Genome Biology and Evolution Advance Access published June 4, 2015 doi:10.1093/gbe/evv105

Article (Discoveries)

Polymorphism Analysis Reveals Reduced Negative Selection and Elevated Rate of Insertions and Deletions in Intrinsically Disordered Protein Regions

Tahsin Khan[1], Gavin M Douglas[2], Priyenbhai Patel[1], Alex N Nguyen Ba[1], Alan M Moses[1,2,3]

[1]Department of Cell & Systems Biology, University of Toronto, Toronto, Canada

[2]Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Canada

[3]Center for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Canada

Corresponding Author: Alan M Moses

Email Address: alan.moses@utoronto.ca

Abstract

Intrinsically disordered protein regions are abundant in eukaryotic proteins and lack stable tertiary structures and enzymatic functions. Previous studies of disordered region evolution based on interspecific alignments have revealed increased propensity for indels and rapid rates of amino acid substitution. How disordered regions are maintained at high abundance in the proteome and across taxa, despite apparently weak evolutionary constraints remains unclear. Here, we use single-nucleotide and indel polymorphism data in yeast and human populations to survey the population variation within disordered regions. First, we show that single-nucleotide polymorphisms in disordered regions are under weaker negative selection compared to more structured protein regions and have a higher proportion of neutral sites. We also confirm previous findings that non-frameshifting indels are much more abundant in disordered regions relative to structured regions. We find that the rate of non-frameshifting indel polymorphism in intrinsically disordered regions resembles that of non-coding DNA and pseudogenes, and that large indels segregate in disordered regions in the human population. Our survey of polymorphism confirms patterns of evolution in disordered regions inferred based on longer evolutionary comparisons.

Keywords: SNP, indel

Introduction

It is widely accepted that the three-dimensional structure of a protein determines its molecular structure and affects its biological function (Alberts et al. 2002; Berg et al. 2002). However, growing evidence suggests that protein regions without rigid, tertiary, three-dimensional structures are prevalent in the eukaryotic proteome (Iakoucheva et al. 2002, 2004; Ward et al. 2004; Galea et al. 2006, 2009). These so-called, “intrinsically disordered regions” have different amino acid compositions than globular proteins (Uversky et al. 2000; Singh et al. 2006; Theillet et al. 2013) and are crucial for signaling and protein-protein interactions (Garza et al. 2009; Ren et al. 2008). Disordered regions have been of recent research interest and approximately 30% of human proteins contain disordered regions (Ward et al. 2004), including many disease-associated proteins (Uversky et al. 2008).

Despite the abundance of these regions, the amino acid sequences of most disordered regions are poorly conserved across taxa (Daughdrill et al. 2007). Therefore, it remains unclear how these disordered regions have been maintained through evolutionary time and why they are so predominant in eukaryotic proteins. The amino acid composition of disordered regions is significantly different from random amino acids expected based on the genetic code, which suggests that these regions are unlikely to be entirely randomly evolving "junk" (Szalkowski and Anisimova 2011). Preceding investigations have revealed increased rates of amino acid substitutions (Brown et al. 2002), differences in patterns of substitution (compared to typical ordered protein) (Brown et al. 2010), and increased rates of insertion and deletion in disordered regions (de la Chaux et al. 2007; Light et al. 2013a). Further, indels in disordered regions are under weaker negative selection relative to structured protein regions, and a large degree of variation in protein length is attributable to disordered regions (Light et al. 2013b). Indeed, disordered regions seem to have particularly weak constraints on indels, as the ratio of substitutions to indels in disordered regions is dramatically lower (Toth-Petroczy and Tawfik 2013). All of these patterns are consistent with weak constraints on disordered regions compared to globular proteins. On the other hand, a recent study has identified about 5% of residues in yeast disordered regions as short stretches of conserved amino acids, a subset of which are confirmed functional motifs, though most have still not been examined in detail (Nguyen Ba et al. 2012). It has also been suggested that some disordered regions are conserved over long evolutionary distances even when the specific amino acids are not (Chen et al. 2006; Daughdrill et al. 2007; Toth-Petroczy et al. 2008) and that there is a subset of disordered regions that are conserved at the amino acid level (Bellay et al. 2011, Colak et al. 2013). Further, one study suggested that disordered regions are the primary targets of positive selection in yeast and that disordered regions in Pfam domains contain similar numbers of functional sites as other protein regions (Nilsson et al. 2011). These observations point to the possibility of strong functional constraints on disordered regions.

Most studies that estimated evolutionary constraint on disordered regions have relied on interspecies comparisons using pairwise alignments of disordered regions across species. Due to their apparent rapid rate of evolution (Brown et al. 2002), protein sequences of disordered regions are difficult to align accurately, particularly at long evolutionary distances, which limits the power of classical comparative molecular evolutionary approaches to study these regions.

Here we take advantage of population genomics data to circumvent potential alignment issues to confirm the previous observations, and quantify rates of polymorphism and test for evidence of selection based on site frequency spectra (SFS). Using single-nucleotide polymorphisms (SNPs) from human and yeast populations, and polarized distribution of fitness effects (DFE) analyses, we validate previous reports that amino acid replacements are under negative selection in yeast and humans, although disordered regions appear to be under moderately weaker constraints of selection compared to folding protein regions. We also find up to an order of magnitude increase in the rate of non-frameshifting insertion-deletion (indel) polymorphism in disordered regions, which approaches the rate in non-coding DNA or pseudogenes. We identify examples of large indels in disordered regions segregating in the human population. Our findings also suggest that non-frameshifting indels in disordered regions are largely neutral.

Results

Single-nucleotide polymorphisms in disordered regions show evidence for negative selection

To compare patterns of evolution in disordered regions to those exhibited by structured protein regions, we divided the protein-coding regions of *Saccharomyces cerevisiae* into three parts (see Methods): disordered regions, conserved protein domains (hereafter referred to as "Pfam domains"), and all other proteins regions, which we expect to consist largely of globular, "ordered" proteins. Although computational predictions of disordered regions include errors, we expect them to be strongly enriched for disordered regions. Errors in prediction will lead to our estimates of the differences between types of proteins being conservative. To exclude regions of disorder within Pfam domains (Williams et al. 2013), we excluded from our analysis ~5% of residues that were predicted to be disordered within the Pfam domains.

To determine the relative importance of natural selection in driving amino acid alterations within eukaryotic proteins, we determined the ratio of non-synonymous (amino-acid changing) to synonymous polymorphisms (P_a/P_s) in the *S. cerevisiae* genome and observed P_a/P_s values of 0.24, 0.090, and 0.13, for disordered regions, Pfam domains, and other ordered regions, respectively (Figure 1a). We also determined P_a/P_s ratios and indel polymorphism rates in humans, using data from the 1000 genomes project (The 1000 Genomes Project Consortium 2012). The P_a/P_s ratios in disordered regions, Pfam domains, and other ordered regions were 0.48, 0.36, and 0.40, respectively (Figure 1a). As expected, disordered regions had the highest

P_a/P_s ratio, while Pfam domains showed the lowest value, a trend that is consistent with our findings in the *S. cerevisiae* genome. Since all these ratios are much lower than 1, it can be inferred that all three regions are under strong negative selection, albeit weaker in disordered regions relative to structured protein regions. To get a sense of the variability, we also computed these ratios for each gene and plotted the distribution (Figure 1b). The results were consistent with the pooled result above.

