

1 **Title:**
2 Identifying allosteric hotspots with dynamics: application to inter- and intra-species
3 conservation
4

5 **Authors & associated information:**

6 Declan Clarke^{a,1}, Anurag Sethi^{b,c,1}, Shantao Li^{b,d}, Sushant Kumar^{b,c}, Richard W.F.
7 Chang^e, Jieming Chen^{b,f}, and Mark Gerstein^{b,c,d,2}
8

9 ^a Department of Chemistry, Yale University, 260/266 Whitney Avenue PO Box 208114,
10 New Haven, CT 06520 USA

11 ^b Program in Computational Biology and Bioinformatics, Yale University, 260/266
12 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

13 ^c Department of Molecular Biophysics and Biochemistry, Yale University, 260/266
14 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

15 ^d Department of Computer Science, Yale University, 260/266 Whitney Avenue PO Box
16 208114, New Haven, CT 06520, USA

17 ^e Yale College, 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA

18 ^f Integrated Graduate Program in Physical and Engineering Biology, Yale University,
19 260/266 Whitney Avenue PO Box 208114, New Haven, CT 06520, USA
20

21 ¹ D.C. and A.S. contributed equally to this work.

22 ² Correspondence should be addressed to M.G. (pi@gersteinlab.org)
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46

47 ABSTRACT

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

61

62

63

64

65

66

67

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: missing

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: - ... [1]

71 INTRODUCTION

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

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

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: between

94 Given the importance of allosteric regulation, as well as its role in imparting
95 efficient functionality, several methods have been devised for the identification of likely
96 allosteric residues. Conservation itself has been used, either in the context of conserved
97 residues (Panjkovich and Daura, 2012), networks of co-evolving residues (Halabi *et al.*,
98 2009; Lee *et al.*, 2008; Lockless *et al.*, 1999; Reynolds *et al.*, 2011; Shulman *et al.*, 2004;
99 Süel *et al.*, 2003), or local conservation in structure (Panjkovich and Daura, 2010). In
100 related studies, both conservation and geometric-based searches for allosteric sites have
101 been successfully applied to several systems (Capra *et al.*, 2009).

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

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

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

120 With the objective of identifying allosteric residues within the interior, molecular
121 dynamics (MD) simulations and network analyses have been used to identify residues
122 that may function as internal allosteric bottlenecks (Csermely *et al.*, 2013; Gasper *et al.*,
123 2012; Rousseau and Schymkowitz, 2005; Sethi *et al.*, 2009; Vanwart *et al.*, 2012). Ghosh
124 *et al.* (2008) have taken a novel approach of combining MD and network principles to
125 characterize allosterically important communication between domains in methionyl
126 tRNA synthetase. In conjunction with NMR, Rivalta *et al.* have use MD and network
127 analysis to identify important regions in imidazole glycerol phosphate synthase (Rivalta
128 *et al.*, 2012).

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

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: . The

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: between

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

147

148 RESULTS

149 Identifying Potential Allosteric Residues

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

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

161

162 Identifying Surface-Critical Residues

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: may

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: Critical residues are

DECLAN CLARKE 12/18/15 5:57 PM

Deleted:), and they are

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: investigated

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

177 This approach is a modified version of the binding leverage framework
178 introduced by Mitternacht and Berezovsky (Mitternacht and Berezovsky, 2011). The
179 main modifications implemented here include the use of heavy atoms in the protein
180 during the Monte Carlo search, in addition to an automated means of thresholding the list
181 of ranked scores. These modifications were implemented to provide a more selective set
182 of sites; without them, a very large fraction of the protein surface would be occupied by
183 critical sites (Figure S2C). Within our dataset of proteins exhibiting alternative
184 conformations, we find that this modified approach results in an average of ~2 distinct
185 sites per domain (Figure S2A; see Figure S2B, for the distribution for distinct sites within
186 entire complexes).

187 Within the canonical set of 12 proteins, we positively identify an average of 56%
188 of the sites known to be directly involved in ligand or substrate binding (see Table 1,
189 Figure S1, and SI Methods section 3.1-a-iv). Some of the sites identified do not directly

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: an exceedingly

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: captured

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: 2C

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: 2A

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: 2B

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

DECLAN CLARKE 12/18/15 5:57 PM

Formatted: Not Highlight

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: [[DC2MG(12/11): I actually don't know if I fully agree with this change that was introduced: when we talk about latent allosteric sites, the thing this was previously unfound is not the motions themselves, but rather the pockets which were not previously known to disrupt already-known motions. We can discuss during P2 struct]].

