RESPONSE TO REVIEWERS FOR "INTENSIFICATION: A RESOURCE FOR AMPLIFYING POPULATION-GENETIC SIGNALS WITH PROTEIN REPEATS"

RESPONSE LETTER

Overall comment

We want to thank the reviewers for endorsing our manuscript for publication, recognizing the novelty and importance of our resource and study, and offering insightful comments. We have majorly revised the manuscript to address their concerns. In particular, we have made the web resource more accessible to the less technical users and included more analyses of the motif-MSAs of the 12 RPDs, to make the manuscript more informative and complete. Additionally, in order to better portray the idea of variant amplification, we have also changed the name of the resource from "MotifVar" to "Intensification".

The specific reviewers' comments are further addressed below.

Reviewer #1

-- Ref1.1 – Endorsement for publication --

Reviewer Comment	This MS shows a new way of increasing the variant statistics for a specific type of protein structure called repeat protein domain. While recommend its publication,
Author Response	We thank the reviewer for acknowledging the novelty of our study, recommending it for publication, and for his/her thorough examination of our manuscript

-- Ref1.2 – Variations in motif-MSA and species-MSA --

Reviewer Comment	I have a fundamental question regarding the justification of obtaining variations from motif-MSA. The usual species-MSA has an underlying assumption is that one species' variations are independent of other species' variations and the aligned proteins perform the same function, whereas in this MS, the repeated motifs are not necessarily mutated independently and their functions inside the same protein might not be exactly the same (thus requires a slight variation).
Author Response	We thank the reviewer for the comment. We would first like to clarify that for the purpose of aggregating variants in the human population, the motif-MSA was only used as a coordinate system to identify the positions of the variants in the motif sequences; the variations observed in the motifs do not represent or reflect the

Formatted: Font color: Auto

Deleted: We thank the reviewer for the comment. Indeed, the variants occurring in a species-MSA are based on the interrogation of consensus protein sequences from multiple, independent species over a long evolutionary time-scale (orthologs). On the other hand, the variants occurring in a motif-MSA are based on a shorter evolutionary time-scale, by observing the polymorphisms of multiple individuals within the protein sequence of a *single* species, in our case, the human population. There are two categories of 'variations' – (1) variations stemming from the functionally distinct repeat motif sequences in the human reference genome, and (2) genetic polymorphisms found in the collection of individuals representing the human population. Since they are found in a single species, the repeated motif sequences in the same protein within the single species would have been stably conserved across the individuals. Consequently, most polymorphisms, including those that might co-occur in certain individuals within a population, would be at very low frequencies, driven mainly by negative selection and/or random drift; or polymorphic, driven mainly by adaptive and/or balancing selection. Since our main aim is to identify important sites that may or may not be independent, we can analyze, for each motif position, the distributions of frequencies of aggregated SNVs in the human population. For example, our ΔDAF analysis was meant to identify sites that have an accumulation of highly polymorphic variants between human sub-populations, and the rare-to-common-ratio (R/C) analysis was meant to calculate the enrichment of rare variants relative to common ones in the human population. Thus, even though some variants might be co-dependent or co-evolving in two or more aligned motifs, they can still be used by motif-MSA to improve population-genetic statistics and signal-to-noise ratio, to identify important sites within the motif, which may or may not be independent.¶
¶
Perhaps we were not clear in our discussion.

variants found in the human population. There are two levels of ‘variations’ here – (1) amino acid sequence variation stemming from the motif sequences from the human reference genome (motif-MSA), and (2) genetic polymorphisms found in the collection of multiple genomes from the human population (accumulated variants). In other words, the variants in the human population are distinct from the amino acid sequence variation observed in the motifs of the motif-MSA, which is constructed from the human reference genome, and are accumulated by matching their genomic coordinates to the corresponding genomic positions of the codons that represent each amino acid in the motif used in the motif-MSA.

++

We can observe non-independent mutations at two levels of variations – co-evolution of amino acid mutations in protein sequences, and linkage disequilibrium in genomic (variants or polymorphisms). Co-evolution of protein sequences within the human reference genome is a result of a longer evolutionary time than the linkage observed between genomic variants 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.

TEXT ?