One of the strongest predictors of protein evolutionary rate is the expression level of the protein, due in part to stronger natural selection to reduce the negative consequences of protein misfolding in highly-expressed proteins (Drummond et al. 2005). Hence, we sought to rule out the possibility that differences in observed SNP polymorphisms in yeast were due to differences in expression levels between the different types of proteins regions we studied by determining P_a/P_s ratios as a function of protein expression levels (Figure 2a). This analysis revealed consistent differences in P_a/P_s values between disordered regions, Pfam domains, and other ordered regions at all expression levels. Interestingly, a clear negative correlation (R^2 = 0.73, P=0.002) was observed for disordered regions. Although disordered regions might not be expected to show a correlation between the rate of evolution and expression level under the model of selection against protein misfolding (Drummond et al. 2005), there are several reasons why this might be expected, such as spurious protein-protein interactions (Yang et al. 2012), amyloid formation (Knowles et al. 2014), and folding upon target binding (Love et al. 1995, Bowers et al. 1999, Young et al. 2000). Once again, to get an idea of the variability in these values we averaged the per gene Pa/Ps ratios and plotted the mean and three times the standard error (Figure 2b). Once again we found that disordered regions have higher Pa/Ps ratios across the whole range of expression levels. Taken together this analysis indicate that reduced constraints in disordered regions are not likely to be due to overall expression differences between ordered and disordered protein regions.

Because inferences based on non-synonymous to synonymous substitution ratios from polymorphism data may be less sensitive to selection pressures (Kryazhimskiy and Plotkin 2008), we analyzed derived allele frequency (DAF) spectra in yeast (Figure 3a), which is an alternative method to infer selection. A derived allele is an allele that arises due to mutation during evolution, whose frequency and distribution changes due to natural selection (Nielsen, 2005). The DAF spectra of non-synonymous SNPs for all three regions were skewed towards low frequency SNPs relative to synonymous sites, which supports the finding that all three regions are under negative selection, although weaker in disordered regions. The same trend is also observed for SNPs in humans (Figure 3b). It is important to note that population structure can also influence the DAF spectra, which we believe accounts for the minor peaks between allele frequencies of 0.5 and 0.7 (Figure 3a). Regardless, our conclusions based on comparisons between classes of SNPs in the genome remain unaffected because these factors are expected to equally influence all three regions.

Alleles at low frequencies are representative of new, random mutations, which are most likely to be deleterious, while alleles at higher frequencies reflect mutations that segregate and persist in the population either due to genetic drift or due to positive selection. As such, we analyzed the behaviour of non-synonymous to synonymous mutation ratios as a function of the derived allele frequencies of the respective SNPs (Figure S1). For SNPs with DAF < 10%, there were 1.36, 0.78, and 0.57 amino-acid-changing polymorphisms for every synonymous one in disordered regions, other ordered regions, and Pfam domains, respectively. In contrast, this ratio respectively decreased to 0.77, 0.39, and 0.25 for SNPs with DAF > 10% (high frequency). Assuming that the differences between these ratios are due to the removal of deleterious mutations by natural selection, we determined that 43% (1-0.77/1.36) of the nsSNPs (n=4,277) in disordered regions with DAFs <10%, were deleterious. This fraction increased to 51% (1 – 0.39/0.78) in other ordered regions (n = 7,242) and 56% (1- 0.25/0.57) in Pfam domains (n = 959). As expected, this suggests that disordered regions are more tolerant of new mutations compared to Pfam domains and other ordered regions. Taken together, we can infer that disordered regions evolve under similar, albeit weaker constraints relative to structured protein regions.

Increased Proportion of Nearly Neutral Sites in Disordered Regions Indicates Weaker Negative Selection

To assess whether negative selection is relaxed in disordered regions relative to structured regions in humans, the distribution of fitness effects (DFE) of new mutations in these regions was estimated using DFE-alpha (Keightley and Eyre-Walker 2007), which models demographic changes explicitly, unlike P_a/P_s ratios. This approach uses the site frequency spectrum to infer potential fitness consequences of new point mutations in particular regions. The site frequency spectrum of SNPs segregating in 105 unrelated individuals of the Yoruba (YRI) population of the 1000 genomes project (1000 Genomes Project Consortium 2012) was used for this analysis. By comparing the site frequency spectra of non-synonymous SNPs with synonymous SNPs, the strength of negative selection acting upon non-synonymous changes can be estimated. The DFE of non-synonymous mutations is shown in Figure 4 in terms of N_es, a measure of the efficacy of negative selection, where N_e is the effective population size and s is the selective coefficient. If their product, N_es is less than 1, mutations segregate like neutral mutations. The DFE estimated for non-synonymous mutations within all regions (Figure 4) is consistent with previous reports for humans (Haerty and Ponting 2013) and other mammals (Halligan et al. 2010). If only about 5% of disordered regions, representing short linear motifs (Nguyen Ba et al. 2012) were under negative selection, then 95% of sites would be found in the nearly neutral bin ($N_es < 1$). This is clearly not the case, since only 24% of non-synonymous sites were identified as nearly neutral. This finding is instead consistent with moderately weaker negative selection acting in these regions compared to structured regions with the majority of non-synonymous changes still being selected against.

To test whether these patterns were caused by the specific biochemical properties of disordered regions, or simply due to overall relaxed selection in disordered regions, the DFE of non-synonymous substitutions to amino acids that are more or less predominant in particular regions was also assessed. Disordered regions contain proportionately more charged and Proline residues and have a smaller proportion of hydrophobic residues compared to globular protein regions (Xie et al. 1998, Uversky et al. 2000; Singh et al. 2006; Theillet et al. 2013). Based upon these observations we chose L, I, V, F, Y to represent amino acids that are more abundant in ordered regions ("O" residues) and we chose D, E, P, S, N to represent amino acids more common in disordered regions ("D" residues). Non-synonymous substitutions involving amino acids in these two sets were polarized and classified based on whether they changed the residue from "O" to "D" or not. For example, a non-synonymous change from L to P was classified as O->D, a change from an ordered residue to a disordered residue. Each invariable site was counted as a different polarized class based upon what substitution would be caused by every possible point mutation. As expected, a higher proportion of substitutions causing O->O and D-

>D changes are effectively neutral, meaning they are under less negative selection on average, than either O->D and D->O changes respectively over all regions (Figure S2). In other words, changes between these residue groupings are less likely to be detrimental to fitness than changes across the groupings, as proteins tend to preserve the biochemical type of their amino acids.

We sought to test whether the efficacy of selection to retain the biochemical type of amino acid was similar in ordered regions and disordered regions. To do so, we computed the fraction of the total nearly neutral sites ($N_es < 1$) that change the biochemical type. For example,

$$F_{O\to D} = \frac{neutral_{O\to D}}{neutral_{O\to D} + neutral_{O\to O}}$$

is the fraction of neutral biochemically changing sites for the "O" residues. In defining this fraction, $F_{O\text{->}D}$, we are controlling for the total number of neutral sites, which we found above to be higher in disordered regions (Figure 4). We find that the fraction of neutral biochemically changing sites is higher in disordered regions than in other more structured regions (both $F_{O\text{->}D}$, Figure 5A and $F_{D\text{->}O}$, Figure 5B; Wilcoxon test $P < 10^{-6}$). This confirms that negative selection acts to preserve these biochemical types of residues more strongly within ordered regions, as opposed to simply being stronger proportionately over all residues.

