution values poorer than 2.8, as well as any PDB files with R-Free values poorer than 0.28. STAMP (Russell and Barton, 1992) and MultiSeq (Roberts et al., 2006) were used to generate the multiple structure alignments (MSAs). For each MSA, the final output is a symmetric matrix representing all pairwise RMSD values, which are then used as input to the K-means module (below).

3.2-b Identifying Distinct Conformations within an MSA

For each MSA produced in the previous step, the corresponding matrix of pairwise RMSD values describes the degree and nature of structural heterogeneity among the crystal structures. The objective is to use this data in order to identify the biologically distinct conformations represented by an ensemble of structures. Our framework relies on a modified version of the K-means clustering algorithm, termed K-means clustering with the gap statistic (Tibshirani et al., 2001). *A priori*, performing K-means clustering assumes prior knowledge of the number of clusters (i.e., "K") to describe a dataset, and the gap statistic enables one to identify the optimal number of clusters intrinsic to a complex or noisy dataset. Given multiple resolved crystal structures for a given domain, this method estimates the number of conformational states represented in the ensemble of structures.

As a first step toward clustering the structure ensemble of N structures, we use multidimensional scaling (MDS) to convert an N-by-N matrix of pairwise RMSD values into a set of N distinct points, with each point representing a structure in (N-1)-dimensional space. The values of the N-1 coordinates assigned to each of these N points are such that the Euclidean distance between each pair of points is the same as that corresponding pair's RMSD value in the original matrix.

We point the reader to the work by Tibshirani *et al* for details governing how we perform K-means clustering with the gap statistic, as well as more details on the theoretical justifications of this approach (Tibshirani et al., 2001). However, an overview is provided here. Assume that the data takes the form of 60 data points, with each point represented in 2D space.

1) Start by assuming that the data can be represented with K clusters. Perform standard K-means clustering (i.e., Lloyd's algorithm) on the data to assign each point to one of K clusters. Then, for each cluster k (which contains data points in the set C_k) measure D_k, which describes the 'density' of points within cluster k:

$$D_k = \sum_{x_i \in C_k} \sum_{x_j \in C_k} |x_i - x_j|^2$$

2) Calculate an overall normalized score W to describe how well-clustered the resultant system has become when assigning all 60 data points to the K clusters (n_k denotes the number of points in cluster k):

$$W = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

3) Given our data, how well does this number of assigned clusters K actually represent the 'true' number of clusters, relative to a null model without any apparent clustering? To address this, produce a null distribution of 60 randomly (i.e., uniformly) distributed data points that lack any clear clustering such that the randomly placed points lie within the same bounding box of the observed data.

4) Repeat step (3) M times, and each time a random null distribution is produced, calculate $W_{null(K)}$ (assuming K clusters), just as W is calculated for the observed data. Then calculate the $\text{mean}_M\{\log(W_{null(K)})\}$ for these M null distributions. The $\text{mean}_M\{\log(W_{null(K)})\}$ measures how well *random* systems (with the same number of data points and within the same variable ranges as the observed data) can be described by K clusters. The M $\log(W_{null(K)})$ values produced by the null models have a standard deviation that is ultimately converted to s_k; see (Tibshirani et al., 2001) for details:

$$s_k = \sigma(k)\sqrt{(1 + 1/B)}$$

5) Calculate the gap statistic δ(K), given K clusters. This measures how well our observed data may be described by K clusters relative to null models containing the same number of points and within the same variable ranges. A high δ(K) signifies that our data is well-described using K clusters. Assuming K clusters, the gap statistic is given as:

$$\delta(K) = \text{mean}_M\{\log(W_{null(K)})\} - \log(W)$$

6) Obtain successive values δ(K+1), δ(K+2), δ(K+3), etc. by incrementing the value for K and repeating the steps (1) - (5). The optimal K is the first (i.e., lowest) K such that $\delta(K) >= \delta(K+1) - s_{k+1}$:

