MotifVar: A resource for amplifying population genetic signal by using repeat protein domains

Jieming Chen[1,2,3], Lynne Regan[1,2,3]*, Mark Gerstein[1,2,3,4]*#

[1]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA.
[2]Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT 06520, USA.
[3]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA.
[4]Department of Computer Science, Yale University, New Haven, CT 06520, USA.
*These authors co-directed the work
#Corresponding author

Abstract

Large-scale whole genome and exome sequencing holds great promise for the interpretation of protein structures. However, because protein-coding regions are under high selective constraints, their sequences are extremely conserved and variants occur at low frequencies, such that there is often insufficient statistics for downstream population genetics (popgen) calculations. To address this problem, we develop the MotifVar approach, which uses the modular structure of repeat protein domains (RPDs) to amplify popgen signals. In particular, we are able to aggregate variants at the codon level within the human population and compute popgen metrics to identify important positions of a RPD that show stronger conservation signals. This allows us to compare conservation directly over different evolutionary timescales. It also enables us to readily visualize popgen measures on protein structures. We make available the MotifVar results for RPDs as an online resource (http://motifvar.gersteinlab.org) and illustrate its applicability through a case study on the tetratricopeptide repeat.

Introduction

The combined efforts from large-scale human sequencing projects and clinical sequencing have given rise to an exponentially increasing number of human sequences in recent years.[1–3] With substantial drop in the sequencing cost and improvement in sequencing technologies and data processing capabilities, we now have the ability to generate a huge catalog of variants that exist in the human population in a fairly rapid and high-throughput fashion. One of the challenges is to provide functional annotations for these variants efficiently and accurately.

Much of the variant annotation work has been performed in the protein-coding regions. A non-synonymous mutation is considered functionally disruptive if it occurs in a region of high conservation, which are considered to be important evolutionarily.[4] Evolutionary conservation can be observed at different levels. Inter-species comparison can pick out fixed differences between the dominant homologous sequences of the chosen species across their phylogeny over a long evolutionary time.[5–7] At a more recent timescale, intra-species conservation (across a population) has been observed over specific sites in a few large-scale sequencing studies, by aggregating variants over a region or site within the human population.[8–10] However, all protein-coding regions are, in general, under high selection pressure. As such, almost all positions in

high-impact protein domains tend to be extremely conserved, making it tricky to pinpoint specific positions. Variants also occur sparsely across the coding region and at very low frequencies within a population. Consequently, it is difficult to increase the number of variants for population analyses without increasing the pool of sequenced individuals. To this end, we devise an "intra-genome conservation" approach that is able to "amplify" the variant signal in protein-coding regions within a population.

We focus on a functional category of protein domains that explicitly mediates protein-protein interactions (PPI), known as repeat protein domains (RPDs).[11,12] RPDs have been found to be present in almost one in every three human protein.[13] As a result, many classes of RPDs have also been studied extensively.[14–16] Each RPD is made up of modular repeat motifs of the same class. For example, tetratricopeptide repeat (TPR) domains are made up of only TPR motifs and Ankyrin repeat (ANK) domains of ANK repeat motifs. This modularity gives rise to a strategy that was first introduced in the field of protein engineering to generate protein design templates to create synthetic proteins with desired specificities and affinities.[17–19] We adapted the strategy to build a multiple sequence alignment (MSA) profile, which we term a 'motif-MSA' profile, for each class of RPD. Using the TPR as an example of a class of PPI RPD, we demonstrate that the motif-MSA strategy can "amplify" variant signal by aggregating the variants from all homologous motifs for each class of RPD within the human genome. Interestingly, we note that such analyses of intra-genome conservation can only be performed using a dataset as large as those from ExAC. Our MotifVar database (http://motifvar.gersteinlab.org) contains our results as a resource for annotating variants in 17 PPI RPDs.