Disordered regions show greater indel polymorphism than ordered regions

Given that disordered regions have been predicted to be locations of increased rates of insertions and deletions (Toth-Petroczy and Tawfik 2013) we sought to test for a difference in the abundance indel polymorphism in disordered regions in comparison to structured protein regions. We computed the rate of indel polymorphisms in disordered regions, Pfam domains, other ordered regions, and non-coding DNA in *S. cerevisiae* and in humans (See Methods). As expected, we find periodic variation (multiples of 3) in the rate of indels in protein coding regions of all types. However, we observed a greater than 10-fold increase in the rate of non-frameshifting indels in disordered regions compared to Pfam domains (Figure 6a). Interestingly, frameshifting indels in disordered regions are similar in rate to other protein coding regions (Figure 6b), but the rate of non-frameshifting indels is similar between disordered regions and indels sizes that are multiples of three in non-coding DNA. To illustrate the remarkable difference in rates between disordered regions and Pfam domains, we note that the rate of non-frameshifting (in frame) indels in yeast disordered regions is approximately 8-fold higher than frameshifting indels, whereas in Pfam domains, the rate of non-frameshifting (in frame) indels is less than half of that of frameshifting indels. The indel polymorphism patterns in human protein coding regions also show a periodic pattern as expected (Figure 6c). As in yeast, we observed that non-frameshifting indels in disordered regions are much more frequent relative to Pfam domains and other ordered regions, and show a similar rate to indels in pseudogenes (and pseudogene introns) that are multiples of three. As with the yeast indels, the differences between regions are dramatic: the rate of non-frameshifting (in frame) indels in disordered regions is 2.8 times higher than frameshifting indels, while the frequency of non-frameshifting indels in Pfam domains is less than one half of that of frameshifting indels. Moreover, the frequency of non-frameshifting (in frame) indels in disordered regions is approximately 9-fold higher than that of frameshifting indels in Pfam domains (Figure 6d). Collectively, our analyses on human indel polymorphisms show similar trends to those observed in *S. cerevisiae*, suggesting that the observed patterns may be universal to eukaryotic proteins. The polymorphism frequency for non-frameshifting (in frame) indels, which reaches the frequency expected based on non-coding sequences, suggests that they are under much weaker constraints in disordered regions than in other protein regions.

Segregating frameshift casing indels in the yeast population are enriched near the C-termini of proteins (Liti et al. 2009), presumably because frameshifts at the C-terminus are less likely to disrupt the protein function, and may simply result in the addition of additional residues and a new stop codon. We sought to rule out the possibility that the elevated rate of indels we observed in disordered regions was related to the enrichment of disordered proteins near the termini of proteins. We computed the fraction of indels that are out of frame as a function of position along the gene for indels that fall in disordered vs. other protein regions (there were too few indels in Pfam domains to compute these fractions reliably across the protein length). We found that both disordered and other ordered protein regions show higher proportions of frameshifting indels at both N- and C- termini, probably due to the availability of alternative start and stop codons for many genes. Nevertheless, the proportion of frame-shifting indels is dramatically lower for disordered regions across the whole length of the protein (Figure 6E). This is due to the elevated rate of in-frame indel polymorphism reported above, and consistent with the model that disordered regions are much more tolerant of in-frame indels, irrespective of the position in which these indels occur in protein.

The dramatic increase in rate of non-frameshifting indel polymorphism suggests that most of the large protein coding indels segregating in the human population will be found in disordered protein regions. In Figure 7, we show examples of large indels segregating at high-frequency in two important human proteins, interferon regulator factor (IRF5, Fan et al. 2010) and glutamate receptor (GRIN3B, Niemann et al 2008). In the case of IRF5, an insertion seems to have appeared in the human-chimp ancestor and reached a frequency of 54.6% in the overall 1000 genomes population. This region is not of low complexity, but repeating codons could be increasing the region's propensity for indels. Different length indels of similar sequence in orangutan, marmoset and squirrel monkey support this idea. Interestingly, the orangutan genome appears to contain a similar, albeit independent, insertion in this region. In the case of GRIN3B, the deletion likely represents the derived state, and removes nine amino acids in around 16% of the 1000 genomes population. These examples also illustrate the difficulty in properly aligning rapidly evolving disordered regions over long evolutionary distances.

We considered whether the dramatic increase in indel rates we observed in disordered regions could cause the difficulty in aligning disordered regions, and speculated that errors in alignment could explain the rapid rates of substitutions observed in disordered regions over long evolutionary distnaces. We tested this hypothesis using simulations of molecular evolution under standard models of molecular evolution (see methods), but we found that at the range of indel to substitution ratios consistent with the yeast polymorphism data (0.05-0.1) estimates of substitution rate are still accurate to long evolutionary distances (Figure S3).

Discussion

We used two population-based methods to study the strength of selection on intrinsically disordered protein regions. Both P_a/P_s ratios and the distribution of fitness effects (DFE) support weaker negative selection on substitutions in disordered regions. Prior analysis using a sequence alignment approach coincides with this interpretation (Brown et al. 2002).

Using the DFE approach we also exploited polarized non-synonymous changes to compare changes from amino acid residues that are more predominant in ordered regions ("O" residues) to those that are more predominant in disordered regions ("D" residues) and vice versa. While there are more nearly neutral potential sites for all of these "biochemical" changes in disordered regions relative to other regions, the magnitude of increase in nearly neutral sites is not equal across the different categories of changes. This finding suggests not only overall weaker negative selection in disordered regions, but also subsets of non-synonymous changes having fundamentally different fitness consequences in disordered regions compared to more structured regions. Switching states between "D" and "O" amino acids could be potentially more deleterious in structured regions, since they would more likely disrupt secondary structures of the peptide sequence. In contrast, the observation that switching residues in disordered regions is less deleterious compared to structured regions suggests that the composition of residues is more important for the functions of disordered regions, rather than their primary sequence. This is in accordance with previous evidence that the type of residue found in a disordered region is more strongly conserved than the amino acid sequence in that region (Moesa et al. 2012).

Although disordered regions showed reduced constraints on amino acid changing polymorphisms relative to structured proteins, the rate of amino acid polymorphism was not dramatically different. In contrast, although rates of frameshifting mutations in disordered regions and Pfam domains are similar, non-frameshifting indels in disordered regions are much more abundant, approaching the frequencies observed in pseudogenes and non-coding DNA. This is consistent with other recent studies that associate indels with disordered regions (de la Chaux et al. 2013, Toth-Petroczy and Tawfik 2013, Light et al. 2013a, Light et al. 2013b).

We suggest that segregating non-frameshifting indels in disordered regions are largely neutral, allowing these regions to change their size with minimal functional consequences. This inherent plasticity of disordered regions, which is not evident in ordered regions, is consistent with models suggesting that the lack of a stable protein conformation might be functionally advantageous by allowing the flexibility to interact with many different targets (Dyson and Wright 2002, 2005). Although the consistency of these observations with models of disordered region function is appealing, it is also possible that mutational biases may play a role in preferentially producing non-frameshifting indels in disordered regions (due, for example, to repetitive or low-complexity amino acid sequences, Figure 6A). If natural selection is not strong enough to counteract this elevated mutation rate, increased rates of indel polymorphism and divergence are expected under this mutation bias model as well. Although preceding investigations have suggested the former, recent systematic analysis of selection on protein coding indels suggests complex interactions between selection and mutation (Chong et al. 2013). We speculated that poor quality alignments and large rates of substitution typically observed in disordered might be due simply to the increased rate of indels. However, standard simulations of molecular evolution indicate that elevated indel rates alone are insufficient to bias the estimation of evolutionary rates. In the context of more realistic mutation processes (for example indels that are the result of small duplications, rather than random amino acids, and substitution models that are yield the equilibrium distribution of disordered regions) alignment errors could still account for some of the patterns of evolution observed in disordered regions. Nevertheless, it is reassuring that our study of polymorphism and simulations of molecular evolution generally corroborate previous work based on more distant evolutionary comparisons.

While our SNP and indel polymorphism analyses of human and yeast proteins identified trends that are consistent across both species, the absolute values of the P_a/P_s ratios and indel frequencies were different. The P_a/P_s ratio in the human population is closer to a value of 1 for two possible reasons: there is either less negative selection in humans, or human samples have diverged more recently from their most recent common ancestor. Indel mutations on the other hand occur at much lower frequencies in humans relative to yeast (Lynch et al. 2008), which may partially account for the overall lower frequency of indels in humans as shown in Figure 5.

Conclusion

Our investigation suggests that amino acid polymorphisms in disordered regions of both humans and yeast are under slightly weaker negative selection compared to structured protein regions. In contrast, the rate of indel polymorphisms in disordered regions are dramatically elevated, similar to that of non-coding DNA. Our findings using population genomics confirm recent observations regarding the evolution of disordered regions based on interspecific comparisons, and indicate that disordered regions are probably the major source of segregating protein length variation in the human population.