On the other hand, linkage of 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 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 variants found in a population of human individuals (not the motif sequences from the human reference genome) and amplify their population-genetic signals, such as population allele frequencies, or the nature of the mutation. Thus, even if the mutations are non-independent, we can still broadly identify potentially important positions on the motifs, since these positions will have boosted signal-to-noise ratios.

	Further, we are precisely utilizing the fact that the motifs potentially do not have the same function. 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 class define the structural folds of the same RPD class. This is indeed different from the species-MSA, where functional and structural positions are both conserved. Ultimately, the motif-MSA approach uses the genomic coordinate system of the motifs to integrate two levels of variation, and various associated information, in order to help us broadly identify important positions in these motifs. We have modified the text to better clarify this.
Excerpt From Revised Manuscript	Please refer to the 'Discussion' section. *"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 co-evolution, before or after speciation. When we aggregate motifs that are structurally similar but functionally dissimilar in motif-MSA, we are averaging out these evolutionary signals (presented as amino acid sequence variations), such that positions that show high conservation across motifs are structural, and functional positions become potentially diverse. This is also different from the species-MSA, where functional and structural positions are both conserved. On the other hand, genetic variation, or polymorphisms, within the human population happen on a shorter evolutionary timescale, allowing us to examine these variants at the population level. They are a consequence of different sets of mutational and evolutionary processes that act on the individual (such as recombination, and DNA damage), and at the population level (such as natural selection, and random drift). Hence, in addition to protein sequences, we can also make use of population-genetic characteristics of these genetic polymorphisms (such as population allele frequencies, or the nature of the mutation) in our motif-MSA approach."*

Deleted: *"While there is independence for each aligned orthologous sequence, the functional similarity of the sequences gives rise to widespread conservation across the species-MSA. On the other hand, in motif-MSA, while aligned motif sequences can be co-dependent because they come from the same protein, the functional dissimilarity and structural similarity give rise to differential conservation across the MSA. Moreover, we can systematically aggregate variants from similar protein regions within the genome of a single species in a reasonable manner to identify important sites, regardless of whether the sites are independent of one another. 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 7,600 exomes, was not sufficient to yield immediately-interpretable results (Figure 2c and Supplementary Table 1). At this point, it is also important to note that the motif-MSA contains two categories of 'variations', namely variations found in the repeat motif sequences of the human reference genome and genomic variant information from a representative human population. Motif sequence variation can stem from the duplication and divergence of the same class of repeat motifs within the genome, and can be of long and short evolutionary timescales (before and after speciation). In contrast, the genomic variant catalogue corresponds to the possible polymorphisms found in the human population, representing a shorter evolutionary timescale of a single species. Thus, the biological interpretation of selective constraints in metrics such as log(NS/S) is a confluence of evolutionary timescales and mutation processes."*

-- Ref1.3 – Clarification for repeat protein domains --

Reviewer Comment	The authors claim there is one RPD in every three human proteins. What is the reason their data only covers < 1000

	`proteins and what are the qualitative criteria in their manual selection of data?`
Author Response	We agree with the reviewer that we were not sufficiently clear in our description. The one-in-three statistic was derived from a previous publication by Pellegrini *et al.* [1], which included a wide range of classes of repeat protein domains (RPDs), such as the highly degenerate homopolymeric repeat proteins like polyglutamine, and RPDs with repeat structures so large that they can fold independently like titin [2]. In this work, we want to demonstrate an initial proof-of-concept of this novel amplification approach. Thus, we have specifically chosen a category of RPDs in which the motif-MSA has previously been successfully implemented [3], and also in which there is an additional advantage in visualization, with relatively manageable lengths for each repeat unit of about 12-100 amino acids. The approach can be further developed and expanded in subsequent work, to include more challenging RPDs such as homopolymers and even non-repeat protein domains, such as short linear protein motifs. We have removed the statement to prevent confusion, and clarified our selection criteria in the manuscript. [1] Pellegrini M. *et al*. (1999). *Proteins*, 35(4):440-6 [2] Kajava A. (2012). *J Struct Biol.*, 179(3):279-88 [3] Main *et al*. (2003). *Curr Opin Struct Biol*., 13(4):482-9
Excerpt From Revised Manuscript	Please refer to the ‘Introduction’ and ‘Methods’ sections respectively. *“There is a wide range of repeat protein domains (RPDs).[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.[17–19] We adapted the strategy to build a multiple sequence alignment (MSA) profile, which we term a ‘motif-MSA’ profile, for each class of RPD. 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.”* *“The 12 RPDs were semi-manually curated from the domains found in the SMART database for species, Homo sapiens (downloaded Oct 25, 2013),[40] 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 (Supplementary Table 1).”*

