Intensification: A resource for amplifying population-genetic signals with protein repeats

Jieming Chen[1,2,3], Bo Wang[5], Lynne Regan[1,2,3,5]*, 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.
[5]Department of Chemistry, Yale University, New Haven, CT 06520, USA.
*These authors co-directed the work
#Corresponding author

Abstract
Large-scale genome sequencing holds great promise for the interpretation of protein structures through the discovery of many rare, functional variants in the human population. However, because protein-coding regions are under high selective constraints, these variants occur at low frequencies, such that there is often insufficient statistics for downstream calculations. To address this problem, we develop the Intensification approach, which uses the modular structure of repeat protein domains to amplify signals of selection from population genetics and traditional inter-species conservation. In particular, we are able to aggregate variants at the codon level to identify important positions in repeat domains that show strong conservation signals. This allows us to compare conservation over different evolutionary timescales. It also enables us to visualize population-genetic measures on protein structures. We make available the Intensification results as an online resource (http://intensification.gersteinlab.org) and illustrate the approach 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.[2,8,9] However, all protein-coding regions are, in general, under high selection pressure. As such, almost all positions in

Deleted: [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.

Deleted: [8–10]

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.

There is a wide range of repeat protein domains (RPDs).[10,11] Each RPD is made up of modular repeat motifs of the same class. This modularity gives rise to a strategy for a particular class of RPDs that was first introduced in the field of protein engineering to generate protein design templates to create synthetic proteins with desired specificities and affinities.[12–14] We adapted the strategy to build a multiple sequence alignment (MSA) profile, which we term a 'motif-MSA' profile, for each class of RPD. As an initial proof-of-concept for our novel approach, we focus on this category of RPDs that has been shown to be amenable to the motif-MSA approach. This category of RPDs explicitly mediates protein-protein interactions (PPI), and their repeat motifs in each RPD require each other to maintain their structural fold. Each repeat unit is also relatively short with length of 12-100 amino acids. Many of these classes of RPDs have been studied extensively.[15–17] For example, tetratricopeptide repeat (TPR) domains are made up of only TPR motifs and Ankyrin repeat (ANK) domains of ANK repeat motifs. 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 the Exome Aggregation Consortium (ExAC) database.[1] Our Intensification database contains our results as a resource for annotating variants in 12 PPI RPDs (see 'Methods' for selection criteria).

Deleted: [11,12] Each RPD is made up of modular repeat motifs of the same class. This modularity gives rise to a strategy for a particular class of RPDs that was first introduced in the field of protein engineering to generate protein design templates to create synthetic proteins with desired specificities and affinities.[13–15] We adapted the strategy to build a multiple sequence alignment (MSA) profile, which we term a 'motif-MSA' profile, for each class of RPD. As an initial proof-of-concept for our novel approach, we focus on this category of RPDs that has been shown to be amenable to the motif-MSA approach. This category of RPDs explicitly mediates protein-protein interactions (PPI), and their repeat motifs in each RPD require each other to maintain their structural fold. Each repeat unit is also relatively short with length of 12-100 amino acids. Many of these classes of RPDs have been studied extensively.[16–18] For example, tetratricopeptide repeat (TPR) domains are made up of only TPR motifs and Ankyrin repeat (ANK) domains of ANK repeat motifs. 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.

Results

Intensification database

Figure 1a shows our strategy that is used to build up the resources in our publicly available Intensification database (http://intensification.gersteinlab.org) that relates protein residue to genomic information in 12 RPDs, which encompass 5,508 motifs and 971 proteins in *Homo sapiens* (Supplementary Table 1). Our strategy first produces a motif sequence alignment profile for a class of repeat domain. We obtain every repeat motif of a given amino acid length in the human proteome (typically the length with the most number of available motifs). We then perform an MSA of all the 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 Intensification database. Here, we

use the 34-amino-acid TPR repeat motif as an example (see 'Methods' for details; Figure 1 and Supplementary Figure 1).

