Variant annotation by integrating protein and genomic information using repeat domains

Using repeat domains to integrate protein and genomic information and annotate genomic variants

Using repeat domains to integrate protein and genomic information

Abstract

Large-scale exome sequencing has revealed a preponderance of single nucleotide variants (SNVs) in protein-coding regions of the human genome. Here, we identify SNVs that are important to protein-protein interactions (PPI) by focusing on a class of high impact protein domains that specifically mediate PPI – repeat protein domains (RPDs). Particularly, we develop a strategy to build a multiple sequence alignment profile based on a repeat protein motif. This approach allows a codon-level accumulation of SNVs, which serves as an ideal platform to utilize the copious amount of available variant data from sequencing projects. We use the SNVs to calculate genetic metrics that estimate selective constraints. The combination of protein and genomic information enables the identification of potentially functional positions in RPDs that are under high selective constraints. [redacted] we use the RPD, tetratricopeptide repeat, as an illustrating example, we also provide our results for all RPDs as an online resource, MotifVar (motifvar.gersteinlab.org).

Introduction

The combined efforts from large-scale sequencing projects and clinical sequencing have given rise to an exponentially increasing number of human sequences in recent years [cite ExAC].[1,2] 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 ensuing challenges is then to provide functional annotations for these variants efficiently and accurately [cite].

Much of the variant annotation work in the protein-coding regions have been focused on first identifying amino acid residues that might play important roles in proteins, mainly via protein sequence and structural conservation across species, or a combination of both.[3–5] The key idea is that if a non-synonymous mutation occurs on amino acid residue positions that are highly conserved over a long evolutionary timescale (structurally and/or sequence-wise), it is more likely to be functionally disruptive, especially when it changes an electrochemical property of the amino acid drastically.[6] For example, at a residue position that is a highly conserved alanine across multiple species, a mutation to a larger aromatic tryptophan should be considered more disruptive than a mutation to a glycine, which is more comparable to alanine electrochemically and size-wise [cite an concrete example?]. Sequence conservation has been shown to be one of the most powerful predictors of deleteriousness.[6]

However, protein-coding regions are, in general, under high selection pressure. This is particularly the case in highly functional protein domains, so that they are preserved over a long evolutionary timescale. As such, almost all positions in these high-impact domains tend to be

extremely conserved in their phylogenetic sequences across multiple species, making it hard to pinpoint more important positions. Moreover, since all positions are conserved, we can only determine the general importance of the mutation, without necessarily knowing what specific function it might affect. One way to elucidate functional roles is via 'guilt-by-association', where we can infer the role if we know the function of the protein sub-domain that the mutation resides. Even then though, there is still no clear implication as to how the non-synonymous mutation might more specifically affect the domain structure, or the function. [cite]

Consequently, it becomes very challenging to both identify important residue positions and annotate the explicit function that they perform. This is particularly true in protein-protein interactions (PPI). Many structural and functional studies have shown that there exists only a subset of residues that are important to protein-protein binding, either as 'hotspot' residues that directly participates in the protein interaction,[7] or as structural residues that maintain the stability of protein-binding domains [cite motif/domain papers and disease-causing papers]. To this end, we focus on a category of protein domains that explicitly performs the function of mediating PPI, known as repeat protein domains (RPD).[8,9] RPDs have been found to be present in almost one in every three human protein.[10] As a result, many classes of RPDs have also been studied extensively.[11–13] 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 create protein design templates so that non-template features can be grafted into these protein scaffolds to design synthetic proteins with desired specificities and affinities.[14–16] We adapted the strategy to create a multiple sequence alignment (MSA) profile, which we term a 'motif-MSA' profile, for each class of RPD. Using the tetratricopeptide repeat as an example of a PPI RPD, we demonstrate that the motif-MSA strategy can identify a subset of residues that are important to these PPI domains, such that the occurrence of non-synonymous variants at these particular residue sites have a higher incidence of known clinically-related variants, especially in diseases that are a direct result of an ablation of vital protein-protein interactions.

We further show that the motif-MSA approach enables the accumulation of variants, thereby serving as an ideal platform to combine residue information from PPI domains with the wealth of single nucleotide variant (SNV) information obtained from large-scale exome sequencing. Interestingly, we note that such analyses can only be performed currently using a dataset as large as those from the Exome Aggregation Consortium (ExAC). The motif-MSA approach can complement protein analyses with genetic measures of natural selection. Finally, we illustrate the generalizability of the approach by applying it to repeat and non-repeat PPI domains.