202 **Dynamical Network Analysis to Identify Interior-Critical Residues**

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

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

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

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

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

242

243 **Software Tool: STRESS (STRucturally-identified ESSential residues)**

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

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: A community

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: Figures S2

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: Finally, the betweenness of each edge is calculated. The betweenness of an edge is defined as the number of shortest paths between all pairs of residues that pass through that edge, with each path representing the sum of effective node-node distances assigned in the weighting scheme above.

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

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: Obviating the need for long wait times, the algorithmic implementation of our software is highly efficient (Figures 3B and 3C).

DECLAN CLARKE 12/18/15 5:57 PM

Moved (insertion) [1]

268

269 High-Throughput Identification of Alternative 270 Conformations

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

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

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: also

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: .

DECLAN CLARKE 12/18/15 5:57 PM

Moved up [1]: 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.

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: .

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

298

299 **Evaluating Conservation of Critical Residues**

300 **Using Various Metrics and Sources of Data**

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

311

312 **Conservation Across Species**

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

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: within the

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: identified

319

320 | **Leveraging Next-Generation Sequencing to Measure Conservation Amongst**
321 | **Humans**

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

329 | Non-synonymous single-nucleotide variants (SNVs) from the 1000 Genomes

330 | dataset (McVean *et al.*, 2012) that ~~intersect~~, surface-critical residues tend to occur at
331 | lower DAF values ~~than do SNVs that intersect non-critical residues~~ (Figure 4C). Though
332 | this ~~difference~~, is not observed to be significant, the significance improves when
333 | examining the shift in DAF distributions, as evaluated with a KS test (p= 0.159, Figure
334 | S4A), and we point out ~~only a~~ limited number of proteins (thirty-two) for which these
335 | 1000 Genomes SNVs ~~intersect~~ with surface-critical sites. Furthermore, the long tail
336 | extending to lower DAF values for surface-critical residues may suggest that only a
337 | subset of the residues in our prioritized binding sites is essential. In contrast to surface-
338 | critical residues, however, interior-critical residues ~~intersect~~ 1000 Genomes SNVs with
339 | significantly lower DAF values than ~~do~~ non-critical residues (Figure 4G; see also Figure
340 | S4B).

341 | **[[DC2MG(dec18): The paragraph added below was introduced after you**
342 | **suggested that we discuss the stats issue in the most recent annotated PDF. Another**

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: Between

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: hit

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: trend

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: the

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: coincide

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: are hit by

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: . Given

DECLAN CLARKE 12/18/15 5:57 PM
Formatted: Highlight

350 option might be to put this text in the Discussion instead of the Results, but I don't feel
351 too strongly either way]] When analyzing human polymorphism data, a variety of
352 statistical measures relating SNVs to selective constraint may be calculated, and the
353 results obtained (along with their associated significance levels) are highly dependent on
354 sample size. 1000 Genomes datasets are attractive partially because of their status as a
355 well-established "blue chip" set of variants in human populations. However, given the
356 relatively limited number of proteins that intersect with, 1000 Genomes SNVs, we also
357 analyzed the larger dataset provided by the Exome Aggregation Consortium (ExAC)
358 (Exome Aggregation Consortium, 2015). Though this dataset has been released much
359 more recently (and is consequently not yet as well established as 1000 Genomes), ExAC
360 provides sequence data from more than 60,000 individuals, and samples are sequenced at
361 much higher coverage, thereby ensuring better data quality. This larger dataset enables us
362 to more easily examine trends in the data as they relate to critical and non-critical
363 residues.

364 Using MAF as a conservation metric, we performed a similar analysis using this
365 data. MAF distributions for surface- and non-critical residues in the same set of proteins
366 are given in Figure 4D. Although the mean value of the MAF distribution for surface-
367 critical residues is slightly higher than that of non-critical residues, the median for
368 surface-critical residues is substantially lower than that for non-critical residues,
369 demonstrating that the majority of proteins are such that MAF values are lower in
370 surface- than in non-critical residues. In addition, the overall shifts of these distributions
371 also point to a trend of lower MAF values in surface-critical residues (Figure S5A, KS
372 test $p=9.49e-2$).