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.[18] 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). These observations are highly reproducible across all 12 RPD classes in our database (Supplementary Table 2). The results in Supplementary Table 2 show that, indeed, for all 12 RPD classes, there are higher proportions of sites in species-MSA that are highly conserved (>1.5 bits) as compared to those in motif-MSA. For 11 (out of 12) RPD classes, >80% of sites in species-MSA have high relative entropy (>1.5 bits); for 9 RPD classes, we observe that >70% of sites in species-MSA have relative entropy >2.0 bits. On the contrary, there are no RPD classes in motif-MSA that have at least 80% of sites with relative entropy >1.5 bits. For example, within the TPR repeat motif, there were only six positions with relative entropy > 1 bit and two positions with relative entropy > 1.5; we were able to easily identify positions 8, 11, 20, 24 and 27 as the top five most conserved positions.

Computing population genetic metrics and amplification by motif-MSA

When we focus only on the human species, variant positions in a conventional species-MSA profile are 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 three variant positions can occur for each codon (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 571 34-amino-acid TPR motifs within the human genome. This accumulation of variants amplifies the signal, thereby facilitating the computation of various population genetic metrics to investigate selective constraints in the protein domains. At this juncture, we note that our results were most apparent with the largest ExAC dataset (60,706 exomes) (Supplementary Table 1). At evidently conserved positions such as position 8, 20 and 24, log(NS/S) and motif conservation are reasonable proxies of each other. This is consistent across all three datasets. However, in the smallest dataset of the 1000 Genomes Project Phase 1 data (1000GP; 1,092 whole genomes), we observe at least 10 other positions across the motif-MSA that have similar logNS/S profiles (near-zero or negative), making interpretations using just this dataset difficult. The number of positions with low logNS/S decreases as the number of exomes increases by 6,500 with the Exome Sequencing Project (ESP6500). Finally with ExAC, we are able to more firmly identify the positions in which both the logNS/S and motif conservation profiles agree, where positions with the lowest logNS/S profiles correspond to positions of high sequence conservation in the motif. This further underscores the fact that more genomes are indeed necessary to yield better statistics for such analyses.

Deleted: [19] 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). These observations are highly reproducible across all 12 RPD classes in our database (Supplementary Table 2). The results in Supplementary Table 2 show that, indeed, for all 12 RPD classes, there are higher proportions of sites in species-MSA that are highly conserved (>1.5 bits) as compared to those in motif-MSA. Also for 11 RPD classes, >80% of sites have high relative entropy (>1.5 bits). Within the TPR repeat motif, we were able to easily identify positions 8, 11, 20, 24 and 27 as more conserved

Deleted: less negative;

Deleted: positive

Deleted: high –

Deleted: to

Deleted: least negative

We use four evolutionary measures derived from the accumulation of genomic variants on the motif-MSAs of all 12 RPD classes. We use the TPR domains as an example to illustrate them.

* **SIFT** – For inter-species conservation, we use the SIFT score of a non-synonymous SNV, which is 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 mean SIFT scores across the motif. In the TPR motif-MSA, the most highly conserved position 20 exemplified this observation (Figure 3a).

* **R/C** – As a proxy for intra-species conservation within the human population, we compute a population genetic measure used in the 1000 Genomes Project, 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.[2,8,9] 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 (Supplementary Figure 2).

* **NS/S** – We further compute the NS/S for each position in the motif-MSA profile (Figure 3b). 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.[19] Here, rather than at the gene level, the accumulation of variants enables NS/S to be calculated at the codon level (Figure 3b). 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 over the human population, four of them correspond to 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 3b).

* **ΔDAF (pop)** – The difference of derived (population) allele frequencies, or ΔDAF, has been used in the 1000 Genomes Project to quantify population differentiation and positive selection and identify highly differentiated sites between pairs of populations.[2,20] Because the majority of the variants are rare even within sub-populations, we observe that most SNVs have low ΔDAF < 0.5. More interestingly, when we constrained to only the non-synonymous variants, we were able to identify some TPR residue positions that seem to harbor more non-synonymous variants that are highly differentiated between populations than other positions (Figure 3f). High differentiation can be indicative of positive selection and adaptive evolution among the human populations.

