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. 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 ∆DAF 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. We have modified the text to better clarify this.
Excerpt From Revised Manuscript	Please refer to the ‘Discussion’ section. *“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]. We specifically chose a category of RPDs on which motif-MSA has previously been successfully implemented [3]. These classes of RPDs mediate protein-protein interactions, and the repeat units in each RPD require one another to maintain their structural fold. Each repeat unit is also relatively short with length of 12-60 amino acids. 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' section. *"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. We focus on a category of RPDs that explicitly mediates protein-protein interactions (PPI), and in which the repeat motifs in each RPD require each other to maintain their structural fold. Each repeat unit is also relatively short with length of 12-60 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 and have at least 20 unique repeat motifs in the human genome as annotated by SMART database (Supplementary Table 1)."*

-- 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 of using other annotation approaches. SIFT is not meant to be a fixture, rather an example, to demonstrate variant aggregation in motif-MSA. In fact, all the population-genetic metrics shown in this study are meant to be examples. Other similar variant approaches can definitely be implemented with motif-MSA. We have made this 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)."*

-- 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. It has been shown previously that many hypervariable sites in motif-MSA are associated with peptide or protein binding, due to the fact that the motifs in motif-MSA bind to different partners [1]. However, hypervariable sites can be confounded by unimportant sites that can better accommodate random mutations. 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 ∆DAF 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. We have modified part of the 'Discussion' section to better illustrate this. [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 ?DAF 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.”

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 --

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, Supplementary Table xxx now shows an overview of the characteristics of the motif-MSA across 12 motifs, including the correlation of the conserved and disease-associated sites in motif-MSA. We will address the individual points in detail in the next few sections. 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.
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 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.
Excerpt From Revised Manuscript	’. “”

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

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 have performed correlation analyses of the population-genetic metrics with the sequence conservation for all 12 motif-MSAs (Supplementary Table xxx). In order to show the utility of the resource, we have also used the results in the database to identify important positions across the 12 motif-MSAs using a similar approach implemented on the TPRs.
Excerpt From Revised Manuscript	’ ”

-- 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 - 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. The comparison was meant to show that the ExAC variant catalog made the log(NS/S) ratio more *apparent*. This is because, owing to smaller numbers of SNVs in 1000G and 1000G+ESP6500 datasets, the log ratios of the smaller datasets are largely skewed

	by a large denominator, leading greater variation. Consequently, in these smaller datasets, while highly conserved positions in motif-MSA have consistently low log ratios, most other positions also have very low or negative log ratios in the motif, making interpretations difficult. However, with the ExAC database, and an almost four-fold increase in the number of SNVs in TPRs, there is less noise as log ratios in the other positions become 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 (Supplementary Table yy).
Excerpt From Revised Manuscript	. ”

-- 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 have defined a threshold for conservation and use a Mann-Whitney test to compare the mean number of clinically-relevant and disease-related mutations between sites that are conserved and non-conserved (Supplementary Figure xx). Because most sites in species-MSA are highly conserved, it is not amenable to such an analysis.
Excerpt From Revised Manuscript	. “ ”

-- 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 have taken this comment to heart and revamp our web resource to include a query tool for the non-specialists, who may

	be interested in specific variants, proteins or motifs. The query page now allows the user to input an SNV position or PDB ID or Ensembl Protein ID, or choose from the 12 motifs available to view our results. This will indeed accommodate a wider audience, and increase the usability of the web resource.
Excerpt From Revised Manuscript	. “ ”

-- 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 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.

-- Ref3.3 – Compare AS calls with existing studies --

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 discussed the different timescales that the species- and motif-MSA operate on. We further included a short discussion about the different levels of variations that are being considered in motif-MSA, namely variation from motif sequences and variation information from aggregating genetic polymorphisms in the human population. Perhaps we were not clear in our discussion. We have added more text to bolster the 'Discussion' section about this.
Excerpt From Revised Manuscript	Please refer to the 'Discussion' section. *"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."*

-- 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 have performed a series of correlation analyses of the population-genetic metrics with the sequence conservation for all 12 motif-MSAs (Supplementary Table xxx) in the revised manuscript and summarized the results in a new Table xxx. 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.
Excerpt From Revised Manuscript	Please refer to the ‘Results’ section. “ ”

-- 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.”*