Results
MotifVar *database*
Figure 1a shows our strategy that is used to build up the resources in our publicly available MotifVar database (http://motifvar.gersteinlab.org) that relates protein residue to genomic information in 17 RPDs. Our strategy first produces a motif sequence alignment profile for a class of repeat domain. Using the TPR repeat domain as an example, we obtain every TPR repeat motif of a given amino acid length in the human proteome (typically the length with the most number of available motifs); in this case, the length is 34 amino acids (see 'Methods' for details; Supplementary Figure 1). We then perform an MSA of all the TPR motifs (we term 'motif-MSA') to obtain a residue frequency table, which shows the percentage occurrence of each amino acid at each position in the motif. This table can then be translated into a sequence logo for better visualization. For each repeat motif, we then locate its genomic positions in the human genome. Subsequently, we map SNVs onto the genomic coordinates of the repeat motifs. This allows us to obtain aggregate counts of variants at each residue positions for each class of repeat domain based on SNV allele frequencies and the functional impact, namely whether the SNV is rare (R) or common (C) in the human population and whether the SNV causes a synonymous (S) or non-synonymous (NS) change. From these statistics, we can subsequently derive more meaningful metrics such as ratio of NS-to-S-SNV profile (NS/S) and enrichment of rare variants (R/C) for interpretation of each residue position. We provide these results for the users in our MotifVar database (http://motifvar.gersteinlab.org).

Comparing species- and motif-MSA

An MSA is more typically performed using homologous sequences from multiple species (Figure 1b; we term 'species-MSA'). Here, we perform species-MSA for the first three TPR motif sequences in the TPR-containing protein TTC21B, using orthologous sequences from 66 species (see 'Methods' for details) (Figure 2a). TTC21B contains about 16-19 TPR motifs, with almost all of them having a length of 34 amino acids and is a cilia-specific protein that is necessary for retrograde intra-flagellar transport.[20] Expectantly, most positions are comparably high in sequence conservation. In contrast, the motif-MSA profile exhibits substantially differential sequence conservation among the motif positions (Figure 2b). We were able to easily identify positions 8, 11, 20, 24 and 27 as more conserved within the TPR repeat motif.

Motif-MSA amplifies variant signals to compute population genetic metrics
The conventional species-MSA profile is restricted to the sequence of a single human protein (since the alignment is based on orthologs), hence even with a large catalog of human exonic variants, only a maximum of three human variants can occur for each residue's codon position (Figure 1b). As such, the variant signal is extremely tenuous for any meaningful downstream population genetics analyses. However, in the TPR motif-MSA, variants are aggregated from all 34-amino-acid TPR motifs within the human genome. This accumulation of variants amplifies the variant signal, thereby facilitating the computation of various population genetic metrics to investigate selective constraints in the protein domains. At this juncture, we note that even with amplification, our results were observable only with the ExAC dataset (60,706 exomes), but not when solely with the 1000 Genomes Project Phase 1 data (1000GP; 1,092 whole genomes)[8] nor its combination with the Exome Sequencing Project (ESP; 6,500 exomes)[10], which total more than 7,500 protein-coding exomes (Figure 2c and Supplementary Table 1). The fact that only the largest dataset with more than 60K exomes and 7M SNVs yields interpretable results underscores the importance of still having more genome sequences and rare variants.

We use the TPR domains as an example to show the results of four evolutionary measures derived from the accumulation of genomic variants on the motif-MSA. For inter-species conservation, we use the SIFT score of a non-synonymous SNV, which is directly computed from a species-MSA, such that a lower SIFT score denotes a greater likelihood of an SNV being deleterious (most likely due to high residue conservation).[4] Since protein-coding regions are generally under high selective constraints across species, almost all positions of highly functional PPI domains tend to have very low median SIFT scores across the motif. In the TPR motif-MSA, the most highly conserved position 20 exemplified this observation (Figure 3a and 3d). As a proxy for intra-species conservation within the human population, we compute a population genetic measure introduced in the 1000 Genomes Project as a result of an influx of rare variants, the rare-to-common-variant ratio (R/C), where an enrichment of rare variants (or depletion of common variants) signifies high conservation over a shorter evolutionary timescale.[8–10] We find high rare variant enrichments across the motif-MSA profiles of all classes of RPDs, regardless of residue or positional conservation within the repeat motifs (Figure 3b).