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: to be hit by

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: , Cambridge MA 2015).

DECLAN CLARKE 12/18/15 5:57 PM

Formatted: Indent: First line: 0.5"

375 Interior-critical residues exhibit significantly lower MAF values than do non-
376 critical residues in the same set of proteins. MAF distributions for interior- and non-
377 critical residues are given in Figure 4H (see also Figure S5B).

378 In addition to analyzing overall allele frequency distributions, we also evaluate
379 the *fraction* of rare alleles as a metric for measuring selective pressure. This fraction is
380 defined as the ratio of the number of rare (i.e., low-DAF or low-MAF) non-synonymous
381 SNVs to the number of all non-synonymous SNVs in a given protein annotation (such as
382 all surface-critical residues of the protein, for example; see SI Methods section 3.3-b). A
383 higher fraction is interpreted as a proxy for greater conservation (Khurana *et al.*, 2013;
384 Sethi *et al.*, 2015). Using variable DAF (MAF) cutoffs to define rarity for 1000 Genomes
385 (ExAC) SNVs, both surface- and interior-critical residues are shown to harbor a higher
386 fraction of rare alleles than do non-critical residues, further suggesting a greater degree of
387 evolutionary constraint on critical residues (See Figure 5).

388

389 **Comparisons Between Different Models of Protein Motions**

390 The identification of surface- and interior-critical residues entails using sets of
391 vectors (on each protein residue) to describe conformational change. Notably, our
392 framework enables one to determine these vectors in multiple ways. Conformational
393 changes may be modeled using vectors connecting residues in crystal structures from
394 alternative conformations. We term this approach “ACT”, for “absolute conformational
395 transitions” (see SI Methods section 3.2-c). The crystal structures of such paired
396 conformations may be obtained using the framework discussed above. The protein
397 motions may also be inferred from anisotropic network models (ANMs) (Atilgan *et al.*,
398 2001). ANMs entail modeling interacting residues as nodes linked by flexible springs, in

399 a manner similar to elastic network models (Fuglebakk *et al.*, 2015; Tirion, 1996) or
400 normal modes analysis (Figure 1B). ANMs are not only simple and straightforward to
401 apply on a database scale, but unlike using alternative crystal structures, the motion
402 vectors inferred may be generated using a single structure.

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

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: Modeling

408

409 **Critical Residues in the Context of Human Disease Variants**

410 | Directly related to conservation is confidence with which an SNV is believed to
411 be disease-associated. SIFT (Ng and Henikoff, 2001) and PolyPhen (Adzhubei *et al.*,
412 | 2010) are two tools for predicting SNV deleteriousness. ExAC SNVs ~~that intersect~~
413 critical residues exhibit significantly higher PolyPhen scores relative to non-critical
414 residues, suggesting the potentially higher disease susceptibility at critical residues
415 (Figure S6). Significant disparities were not observed in SIFT scores (Figure S7).

DECLAN CLARKE 12/18/15 5:57 PM
Deleted: hitting

416 | Using HGMD (Stenson *et al.*, 2014) and ClinVar (Landrum *et al.*, 2014), we
417 identify proteins with critical residues that coincide with disease-associated SNVs (Figure
418 7A and File S2). Several critical residues coincide with known disease loci for which the
419 mechanism of pathogenicity is otherwise unclear (File S3). The fibroblast growth factor
420 receptor (FGFR) is a case-in-point (Figure 7). SNVs in FGFR have been linked to
421 craniofacial defects. Dotted lines in Figure 7B highlight poorly understood disease SNVs
422 that coincide with critical residues. In addition, we identify Y328 as a surface-critical

425 residue, which coincides with a disease-associated SNV from HGMD, despite the lack of
426 confident predictions of deleteriousness by several widely used tools for predicting
427 disease-associated SNVs, including PolyPhen (Adzhubei *et al.*, 2010), SIFT (Ng and
428 Henikoff, 2001), and SNPs&GO (Calabrese *et al.*, 2009). Together, these results suggest
429 that the incorporation of surface- and interior-critical residues introduces a valuable layer
430 of annotation to the protein sequence, and may help to explain otherwise poorly
431 understood disease-associated SNVs.