Materials and Methods

Predictions of Disordered Regions and Pfam domains: Protein coding regions for yeast were obtained from SGD (Cherry et al. 2012) and for human from Ensembl (v62, Flicek et al. 2013). For human proteins the longest splice form was used. Disordered regions were predicted using Disopred v3 (Jones & Cozzetto 2015) with default settings using the uniref90 database (http://www.uniprot.org/help/, The UniProt Consortium 2014) that had been filtered for repetitive regions, although we also performed the analysis using Disopred v2 (Ward et al 2004) and found similar results. For yeast, 28560 disordered regions were predicted, and for human, 409044 disordered regions were predicted in all protein isoforms (multiple transcripts per gene). Pfam domains were predicted by running HMMer 3.0 (http://hmmer.janelia.org/, Finn et al. 2011) (using default settings) on PFAM v.24 (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam24.0/relnotes.txt, Finn et al. 2008) on the yeast and human proteomes with the E-value of 0.001 as threshold. For both yeast and human Pfam domains were required to have at least 10 occurrences in the proteome, to ensure they were likely to be independently folding domains, and not highly conserved entire proteins. This left 3258 (of 8482) domains in yeast and 27264 (of 36561) domains in human. We also excluded from the analysis polymorphisms in Pfam domains that were predicted to be disordered, as Pfam domains have been reported to contain disordered regions (Williams et al. 2013).

Determination of P_a/P_s Ratios: 99,656 protein coding yeast SNPs were obtained from SGRP (Bergstrom et al. 2014) and 427,282 human SNPs were obtained from the 1000 genomes project (The Thousand Genomes Project Consortium 2012) website from the phase 1 release v3, which is based on the gencode v7 (Derrien et al. 2012) annotations of genes and pseudogenes. The total number of SNPs in each region in humans include: 132661 (disordered regions), 188393 (other ordered regions), and 100337 (Pfam domains). Rates of amino acid (non-synonymous) polymorphism were computed by dividing the total number of amino acid changing (non-synonymous) polymorphisms by the total number of amino acid changing (synonymous) sites (calculated using the method of Nei & Gojoburi 1986).

Analysis of yeast protein expression levels: Yeast protein expression levels were obtained from PaxDB (Wang et al. 2012). Proteins were binned by protein expression levels and Pa/Ps ratios were calculated as above. Polymorphisms in proteins that had no expression information were excluded from the analysis. P-value for the Figure 1c was determined using t-approximation for the distribution of the correlation co-efficient. 96,373 yeast SNPs were found in the regions of interest and more than 95% of yeast genes with SNPs had protein expression data from PaxDB (http://pax-db.org/#!species/4932).

Yeast Derived Allele Frequency (DAF) Spectra: The ancestral state of each SNP for the determination of derived allele frequencies was inferred by using *S. paradoxus* as the reference outgroup. The SNPs were assorted according to their predicted region in *S. cerevisiae* (disordered regions, other ordered regions, Pfam domains). The assorted data can be downloaded from: http://www.moseslab.csb.utoronto.ca/alan/snps_diso_pfam.txt. Lastly, the SNPs in each protein region were assorted into 10 bins (0-0.9) according to their DAF values.

Human Site Frequency Spectra: SNPs from the Yoruba (YRI) population of the 1000 genomes project were used to estimate the DFE while the full data-set was used to analyze P_a/P_S ratios and indel polymorphisms (1000 Genomes Project Consortium 2012). Relatives of the third order (1st cousins) and closer were removed from the DFE analysis, leaving 105 YRI individuals. Inferred ancestral alleles reported in this dataset were used to determine the derived allele frequency for each SNP.

Distribution of Fitness Effects: DFE-alpha v.203 (Keightley and Eyre-Walker 2007) was implemented to estimate the distribution of effects of new mutations in focal protein regions, with the default options (Keightley and Eyre-Walker 2012) and custom scripts courtesy of Dan Halligan. The input folded SFS were used from the YRI dataset described above and sites divergent with macaque for each focal site-type and region were counted according to the Enredo-Pecan-Ortheus (EPO) 6 primate alignment. This method assumes independence between sites (i.e. no linkage disequilibrium). This assumption was tested by dividing the genome into 50kb windows and generating 200 bootstrap replicates for each site category, which gives a sense of the estimate error due to non-independence. Significance between site categories was determined through a randomization test as described in Keightley and Eyre-Walker 2007. The total number of sites used in the DFE analysis are 5,820,297 (Pfam domains), 7,202,040 (disordered regions), and 9,167,914 (other ordered regions).

Indel Polymorphisms: Indel polymorphisms were obtained from SGRP (Bergstrom et al. 2014) and 1000 genomes project (The Thousand Genomes Project Consortium 2012). Dubious genes were excluded from the analysis. The total number of *S. cerevisiae* indels (less than or equal to

40 base pairs) is 415, 48, 800, and 8202, for disordered regions, Pfam domains, other ordered regions, and non-coding regions, respectively (Figure 5a). Moreover, the number of human indels (less than or equal to 30 base pairs) is 534, 125, 320, and 15601 for disordered regions, Pfam domains, other ordered regions, and pseudogenes, respectively. In order to obtain a sufficient sample size, indels in pseudogene introns were included in the "pseudogene" indel set. Although introns are not expected to have similar patterns of indels to bona fide protein coding genes, because the idea of the psuedogenes was to have a set of indels polymorphism that reflects the mutation spectrum with as few constraints as possible, including introns was acceptable, as they are expected to have few constraints as well. To compute the frequency of indel polymorphisms in each region, indels of each size were reported per kilobase pair of nucleotides in the corresponding region. For analysis of indel positions in protein coding genes, intron containing genes were not included.

DNA and Protein Sequence Alignments: Vertebrate coding sequence alignments were downloaded from the UCSC genome browser (dbSNP 138) as subsets of the Multiz 100 vertebrate multiple sequence alignment. The UCSC versions of the sequence alignments of each species include: hg19 (human), canfam3 (dog), nomLeu3 (gibbon), gorGor3 (gorilla), mm10 (mouse), rheMac3 (rhesus), ponAbe2 (orangutan), calJac3 (marmoset), saiBol1 (squirrel monkey), panTro4 (chimp) (Kent et al. 2002). The dbSNP (Sherry et al. 2001) IDs for IRF5 and GRIN3B are rs199508964 and rs142516571, respectively. The allele frequencies were also obtained from the UCSC genome browser (Kent et al. 2002).

Simulations of Molecular Evolution

To test our ability to estimate protein substitution rates from sequence alignments as a function of indel to substitution ratio, we used indelible (Fletcher & Yang 2009) to simulate 100 amino acid proteins randomly generated from the amino acid frequencies found in yeast disordered regions under the WAG model with the default indel model. We let the ancestral protein evolve into two extant proteins at the evolutionary distance indicated and realigned them using MAFFT (Katoh et al. 2002). We estimated rates of evolution using aaml for pairwise comparison from the PAML package (Yang 2007) assuming WAG evolution and filtering gapped columns on both the true alignments (obtained from indelible) and the MAFFT alignments. For each evolutionary distance we did 100 replicates.

Acknowledgement and Funding Information

Thanks to Dan Halligan for use of his custom DFE-alpha script. Thanks to Keith Dunker, Julie Forman-Kay and Robert Vernon for reading earlier drafts of the manuscript. ANNB and GMD are funded by postgraduate scholarships from the Natural Sciences and Engineering Research Council of Canada (NSERC). AMM is supported by a NSERC Discovery grant and Canadian Institutes of Health research (grant MOP-119579). This research was supported by infrastructure grants from the Canadian Foundation for Innovation to AMM.

References

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walker W. 2002. Molecular biology of the cell. 4th edition. New York: Garland Science.

Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, Boone C, Bader GD, Myers CL, Kim PM. 2011 Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol.* 12(2):R14.