Combining protein and genomic information to identify important residues

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. In order to demonstrate the utility of the resource, we combine positions with the highest five sequence conservation in the TPR motif-MSA and the lowest five mean SIFT scores and NS/S ratio. Collectively, the four metrics complement each other, and 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 3c). We also note that in TPR, the differences in R/C

Deleted: [10]

Field Code Changed

Deleted: [20]

Deleted: [8,21] Because the majority of the variants are rare even within sub-populations, we observe that most SNVs have low ΔDAF < 0.5.

Formatted: Font color: Red

between positions within the TPR motif-MSA are too subtle to be used. We further analyze the Spearman correlation between the conservation profile of the positions in motif-MSA (relative entropy) and each of the four metrics, for all the 12 RPD motifs (Supplementary Table 3). The varying correlations of each metric with motif conservation indicate that there are non-overlapping information content in each metric, in relation to the conservation profile of the motif-MSA. This suggests that the metrics can be useful in identifying important positions that cannot be picked out by using just motif conservation (motif-MSA) alone.

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).[21] We map the eight positions derived from Figure 3c onto all three motifs of the protein structure, identifying 24 residues in total (Figure 3d). In each TPR motif, except for position 17, we find that all the other seven residue positions with high selective constraints – from either low mean SIFT scores, low log (NS/S) or high motif sequence conservation – are buried residues in the PPI domain (Figure 3d), in line with a previous study.[22]

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

We further validate our findings using two databases, ClinVar[23] and the proprietary Human Gene Mutation Database (HGMD)[24]. We found that the highly constrained positions have some of the most occurrences of clinically-relevant or disease-related mutations (hereafter referred to as 'disease SNVs') along the TPR motif-MSA profile. This is generally observed across all the positions in 7 RPD classes that have at least 1 disease SNV on each motif position (Supplementary Figure 3). 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.[25,26] However, the highest numbers of disease mutation do not necessarily always occur at positions with high sequence conservation in the motif-MSA profile. For example, in TPR, the highest numbers of disease mutation occur at two positions, positions 6 and 7, which will not be detected if only motif-MSA or inter-species conservation was used (Figure 3e). In fact, modest correlations are observed between the number of disease mutations and the conservation profile of motif positions in 7 RPD classes (Supplementary Table 3). This highlights the need to integrate multiple layers of information to better identify important positions.

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,[27,28] investigate interfacial properties[29,30] or to predict interacting 'hotspots'[31–33]. 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.[34] This huge trove of sequence information should be leveraged upon for variant annotation in protein-coding regions,

Deleted: [22] We map the eight positions derived from Figure 3c onto all three motifs of the protein structure, identifying 24 residues in total (Figure 3d).

Field Code Changed

Deleted: [23]

Deleted: [24] and the proprietary Human Gene Mutation Database (HGMD)[25]. We found that the highly constrained positions have some of the most occurrences of clinically-relevant or disease-related mutations (hereafter referred to as 'disease SNVs') along the TPR motif-MSA profile. This is generally observed across all the positions in 7 RPD classes that have at least 1 disease SNV on each motif position (Supplementary Figure 3). 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.[26,27] However, the highest numbers of disease mutation do not necessarily always occur at positions with high sequence conservation in the motif-MSA profile.