$$K_{optimal} = \{K|\ \delta(K) >= \delta(K+1) - s_{k+1}\}$$

We confirmed that these $K_{optimal}$ values accurately reflect the number of clusters by manually studying dozens of MSAs. We also examined several negative controls, such as CAP, an allosteric protein that does not undergo conformational change. We identified a vast array well-studied allosteric domains and proteins. There may be many factors driving conformational change, and those cases for which the change is induced by the binding to a simple ligand (i.e., a simple consideration of *apo* or *holo* states) constitute only a very small subset of the conformational shifts observed in the PDB. The gap statistic performed well in discriminating crystal structures that constitute such a diverse set.

Each structure is assigned to its respective cluster using the assigned optimal K-values as input to Lloyd's algorithm (i.e., standard K-means clustering). For each sequence group, we perform 1000 K-means clustering simulations on the MDS coordinates, and take the most common partition generated in these simulations to assign each structure to its respective cluster. We then select a "representative structure" from each of the assigned clusters. This representative is the member with the lowest Euclidean distance to the cluster mean, using the coordinates obtained by MDS (see description above). These cluster representatives are then taken as the distinct conformations for this protein, and they are used for the binding leverage calculations and networks analyses (below).

3.2-c Models of Conformational Change via Displacement Vectors from Alternative Conformations

Unless otherwise specified, we use normal modes analysis to model conformational change. However, one potential concern with this approach is that normal modes may not faithfully represent plausible conformational changes. To evaluate the robustness of different means for inferring motions (especially those results relevant to the conservation of critical residues), we also model conformational change using vectors connecting pairs of corresponding residues in crystal structures of alternative conformations. We term this approach "absolute conformational transitioning" (ACT). This more direct

model of conformational change is especially straightforward to apply to single-chain proteins (applying ACT on a database scale to multi-chain complexes would introduce confounding factors related to chain-chain correspondence between such complexes when each complex has multiple copies of a given chain).

3.2-c-i Inferring Protein Conformational Change Using Displacement Vectors from Alternative Conformations

Given a particular protein, how are these ACT vectors defined to find critical residues? We discuss a hypothetical example consisting of a multiple structure alignment of 8 sequence-identical structures. Starting with the protein's alignment using all 8 structures, we determine the optimal number of clusters represented by the alignment using the K-means algorithm with the gap statistic (see the above SI Methods section 3.2-b). Suppose that these 8 structures may be grouped into 2 distinct clusters (4 structures in *cluster A*, and 4 in *cluster B*). As discussed in SI Methods section 3.2-b, a representative structure is taken from each of these two clusters (*structure A* and *structure B*). These two representatives are taken to constitute the alternative conformations for the protein. As an alternative to using ANMs, we may use *structure A* and *structure B* to infer information about the protein's global conformational shifts by assigning a displacement vector to each residue (for instance, residue Y140), where the displacement vector is simply defined by the two corresponding residues in the different structures within the structure alignment (i.e., Y140 within *structure A* of the structure alignment and Y140 within *structure B* of the structure alignment). Because the structures are sequence-identical, each residue in one of these two representative structures matches a residue on the other representative. If each structure represents a sequence-identical 200-residue protein, then 200 ACT vectors represent the conformational change. These 200 ACT vectors for the protein may then be used to identify surface- and interior-critical residues (see below), and downstream analysis on these residues is then performed.