Results

Strategy to relate protein to genomic information

Figure 1 shows our strategy to relate protein residue to genomic information. We first produce 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, Supp 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 genomic variants from the ExAC catalog onto the genomic coordinates of the repeat motifs. Finally, this allows us to obtain aggregate statistics of variants at each residue positions for each class of repeat domain, namely ratio of the number of non-synonymous SNVs to synonymous ratio, ΔDAF, enrichment of rare variants and the distribution of SIFT scores. We provide these data for 34 protein domains – 17 RPDs and 17 non-RPDs – and host the data on a publicly available repository, *MotifVar* (motifvar.gersteinlab.org).

Identifying conserved positions in PPI domains using motif-MSA profile

An MSA is more typically performed tusing homologous sequences from multiple species (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.[cite] Expectantly, most positions are comparably high in sequence conservation (Figure 2a). In contrast, the motif-MSA profile exhibits substantially differential sequence conservation among the motif positions (Figure 2b). Because we are aligning motifs of the same class of repeat domain, residue positions that are characteristic of, and thereby important to, the structural fold of the repeat motif will be more conserved than other positions. We were able to easily identify positions 8, 11, 20, 24 and 27 as more conserved within the TPR repeat motif, thereby facilitating the identification of more conserved residue positions that might be important in the TPR domains.

Mapping genomic variants to the motif-MSA profile

With a large catalog of human exonic variants, we can allocate each variant to a position on the MSA, based on its genomic and translated protein location. The conventional species-MSA profile is restricted to the sequence of a single human protein (since the alignment is based on orthologs), hence only a maximum of three human variants can occur for each residue's codon position (Figure 2c). However, in motif-MSA, for a variant to occur on a particular position of the TPR motif, it can occur on any of the codons that correspond to that position on any 34-amino-acid TPR motif on any TPR-containing proteins within the human proteome. Consequently, the number of variants at each amino acid position in a motif-MSA is not limited by the codon size, but by the number of repeat motifs (from the same class) within the human proteome. This accumulation of variants using motif-MSA enables the computation of various metrics to investigate the selective constraints in the protein domains using genomic information. In Figure 3, we use the TPR domains as an example to show the results of three aggregate statistics derived from the accumulation of genomic variants on the motif-MSA, namely the distribution of SIFT scores (Figure 3a), rare-to-common-SNVs ratio (R/C) (Figure 3b), non-synonymous-to-synonymous-SNVs ratio (NS/S) (Figure 3c), and the distribution of the difference in derived allele frequencies between the different human ethnic populations (ΔDAF) (Supplementary Table XX).

Each individual SIFT score of a non-synonymous SNV is mostly a reflection of its inter-species conservation and the impact of its represented amino acid change in the context of a long

evolutionary timescale, with lower SIFT score denoting a greater likelihood of an SNV being deleterious.[6] The collective SIFT scores, perceived as a distribution at each position in the motif, gives a quantitative perspective on how conserved the position is across multiple species and motifs of the same class. As we have seen from a species-MSA, protein-coding regions are generally under high selective constraints. This is also reflected in the SIFT score distributions, such that almost all positions of highly functional PPI domains tend to have very low median SIFT scores across the motif.[17] Position 20, the most highly conserved position in the TPR motif-MSA, exemplified this observation, by not only exhibiting the lowest median SIFT score, but also a very sharp distribution with a very long tail (Figure 3a and 3d).

To look at selection within the human population, we can compute R/C, with an enrichment of rare variants (or depletion of common variants) signifying high conservation (Figure 3b). The idea is that the accumulation of rare variants (from a single population) reflects a low tolerance for a site to harbor deleterious mutations, a property similar to sites of high functional impact and conservation. This has been recently shown to give a good proxy for intra-species conservation.[17–19] In general, we find that there is a high rare variant load across the motif-MSA profiles, regardless of residue or positional conservation within the repeat motif; a representative example is shown for the TPR domains (Figure 3b). The differences in R/C between positions within the motif are too subtle for identifying important positions or residues. This is strongly reflective of our focus on high impact domains involved in PPI.