Deleted: We

Deleted: chose

Deleted: on

Deleted:]. These classes of RPDs mediate protein-protein interactions,

Deleted: the repeat units in each RPD require one another to maintain their structural fold. Each repeat unit is

Deleted: short with length of

Deleted: 60

Deleted: section

Deleted: *a*

Deleted: *in which the*

Deleted: *60*

Formatted: Normal

Deleted: *and*

Deleted: *at least*

-- Ref1.4 – SIFT --

Reviewer Comment	SIFT as well as many other annotation approaches has very high false positive rate (SIFT has ~ 40% false positive rate), it might be better using approaches such as FATHMM, ENTPRISE methods that have much lower false positive rate.
Author Response	We thank the reviewer for the suggestion. We have previously included our results from other approaches, specifically Condel and PolyPhen2, which are already available for download in our online data resource. We have chosen SIFT for analyses in the manuscript because it is one of the most well-known tools and its score is known to be derived from species conservation. We would like to re-iterate that SIFT is meant to be an example, not a fixture, in the motif-MSA approach. In fact, all the population-genetic metrics shown in this study are meant to be examples. The motif-MSA approach integrates variant information, so any other similar variant metrics or approaches can definitely be implemented with motif-MSA. We have edited the text to make this point clearer in the manuscript.
Excerpt From Revised Manuscript	Please refer to the 'Discussion' section. *"Potentially, motif-MSA is amenable to the entire repertoire of genomic metrics. We used four metrics as 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)."*

Deleted: of using

Deleted: annotation

Deleted: .

Deleted: is not meant

Deleted: rather an example, to demonstrate variant aggregation in

Deleted: Other

Deleted: made

-- Ref1.5 – Interface residues --

Reviewer Comment	Can the authors also show the interface residues participating protein-protein interactions?
Author Response	We thank the reviewer for this question. While it would be interesting to show the interface residues involved in protein-protein interactions, our emphasis is not in the explicit identification of these residues. It has been shown previously that many hypervariable sites in motif-MSA are associated with peptide or protein binding, due to the diversity of binding partners associated with all the motifs in a motif-MSA [1]. However, hypervariable sites can be confounded by unimportant sites that can better accommodate random mutations. Hence, while motif-MSA do hold potential for identifying these positions, more in-depth exploration and analyses are outside the scope of this manuscript. We have modified part of the 'Discussion' section to better illustrate this.

Deleted: fact that

Deleted: bind to different partners

Deleted: Hence, in this study, we have used several layers of population genetic information to complement the identification of potentially important sites, including among hypervariable sites. Unfortunately, the combination of population genetic information and motif-MSA does not seem to identify hypervariable positions very well, even though the most hypervariable site of position 2 was picked out by the ΔDAF analysis. Thus, while we cannot definitively inform the reader of interface residues participating in protein-protein interactions, the motif-MSA still holds potential for identifying these positions.

	[1] Magliery T. and Regan L. (2005). *BMC Bioinformatics*, 6:240.
Excerpt From Revised Manuscript	Please refer to ‘Discussion’ section. *“In addition, it has been suggested that because motifs in motif-MSA are from a myriad of proteins with diverse binding partners, positions that are low in sequence conservation, or ‘hypervariable’, are found in the binding pockets of the corresponding domains.*[24,38] *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. 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.”*

Formatted: Superscript

Deleted: *?*

Reviewer #2

-- Ref2.1 – Positive comment --

Reviewer Comment	This manuscript presents a very interesting idea to generate multiple alignments of protein motifs (particularly those involved in Protein-protein interactions) to identify positions that are conserved within the motifs that may not be identified from using full length sequences, with the aim of identifying positions where variants are likely to be associated with disease. Overall the research is well thought out and an elegant idea for considering the effect of variants present in motifs. However, I have a number of comments for the authors to address.
Author Response	We thank the reviewer for the thorough examination of our manuscript. We have provided additional analyses and updated the website to address the reviewer's comments.