Deleted: ¶
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,[28,29] investigate interfacial properties[30,31] or to predict interacting 'hotspots'[32–34]. 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.[35] This huge trove of sequence information should be leveraged upon for variant annotation in protein-coding regions, especially in the interpretation of protein data. Our introduction of the motif-MSA facilitates genomic analyses with protein information (and vice versa) in several ways.¶
¶
Firstly, motif-MSA extends the utility of protein sequences, which has been largely focused on species-MSA, that is, the more traditional perspective of sequence conservation across multiple species based on homology.[5,6,36] Beyond mere sequences, the motifs in motif-MSA are also used for their genomic coordinate system to identify the corresponding genomic positions of the variants within the motif sequences. This is then coupled with the repeat nature of the repeat motifs in RPDs to integrate heterogeneous layers of variation information. In our analyses, two levels of 'variations' are being integrated – (1) amino acid variations stemming from the motif sequences from the human reference genome (motif-MSA), and (2) genetic polymorphisms found in the collection of individuals representing the human population (accumulated variants). At this juncture, it might also be important to mention that the two levels of variations occur as a result of different evolutionary timescales and mutational processes. In motif-MSA, amino acid variation observed by comparing motifs within the human reference genome can happen on short- or long evolutionary timescales, due to duplication, functional divergence and c...

especially in the interpretation of protein data. Our introduction of the motif-MSA facilitates genomic analyses with protein information (and vice versa) in several ways.

Firstly, motif-MSA extends the utility of protein sequences, which has been largely focused on species-MSA, that is, the more traditional perspective of sequence conservation across multiple species based on homology.[5,6,35] Beyond mere sequences, the motifs in motif-MSA are also used for their genomic coordinate system to identify the corresponding genomic positions of the variants within the motif sequences. This is then coupled with the repeat nature of the repeat motifs in RPDs to integrate heterogeneous layers of variation information. In our analyses, there are two distinct levels of 'variations' being integrated – (1) amino acid variations stemming from the motif sequences from the human reference genome (motif-MSA), and (2) genetic polymorphisms found in the collection of individuals representing the human population (accumulated variants). At this juncture, it might also be important to mention that the two levels of variations occur as a result of different evolutionary timescales and mutational processes. In motif-MSA, amino acid variation observed by comparing motifs within the human reference genome is a result of a longer evolutionary time than the genomic variants observed within multiple genomes in a human population. The former is often discussed in the context of phylogeny and can occur before or after speciation events, due to duplication, functional divergence (e.g. functions of motifs inside the same protein might not be the same) and co-evolution (e.g. motifs are not mutated independently). When observed in the context of a single species, co-evolutionary signals of protein motif sequences are therefore comparatively more stable. On the other hand, genomic variation within the human population happens on a shorter evolutionary timescale, since it mostly occurs within the genetic history of a single species. They are a consequence of different sets of mutational and evolutionary processes that act on the individual (such as recombination, and DNA damage), and the human population (such as linkage disequilibrium, natural selection, and random drift). Thus, we can describe these genetic variants at the population level. By separately making use of the coordinate system of repeat protein motifs in motif-MSA, we can reasonably accumulate the genomic variants found in a population of human individuals and amplify their associated population-genetic signals, such as population allele frequencies, or the nature of the mutation.

Secondly, motif-MSA is 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, when we align motifs that are structurally similar but functionally divergent in motif-MSA, we are essentially 'averaging' out evolutionary signals (presented as amino acid sequence variations), such that functionally diverse positions have high sequence entropy while positions that show high conservation across motifs of the same RPD class are important structural features that determine the folds of these PPI domains. These features are observed as buried residues within the interior of PPI domains (Figure 3d). 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.[22,36] We noticed few hypervariable positions harbor a large number of disease-related variants, for example, position 2 in TPR motifs, which has been identified by the ΔDAF analysis.

Deleted: sequence features

Deleted: a

Deleted: the

Deleted: [23,37]

Hence, while we cannot definitively identify interface residues that participate in protein interactions, motif-MSA does still hold potential in facilitating such an endeavor in the future.

Lastly, the ability to gain statistical power from variant aggregation makes motif-MSA an extremely powerful platform in investigating evolutionary constraints using genomic information. We only used four metrics as main examples to demonstrate how motif positions and residues that show evidence for clinical and disease relevance can be identified beyond the use of the more conventional species conservation (Figure 3). Our motif-MSA approach is amenable to the entire repertoire of SNV-associated genomic metrics.