432

433 **DISCUSSION & CONCLUSIONS**

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

445 When applied to many proteins with distinct conformational changes in the PDB,
446 we investigate the conservation of potential allosteric residues in both inter-species and
447 intra-human genomes contexts, and find that these residues tend to exhibit greater

448 conservation in both cases. In addition, we identify several disease-associated variants for
449 which plausible mechanisms had been unknown, but for which allosteric mechanisms
450 provide a ~~reasonable~~ rationale.

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

462 Despite these limitations in MD and NMR, allosteric mechanisms and signaling
463 pathways may be conserved across many different but related proteins within the same
464 family, suggesting that such computationally- or labor-intensive approaches for all
465 proteins may not be entirely essential. Flock *et al.* have carefully demonstrated that the
466 allosteric mechanisms responsible for regulating G proteins through GPCRs tend to be
467 conserved (Flock *et al.*, 2015). Investigations into representative families have also been
468 enlightening in other contexts. In one of the early studies employing network analysis,
469 del Sol *et al.* conduct a detailed study of several allosteric protein families (including
470 GPCRs) to demonstrate that residues important for maintaining the integrity of short

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: plausible

472 paths within residue contact networks are essential to enabling signal transmission
473 between distant sites (del Sol *et al.*, 2006). Another notable result in the same work is that
474 these key residues (which match experimental results) may become redistributed when
475 the protein undergoes conformational change, thereby changing optimal communication
476 routes as a means of conferring different regulatory properties.

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

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

490 We anticipate that, within the next decade, deep sequencing will enable structural
491 biologists to study evolutionary conservation using sequenced human exomes just as
492 routinely as cross-species alignments. Furthermore, intra-species metrics for conservation
493 provide added value in that the confounding factors of cross-species comparisons are

494 | removed: different ~~species~~, evolve in ~~various~~ evolutionary contexts ~~and at different rates~~,

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: organisms

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: different cellular and

497 and it can be difficult to decouple these different effects from one another. Cross-species
498 metrics of protein conservation entail comparisons between proteins that may be very
499 different in structure and function. Sequence-variable regions across species may not be
500 conserved, but nevertheless impart essential functionality. Intra-species comparisons,
501 however, can often provide a more direct and sensitive evaluation of constraint.

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

508

509 **METHODS**

510 An overview of the framework for finding surface- and interior-critical residues is
511 given in Figure 1. Figure S3 provides a schematic of our pipeline for identifying
512 alternative conformations throughout the PDB. Cross-species conservation scores were
513 analyzed in those PDBs for which full ConSurf files are available through the ConSurf
514 server. 1000 Genomes SNVs were taken from the Phase 3 release, and ExAC SNVs were
515 downloaded in May 2015. Further details on all protocols are provided in SI Methods.

516

517 **ACKNOWLEDGMENTS**

518 DC acknowledges the support of the NIH Predoctoral Program in Biophysics (T32
519 GM008283-24). We thank Simon Mitternacht for sharing the original source code for
520 binding leverage calculations, as well as Koon-Kiu Yan for helpful discussions and
521 feedback. The authors would like to thank the Exome Aggregation Consortium and the
522 groups that provided exome variant data for comparison. A full list of contributing groups
523 can be found at <http://exac.broadinstitute.org/about>
524

525 REFERENCES

- 526 Adzhubei, I. Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P.,
527 Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting
528 damaging missense mutations. *Nat. Methods.* *7*, 248–249
- 529 Ansari, A., Berendzen, J., Bowne, S., Frauenfelder, H., Iben, I.E.T., Sauke, T.B.,
530 Shyamsunder, E., and Young, R.D. (1985). Protein states and protein quakes. *Proc.*
531 *Natl. Acad. Sci. U.S.A.* *82*, 5000–5004.
- 532 Arnlund, D., Johansson, L.C., Wickstrand, C., Barty, A., Williams, G.J., Malmerberg, E.,
533 Davidsson, J., Milathianaki, D., DePonte, D.P., Shoeman, R.L., *et al.* (2014).
534 Visualizing a protein quake with time-resolved X-ray scattering at a free-electron
535 laser. *Nat. Methods.* *11*, 923–6.
- 536 Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., and Bahar, I.
537 (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network
538 Model. *Biophys. J.* *80*, 505–515.
- 539 Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H.,
540 Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids*

541 Res. 28, 235–242.

542 Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L. and Casadio, R. (2009).

543 Functional annotations improve the predictive score of human disease-related

544 mutations in proteins. *Hum. Mutat.* 30, 1237–1244.