We can also 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, especially for inter- and intra-species comparisons of protein-coding regions typically at the gene level, where dN/dS, pN/pS and the McDonald Kreitman's test have been extensively adapted and used [cite dN/dS, pN/pS, McDonald Kreitman]. In estimating selective constraints, NS/S additionally takes into account the functional significance of the gene or genomic region based on the premise that NS amino acid changes would tend to be more disruptive in the function of a protein, resulting in higher selective constraints against NS mutations in functionally-important regions. Thus, a lower NS/S ratio suggests that the gene or genomic region is undergoing more selective constraints and potentially more functionally important. Here, rather than at the gene level, NS/S is 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 the positions with the four most conserved position in the TPR motif-MSA, reinforcing the utility of motif-MSA in picking out functionally important residue positions (Figure 3c).

At this juncture, we also note that this result was observable only with the ExAC data (60,706 exomes)[cite], but not when solely with the 1000 Genomes Project Phase 1 data (1000GP; 1,092 whole genomes)[18] nor its combination with the Exome Sequencing Project (ESP; 6,500 exomes)[17], which total more than 7,500 protein-coding exome data (Supplementary Figure 2 and Supplementary Table 1). The combined dataset of 1000GP and ESP is about 7 folds and 3 folds the size of the 1000GP data in terms of the number of individuals and autosomal SNVs respectively, and the ExAC dataset is about 8 folds and 5 folds the size of the combined set. Even though the rate of increase in sizes in both cases are comparable, the fact that only the

largest dataset with more than 60K exomes and 7M SNVs yields interpretable results underscores the importance of having more genome sequences and rare variants.

Finally, using the motif-MSA, we are able to integrate both protein and genomic information to better pinpoint positions that might be more functionally important. By combining positions with the highest five sequence conservation in the 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).

Mapping genomic information onto protein structures

In order to visualize the eight residue positions in a spatial context, we further integrated genomic information with protein structures. We use the X-ray crystal structure of the 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). We then map the eight positions derived from Figure 3d onto the protein structure (Figure 4). HOP is a scaffold protein with three domains of three TPR motifs that bring together two critical molecular chaperones, Hsp70 and Hsp90. TPR1 contains three 34-amino-acid TPR motifs and mediates the interaction with Hsp70, by binding to the latter's C-terminal peptide sequence of PTIEEVD.[20] Except for position 17 in each of the three TPR motifs in TPR1, we found that all the other seven residue positions with high selective constraints (from either low median SIFT scores, low log (NS/S) or high motif sequence conservation) are buried residues in the PPI domain (Figure 4a). Buried residues are known to be essential in maintaining the stability of globular proteins.[cite] For short and non-globular PPI domains, these buried residues become critical for maintaining the structural folds of the domains in order to perform their roles in mediating protein-protein interactions.[cite]

Relating residues positions to clinically-relevant and disease-related mutation data

Previous studies have shown that buried residues are highly constrained and have a higher tendency to be disease-causing.[cite] Using two databases, ClinVar[21] and the proprietary Human Gene Mutation Database (HGMD)[22], 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 (Figure 4b). 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. For example, on position 8 of the TPR motif, an NS mutation, A128V, on the protein p67phox (neutrophil cytosolic factor 2), which is part of the enzyme complex NADPH oxidase found in neutrophils, disrupts its interaction with a membrane-bound cytochrome and GTPase b558, thereby giving rise to an immunodeficiency disease known as chronic granulomatous disease.[23] Another mutation, A197P, on the seemingly most conserved position 20 along the TPR motif, occurring in the protein AIPL1 (aryl hydrocarbon receptor interacting protein-like 1) results in a loss of interaction with farnesylated proteins, due to a compromise in the stability of the TPR domain and AIPL1, thereby giving rise to retinal dystrophy and blindness. It is perhaps worth mentioning that the protein *per se* is able to fold and only the PPI is lost.[24] On the TPR motif profile, it might also be interesting to note that position 2, even though the most variable position, has a comparable number of disease-related variants as the most conserved positions 8 and 20

and there is a comparatively fair number of disease-related variants at some of the most variable positions in the TPR motif profile (Figure 4b).