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

Intensification database

Our publicly available Intensification database (http://intensification.gersteinlab.org) provides data files for 12 RPDs, namely ankyrins (ANK), annexins (ANX), armadillos (ARM), cadherin repeats (CA), fibronectin type 2 domains (FN2), fibronectin type 3 domains (FN3), leucine-rich repeats (LRR_TYP), spectrin repeats (SPEC), tetratricopeptide repeats (TPR), ubiquitin-interacting motifs (UIM), WD40 repeats (WD40), and WW domains (WW). The 12 RPDs were semi-manually curated from the domains found in the SMART database for species, *Homo sapiens* (downloaded Oct 25, 2013),[37] and selected for those that are known to mediate protein-protein interactions. We also filter out classes of RPDs that have less than 20 unique repeat motifs in the human genome as annotated by SMART database, to remove classes of RPDs that do not have sufficient statistics for analyses. The results for each class of RPD is a tarball, which contains residue frequency tables (to rebuild the sequence logo), the SIFT score distributions, mean SIFT scores, log(NS/S), log(R/C), ΔDAF and relative entropy values for each position along each RPD motif to allow versatile thresholding by the users. The database also provide links to the scripts used in the pipeline on Github, and the resource is freely downloadable as flat files.

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).[37] The 12 PPI repeat domains are manually selected based on their availability in the SMART database, repeat nature and involvement in PPI.

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

Deleted: [38]

Deleted: resource and

Deleted: are

Deleted: at the database

Deleted: ¶
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).[38] The 12 PPI repeat domains are manually selected based on their availability in the SMART database, repeat nature and involvement in PPI.¶

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.[38]

All species-MSA (Supplementary Table 2) are obtained by aligning a single set of homologous protein sequences for each class of RPD (e.g. TPR-containing TTC21B orthologs from 43 species) using UniProt.[39] Using the MEGA5 software[40], we extract the relevant domain from the alignment. For example, based on the 45-ortholog alignment of TTC21B, we extracted all 16 TPR motifs in TTC21B found in the SMART database. Finally, we construct the sequence logo of 16 TPRs in TTC21B using WebLogo 3.2.[38] We show the alignment of only the first three TPR motifs in Figure 2a.

In order to compare the percentage of positions that are highly conserved in motif- and species-MSA. We have defined two and three thresholds arbitrarily as metrics of increasing sequence conservation, based on the relative entropy at each position, for motif-MSA and species-MSA respectively, namely: 1 and 1.5 bits of information for motif-MSA, and 1, 1.5 and 2 bits for species-MSA. We then count the number and percentage of residues that exceeded these thresholds for each MSA.

Sequence logo visualization

All sequence logos are created by WebLogo 3.2[38], 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.

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 February 1, 2015), after removing the variants from the sex and mitochondrial chromosomes and singletons (those variants that only occur in one chromosome in the entire ExAC dataset). This ends up with

Deleted: [39]

Field Code Changed

Deleted: ¶
All species-MSA (Supplementary Table 2) are obtained by aligning a single set of homologous protein sequences for each class of RPD (e.g. TPR-containing TTC21B orthologs from 43 species) using UniProt.[40] Using the MEGA5 software[41], we extract the relevant domain from the alignment. For example, based on the 45-ortholog alignment of TTC21B, we extracted all 16 TPR motifs in TTC21B found in the SMART database. Finally, we construct the sequence logo of 16 TPRs in TTC21B using WebLogo 3.2.[39] We show the alignment of only the first three TPR motifs in Figure 2a. ¶

Deleted: [39], using the following parameters:

Deleted: [3]

7,202,445 exonic, 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.[41] DAF information is derived from 1000 Genomes Phase 1 SNVs with ancestral alleles and ExAC population frequencies. ΔDAF (pop) values are computed from the pairwise differences in population DAF for each SNV among five populations, namely Africans/African Americans (AFR), Latino (AMR), East Asians (EAS), South Asians (SAS) and Europeans (EUR), which are combined from the Finnish (FIN) and non-Finnish European (NFE) populations in ExAC; we excluded SNVs in the 'others' (OTH) category. Also, we note that not all SNVs have known ancestral alleles in the 1000 Genomes Project, thus DAFs are not be known for all SNVs. For Figure 3f, only non-synonymous SNVs are used, and in addition, any SNVs with < 100 subjects in a population that is being compared is removed, because they will skew ΔDAF.