We further compute the NS/S for each position in the motif-MSA profile (Figure 3c). The use of NS/S has been traditionally useful in the estimation of selection pressures in the protein-coding regions typically at the gene level.[21] Here, rather than at the gene level, the accumulation of variants enables NS/S to be calculated at the codon level (Figure 3c). We observe that most of the positions in the TPR motif with very low NS/S coincide very well with positions of high

sequence conservation in the motif-MSA profile. In fact, if we arbitrarily take the top five positions with the lowest NS/S, four of them are with four of the most conserved positions in the TPR motif-MSA, reinforcing the utility of motif-MSA in picking out functionally important residue positions (Figure 3c).

The difference of derived (population) allele frequencies, or ΔDAF, has been introduced in the 1000 Genomes Project to quantify population differentiation (between pairs of populations) and identify highly differentiated (highD) sites, in view of the large number of rare variants.[8,22] Because the majority of the variants are rare even within sub-populations, we observe that all positions have low ΔDAF medians. More interestingly, we can identify some residue positions that are differentiated between certain populations (Figure 3xxx).

Combining protein and genomic information to identify important residues
Finally, using the motif-MSA, we are able to integrate both protein (from MSA) and genomic information (SNVs) to better pinpoint positions that might be more functionally important. By combining positions with the highest five sequence conservation in the TPR motif-MSA and the lowest five median SIFT scores and NS/S ratio, we are able to identify eight positions (out of 34 positions on the TPR motif), with four positions that fulfil at least two of the three selective constraint conditions (Figure 3d). The differences in R/C between positions within the TPR motif-MSA are too subtle to be used.

Mapping genomic information onto protein structures
Because the motif-MSA identifies important residues in the simplest unit of an RPD, we can visualize the residues in 3D structures of the same class of RPDs with any number of motifs. As an example, we use the X-ray crystal structure of a three-motif TPR domain (TPR1) from the human protein Hsp-organizing protein (HOP) bound to its cognate ligand, a short peptide sequence consisting of seven amino acids, PTIEEVD (PDB ID: 1ELW).[23] We map the eight positions derived from Figure 3d onto all three motifs of the protein structure, identifying 24 residues in total (Figure 3e). In each TPR motif, except for position 17, we find that all the other seven residue positions with high selective constraints – from either low median SIFT scores (labeled as red numbers), low log (NS/S) (labeled as blue numbers) or high motif sequence conservation (highlighted orange) – are buried residues in the PPI domain (Figure 3e), in line with a previous study.[24]

Relating residues positions to clinically-relevant and disease-related mutation data
Using two databases, ClinVar[25] and the proprietary Human Gene Mutation Database (HGMD)[26], we found that the highly constrained positions have some of the most occurrences of clinically-relevant or disease-related mutations along the TPR motif-MSA profile, including the highest two at positions 6 and 7, which would otherwise not be detected if only motif-MSA or inter-species conservation was used (Figure 3f). In fact, mechanistic studies of a number of these mutations show that the occurrence of certain NS mutations on these positions give rise to diseases precisely as a result of ablation of protein-protein interactions.[27,28]

Discussion
For decades, the focus in research on PPI has typically been the investigation of protein interfaces that directly take part in the protein interaction. Most studies involved the use of 3D

protein structures, for instance, to identify protein-protein interfaces,[29,30] investigate interfacial properties[31,32] or to predict interacting 'hotspots'[33–35]. While extremely useful in protein engineering and drug design, it is also very limited by the number of available protein structures. On the other hand, the amount of human sequencing data has been growing dramatically over the past decade, in particular, the number of protein-coding exome sequences.[36] This huge trove of sequence information should be leveraged upon for variant annotation in protein-coding regions, especially in complementing protein data with the copious amount of human genomic data. Our introduction of the motif-MSA facilitates genomic analyses with protein information (and vice versa) in several ways.

Firstly, motif-MSA removes the limitation imposed by species-MSA. Thus far, the utility of protein sequences has been largely focused on the more traditional perspective of sequence conservation across multiple species based on homology.[5,6,37] By using information from the same motif class, we can systematically aggregate variants from similar protein regions within the genome of a single species in a reasonable manner. This aggregation is key to achieving the variant statistics required to perform analyses that are meaningful, especially in light of the observation that even a combined set of 1000GP and ESP6500 variant data, derived from almost 7600 exomes, was not sufficient to yield immediately-interpretable results (Figure 2c and Supplementary Table 1). At this point, it is also important to note that intra-genome conservation, while allowing amplification, combines genomic variant information not only from long and short evolutionary time scales, but also from the evolution of the same class of repeat motifs within the genome. Thus, the interpretation of selective constraints in metrics such as log(NS/S) is a confluence of evolutionary timescales and mutation processes.