Extending strategy to other repeat and non-repeat PPI domains

To demonstrate the generalizability of our approach, we extend the computation of the motif-MSA and selective constraint metrics to 17 repeat and 17 non-repeat PPI domains. Here, we use the Ankyrin domains as an additional example (Supplementary Figure 3). To allow researchers to easily identify positions for RPDs and non-RPDs, we provide the residue frequency tables for their motif-MSA profiles for the most frequent motif size for each domain. Also, we provide the SIFT score distributions, median SIFT scores, log(NS/S), log(R/C) and ΔDAF values for each position along the motif to allow versatile thresholding by the users. We host these data as flat files in a publicly available repository, *MotifVar* (motifvar.gersteinlab.org).

Discussion

For decades, focus in research on PPI has typically been the investigation of protein interfaces that directly take part in the protein interaction. Historically, most studies involved the use of 3D protein structures, for instance, to identify protein-protein interfaces,[25,26] investigate interfacial properties[27,28] or to predict interacting 'hotspots'[29–31] [barry honig PrePPI, 1998 hotspots papers, interface propoerties]. 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.[32] As a result, there is also an increased urgency in the endeavor for variant annotation in protein-coding regions capitalizing on sequence information. Hence, there is great value 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 broke the limitation imposed by species-MSA, by enabling the 'amplification' of variant information for large-scale genomic analyses. 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.[3,4,33] This limits the use of the bulk of the available sequencing data, which is focused on a single species, *Homo sapiens*. 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 (Supplementary Figure 2 and Supplementary Table 1).

Secondly, the ability to gain statistical power from variant aggregation makes motif-MSA an extremely powerful platform in investigating selective constraints in protein-coding regions using genomic information. We only used three metrics derived from the genomic data to identify highly constrained motif positions and residues. It is encouraging to see that many of the top five positions for each metric overlap and show clear evidence of structural importance and disease implications. More importantly, these metrics also uncover complementing sites that show evidence for clinical and disease relevance, which would have been missed otherwise.

Potentially, motif-MSA is amenable to the entire repertoire of genomic metrics such as ΔDAF[34], and F_{ST}[35], which provides a quantitative means for thresholding. In turn, this enables computational tractability and the incorporation of the approach into variant annotation pipelines, which is critical in efficiently triaging variants for more resource-consuming experimental validation amidst high volumes of sequencing data.[19]

Thirdly, motif-MSA is also able to reflect protein structural properties and clarify the roles of the positions or residues 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 with great but unknown significance to the specific proteins. On the other hand, because the protein motifs are classified by their structural folds, sequence features in a motif-MSA are the manifestations of important structural properties that determine the folds of the domains. Highly conserved residues or positions in motif-MSA have the proclivity to be structural residues that maintain the integrity and stability of the protein domains. This is exemplified by the observation that most of the highly conserved positions in motif-MSA correspond to buried residues within the interior of PPI domains, when we visualize the residues in the context of 3D protein structures (Figure 4a) – an observation that is in line with a previous study.[36] Also, when we restrict the protein domains in question to those specifically involved in PPI, we are picking out structural residues or positions that are involved in specific PPIs. This is exemplified in the TPR mutation at position 20 in AIPL1 that causes retinal dystrophy. With the mutation, there is a specific ablation of interaction with farnesylated proteins due to the unfolded TPR domain, but AIPL1 still folds and binds to another cognate protein (that does not interact with the TPR domain on AIPL1).[24] In addition, because motifs in motif-MSA are also derived from a myriad of proteins with diverse binding partners, it has been demonstrated and suggested that positions, which are low in sequence conservation, or 'hypervariable', are found in the binding pockets of the corresponding domains and are thus involved directly in protein-binding.[36,37] We also noticed hypervariable positions, such as position 2 in TPR motifs, also harbor a good number of disease-related variants. This observation bolsters the proposal from the previous study and further hints at the utility of motif-MSA in annotating and clarifying the roles of variants directly involved in PPI.

Fourthly, the motif-MSA strategy presents an opportunity for its application beyond protein motifs, to whole domains (domain-MSA). For example, a domain-MSA for RPDs has been shown to be very informative in uncovering domain-specific protein features that are not observed in a motif-MSA.[38] It will be helpful to extend the strategy presented here to relate genomic information with these domain-specific features. However, a foreseeable challenge is the availability of RPDs of a given type. In a domain-MSA, because each type of RPD is characterized by the sizes of both domain (how many motifs) and motif (how many amino acids), the frequency of a particular RPD for MSA may decrease as the definition of the RPD changes (Supplementary Figure 1). For example, the most common type of TPR domains are those with three TPR motifs, but the number declines as one limits the size of the TPR motifs to 5 TPR motifs, and/or 33 amino acids. A domain-MSA, while limited by numbers, can be extremely useful in uncovering domain-specific features important for PPI.