Berg JM, Tymoczko JL, Stryer L. 2002. Biochemistry. 5th Edition. New York: W H Freeman. Chapter 3, Protein Structure and Function.

Bergstrom A, Simpson JT, Salinas F, Barre B, Parts L, Zia A, Nguyen Ba AN, Moses AM, Louis EJ, Mustonen V, et al. 2014. A high-definition view of functional genetic variation from natural yeast genomes. *Mol Biol Evol.* 31(4): 872-88.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol.* 55(1):104–10.

Brown CJ, Johnson AK, Daughdrill GW. 2010. Comparing models of evolution for ordered and disordered proteins. *Mol Biol Evol.* 27(30): 609-21.

Bowers PW, Schaufler LE, Klevit RE. 1999. A folding transition and novel zinc finger accessory domain in the transcription factor ARD1. *Nat Struct Biol.* 6:478-85.

de la Chaux N, Meser PW, Arndt PF. 2007. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol.* 7:191.

Chen JW, Romero P, Uversky VN, Dunker AK. 2006. Conservation of intrinsic disorder in protein domains and families: I. A database database of conserved predicted disordered regions. *J Proteome Res.* 5:879-87.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012 Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40(Database issue):D700-5.

Chong Z, Zhai W, Li C, Gao M, Gong Q, Ruan J, Li J, Jiang L, Lv X, Hungate E, Wu CI. 2013. The evolution of small insertions and deletions in the coding genes of Drosophila melanogaster. *Mol Biol Evol.* 30(12):2699-708.

Colak R, Kim T, Michaut M, Sun M, Irimia M, Bellay J, Myers CL, Blencowe BJ, Kim PM. 2013 Distinct types of disorder in the human proteome: functional implications for alternative splicing. *PLoS Comput Biol.* 9(4)

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genom Res.* 22(9):1775-89.

Daughdrill GW, Narayanaswami P, Gilmore SH, Belczyk A, Brown CJ. 2007. Dynamic behaviour of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. *J Mol Evol.* 65(3):277–98.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci.* 102(40): 14338-43

Dunker AK, Lawson JD, Brown CJ, Williams EM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, et al. 2001. Intrinsically disordered protein. *J Mol Graph Model*. 19: 26–59.

Dyson H, Wright P. 2002. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol*. 12 (1): 54–60.

Dyson H, Wright P. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol.* 6 (3): 197–208.

Fan JH, Gao LB, Pan XM, Li C, Liang WB, Liu J, Li Y, Zhang L. 2010. Association between IFR-5 polymorphisms and risk of acute coronary syndrome. *DNA Cell Biol.* 29(1):19-23.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* Web Server Issue 39:W29-W37.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. 2008. The Pfam protein families database. *Nucleic Acids Res.* 36(Database issue):D281-288.

Fletcher W, Yang Z. 2007 INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 2009 Aug;26(8):1879-88.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013 Ensembl 2013. *Nucleic Acids Res.* 41(Database issue):D48-55.

Garza AS, Ahmad N, Kumar R. 2009. Role of intrinsically disordered protein regions/domains in transcriptional regulation. *Life Sci.* 84(7-8):189-93.

Galea CA, Pagala VR, Obenauer JC, Park CG, Slaughter CA, Kriwacki RW. 2006. Proteomic studies of the intrinsically unstructured mammalian proteome. *J Proteome Res.* 5:2839-48.

Galea CA, High AA, Obenauer JC, Mishra A, Park CG, Punta M, Schlessinger A, Ma J, Rost B, Slaughter CA, et al. 2009. Large-scale analysis of thermostable, mammalian proteins provides insights into the intrinsically disordered proteome. *J Proteome Res.* 8:211-26.

Haerty W and Ponting CP. 2013. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Evolution.* 74: 61-68.

Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for Pervasive Adaptive Protein Evolution in Wild Mice. *PLoS Genet* 6(1): e1000825. doi:10.1371/journal.pgen.1000825.

Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol.* 323:573-84.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32:1037-49.

Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 2015 Mar 15;31(6):857-63.

Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 2002 Jul 15;30(14):3059-66.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* 177: 2251-61.

Keightley PD, Eyre-Walker A. 2012. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J Mol Evol.* 74: 61-68.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* 12(6):996-1006.

Knowles TP, Vendruscolo M, Dobson CM. 2014. The amyloid state and its association with protein misfolding diseases. *Nat Rev Mol Cell Biol.* 15(6):384-96.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet*. 4(12): e1000304.

Light S, Sagit R, Ekman D, Elofsson A. 2013a. Long indels are disordered: a study of disorder and indels in homologous eukaryotic proteins. *Biochim Biophys Acta*. 1834(5):890-97.

Light S, Sagit R, Sachenkova O, Ekman D, Elofsson A. 2013b. Protein expansion is primarily due to indels in intrinsically disordered regions. *Mol Biol Evol.* 30(12):2645-53

Love JJ, Li XA, Case DA, Giese K, Grosschedl R, Wright PE. 1995. Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature*. 376:791-95.

Lynch M, Sung W, Morris K, Coffey N, Landry Cr, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*. 105(27):9272-77.

Moesa HA, Wakabayashi S, Nakai K, Patil A. 2012. Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Mol Biosyst.* 8(12): 3262-73.

Nielsen, R. Molecular signatures of natural selection. 2005. *Annu Rev Genet.* 39:197-218.

Niemann S, Landers JE, Churchill MJ, Hosler B, Sapp P, Speed WC, Lahn BT, Kidd KK, Brown RH Jr. 2008. Motoneuron-specific NR3B gene: no association with ALS and evidence for a common null allele. 70(9):666-76.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 5:418-26.

Nilsson J, Grahn M, Wright AP. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. Genome Biol. 2011 Jul 19;12(7):R65.

Nguyen Ba AN, Yeh BJ, va Dyk D, Davidson AR, Andrews BJ, Weiss EL, Moses AM. 2012. Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Science Signalling* 5(215): DOI: 10.1126/scisignal.2002515

Singh GP, Ganapathi M, Sandhu KS, Dash D. 2006. Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes. *Proteins*. 62:309–15.

Ren S, Uversky VN, Chen Z, Dunker AK, Obradovic Z. 2008. Short linear motifs recognized by SH2, SH3, and Ser/Thr kinase domains are conserved in disordered protein regions. *BMC Genomics*. 9, Suppl 2: S26.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 29(1):308-11.

Szalkowski AM, Anisimova M. 2011. Markov models of amino acid substitution to study proteins with intrinsically disordered regions. *PLoS One*. 6(5): e20488.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 491: 56-65.

The UniProt Consortium. 2014. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. 42: D191-D198.

Theillet FX, Kalmar L, Tompa P, Han KH, Selenko P, Dunker AK, Daughdrill GW, Uversky VN. 2013. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. *Intrinsically Disordered Proteins*. 1:5-17.

Toth-Petroczy A, Oldfield CJ, Simon I, Takagi Y, Dunker AK, Uversky VN, Fuxreiter M. 2008. Malleable machines in transcription regulation: the mediator complex. *PLoS Comput Biol*. 4:e1000243.

Toth-Petroczy A, Tawfik DS. 2013. Protein insertions and deletions enabled by neutral roaming in sequence space. *Mol Biol Evol*. 30(4):761–71.

Uversky VN, Gillespie JR, Fink AL. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*. 41(3):415-427.

Uversky VN, Oldfield CJ, Dunker AK. 2008. Intrinsically disordered proteins in human diseases: Introducting the D2 concept. *Annu Rev Biophys*. 37: 215-46.

Wang M, Weiss M, Simonovic M, Haertinger G, Schrimpf SP, Hengartner MO, von Mering C. 2012. PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics*. 11(8):492-500.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 337:635-45.