3.2-c-ii Identifying Surface-Critical Residues Using Vectors from Alternative Conformations

All preliminary steps performed when identifying surface-critical residues using normal modes (such as the MC search) are the same as those when using ACT vectors, with the important difference, of course, being the use of these ACT vectors as oppose to using eigenvectors when inferring motion. Thus, when using ACT vectors, the binding leverage score for a given site is simply calculated as:

$$binding\ leverage\ =\ \sum_i \sum_j \Delta d_{ij}^2$$

where the sum is taken over all pairs of residues (i,j) such that the line connecting the pair lies within 3.0 Angstroms of any atom within the simulated ligand, and the value Δd_{ij} for each residue pair (i,j) represents the change in the distance between residues i and j when this distance is calculated in alternative crystal structure. Thus, for each residue, the 10 vectors provided by the normal modes are simply replaced by the single ACT vector that defines the change in position of that residue when going from the protein conformation given by one representative structure to the conformation given by the other representative.

3.2-c-iii Identifying Interior-Critical Residues Using Vectors from Alternative Conformations

When identifying interior-critical residues, ACT vectors may be produced in the exact same way that they are produced when identifying surface-critical residues. When identifying interior-critical residues, the inferred conformational changes are used in order to assign weights within the residue contact maps. In the scheme in which normal modes are used, these weights are assigned by averaging over to 10 sets of vectors given by the 10 modes. However, when using ACT vectors, there is only one vector for each residue (i.e., the vector defining the "displacement" defined by two structures). Thus, when using ACT vectors, the weight parameters are calculated as

$$C_{ij}\ =\ Cov_{ij}\ /\ \sqrt{(|\mathbf{r}_i|^2 * |\mathbf{r}_j|^2)}$$

where

$$Cov_{ij}\ =\ \mathbf{r}_i \bullet \mathbf{r}_j$$

Here, $\mathbf{r}_i$ denotes the vector that defines the change in position for residue i when going from one representative conformation to the other.

3.2-c-iv Using Vectors from Alternative Conformations Recapitulates Results Using Normal Modes

When we use ACT vectors to apply the modified binding leverage framework for these proteins, we again observe that our surface-critical residues are significantly more conserved than are non-critical residues (Figure 6A), and the same trend is also observed when ACT vectors are applied in our dynamical network analysis for identifying interior-critical residues (Figure 6B). The fact that ACT vectors produce a similar set of results to those obtained using normal modes analysis suggests that our approach is robust to different methods for inferring protein conformational change. We note that there are too few human single-chain proteins to perform a reliable analysis in which conservation is evaluated using 1000 Genomes or ExAC data – for instance, only 9 (16) structures are such that 1000 Genomes (ExAC) SNVs overlap with interior-critical residues.

3.3 Evaluating Conservation of Critical Residues Using Various Metrics and Sources of Data

How conserved are the surface- and interior-critical residues identified, relative to other residues in the protein? Certainly, allosteric residues are known to exhibit conservation, and we should expect that the critical residues identified exhibit strong conservation. Conservation may be measured across diverse evolutionary time scales. Metrics for selective constraint that correspond to long evolutionary time scales entail sequence comparisons across diverse species. At the other extreme, metrics for short-term evolutionary conservation entail analyzing multiple genomes from within the same species (e.g., multiple human genomes). In order to evaluate the relative conservation of the critical residues identified in this study, we measure conservation using both types of measures, and demonstrate that, as expected, critical residues are under stronger evolutionary constraint relative to other regions of the protein.

3.3-a Conservation Across Species

All cross-species conservation scores represent the ConSurf scores, as downloaded from the ConSurf Server (Ashkenazy et al., 2010; Celniker et al., 2013; Glaser et al., 2003; Landau et al., 2005), in which ConSurf scores for each protein chain are normalized to have a mean ConSurf score of 0 (the ConSurf score variance is 1 for each chain). Low (i.e., negative) ConSurf scores represent a stronger degree of conservation, and high (i.e., positive) scores designate weaker conservation. We perform cross-species conservation analysis on those proteins for which ConSurf files are available from the ConSurf server, and all ConSurf scores were calculated using default parameters, listed here:

```
Homolog search algorithm: CSI-BLAST
Number of iterations: 3
E-value cutoff: 0.0001
Proteins database: UniRef-90
Maximum homologs to collect: 150
Maximal %ID between sequences: 95
Minimal %ID for homologs: 35
Alignment method: MAFT-L-INS-i
Calculation method: Baysian
Calculation method: JTT
```