Talk about generalizability to non-RPD?

At this point, it is also important to note that aggregating variants on the motif-MSA *per se* conflates genomic variant information not only from long and short evolutionary time scales, but also from the evolution of the same class of repeat motifs. Even though we have used SIFT scores as a proxy for inter-species comparison and log(R/C) for intra-species comparison to tease apart contributions from selective constraints over long and short evolutionary timescales respectively, the interpretation of selective constraints in more generic metrics such as log(NS/S) is a confluence of evolutionary timescales and mutation processes. Hence, for such metrics, it is imperative to be aware that they might demonstrate selective constraints in a broader sense, rather than a specific timescale or mutation process.

Much knowledge can be gleaned from the integration of heterogeneous data in a meaningful and computationally-tractable way. The motif-MSA approach provides a powerful and versatile platform to facilitate the melding of protein and genome information. It will add to the existing arsenal of tools in identifying residues that are involved in protein-protein interactions, which has a direct and long-standing significance in the fields of protein engineering and drug design. Perhaps more pertinently, this methodology will serve well, both as a way to leverage the vast amount of sequencing data currently available, and as a complementary perspective in the genomic variant annotation endeavor – an activity that will increase in importance and urgency in the near future, as human genome sequencing becomes more clinical and personal genome interpretation takes center stage.

Methods

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).[39] The 34 PPI domains, repeat and non-repeat, 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.[40]

The TPR species-MSA is obtained by aligning the homologous protein TTC21B from 43 species. (1) First, we perform an online BLASTp search [cite; http://blast.ncbi.nlm.nih.gov/] for the human TTC21B protein sequence obtained from UniProt, with ID Q7Z4L5, using the UniProtKB database, BLOSUM62 matrix, allowing gapped alignment, and a minimum alignment score threshold of 10. (2) We obtain the top 250 sequences based on the alignment scores. (3) Subsequently, we impose a series of filters manually to pare down unwanted and redundant sequences. (3a) We remove all non-TTC21B (based on the gene names provided in the results of BLASTp). (3b) We also eliminated redundant entries or isoforms from the same species, strictly retaining only one form of the protein from each species, based on first whether it is stated as 'characterized' or 'reviewed', then the highest alignment score, followed by the lowest E-value in this order. We find that those isoforms that are 'characterized', 'reviewed' or have the highest alignment scores are typically long and hence, having a less likelihood of obtaining non-functional protein fragments. At this point, the number of sequences is 45. (4) Using the MEGA software [cite], 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, resulting in the final number of 43 homologous sequences with ungapped alignment. (5) Finally, we construct the sequence logo of all 16 TPRs using WebLogo 3.2.[40] We show the alignment of only the first three TPR motifs of TTC21B in Figure 2.

MORE ON PRESENTATION IN DB

Sequence logo visualization

All sequence logos are created by WebLogo 3.2[40], 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 (Version 0.3, downloaded Feb 1 2015)[cite], 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 77) from Ensembl.[41]

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)[18] and Exome

Sequencing Project data (6,500 exomes)[17], to produce Supplementary Figure 2 and Supplementary Table 1.

All coordinates are based on the human reference genome assembly version of GRCh37.

Relating genomic and protein information
Custom scripts are written to relate genomic to protein information. The key idea 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 available upon request.

Protein structure visualization
The X-ray crystal structures from Protein Data Bank (PDB) are created using Pymol 1.3.[42]

Clinically-relevant and disease-related variants
Clinically-relevant and disease-related variants in GRCh37 were downloaded from ClinVar[21] on July 8, 2015 and the proprietary HGMD Professional Database downloaded on July 27, 2015.[22]

Acknowledgements

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. 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).
4. 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).
5. 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).
6. Ng, P. C. & Henikoff, S. Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7,** 61–80 (2006).
7. Clackson, T. & Wells, J. A. A hot spot of binding energy in a hormone-receptor interface. *Science* **267,** 383–6 (1995).
8. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. A census of protein repeats.