To produce Figure 2c and Supplementary Table 1, we have used a combined number of 1,328,447 unique, non-singleton, exonic and autosomal SNVs from the 1000 Genomes Project Phase 1 (1,092 whole genomes)[2] and Exome Sequencing Project data (6,500 exomes)[9].

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 on Github.

Correlations between disease SNVs, population-genetic metrics and motif conservation profiles (from motif-MSA) are computed using Spearman correlation. For computing the Spearman correlation, a mean ΔDAF is also calculated at each position of the motif-MSA in each RPD class.

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[23] on July 8, 2015 and the proprietary HGMD Professional Database downloaded on July 27, 2015.[24] In ClinVar, we only extracted those variants that are annotated to be "Pathogenic". For Figure 2e, we performed analyses on the ClinVar and HGMD variants, separately and their union, with the genomic codon positions corresponding to the residue positions in the motifs.

For disease SNV analyses, only disease SNVs from 7 RPD classes are used, since they have at least 1 SNV on each position along its motif-MSA profile. In order to examine the correlation between the number of disease SNVs and the relative entropy of the motif profile, we use the

Deleted: [42]

Deleted: [8]

Deleted: [10]

Field Code Changed

Deleted: in the Intensification resource

Field Code Changed

Deleted: [43]

Deleted: [24]

Deleted: [25] In ClinVar, we only extracted those variants that are annotated to be "Pathogenic".

Spearman correlation. In order to investigate the enrichment of disease SNVs in 'conserved' sites, we first define conserved sites to be those with ≥1 or ≥ 1.5 bits of sequence relative entropy, and then we compare the distributions of number of disease SNVs in these two categories, using the Wilcoxon-Mann-Whitney U test.

Acknowledgements

We would like to thank Mr. Everett Sussman and Mr. Rahim Hashim for background work related to this project. We acknowledge support from the Raymond and Beverly Sackler Institute for Biological, Physical and Engineering Sciences, NIH and from the A.L. Williams Professorship funds. This work was also supported in part by the Yale University Faculty of Arts and Sciences High Performance Computing Center.

Conflicts of interest statement

The authors declare that there is no conflict of interest.

References

1. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–91 (2016).
2. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
3. Muddyman, D., Smee, C., Griffin, H. & Kaye, J. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med.* **5,** 100 (2013).
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. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342,** 1235587 (2013).
9. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–9 (2012).
10. Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. A census of protein repeats. *J. Mol. Biol.* **293,** 151–60 (1999).
11. Kajava, A. V. Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.* **179,** 279–88 (2012).
12. Lehmann, M., Pasamontes, L., Lassen, S. F. & Wyss, M. The consensus concept for thermostability engineering of proteins. *Biochim. Biophys. Acta* **1543,** 408–415 (2000).
13. 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).

Deleted: 1000 Genomes Project Consortium

Deleted: A global reference for human

Deleted: .

Deleted: **526,** 68–74 (2015

Moved (insertion) [1]

Deleted: 2

Deleted: 3. Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *BioRxiv* (2015). doi:10.1101/030338¶

Moved up [1]: 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).¶

Deleted: 9. .