Williams RW, Xue B, Uversky VN, Dunker AK. 2013 Distribution and cluster analysis of predicted intrinsically disordered protein Pfam domains. *Intrinsically Disordered* Proteins Vol. 1, Iss. 1,

Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK. 1998 The Sequence Attribute Method for Determining Relationships Between Sequence and Protein Disorder. Genome Inform Ser Workshop Genome Inform.;9:193-200.

Yang JR, Liao BY, Zhuang SM, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci*. 109(14): E831-40.

Yang Z. 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. Aug;24(8):1586-91.

Young ET, Kacherovsky N, Cheng C. 2000. An accessory DNA binding motif in the zinc finger protein Adr1 assists stable binding to DNA and can be replaced by a third finger. *Biochemistry*. 39:567-74.

Figure Legends

Figure 1. Non-synonymous to synonymous polymorphism ratios in disordered regions, Pfam domains, and other ordered regions in the *S. cerevisiae* genome. **A)** The P_a/P_s ratio is highest in disordered regions (white) and lowest in Pfam domains (black) in *S. cerevisiae* and human proteins. **B)** P_a/P_s ratios computed for each gene show a wide distribution, and there is significant overlap between the disordered regions (unfilled squares) and the more structured regions (unfilled triangles and filled circles).

Figure 2. A) The P_a/P_s ratio decreases for all three regions with increased protein expression levels. The difference in polymorphism ratio between the three regions remains approximately constant, independent of protein expression level. Disordered regions (white) shower higher P_a/P_s ratios across the whole range of expression levels compared to Pfam domains (black) or other protein regions (white). Protein abundance estimates were measured in ppm (parts per million). **B)** Similar results are found for the per gene estimates of the P_a/P_s ratio. Error bars represent three times the standard error of the mean.

Figure 3. Frequency spectra of single-nucleotide polymorphisms in the *S. cerevisiae* and human genome. Non-synonymous and synonymous SNPs are represented by black and white bars, respectively. **A)** Allele frequencies of SNPs in yeast regions indicated are all skewed towards lower frequencies. **B)** Site frequency of spectra of non-synonymous and synonymous SNPs in human are also skewed to the right. Only the first 10 bins of the spectra are shown for illustrative purposes.

Figure 4. Distribution of fitness effects of non-synonymous sites in disordered regions (white), other ordered regions (light grey), and Pfam domains (dark grey) in humans. Synonymous sites within each region were used as the neutral reference. Error bars correspond to bootstrapped (n=200) 95% confidence interval. * and ** indicate that the proportion of sites in disordered regions are significantly different from other ordered regions at $P < 0.05$ and $P < 0.005$ respectively, based on a randomization test of the bootstrap replicates

Figure 5. Boxplots of the fraction of nearly neutral sites where biochemical changes can occur. Disordered regions show significantly greater fraction of nearly neutral sites for positions that change disordered to ordered amino acids (**A**), as well as ordered to disordered (**B**). Distribution is based on bootstrapped replicates as described in Methods. O and D refer to amino acids found at relatively higher proportions in ordered and disordered regions, respectively.

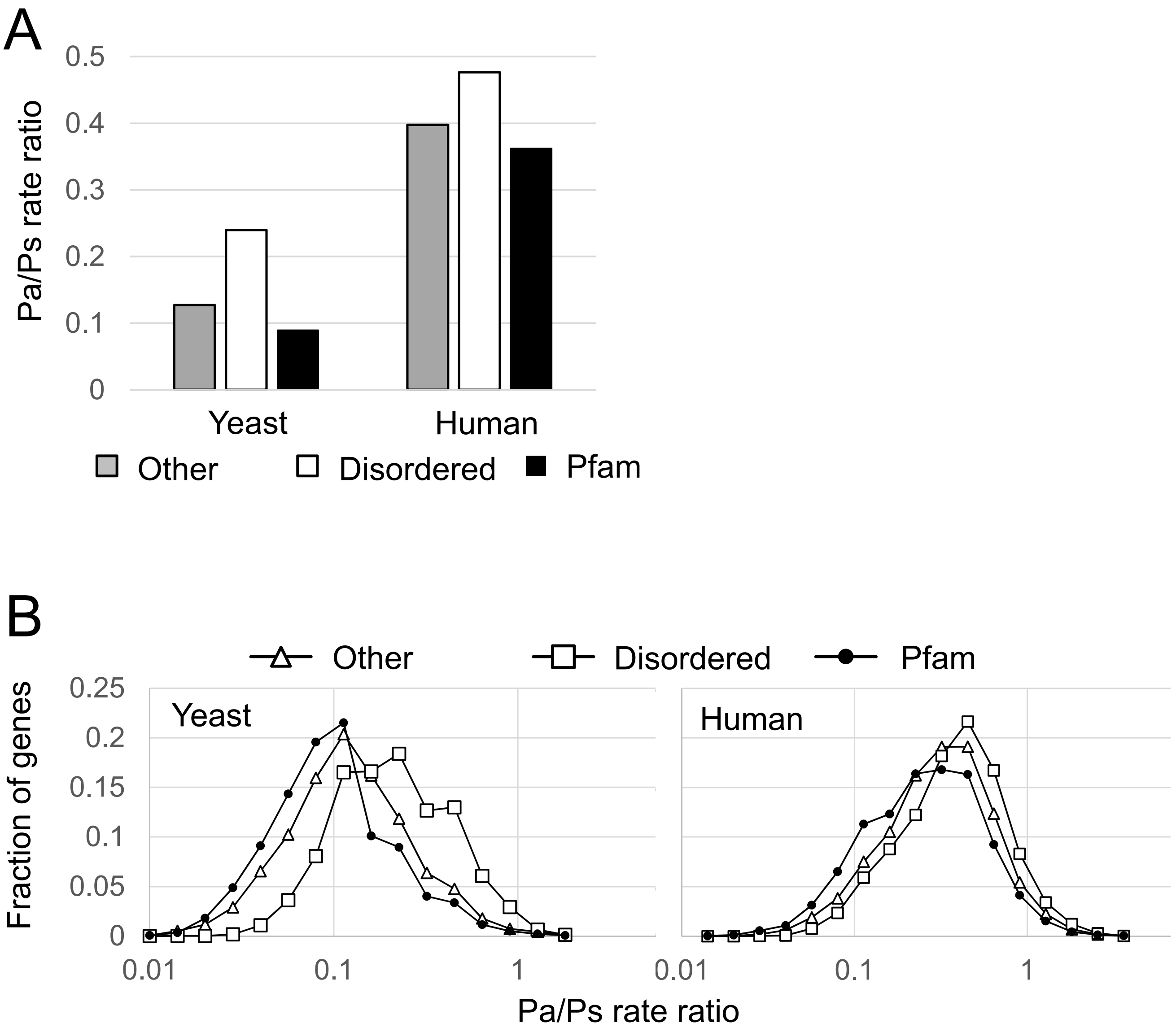
Figure 6. Frequency of insertions/deletions per kilobase pair (kbp) in disordered regions, Pfam domains, other ordered regions, and non-coding DNA in *S. cerevisiae* and humans. **A)** Grey bars represent indels that were observed in multiples of three (non-frameshifting indels). White bars represent indels that were not found as multiples of three (frameshifting indels). **B)** The frequency of frameshifting indels in disordered regions is similar to that of Pfam domains. The frequency of non-frameshifting indels in disordered regions is similar to that of non-coding DNA, and much higher than the frequency of frame-shifting indels. **C-D)** Indels in the human genome display similar patterns as in yeast. E) The fraction of out of frame indels is lower in proteins than expected based on non-coding regions (dashed line), but is higher at the termini of proteins, consistent with reduced selection on indels in the termini. This effect does not explain the difference between disordered regions (squares) and other protein regions (triangles).