Each individual point within the cross-species conservation plots (e.g., Figures 4B, 4F, and 6) represents data from one structure: the value of the point for any given structure represents the mean conservation score for all residues within one of two classes: the set of N critical residues within a protein structure (surface or interior) or a randomly-selected set of N non-critical residues (with the same “degree”, see below) within the same structure. The randomly selected non-critical set of residues was chosen in a way such that, for each critical residue with degree k (k being the number of non-adjacent residues with which the critical residue is in contact, see below), a randomly selected non-critical residue with the same degree k was included in the set. The distributions of non-critical residues shown are very much representative of the distributions observed when re-building the random set many times.

Note that the degree (i.e., k) of residue *j* is defined as the number of residues which interact with residue *j*, where residues adjacent to residue *j* in sequence are not considered, and an interaction is defined whenever any heavy atom in an interacting residue is within 4.5 Angstroms of any heavy atom in the residue *j*. We use degree as a measure of residue burial for several reasons. This metric for burial is consistent with our networks-based analysis for identifying interior-critical residues, as well as our use of residue-residue contacts in building networks for producing the ANMs. In addition, degree is also an attractive metric because it is discrete in nature, thereby allowing us to generate null distributions of non-critical residues with the exact same degree distribution.

3.3-b Measures of Conservation Amongst Humans from Next-Generation Sequencing

All SNVs intersecting protein-coding regions that result in amino acids changes (i.e., nonsynonymous SNVs) were collected from the phase 3 release of The 1000 Genomes Project (McVean et al., 2012). VCF files containing the annotated variants were generated using VAT (Habegger et al., 2012). For nonsynonymous SNVs, the VCF files included the residue ID of the affected residue, as well as additional information (such as the corresponding allele frequency, the ancestral allele, and the residue type). To map the 1000 Genomes SNVs on to protein structures, FASTA files corresponding to the translated chain(s) of the respective transcript ID(s) were obtained using BioMart (Smedley et al., 2015). FASTA files for each of the PDB structures associated with these transcript IDs (the PDB ID-transcript ID correspondence was also obtained using BioMart) were generated based on the ATOM records of the PDB files. For each given protein chain, BLAST was used to align the FASTA file obtained from BioMart with that generated from the PDB structure. The residue-residue correspondence obtained from these alignments was then used in order to map each SNV to specific residues within the PDB. As a quality assurance mechanism, we confirmed that the residue type reported in the VCF file matched that specified in the PDB file.

ExAC SNVs were downloaded from the ExAC Browser (Beta), as hosted at the Broad Institute. SNVs were mapped to all PDBs following the same protocol as that used to map 1000G SNVs, and only non-synonymous SNVs in ExAC were analyzed. When evaluating SNVs from the ExAC dataset, minor allele frequencies (MAF) were used instead of DAF values. The ancestral allele is not provided in the ExAC dataset – thus, analysis is performed for MAF rather than DAF. However, we note that little difference was observed when using AF or DAF values with 1000 Genomes data, and we believe that the results with MAF values would generally be the same as those with DAF values. We also highlight the attractive feature of recapitulating the general conservation trends observed using a separate matric.

When analyzing both 1000 Genomes and ExAC data, we consider only those structures in which at least one critical and one non-critical residue intersect a non-synonymous SNV. This enables a more direct comparison between critical and non-critical residues, as comparisons between two different proteins would rely on the assumption of equal degrees of selection between such proteins.

Each individual point within the intra-human conservation plots (e.g., Figures 4C, 4D, 4G, and 4H) represents data from one structure: the value of the point for any given structure represents the mean score (DAF or MAF, for 1000 Genomes or ExAC SNVs, respectively) for all critical (red bars) or non-critical (blue bars) residues to intersect SNVs.