Secondly, the ability to gain statistical power from variant aggregation makes motif-MSA an extremely powerful platform in investigating selective constraints using genomic information. Potentially, motif-MSA is amenable to the entire repertoire of genomic metrics. We used four metrics to demonstrate how motif positions and residues that show evidence for clinical and disease relevance can be identified, and would have been missed otherwise.[9]

Lastly, motif-MSA is also able to reflect protein structural properties and their roles in PPI. Conventional species-MSA aligns sequence orthologs that are similar in function and structure. Hence, highly conserved residues or positions are a mix of structural and functional residues. On the other hand, because the protein motifs are classified by their structural folds, sequence features in a motif-MSA are important structural features that determine the folds of the PPI domains. These features are observed as buried residues within the interior of PPI domains (Figure 3e). In addition, it has been suggested that because motifs in motif-MSA are from a myriad of proteins with diverse binding partners, positions that are low in sequence conservation, or 'hypervariable', are found in the binding pockets of the corresponding domains.[24,38] Similarly, we noticed hypervariable positions, such as position 2 in TPR motifs, harbor a good number of disease-related variants.

The motif-MSA approach provides a powerful and versatile platform to facilitate the combination of protein and genome information for use in the annotation of protein structures. It enables the leveraging of the vast amount of human sequencing data currently available. This

will become increasingly more imperative and urgent in the future as human genome sequencing becomes more commonplace and personal genome interpretation takes center stage.

Methods

MotifVar *database*
Our publicly available MotifVar database (http://motifvar.gersteinlab.org) provides data files for for 17 RPDs, including TPRs. Each class of RPD is a tarball, which contains residue frequency tables (to rebuild the sequence logo), the SIFT score distributions, median SIFT scores, log(NS/S), log(R/C) and values for each position along each RPD motif to allow versatile thresholding by the users. The resource and scripts used in the pipeline are freely downloadable at the database.

Multiple sequence alignment (MSA)
All protein, motif and domain information are extracted from Ensembl database version 73 and SMART database, under the 'genomic' mode, for species, *Homo sapiens* (downloaded Oct 25, 2013).[40] The 17 PPI repeat domains are manually selected based on their availability in the SMART database.

We will use the TPR domains as an example to illustrate the process of motif- and species-MSA in our study.

To obtain a motif-MSA sequence profile, (1) we first extract all TPR domains in the human proteome and break them up into its constituent motifs. (2) Here, the motif-MSA is performed based on the most representative size of the motif. Hence, in order to select the motif size, a histogram of all sizes of TPR motifs is constructed (Supplementary Figure 1) and the most common motif size is selected for motif-MSA alignment; in TPR motifs, the most common motif size is 34 amino acids. There are a total of 114 human proteins (from unique genes) with 571 unique 34-amino-acid TPR motif sequences; we only keep one motif when there are multiple with 100% sequence identity. (3) MSA is then performed on of these 571 TPR motifs with 34 amino acids, with no gaps allowed, i.e. we line up all sequences by position end to end. This 'ungapped' alignment allows the derivation of a 20-by-*n* frequency table for 20 residues and *n* positions on the motif profile, and subsequently, visualization, using a sequence logo constructed by WebLogo 3.2.[41]

The TPR species-MSA is obtained by aligning the homologous protein sequences of TTC21B from 43 species in an 'ungapped' fashion. Using the MEGA5 software[42], we extracted the TPR domain from the 45-sequence alignment, based on the human TTC21B information in SMART database. There are 16 TPR motifs in TTC21B found in the SMART database. We remove two orthologs due to the existence of gaps in at least one of the 16 TPRs. Finally, we construct the sequence logo of all 16 TPRs using WebLogo 3.2.[41] We show the alignment of only the first three TPR motifs of TTC21B in Figure 2.