-- Ref2.2 – High level quantification --

Formatted: Font color: Red

Formatted: Font color: Red

Reviewer Comment	My main concern is that the authors present results solely for a single example. There is a lack of quantification. Users of this resource, may be interested in variants in particular regions of a motif and to have an idea of how strong a correlation there is between the conservation observed in the motif and associated with disease. Quantification of the following form should be included:
Author Response	We agree with the reviewer that it would be useful to provide high-level quantifications of all the 12 motifs. We have included new results and analyses for all 12 motifs. For users to get a better sense of the resource, we have included new Supplementary Tables 1, 2 and 3 to give a more extensive overview of the motif-MSA characteristics across all 12 motifs, including the correlation of conservation and disease-associated sites in motif-MSA. We will address the individual points in detail in the next few sections.

Deleted: Table xxx now shows an

Deleted: of the motif-MSA

Deleted: the conserved

Deleted: ¶
¶
At this point, we would also like to further emphasize that motif-MSA is a good platform to both (1) visualize conserved positions that seem to be more structurally important, and (2) amplify population genetic signals by the accumulation of variants, so that they may be used to help identify, more generally, important positions on the repeat motif. Hence, the approach is not limited to only detecting only conserved sites, but also (hyper)variable sites, which can be potentially important.

Deleted: Excerpt From¶
Revised Manuscript

...

-- Ref2.3 – Conservation in motif-MSA vs species-MSA –

Reviewer Comment	It is proposed that the motif-MSAs are better at revealing conservation that species-MSA (example shown in Figure 2). For example the authors could consider over all of the motifs how many positions are highly conserved in motif-MSAs compared to species-MSAs.
Author Response	We agree with the reviewer's suggestion. Hence, in order to show that motif-MSAs are better at revealing conservation than species-MSA, we have performed an additional analysis in Supplementary Table 2 to compare the percentage of positions that are highly conserved in motif- and species-MSA. We have defined two and

Deleted: We have defined a threshold for conservation and computed the number of positions that are highly conserved in motif-MSA versus species-MSA for all 12 RPDs and added this information to Supplementary Table xxx.

	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. In order to perform similar analyses for species-MSA, we arbitrarily choose 12 human proteins, one within each class of RPDs. Each protein is then aligned to at least 20 other orthologs to produce a species-MSA. The results 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), as summarized in Supplementary Table 2. We have included texts in the manuscript to describe these new analyses.
Excerpt From Revised Manuscript	Please refer to the new Supplementary Table 2, the ‘Results’ section, under ‘Comparing species- and motif-MSA’, and the ‘Methods’ section. *“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).”* *“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.”*

Deleted: ’.¶
¶
“ ”

2

-- Ref2.4 – Correlation analyses for population-genetic metrics --

Deleted:

Reviewer Comment	The authors then consider four population genetic metrics and show data referring to a single motif. The authors should present a rigorous analysis of these metrics with their motif-MSAs compared to show how useful this resource is.
Author Response	We agree with the reviewer’s suggestion and have presented a rigorous analysis of the population-genetic metrics in relation to the sequences conservation for all the 12 motif-MSAs in Supplementary Table 3. Specifically, we have performed correlation analyses for each of the four population-genetic metrics

Deleted: have performed correlation analyses

Deleted: with

Deleted: sequence

Deleted: (

Deleted: xxx). In order to show the utility of the resource

Deleted: also used the results in

	with the relative entropy (conservation) of each of the 12 RPD motif-MSA. In Figure 2, we already show how the four metrics from the resource can be used to complement each other and collectively identify potentially important positions. The newly-added correlation analyses further show the differing correlations with motif-MSA conservation, for the four metrics. This suggests that there are some non-overlapping information content in each metric, in relation to the conservation profile of motif-MSA, and can be used to further identify important positions that may not be picked out by motif-MSA alone. Such secondary analyses of the resource further demonstrate its utility.
Excerpt From Revised Manuscript	Please refer to the new Supplementary Table 3, the 'Results' section, under 'Combining protein and genomic information to identify important residues', and the 'Methods' section. *"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 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."* *"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."*

Deleted: database to identify

Deleted: across the 12 motif-MSAs using a similar approach implemented on the TPRs.