Deleted: 10

Deleted: 11

Deleted: 12

Deleted: 13

Deleted: 14

14. 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).
15. Li, J., Mahajan, A. & Tsai, M.-D. Ankyrin repeat: a unique motif mediating protein-protein interactions. *Biochemistry* **45,** 15168–78 (2006).
16. 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).
17. Allan, R. K. & Ratajczak, T. Versatile TPR domains accommodate different modes of target protein recognition and function. *Cell Stress Chaperones* **16,** 353–67 (2011).
18. 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).
19. Fay, J. C. Weighing the evidence for adaptation at the molecular level. *Trends Genet.* **27,** 343–9 (2011).
20. 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).
21. Schmid, A. B. *et al.* The architecture of functional modules in the Hsp90 co-chaperone Sti1/Hop. *EMBO J.* **31,** 1506–17 (2012).
22. Magliery, T. J. & Regan, L. Sequence variation in ligand binding sites in proteins. *BMC Bioinformatics* **6,** 240 (2005).
23. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980-5 (2014).
24. 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).
25. 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).
26. 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).
27. Valdar, W. S. & Thornton, J. M. Conservation helps to identify biologically relevant crystal contacts. *J. Mol. Biol.* **313,** 399–416 (2001).
28. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490,** 556–60 (2012).
29. 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).
30. Valdar, W. S. & Thornton, J. M. Protein-protein interfaces: analysis of amino acid conservation in homodimers. *Proteins* **42,** 108–24 (2001).
31. 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).
32. 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).
33. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3,**

Deleted: 15
Deleted: 16
Deleted: 17
Deleted: 18
Deleted: 19
Deleted: 20
Deleted: 21
Deleted: 22
Deleted: 23
Deleted: 24
Deleted: 25
Deleted: 26
Deleted: 27
Deleted: 28
Deleted: 29
Deleted: 30
Deleted: 31
Deleted: 32
Deleted: 33
Deleted: 34

e02030 (2014).
34. Sethi, A. *et al.* Reads meet rotamers: structural biology in the age of deep sequencing. *Curr. Opin. Struct. Biol.* **35,** 125–34 (2015).
35. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35,** 3823–35 (2007).
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. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40,** D302-5 (2012).
38. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–90 (2004).
39. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204-12 (2015).
40. 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).
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.

Deleted: 35
Deleted: 36
Deleted: 37
Deleted: 38
Deleted: 39
Deleted: 40
Deleted: 41
Deleted: 42
Deleted: 43

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 (black sequence logo) 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 over 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 Intensification database. For each protein repeat domains, we build a motif-MSA, and compute corresponding SNV profiles, 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 species, only three variant positions can occur at each codon in a species-MSA profile. We illustrate this with the human protein, TTC21B, which contains TPR motifs.

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)** In order to show the utility of motif-MSA and amplification, we compare the results for the log(NS/S) 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 only subtle differences in log(NS/S) for each position along the TPR motif when using variant datasets from 1000GP to 1000GP+ESP6500. We were only able to make meaningful interpretations only when we use variant data from ExAC.

Deleted: negative

Deleted: -

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, where a lower SIFT score means the SNV may be 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 showing frequency density and the black dot denoting the mean SIFT score. The distribution provides an estimation of selective constraints based on inter-species comparison. **(b)** 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. **(c)** The five positions with the least mean SIFT scores are numbered in blue according to their rank. 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 mean scores, log(NS/S) and motif-MSA sequence conservation, with four positions satisfying at least two. **(d)** Using the X-ray crystal structure of the human HOP TPR1 domain (PDB ID: 1ELW), it consists of three TPR motifs as shown as cartoon ribbons in the inset, colored separately in shades of grey (white represents capping helix). 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. Residues 6 and 7 are identified by SIFT scores (blue in (a)). Residues 8, 11, 20, 24 and 27 are identified motif conservation (orange). Residue 17 is identified by log(NS/S). The ligand-binding convex profile of the TPR1 domain (the cognate ligand is represented by the green stick model) is rotated 180^{o} to reveal the concave profile of the same TPR1 domain. **(e)** We also use two databases, ClinVar (dark purple) and HGMD (light purple) and the union of the two sets (purple), to demonstrate which TPR motif positions accumulates more clinically-relevant and disease-related SNVs. We use the same color scheme to number the residue numbers identified in (d). **(f)** The collective ΔDAF values for non-synonymous SNVs found for the TPR motif-MSA show that some motif positions, such as positions 2, 17 and 32, contain more SNVs that are highly differentiated among populations than other positions. We calculated the ΔDAF for pairwise comparisons of five populations, namely individuals with African ancestry (AFR), as well as, Latino (AMR), East Asian (EAS), European (EUR) and South Asian (SAS) ancestries. The DAFs are compared between pairs of populations.