Sequence logo visualization
All sequence logos are created by WebLogo 3.2[41], using the following parameters:

-A protein -U bits --composition
"{'L':9.975,'A':7.013,'S':8.326,'V':5.961,'G':6.577,'K':5.723,'T':5.346,'I':4.332','E':7.096,'P':6.316,'R':5.650,'D':4.728,'F':3.658,'Q':4.758,'N':3.586,'Y':2.653,'C':2.307,'H':2.639,'M':2.131,'W':1.216}
" -n 34 -c chemistry --stack-width 25 --errorbar no

For the 'composition' parameter (used for the relative entropy calculation), we provided manually the background distribution of the amino acids in the entire SMART database ('genomic' mode), in order to be in line with our input data from the SMART database; the values above are in percentages. We separately computed these values from the SMART database. Unless the sequence logos are in monochrome (as in Figure 2), they are colored by amino acid chemistry, where polar residues (G, S, T, Y, C) are colored green, neutral residues (Q, N) purple, basic residues (K, R, H) blue, acidic residues (D, E) red, and hydrophobic residues (A, V, L ,I ,P ,W, F, M) black.

Variant information from exomes
For all the analyses in this study, we use the SNVs and their minor allele frequencies from 60,706 exomes found in the ExAC database[3] (Version 0.3, downloaded February 1 2015), after removing the variants from the sex chromosomes and singletons (those variants that only occur in one chromosome in the entire ExAC dataset). This ends up with 7,202,445 autosomal SNVs. We obtained SIFT scores, and non-synonymous nature of the SNVs on the proteins using the VEP tool (Version 73) from Ensembl release 73.[43]

Similarly, we have also used a combined number of 1,328,447 unique, non-singleton, and autosomal SNVs from the 1000 Genomes Project Phase 1 (1,092 whole genomes)[8] and Exome Sequencing Project data (6,500 exomes)[10], to produce Figure 2c and Supplementary Table 1.

All coordinates are based on the human reference genome assembly version of hg19.

Relating genomic and protein information
Custom scripts are written to relate genomic to protein information. The key portion is in identifying codon coordinates. We first obtain all genomic coordinates and strand information of protein-coding exons and residue coordinates of SMART protein domains from Ensembl 73 and GENCODE 18 on the reference genome, hg19. The exon information will give us the exact genomic coordinates of the codons for each protein-coding gene, using the locations of the exon-intron junctions. This allows mapping of genomic variants to specific codons, enabling positional accumulation of variant information across a motif-MSA profile. These scripts are part of the pipeline available for download in the MotifVar database.

Protein structure visualization
The X-ray crystal structures from Protein Data Bank (PDB) are created using Pymol 1.3.[44]

Clinically-relevant and disease-related variants
Clinically-relevant and disease-related variants in GRCh37 were downloaded from ClinVar[25] on July 8, 2015 and the proprietary HGMD Professional Database downloaded on July 27, 2015.[26]

Conflicts of interest statement
The authors declare that there is no conflict of interest.

References
1. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
2. Muddyman, D., Smee, C., Griffin, H. & Kaye, J. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med.* **5,** 100 (2013).
3. Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *BioRxiv* (2015). doi:10.1101/030338
4. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7,** 61–80 (2006).
5. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–81 (2009).
6. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7,** Unit7.20 (2013).
7. González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **88,** 440–9 (2011).
8. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
9. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
10. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–9 (2012).
11. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. A census of protein repeats. *J. Mol. Biol.* **293,** 151–60 (1999).
12. Kajava, A. V. Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.* **179,** 279–88 (2012).
13. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134,** 117–31
14. Li, J., Mahajan, A. & Tsai, M.-D. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry* **45,** 15168–78 (2006).
15. Andrade, M. A., Petosa, C., O'Donoghue, S. I., Müller, C. W. & Bork, P. Comparison of ARM and HEAT protein repeats. *J. Mol. Biol.* **309,** 1–18 (2001).
16. Allan, R. K. & Ratajczak, T. Versatile TPR domains accommodate different modes of target protein recognition and function. *Cell Stress Chaperones* **16,** 353–67 (2011).
17. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. The consensus concept for

thermostability engineering of proteins. *Biochim. Biophys. Acta* **1543,** 408–415 (2000).

18. Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. Design of stable alpha-helical arrays from an idealized TPR motif. *Structure* **11,** 497–508 (2003).

19. Parizek, P. *et al.* Designed ankyrin repeat proteins (DARPins) as novel isoform-specific intracellular inhibitors of c-Jun N-terminal kinases. *ACS Chem. Biol.* **7,** 1356–66 (2012).

20. Tran, P. V *et al.* THM1 negatively modulates mouse sonic hedgehog signal transduction and affects retrograde intraflagellar transport in cilia. *Nat. Genet.* **40,** 403–10 (2008).

21. Fay, J. C. Weighing the evidence for adaptation at the molecular level. *Trends Genet.* **27,** 343–9 (2011).

22. Colonna, V. *et al.* Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol.* **15,** R88 (2014).

23. Schmid, A. B. *et al.* The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop. *EMBO J.* **31,** 1506–17 (2012).

24. Magliery, T. J. & Regan, L. Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics* **6,** 240 (2005).

25. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–5 (2014).

26. Stenson, P. D. *et al.* 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 (2014).

27. Noack, D. *et al.* Autosomal recessive chronic granulomatous disease caused by novel mutations in NCF-2, the gene encoding the p67-phox component of phagocyte NADPH oxidase. *Hum. Genet.* **105,** 460–7 (1999).

28. Ramamurthy, V. *et al.* AIPL1, a protein implicated in Leber's congenital amaurosis, interacts with and aids in processing of farnesylated proteins. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 12630–5 (2003).

29. Valdar, W. S. & Thornton, J. M. Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **313,** 399–416 (2001).

30. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490,** 556–60 (2012).

31. Chen, J., Sawyer, N. & Regan, L. Protein-protein interactions: general trends in the relationship between binding affinity and interfacial buried surface area. *Protein Sci.* **22,** 510–5 (2013).

32. Valdar, W. S. & Thornton, J. M. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42,** 108–24 (2001).

33. Tuncbag, N., Kar, G., Keskin, O., Gursoy, A. & Nussinov, R. A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief. Bioinform.* **10,** 217–32 (2009).

34. Moreira, I. S., Fernandes, P. A. & Ramos, M. J. Hot spots--a review of the protein-protein interface determinant amino-acid residues. *Proteins* **68,** 803–12 (2007).

35. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3,** e02030 (2014).

36. Sethi, A. *et al.* Reads meet rotamers: structural biology in the age of deep sequencing. *Curr. Opin. Struct. Biol.* **35,** 125–34 (2015).

37. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35,** 3823–35 (2007).

38. Magliery, T. J. & Regan, L. Beyond consensus: statistical free energies reveal hidden interactions in the design of a TPR motif. *J. Mol. Biol.* **343,** 731–45 (2004).

39. Sawyer, N., Chen, J. & Regan, L. All repeats are not equal: a module-based approach to guide repeat protein design. *J. Mol. Biol.* **425,** 1826–38 (2013).

40. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40,** D302–5 (2012).

41. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–90 (2004).

42. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28,** 2731–9 (2011).

43. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–70 (2010).

44. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

Figure Legends

Figure 1. Our motif-MSA approach amplifies variant information as compared to species-MSA. (a) (1) We first query a database and obtain all the proteins with the desired domains or motifs. We use the TPR motifs as an example in this figure. These motifs have to be the same length. Here, we select TPR motifs that are 34 amino acids since they are the most frequently-occurring size. (2) Subsequently, we perform an 'ungapped' multiple sequence alignment (MSA) of the human TPR motifs by lining them up end to end, to obtain a sequence conservation profile. This motif-based MSA typically exhibits differential sequence conservation among the positions across the length of the motif. (3) The third step involves collecting genomic single nucleotide variants (SNVs) for each amino acid position of the motif-based alignment profile. In TPR domains, we obtain the specific genomic coordinates of each codon (in each motif), and then we locate all variants (black diamonds) that fall into each codon, allowing us to aggregate variants across all motifs within the human genome, thereby amplifying variant information sufficiently for further downstream analyses. (4) For each motif-MSA, we then host the results on our MotifVar database, including residue frequency tables, log(NS/S), log(R/C) and SIFT score distributions. **(b)** For species-MSA, we align orthologous sequences across multiple species. However, because we are focusing on proteins and sequencing data only in the human population, the number of variants at each amino acid position or codon in a species-MSA profile will never exceed a maximum of three.