J. Mol. Biol. **293,** 151–60 (1999).
9. Kajava, A. V. Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.* **179,** 279–88 (2012).
10. Andrade, M. A., Perez-Iratxeta, C. & Ponting, C. P. Protein repeats: structures, functions, and evolution. *J. Struct. Biol.* **134,** 117–31
11. Li, J., Mahajan, A. & Tsai, M.-D. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry* **45,** 15168–78 (2006).
12. 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).
13. Allan, R. K. & Ratajczak, T. Versatile TPR domains accommodate different modes of target protein recognition and function. *Cell Stress Chaperones* **16,** 353–67 (2011).
14. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta* **1543,** 408–415 (2000).
15. 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).
16. 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).
17. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–9 (2012).
18. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
19. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
20. Schmid, A. B. *et al.* The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop. *EMBO J.* **31,** 1506–17 (2012).
21. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980–5 (2014).
22. 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).
23. 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).
24. 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).
25. Valdar, W. S. & Thornton, J. M. Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **313,** 399–416 (2001).
26. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490,** 556–60 (2012).
27. 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).
28. Valdar, W. S. & Thornton, J. M. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42,** 108–24 (2001).
29. 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).
30. 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).
31. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3,** e02030 (2014).
32. Sethi, A. *et al.* Reads meet rotamers: structural biology in the age of deep sequencing. *Curr. Opin. Struct. Biol.* **35,** 125–34 (2015).
33. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35,** 3823–35 (2007).
34. 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).
35. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat. Rev. Genet.* **10,** 639–50 (2009).
36. 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).
37. Magliery, T. J. & Regan, L. Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics* **6,** 240 (2005).
38. 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).
39. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40,** D302–5 (2012).
40. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–90 (2004).
41. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–70 (2010).
42. The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC.

Figure Legends

Figure 1. Our motif-MSA approach. (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. For example, 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 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. For TPR domains, we obtain the specific genomic coordinates of each codon of every TPR motif in the human proteome, and then we locate all variants that fall into each codon. (4) An SNV can be non-synonymous or synonymous, common or rare in the human population, functionally disruptive or benign (depending on SIFT scores). Based on the nature of the SNVs, we can describe various statistics based on these SNV properties. We can also

visualize the locations of the variants by pinpointing the amino acid positions that reside in the 3D protein structures.

Figure 2. Motif-MSA can uncover important domain positions missed by species-MSA and it also serves as a "variant information amplifier". 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 (positions are highlighted in yellow). (c) In order to integrate the vast amount of sequencing data, we can directly map genomic variants (black diamonds) onto the coordinates of TPR motifs in protein-coding genes. We can use species-MSA to align orthologous sequences across multiple species, as in (a). However, because we are focusing on proteins and sequencing data in humans, the number of variants at each amino acid position or codon in a species-MSA profile will never exceed a maximum of three. On the contrary, a motif-MSA profile is able to aggregate variants across all motifs within the human genome, thereby amplifying variant information sufficiently for further downstream analyses.

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, based on sequence conservation over multiple species, 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 intra-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 (NS) versus synonymous (S) SNVs. 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 TPR motif are highlighted in yellow. There are seven 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.

Figure 4. Mapping genomic information onto protein structures and disease-related mutation data. (a) We choose the TPR domain, TPR1, found on the Hsp-organizing protein

(HOP; PDB ID: 1ELW), as a basis of mapping candidate positions. TPR1 contains three 34-amino-acid TPR motifs (e.g. there are three position 20s). We find that all positions with high selective constraints are found buried within the PPI domains (red residues on protein structure), except for position 17 on each of the TPR motifs. The colors are overlaid in order: positions with lowest median SIFT scores (light blue numbers and residues in structure), with lowest log(NS/S) (red numbers and residues in structure), then finally positions with highest sequence conservation in the motif-MSA profile (orange highlights in motif sequence and residues in structure). The ligand-binding convex profile of the TPR1 domain (the cognate ligand is represented by the green stick sticks) is rotated 180^{o} to reveal the concave profile of the same TPR1 domain. (b) We also use two databases, ClinVar and HGMD, 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 Figure 2. We compare the utility among three variant sets, namely from 1000 Genomes Project Phase 1 (1000GP; green bars), the combined set of 1000GP and the Exome Sequencing Project (1000GP+ESP6500; blue bars) 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 (grey bars).

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 and non-repeat domains that we performed the motif-MSA approach and are included in the *MotifVar* repository.