Deleted: '¶
¶
"

-- Ref2.5 – ExAC dataset --

Reviewer Comment	The authors state that only the ExAC dataset is sufficient to yield useful data and refer to figure 2C. this should be expanded across all of the 12 motifs in the resource. Additionally the information shown in Figure 2c is not clearly presented, The figure legends states " 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". This needs to be clarified -

Formatted: Font color: Auto

	looking at the figure there seems to be greater variation for the smaller datasets.
Author Response	We agree with the reviewer that the description was unclear. We have modified it to better convey what we mean. We have also included the number of SNVs in the three datasets (1000GP, 1000GP+ESP6500 and ExAC) for all 12 RPD motifs in Supplementary Table 1. We have reworded the text, and replotted the figure as a negative logarithm to provide a better visual representation. This comparison was meant to show that the ExAC variant catalog exhibits more consistent and interpretable signals than the smaller datasets. In Figure 2, in the smaller datasets, while highly conserved positions in motif-MSA have consistently high log ratios, most other positions also have very high or positive log ratios in the motif, making interpretations difficult. For example, in 1000GP, there are at least 10 positions exhibiting positive log ratios. This is because there are smaller numbers of SNVs at each position, easily giving rise to smaller fractions of NS/S and skewing the log ratios of the 1000G and 1000GP+ESP6500 datasets negative. With the ExAC database, and an almost four-fold increase in the number of SNVs in TPRs, there are is a greater signal-to-noise ratio (SNR) as log ratios in the other positions become more robust and less skewed. For example, at position 2, where the log ratio changes from negative to consistently positive in the larger 2 datasets; there are only 7 SNVs (NS/S=2/5) in 1000GP dataset, 32 SNVs (NS/S=19/13) in 1000GP+ESP6500, but 122 SNVs (NS/S=82/40) in ExAC. At position 20, where the log ratio changes to positive only at the largest ExAC dataset: there are only 7 SNVs (NS/S=0/7, a pseudo-count of 0.01 was given to calculate log ratio, hence the very tall bar) in 1000GP set, 13 SNVs (NS/S=2/11) in 1000G+ESP6500 set and 86 SNVs (NS/S=44/42) in ExAC. To illustrate how a general increase in numbers can enhance SNR, we draw analogy from the notion of shot noise, or Poisson noise [1]. We can reasonably model the discrete events of genomic variant occurrence as a Poisson process. SNR thus follows the expression $N/\sqrt{N}$ (where N refers to number of events, or variants in this case, and shot noise is denoted by $\sqrt{N}$). In smaller datasets, with low N, the numerator grows much faster than the denominator, making SNR highly susceptible to fluctuations in N. As the dataset gets larger, the Poisson distribution approaches normality (law of large numbers), the numerator and the denominator are growing comparably, making SNR more robust and stable to relative fluctuations in N. In Supplementary Table 1, we can observe that

D|Sz

2.

Deleted: The

Deleted: made the log(NS/S) ratio

Deleted: *apparent*. This is because, owing to smaller numbers of SNVs in 1000G and 1000G+ESP6500 datasets, the log ratios of the

Deleted: are largely skewed by a large denominator, leading greater variation. Consequently, in these

Deleted: low

Deleted: low

Deleted: negative

Deleted: However, with

Deleted: is less

Deleted: less skewed. As a result, the signals become more apparent and interpretable, with only the conserved positions being prominently lower or negative than the rest of the positions. We have modified the description to better convey what we mean. To make such comparisons, we have also added the numbers of SNVs in all three datasets for all 12 motifs in the Supplementary material (

Deleted: yy).

	the ExAC dataset is consistently the largest dataset at least a 2-fold increase from 1000GP or 1000GP+ESP6500 datasets. [1] Schottky, W. (1918). *Ann. Phys* 57: 541-567.
Excerpt From Revised Manuscript	Please refer to Supplementary Table 1, and the 'Results' section, under 'Computing population genetic metrics and amplification by motif-MSA'. *"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 (less negative; near-zero or positive), making interpretations using this dataset difficult. The number of positions with high –logNS/S decreases as the number of exomes increases to 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 least negative 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."*

Formatted: Font color: Custom Color(RGB(34,34,34)), Pattern: Clear (White)

Deleted: .¶
¶
"