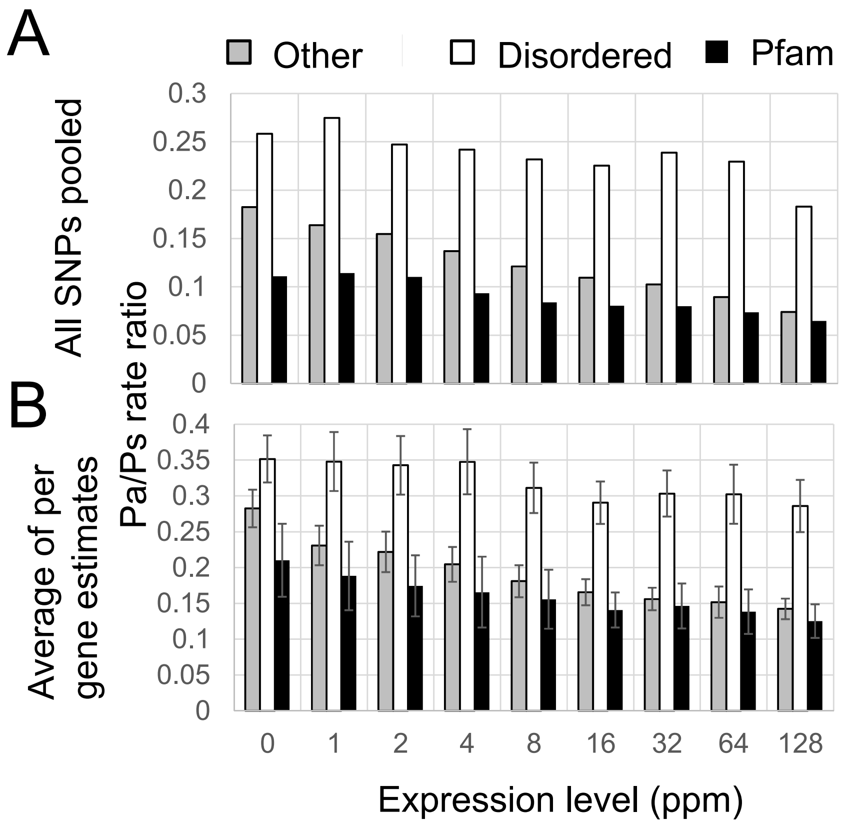
Figure 7. Multiple sequence alignments of indel and subset of Multiz 100 vertebrate alignment. For each species, the first row represents DNA sequence alignments, while the second row of letters represents the amino acid sequence alignments corresponding to each codon. The black boxes highlight the region of insertion/deletion. The percentage value corresponding to each human protein alignment represents the frequency of the respective allele. **A)** Alignment of the insertion TLQPPTLRPP (10 amino acids) in the IRF5 (Interferon Regulatory Factor) protein in humans. **B)** Sequence alignment of the deletion APAEAPPHS (9 amino acids) in the GRIN3B (Glutamate receptor) protein in humans.

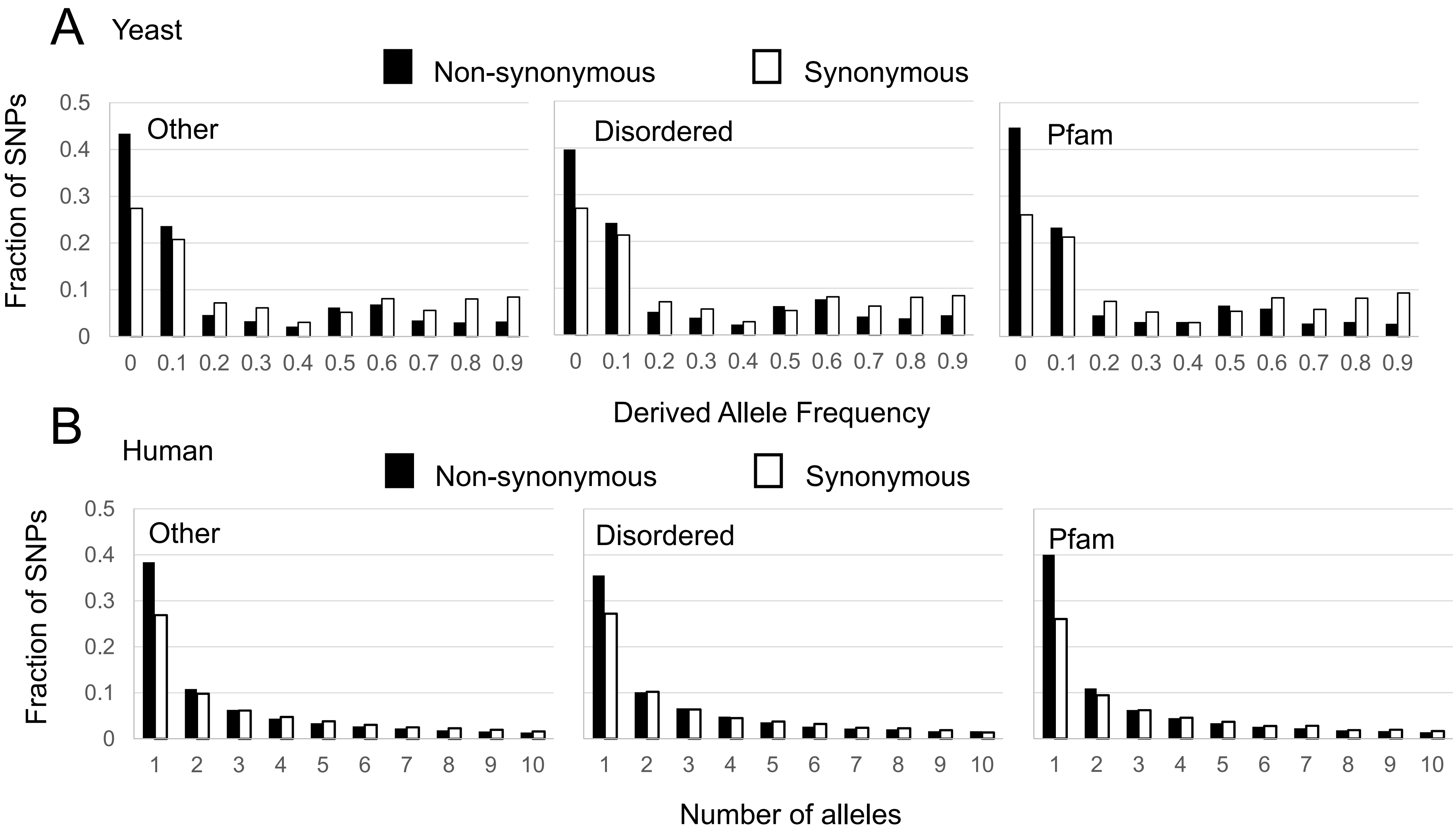
Figure S1. A) Ratio of non-synonymous (ns) to synonymous (s) SNPs as a function of low (0-0.1) and high (>0.1) derived allele frequency in yeast disordered regions (white), other ordered regions (grey), and pfam domains (black).