The *fraction* of rare SNVs intersecting a particular "protein annotation" (described below) is defined to be the ratio of the number of rare non-synonymous SNVs in that annotation to the total number of non-synonymous SNVs intersecting that annotation. An annotation for a given protein is simply the set of residues within a particular category, such as the set of all surface-critical residues (or alternatively the set of all interior-critical residues, or the set of non-critical residues). We define the term "rare" to mean that a 1000 Genomes SNV has a DAF value below a certain threshold – for instance, variable thresholds ranging from DAF = 0.05% to 0.50% are evaluated in Figures 5A and 5C. An SNV in ExAC is defined to be rare if it has a MAF value below a certain threshold – variable thresholds ranging from MAF = 0.05% to 0.50% are evaluated in Figures 5B and 5D.

If a particular annotation, such as the set of surface-critical residues, has a rare SNV, then this rarity may potentially be a consequence of purifying selection acting to remove a deleterious SNV from the population pool (thereby making it rare). Such an annotation may thus be sensitive to sequence changes, and would thus be conserved. If there is a high fraction of such rare SNVs within the annotation, it provides further confidence to the claim that the annotation is conserved. Thus, a high fraction of rare SNVs is used

as a signature for stronger conservation. Supporting this intuition, previous studies have observed that conserved genomic regions within the human population harbor higher fractions of rare SNVs (Khurana et al., 2013; McVean et al., 2012; Tennessen et al., 2012).

4 - Supplemental References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. *38*, W529–W533.

Bryngelson, J.D., Onuchic, J.N., Socci, N.D., and Wolynes, P.G. (1995). Funnels, pathways, and the energy landscape of protein folding: A synthesis. Proteins Struct. Funct. Bioinforma. *21*, 167–195.

Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. (2013). ConSurf: using evolutionary data to raise testable hypotheses about protein function. Isr. Jounral Chem. *13*, 199–206.

Fox, N.K., Brenner, S.E., and Chandonia, J.-M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. *42*, D304–D309.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., and Ben-Tal, N. (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics *19*, 163–4.

Habegger, L., Balasubramanian, S., Chen, D.Z., Khurana, E., Sboner, A., Harmanci, A., Rozowsky, J., Clarke, D., Snyder, M., and Gerstein, M. (2012). VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. Bioinformatics *28*, 2267–2269.

Hinsen, K. (2000). The Molecular Modeling Toolkit: A New Approach to Molecular Simulations. J. Comput. Chem. 21, 79–85.

Hubbard, S. J., and Thornton, J. M. (1993). Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London, 2(1).

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. Phys Rev E Stat Nonlin Soft Matter Phys. *80*, 56117.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., and Ben-Tal, N. (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res. *33*, W299–W302.

Mitternacht, S. and Berezovsky, I.N. (2011b). A geometry-based generic predictor for catalytic and allosteric sites. Protein Eng Des Sel *24*: 405–409.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Sokal, R.R. (1958). A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. *38*, 1409–1438.

Roberts, E., Eargle, J., Wright, D. and Luthey-Schulten, Z. (2006). MultiSeq: unifying sequence and structure data for evolutionary analysis. BMC Bioinformatics *7*, 382.

Rosvall, M. and Bergstrom, C.T. (2007). An information-theoretic framework for resolving community structure in complex networks. Proc. Natl. Acad. Sci. U. S. A. *104*, 7327–7331.

Russell, R.B. and Barton, G.J. (1992). Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. Proteins *14*, 309–323.

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G., et al. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. *43*, W589–W598.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–9.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. *63*, 411–423.

Click here to access/download

Supplemental Movies & Spreadsheets

File_S1.xlsx

Click here to access/download

Supplemental Movies & Spreadsheets

File_S2.xlsx

Click here to access/download

Supplemental Movies & Spreadsheets

File_S3.xlsx