Formatted: Normal

Formatted: Font: 10 pt, Italic

-- Ref2.6 – Clinically-relevant mutations in conserved sites --

Reviewer Comment	The authors also consider clinically relevant and disease-related mutations. Again this should be quantified - are the highly conserved motif-MSA positions enriched in such variants? How does this compare with the species-MSA?
Author Response	We thank the reviewer for the comment and have provided an analysis to examine the correlation of disease SNVs with the relative entropy of the motif in motif-MSA. In addition, we have also checked whether there is an enrichment of disease SNVs in conserved motif-MSA positions. However, because most sites in species-MSA do not have sufficient variants, it is not meaningful to conduct such analyses for species-MSA. For example, there are 155 SNVs in TTC21B, which is about 0.28 SNV per site, and 296 SNVs in ANK1, about 0.47 per site. We have defined a threshold to define a 'conserved' position (a relative entropy of 1 and 1.5 bits) and use a Wilcoxon-Mann-Whitney U test to compare the distributions of disease-related mutations between sites that are conserved and non-conserved (Supplementary Figure 3). Also, we have included a new Supplementary Table 3, with the number of disease SNVs (union of ClinVar and HGMD SNVs) and the Spearman correlation between the relative entropy (conservation) and the number of disease SNVs.

Deleted: for conservation

Deleted: mean number of clinically-relevant and

Deleted: xx). Because most sites in species-MSA are highly conserved, it is not amenable to such an analysis

Formatted: Font color: Red

Excerpt From Revised Manuscript	Please refer to the new Supplementary Figure and Table 3, the 'Results' section, under 'Relating residue positions to clinically-relevant and disease-related mutation data', and the 'Methods' section. *"We further validate our findings using two databases, ClinVar24 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. 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."* *"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 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."*

-- Ref2.7 – Web resource --

Reviewer Comment	Additionally this manuscript has been submitted to a specific biological resource issue of the journal. Reviewing the associated website limited information is available and data is purely available as download of data files for each of the repeats considered. This means that the resource will largely only be used by computational biologists performing analysis or developing methods. While this is useful is makes the resource of limited to use to other non specialists who may be interested in investigating a small set or a particular variant that they have identified in a study.
Author Response	We thank the reviewer for the comment and have revamped our web resource to include a query page for the non-specialists, who may be interested in specific variants or motifs. Now, the query page includes an interactive web interface, with three query options, namely: the user can input and submit (1) one or a range of genomic position(s), (2) choose from the 12 motifs available, and (3) a PDB ID. The results include the sequence logo of the motif-MSA, and a list of SNV(s) with the corresponding SNV information such as SIFT score, ExAC population allele frequency, and the positions the SNVs reside on the motif sequence. Additionally, the

Deleted: .¶
¶
“”

Deleted: have taken this

Deleted: to heart

Deleted: revamp

Deleted: tool

Deleted: , proteins

Deleted: The

Deleted: now allows

Deleted: to

Deleted: an SNV

Deleted: or PDB ID or Ensembl Protein ID, or

Formatted: Font color: Auto

Deleted: to view our

Formatted: Font color: Auto

Deleted: . This will indeed accommodate

SJC2

	user also have the flexibility to download a more complete set of data as flat files, a separate tab-delimited file for the list of SNVs and a PDF for the sequence logo. We believe that in this way, a wider audience will be accommodated, and the usability of the web resource can be increased.
Excerpt From Revised Manuscript	Please refer to the website at http://intensification.gersteinlab.org/.

Deleted: increase

Deleted: .¶
“”

Formatted: Font: Not Italic

-- Ref2.8 – Figure 1b --

Reviewer Comment	Figure 1b is missing.
Author Response	We have made the label and boundary for Figure 1b more evident.
Excerpt From Revised Manuscript	Please refer to Figure 1b.

-- Ref2.9 – names of the 12 RPDs --

Reviewer Comment	It would be useful if the 12 PPI RPDs were listed at least once in the manuscript.
Author Response	We have included the names of the RPDs in the revised manuscript.
Excerpt From Revised Manuscript	Please refer to the ‘Methods’ section under ‘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).”*

Reviewer #3

-- Ref3.1 – Endorsement for publication --

Reviewer Comment	The authors are doing a great job to increase the ability of using large scale genome sequencing data to analyze intra-species population-genetic signals without experimentally increasing the pool of sequenced individuals. Their method can overcome the difficulties of the extremely conservations in high-impact protein domains and the sparsely locations of variants, by selecting and combining useful information together and extracting meaningful signals. I think the article is valuable and suitable for Journal of Molecular Biology after revition.
Author Response	We thank the reviewer for the endorsement for publication and the thorough examination of the manuscript.