Figure 2. Motif-MSA can uncover important domain positions missed by species-MSA. This figure uses TPR as an example. **(a)** We perform a species-MSA using orthologous TTC21B from 66 species (species-MSA). Here, we show the alignment profiles for the first three TPR motifs (red, blue and green sequence logos), out of the possible 16. We observe that almost all the positions are highly conserved. **(b)** In contrast to conventional species-MSA, there is a differential sequence conservation profile across the TPR motif-MSA (black sequence logo), which facilitates the identification of more conserved motif positions that are potentially

important (five positions are highlighted in orange). **(c)** We also compare the utility among three variant sets, namely from 1000 Genomes Project Phase 1 (1000GP), the combined set of 1000GP and the Exome Sequencing Project (1000GP+ESP6500) and the ExAC dataset. We can see that there are subtle differences in log(NS/S) for each position along the TPR motif, when using variant datasets from 1000GP to 1000GP+ESP6500. We were able to make meaningful interpretations only when we use variant data from ExAC.

Figure 3. Using genomic variant information in the motif-MSA profile to investigate selective constraints in PPI motifs. Using SNVs from the ExAC dataset, we use various SNV properties to investigate the extent of selective constraints at each position in the motif-MSA profile. **(a)** For each non-synonymous SNV, a score can be computed from the SIFT tool to approximate its deleteriousness phylogenetically, where a lower SIFT score means more deleterious. Each blue violin plot represents the distribution of SIFT scores at each position in the TPR motif, with the width of the plot approximating frequency density and the black dot denoting the median SIFT score. The distribution provides an estimation of the selective constraints based on inter-species comparison. **(b)** For each SNV, the minor allele frequency (MAF) in the human population can determine whether an SNV is rare (MAF ≤ 0.005) or otherwise, common. The log ratio of the number of rare versus common variants (log R/C) represents the enrichment of rare variants, which has been used as a metric for estimating selective constraints based on intra-species comparison. All positions have an enrichment of rare variants, with position 25 having no common variants (log ratio with a zero denominator is undefined). **(c)** We can also calculate the log ratio of non-synonymous versus synonymous SNVs (log NS/S). A depletion of NS variants with respect to the background of S SNVs suggests a position might be functionally significant. **(d)** The five positions with the least median SIFT scores are numbered in blue according to their rank (there are four positions tied at rank 2). The five positions with the lowest log (NS/S) are ranked in red. The top five most conserved positions in the motif-MSA are highlighted in orange. There are eight candidate positions which fulfil at least one of the above criteria of the lowest SIFT median scores, log(NS/S) and motif-MSA sequence conservation, with four positions satisfying at least two. **(e)** Using the X-ray crystal structure of the human HOP TPR1 domain, which consists of three TPR motifs (PDB ID: 1ELW), we can see the 24 residues (8 residues in each of three motifs) in the spatial context and observe that they are mostly buried residues. The ligand-binding convex profile of the TPR1 domain (the cognate ligand is represented by the green stick model) is rotated 180° to reveal the concave profile of the same TPR1 domain. **(f)** We also use two databases, ClinVar (blue) and HGMD (light blue) and the union of the two sets (dark blue), to demonstrate which TPR motif positions accumulates more clinically-relevant and disease-related SNVs.

Supplementary Figure 1. The most frequent size of the TPR motif is 34 amino acids.

Supplementary Table 1. The 1000 Genomes Project (1000GP) provides the least number of autosomal SNVs, followed by an approximate 6-fold increase in number of exomes in the combined set of 1000GP and Exome Sequencing Project (ESP6500); this is a corresponding ~3-fold increase in the number of autosomal SNVs. Our study uses the dataset from ExAC, with 60,706 individuals, an almost 8-fold increase from the combined set of 1000GP+ESP6500; this is a corresponding ~5-fold increase in the number of autosomal SNVs.

Supplementary Table 2. The lists of repeat domains that we performed the motif-MSA approach and are included in the MotifVar repository.