Deleted: negative

Deleted: (there are four positions tied at rank 2).

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

Supplementary Figure 2. 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/depletion 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).

Supplementary Figure 3. In order to examine the enrichment of disease SNVs in highly conserved sites, we compare whether the distribution of the number of disease SNVs and whether the mean number is higher in 'conserved' (C) positions than non-conserved (N) ones. 'Conservation' is defined by having ≥ 1 (left panel) and 1.5 (right panel) relative entropy of sequence information on the motif-MSA. Only 7 RPD classes were used, as they have at least 1 SNV on each site of the motif. In general, we do observe higher numbers of disease SNVs in conserved positions. But owing to relatively low number of conserved positions in motif-MSA and disease variants, most of these comparisons are not statistically significant, except for WD40 and the entire collection of positions and their corresponding number of disease SNVs in the 7 RPDs (statistical significant categories are marked by '*').

Deleted: .

Supplementary Figure 4. A general introduction of the Intensification website, which is mainly divided into three sections: 'Query', 'Download', and 'Documentation'. The 'Query' page provides three options to explore the database. Users can choose to input a genomic region or the position of a single SNV, choose from our list of 12 RPD classes, or input a PDB ID, which contains at least a domain from one of the 12 RPD classes in our database. The query results is a list of SNVs found in the motifs in our database, accompanied by the motif-MSA sequence logo and SNV information, including SIFT and PolyPhen2 scores, and ExAC alternate allele frequency. The 'Download' page provides all the data files for users to download. We also provide scripts associated with the pipeline on Github. Details on how to use the resource can be found on the 'Documentation' page.

Supplementary Table 1. The lists of 12 repeat domains that we performed the motif-MSA approach and are included in the Intensification repository, with corresponding number of motifs (regardless of size) and proteins in the human proteome. We also included the number of SNVs in each of the 3 datasets: the ExAC database, the Exome Sequencing Project (ESP6500), and the 1000 Genomes Project (1000GP). 1000GP provides the least number of exonic, autosomal SNVs for each RPD, followed by an approximate 2- to 3-fold increase in the combined set of 1000GP and Exome Sequencing Project (ESP6500); this is a corresponding ~6-fold increase in the number of exomes. Our study uses the dataset from ExAC, this is a corresponding ~3- to 5-fold increase in the number of exonic, autosomal SNVs from the combined set; with 60,706 individuals, this is an almost 8-fold increase in exomes from the combined set of 1000GP+ESP6500.

Supplementary Table 2. This tabulates the comparison between the conservation in each MSA, across 12 RPD classes, based on relative entropy. For species-MSA, we arbitrarily choose a protein that contains the RPD, and for motif-MSA, we select the motif size that is the most

common within the human reference genome. We have defined two and three thresholds arbitrarily as metrics of increasing sequence conservation, based on the relative entropy at each position, for motif-MSA and species-MSA respectively, namely: 1 and 1.5 bits of information for motif-MSA, and 1, 1.5 and 2 bits for species-MSA. We then count the number and percentage of residues that exceeded these thresholds for each MSA. We observe that motif-MSA is more discerning at identifying conserved sites, with species-MSA showing >80% of its sites having relatively high conservation (>1.5 bits of information) for 11 out of 12 RPD classes. For all 12 RPD classes, there are higher proportions of sites in species-MSA that are highly conserved (>1.5 bits of information) than motif-MSA.

Supplementary Table 3. This table shows the Spearman's correlation between conservation in motif-MSA and the 4 population-genetic metrics. We also provided the number of residues in the most common motif length and the union number of disease SNVs from ClinVar and HGMD. There is a strong negative correlation between the mean SIFT scores with the relative entropy at each position in the RPD class for all 12 motif classes. Log R/C does not show correlation at all in all 12 RPD classes.