-- Ref3.2 – Increasing the number of proteins --

Reviewer Comment	The MotifVar database encompass 971 proteins in human genome. However, we know that the total human proteome is more than 20,000 proteins. The authors should include more proteins in the analysis to give more universal information and conclusions. Please provide more information and discussion regarding extension of the number of proteins and motifs of the database and generate more concrete results. For example, the newly published SRMatlas database is providing more than 99.7% human protein sequence information.
Author Response	We agree with the reviewer that we were not sufficiently clear in our description. Repeat protein domains (RPDs) can indeed be found in more proteins that include degenerate homopolymeric repeat proteins like polyglutamine, and RPDs with repeat structures so large that they can fold independently like titin [1]. In this work, we want to demonstrate an initial proof-of-concept of this novel approach. Thus, we have specifically chosen a category of RPDs in which the motif-MSA has previously been successfully implemented [2], and also that the length of each repeat unit is relatively manageable with 12-100 amino acids for visualization. We have also the following additional criteria: (1) has at least 20 unique motifs in the human genome, and (2) motifs with the most frequently occurring length. Consequently, we are only restricted to only 971 proteins for our analyses. The approach can be further developed and expanded in later work, to include more challenging RPDs such as homopolymers and even non-repeat protein domains, such as short linear protein motifs. However, this is not within the scope of this study. We have clarified our selection criteria in the manuscript. [1] Kajava A. (2012). *J Struct Biol.*, 179(3):279-88

	[2] Main *et al.* (2003). *Curr Opin Struct Biol.*, 13(4):482-9
Excerpt From Revised Manuscript	Please refer to the 'Introduction' and 'Methods' sections respectively. *"There is a wide range of repeat protein domains (RPDs).[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.[17–19] We adapted the strategy to build a multiple sequence alignment (MSA) profile, which we term a 'motif-MSA' profile, for each class of RPD. 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."* *"The 12 RPDs were semi-manually curated from the domains found in the SMART database for species, Homo sapiens (downloaded Oct 25, 2013),[40] 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 (Supplementary Table 1)."*

-- Ref3.3 – Biological meaning of the differences between inter- and intraspecies MSA --

Reviewer Comment	In Figure 2, the authors compared sequence motif conservations between species-MSA and motif-MSA. We can see clearly that the results are different, and we do believe it is important and holds significant biological mechanism. Please provide some further discussion on the biological meaning of the differences between inter-species and intra-species MSA.
Author Response	We thank the reviewer for his/her comment. We have provided more discussion on the potential biological meaning of the differences between inter- and intra-species MSA. We have also further discussed the different evolutionary timescales and mutational processes that the species- and motif-MSA operate on. We further included descriptions about the different levels of variations that are being integrated in motif-MSA, namely variation from motif sequences and variation information from aggregating genetic polymorphisms in the human population. We have added more text to bolster the 'Discussion' section about these.
Excerpt From Revised Manuscript	Please refer to the 'Discussion' section. *"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*

Deleted: Author¶
Response

Deleted: We thank the reviewer for his/her suggestion on increasing the annotation of proteins in the human proteome. Currently, our resource is meant for identifying important motif positions and annotating variants corresponding to protein positions in 12 classes of RPDs. Motif-MSA is also more appropriately constructed by considering only a single gene product per gene. Hence, while we are limited by the proteins we used, we can definitely annotate any gene product positions (both transcripts and proteins included) that can be back-transcribed or back-translated to their corresponding genomic positions found in our study.

Formatted: new-text

Deleted: Compare AS calls with existing studies

Formatted: Font color: Auto

Deleted: a short discussion

Deleted: considered

Deleted: Perhaps we were not clear in our discussion.