545 Exome Aggregation Consortium (ExAC). (2015) Cambridge, MA.

546 <http://exac.broadinstitute.org>.

547 Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M. and Funkhouser, T.A. (2009).

548 Predicting protein ligand binding sites by combining evolutionary sequence

549 conservation and 3D structure. *PLoS Comput. Biol.* 5, e1000585.

550 Chennubhotla, C. and Bahar, I. (2006). Markov propagation of allosteric effects in

551 biomolecular systems: application to GroEL–GroES. *Mol. Syst. Biol.* 2.

552 del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residues crucial for

553 maintaining short paths in network communication mediate signaling in proteins.

554 *Mol. Syst. Biol.* 2(1).

555 Csermely, P., Korcsmáros, T., Kiss, H.J.M., London, G., and Nussinov, R. (2013).

556 Structure and dynamics of molecular networks: A novel paradigm of drug discovery.

557 *Pharmacol. Ther.* 138, 333–408.

558 Flock, T., Ravarani, C.N.J., Sun, D., Venkatakrishnan, A.J., Kayikci, M., Tate, C.G.,

559 Veprintsev, D.B. and Babu, M.M. (2015). Universal allosteric mechanism for $G\alpha$

560 activation by GPCRs. *Nature* 524, 173–179.

561 Fuglebakk, E., Tiwari, S.P., and Reuter, N. (2015). Comparing the intrinsic dynamics of

562 multiple protein structures using elastic network models. *Biochim. Biophys. Acta -*

563 *Gen. Subj.* 1850, 911–922.

564 Gasper, P.M., Fuglestad, B., Komives, E.A., Markwick, P.R.L., and McCammon, J.A.
565 (2012). Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant
566 activities. *Proc. Natl. Acad. Sci. U. S. A.* *109*, 21216–22.

567 Ghosh, A., and Vishveshwara, S. (2008). Variations in Clique and Community Patterns in
568 Protein Structures during Allosteric Communication: Investigation of Dynamically
569 Equilibrated Structures of Methionyl tRNA Synthetase Complexes. *Biochemistry*.
570 *47*, 11398-11407.

571 Girvan, M., Girvan, M., Newman, M.E.J., and Newman, M.E.J. (2002). Community
572 structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.* *99*, 7821–
573 7826.

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

576 Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A.,
577 Lochovsky, L., Chen, J., Harmanci, A., *et al.* (2013). Integrative Annotation of
578 Variants from 1092 Humans: Application to Cancer Genomics. *Science*. *342*,
579 1235587–1235587.

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

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

585 Lee, J., Natarajan, M., Nashine, V.C., Socolich, M., Vo, T., Russ, W.P., Benkovic, S.J.,
586 and Ranganathan, R. (2008). Surface Sites for Engineering Allosteric Control in

587 Proteins. *Science* 322, 438-442.

588 Lockless, S.W., Ranganathan, R., Kukic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri,
589 P., Pollastri, G., Socolich, M., Lockless, S.W., *et al.* (1999). Evolutionarily
590 conserved pathways of energetic connectivity in protein families. *BMC*
591 *Bioinformatics* 15, 295–299.

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

596 Ming, D. and Wall, M.E. (2005). Quantifying allosteric effects in proteins. *Proteins* 59,
597 697–707.

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

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

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

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

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

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

612 Rivalta, I., Sultan, M.M., Lee, N.-S., Manley, G. a., Loria, J.P., and Batista, V.S. (2012).
613 PNAS Plus: Allosteric pathways in imidazole glycerol phosphate synthase. *Proc.*
614 *Natl. Acad. Sci.* *109*, E1428–E1436.

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

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

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

623 Sethi, A., Clarke, D., Chen, J., Kumar, S., Galeev, T.R., Regan, L., and Gerstein, M.
624 (2015). Reads meet rotamers: structural biology in the age of deep sequencing. *Curr.*
625 *Opin. Struct. Biol.* *35*, 125-134.

626 Shulman, A.I., Larson, C., Mangelsdorf, D.J., and Ranganathan, R. (2004). Structural
627 determinants of allosteric ligand activation in RXR heterodimers. *Cell* *116*, 417–
628 429.

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

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

636 Tirion, M.M. (1996). Large Amplitude Elastic Motions in Proteins from a Single-
637 Parameter, Atomic Analysis. *Phys. Rev. Lett.* *77*, 1905–1908.