Deleted: this

Deleted: *"While there is independence for each aligned orthologous sequence, the functional similarity of the sequences gives rise to widespread conservation across the species-MSA. On the other hand, in motif-MSA, while aligned motif sequences can be co-dependent because they come from the same protein, the functional dissimilarity and structural similarity give rise to differential conservation across the MSA. Moreover, we can systematically aggregate variants from similar protein regions within the genome of a single species in a reasonable manner to identify important sites, regardless of whether the sites are independent of one another. 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 7,600 exomes, was not sufficient to yield immediately-interpretable results (Figure 2c and Supplementary Table 1). At this point, it is also important to note that the motif-MSA contains two categories of 'variations', namely variations found in the repeat motif sequences of the human reference genome and genomic variant information from a representative human population. Motif sequence variation can stem from the duplication and divergence of the same class of repeat motifs within the genome, and can be of long and short evolutionary timescales (before and after speciation). In contrast, the genomic variant catalogue corresponds to the possible polymorphisms found in the human population, representing a shorter evolutionary timescale of a single species. Thus, the biological interpretation of selective constraints in metrics such as log(NS/S) is a confluence of evolutionary timescales and mutation processes."*

	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 co-evolution, before or after speciation. When we aggregate motifs that are structurally similar but functionally dissimilar in motif-MSA, we are averaging out these evolutionary signals (presented as amino acid sequence variations), such that positions that show high conservation across motifs are structural, and functional positions become potentially diverse. This is also different from the species-MSA, where functional and structural positions are both conserved. On the other hand, genetic variation, or polymorphisms, within the human population happen on a shorter evolutionary timescale, allowing us to examine these variants at the population level. They are a consequence of different sets of mutational and evolutionary processes that act on the individual (such as recombination, and DNA damage), and at the population level (such as natural selection, and random drift). Hence, in addition to protein sequences, we can also make use of population-genetic characteristics of these genetic polymorphisms (such as population allele frequencies, or the nature of the mutation) in our motif-MSA approach."

-- Ref3.4 – Correlation analyses for motif-MSA conservation --

Reviewer Comment	The author could do some statistical analysis about the correlation between the occurrences of clinically-relevant and disease-related mutations and the highest sequence conservation motif-MSA combined with lowest median SIFT scores and NS/S ratio, to point out their significant correlated with each other. This will make their conclusion more statistical meaningful.
Author Response	We agree with the reviewer's suggestion and have performed a series of correlation analyses of the population-genetic metrics and disease-related SNVs with the sequence conservation for all 12 motif-MSAs in the revised manuscript and summarized the results in a new Supplementary Table 3. At this point, we would also like to further emphasize that motif-MSA is a platform to both (1) visualize conserved positions that seem to be more structurally important, and (2) amplify population genetic signals by the accumulation of variants, so that they may be used to help identify, more generally, important positions on the repeat motif. Hence, the approach is not limited to only detecting conserved sites.
Excerpt From Revised Manuscript	Please refer to the new Supplementary Table 2, the 'Results' section, under 'Comparing species- and motif-MSA', and the 'Methods' section. *"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*

Deleted: We

Deleted: (Supplementary Table xxx)

Deleted: xxx

Deleted: good

Deleted: only

Deleted: , but also (hyper)variable sites, which can be potentially important

Deleted: .

Deleted: ""

	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).” *“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.”*

Formatted: Font: 10 pt

-- Ref3.5 – Sentence structure --

Reviewer Comment	The authors need to improve their English writing in the article. For example, "The fact that only the largest dataset with more than 60K exomes and 7M SNVs yields interpretable results underscores the importance of amplification and still having more genome sequences." in the first paragraph of page 6 is not correct.
Author Response	We have modified this sentence to better clarify way we mean.
Excerpt From Revised Manuscript	Please refer to ‘Results’ section under ‘Computing population genetic metrics and amplification by motif-MSA’. *“This further underscores the value of amplification, and exemplifies the fact that more genomes are necessary to yield better statistics for such analyses.”*

-- Ref3.6 – Ambiguous parentheses --

Reviewer Comment	There are several ambiguous parentheses in the text, i.e. the first pair in "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)." in line 41 page 7. The author would better use more words to explain whether there were more variants, or more non-synonymous variants, or both.
Author Response	We have altered this sentence to better clarify what we mean.
Excerpt From Revised Manuscript	Please refer to ‘Results’ section under ‘Computing population genetic metrics and amplification by motif-MSA’ and ‘ΔDAF (pop)’. *“More interestingly, we were able to identify some TPR residue positions that seem to harbor more 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.”*