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

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

643 Yang, L.W. and Bahar, I. (2005). Coupling between catalytic site and collective
644 dynamics: A requirement for mechanochemical activity of enzymes. *Structure* *13*,
645 893–904.

646
647

648 CAPTIONS

649 **Figure 1. Schematic overviews of methods for finding surface- and interior-critical**
650 **residues.** (A) A simulated ligand probes the protein surface in a series of Monte Carlo
651 simulations (top-left). The cavities identified may be such that occlusion by the ligand
652 strongly interferes with conformational change (top-right; such a site is likely to be
653 identified as surface-critical, in red), or they may have little effect on conformational
654 change, as in the case of shallow pockets (bottom-left) or pockets in which large-scale
655 motions do not drastically affect pocket volume (bottom-right). (B) Interior-critical

656 residues are identified by weighting residue-residue contacts (edges) on the basis of
657 correlated motions, and then identifying communities within the weighted network.
658 Residues involved in the highest-betweenness interactions between communities (in red)
659 are selected as interior-critical residues.

660

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

668

669 **Figure 3. STRESS web server front page, running times, and server architecture.**

670 (A) The server enables users to either provide PDB IDs or to upload their own PDB files
671 for proteins of interest. Users may opt to identify surface-critical residues, interior-critical
672 residues, or both. (B) Running times are shown for systems of various sizes. Shown in
673 red are the running times without optimizing for speed, and green shows running times
674 with algorithmic optimization. (C) The same data is represented as a log-log plot. The
675 slopes of these two approaches demonstrate that our algorithm reduces the computational
676 complexity by an order of magnitude. Our speed-optimized algorithm scales at $O(n^{1.3})$,
677 where n is the number of residues. (D) A thin front-end server handles incoming user
678 requests, and more powerful back-end servers perform the heavier algorithmic

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: Figure 2. Summary statistics for surface-critical sites. The distributions of the numbers of surface-critical sites per domain and per complex are given in (A) and (B), respectively. Panel (C) gives the distributions of the number of surface-critical sites per complex without thresholding. Complexes are taken from the the PDB biological assembly files. Without applying thresholds to the list of ranked surface-critical sites, the protein is often covered with an excess of identified critical sites. -

691 calculations. The back-end servers are dynamically scalable, making them capable of
692 handling wide fluctuations in user demand. Amazon's Simple Queue Service is used to
693 coordinate between user requests at the front end and the back-end compute nodes: when
694 the front-end server receives a request, it adds the job to the queue, and back-end servers
695 pull that job from the queue when ready. Source code is available through Github
696 (github.com/gersteinlab/STRESS).

697

698 **Figure 4. Multiple metrics and datasets reveal that critical residues tend to be**
699 **conserved.** Surface- and interior-critical residues (red) in phosphofructokinase (PDB
700 3PFK) are given in (A) and (E), respectively. Distributions of cross-species conservation
701 scores, 1000 Genomes SNV DAF averages, and ExAC SNV MAF averages for surface-
702 and non-critical residue sets are given in (B), (C), and (D), respectively. The same
703 distributions corresponding to interior- and non-critical residue sets are given in (F), (G),
704 and (H), respectively. In (B), mean inter-species conservation scores for surface-critical
705 sets are -0.131, whereas non-critical residue sets with the same degree of burial have a
706 mean score of +0.059 ($p < 2.2e-16$). In (F), mean inter-species conservation scores for
707 interior-critical sets are -0.179, whereas non-critical residue sets with the same degree of
708 burial have a mean score of -0.102 ($p=3.67e-11$). In (C), means for surface- and non-
709 critical sets are $9.10e-4$ and $8.34e-4$, respectively ($p=0.309$); corresponding means in (D)
710 are $4.09e-04$ and $2.26e-04$, respectively ($p=1.49e-3$). In (G), means for interior- and non-
711 critical sets are $2.82e-4$ and $3.12e-3$, respectively ($p=1.80e-05$); corresponding means in
712 (H) are $3.08e-05$ and $3.27e-04$, respectively ($p=7.98e-09$). N = 421, 32, 84, 517, 31, and

713 90 structures for panels B, C, D, F, G, and H, respectively. P-values are based on
714 Wilcoxon-rank sum tests. See SI Methods for further details.

715

716 **Figure 5. Critical residues are shown to be more conserved, as measured by the**
717 **fraction of rare alleles.** Protein regions with high fractions of *rare* variants are believed
718 to be more sensitive to sequence variants than other regions, thereby explaining why such
719 variants occur infrequently in the population. Panels (A) and (C) show distributions for
720 rare (low DAF) non-synonymous SNVs (taken from the 1000 Genomes dataset) in which
721 the critical residues are defined to be the surface-critical (A) and interior-critical (C)
722 residues. Panels (B) and (D) show distributions for rare (low MAF) non-synonymous
723 SNVs (taken from the ExAC dataset) in which the critical residues are defined to be the
724 surface-critical (B) and interior-critical (D) residues. For varying thresholds to define
725 rarity, there are more structures in which the fraction of rare variants is higher in critical
726 residues than in non-critical residues. Cases in which the fraction is equal in both
727 categories are not shown. We consider all structures such that at least one critical and at
728 least one non-critical residue intersect a non-synonymous SNV. Panels (A), (B), (C), and
729 (D) represent data from 31, 90, 32, and 84 structures, respectively.

730

731 **Figure 6. Modeling protein conformational change through a direct use of crystal**
732 **structures from alternative conformations using absolute conformational transitions**
733 **(ACT).** (A) Distributions (155 structures) of the mean conservation scores on surface-
734 critical (red) and non-critical residues with the same degree of burial (blue). (B)
735 Distributions (159 structures) of the mean conservation scores for interior-critical (red)

DECLAN CLARKE 12/18/15 5:57 PM

Deleted: are hit by

737 and non-critical residues with the same degree of burial (blue). Mean values are given in
738 parentheses. Results for single-chain proteins are shown, and p-values were calculated
739 using a Wilcoxon rank sum test.

740

741 **Figure 7. Potential allosteric residues add a layer of annotation to structures in the**

742 **context of disease-associated SNVs.** The structure shown (*A*) is that of the fibroblast

743 growth-factor receptor (FGFR) in VMD Surf rendering, with HGMD SNVs shown in

744 orange, bound to FGF2, in ribbon rendering (PDB 1HIL). (*B*) A linear representation of

745 structural annotation for FGFR. Dotted lines highlight loci which correspond to HGMD

746 sites that coincide with critical residues, but for which other annotations fail to coincide.

747 Deeply-buried residues are defined to be those that exhibit a relative solvent-exposed

748 surface area of 5% or less, and binding site residues are defined as those for which at

749 least one heavy atom falls within 4.5 Angstroms of any heavy atom in the binding partner

750 (heparin-binding growth factor 2). The loci of PTM sites were taken from UniProt

751 (accession P21802).

752

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

754 **proteins**

755 For each *apo* structure within the canonical set of proteins, statistics relating surface-

756 critical sites to known ligand-binding sites are reported. The surface of a given structure

757 is defined to be the set of all residues that have a relative solvent accessibility of at least

758 50%, where relative solvent accessibility is evaluated using all heavy atoms in both the

759 main-chain and side-chain of a given residue. Mean values are given in the bottom row.

760 NACCESS is used to calculate relative solvent accessibility (Hubbard and Thornton,
761 1993) . *Column 1*: PDB IDs for each structure; *Column 2*: among these surface residues,
762 the fraction that constitute surface-critical residues; *Column 3*: among surface residues,
763 the fraction that constitute known ligand-binding residues (known ligand-binding
764 residues are taken to be those within 4.5 Angstroms of the ligand in the *holo* structure;
765 Table S1); *Column 4*: the Jaccard similarity between the sets of residues represented in
766 columns 2 and 3 (i.e., surface-critical and known-ligand binding residues), where values
767 given in parentheses represent the expected Jaccard similarity, given a null model in
768 which surface-critical and ligand-binding residues are randomly distributed throughout
769 the surface (for each structure, 10,000 simulations are performed to produce random
770 distributions, and the expected values reported here constitute the mean Jaccard similarity
771 among the 10,000 simulations for each structure); *Column 5*: the number of distinct
772 surface-critical sites identified in each structure; *Column 6*: the number of known ligand-
773 binding sites in each structure; *Column 7*: the number of known ligand-binding sites
774 which are positively identified within the set of surface-critical sites, where a positive
775 match occurs if a majority of the residues in a surface-critical site coincide with the
776 known ligand-binding site; *Column 8*: The fraction of ligand-binding sites captured is
777 simply the ratio of the values in column 7 to those in column 6.
778