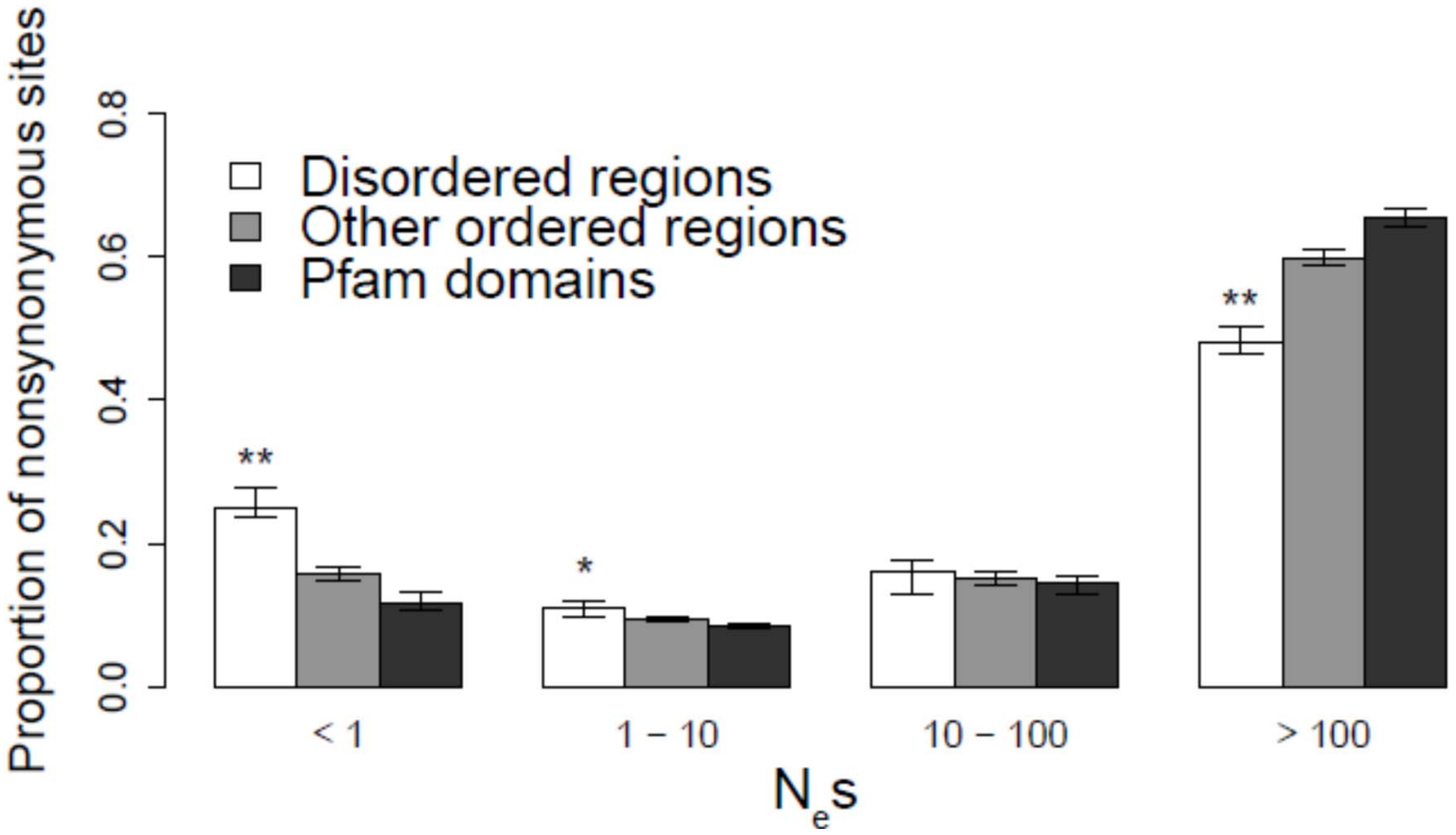
Figure S2. Distribution of fitness effects in humans of region-enriched non-synonymous sites in disordered regions (white), other ordered regions (light grey), and Pfam domains (dark grey). O and D refer to amino acids found at relatively higher proportions in ordered and disordered regions, respectively. Non-synonymous SNPs were distinguished based on whether they caused a substitution **A)** from one disordered residue to another disordered residue (D → D), **B)** from a disordered residue to an ordered residue (D → O), **C)** from one ordered residue to another ordered residue (O → O), or **D)** from an ordered residue to a disordered residue (O → D). Synonymous sites within each region were used as the neutral reference. Error bars correspond to bootstrapped (n= 200) 95% confidence interval for each category.

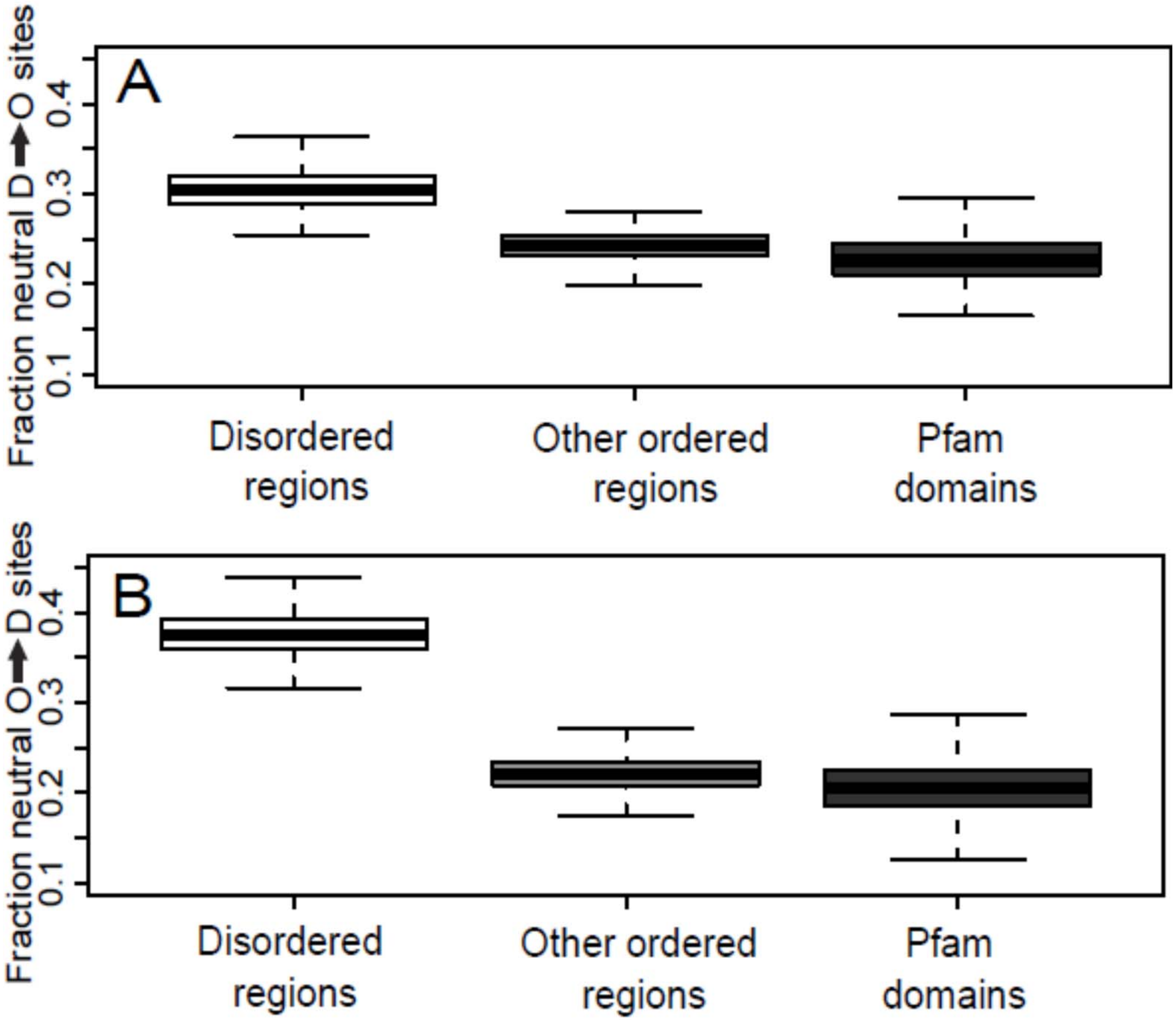
Figure S3. Elevated rate of indels observed in disordered regions is not high enough to cause errors in inference of substitution rate. Simulations of molecular evolution show that at indel to substitution ratios (Indel:Sub ratio) greater than 0.4 (filled triangle and squares) the evolutionary distances measured in substitutions per site are badly overestimated, presumably due to alignment errors. However, in the range of indel rate ratios inferred from yeast polymorphism data (0.045 to 0.1), the estimates of evolutionary distances are largely accurate (circles) up to 10 substitutions per site.

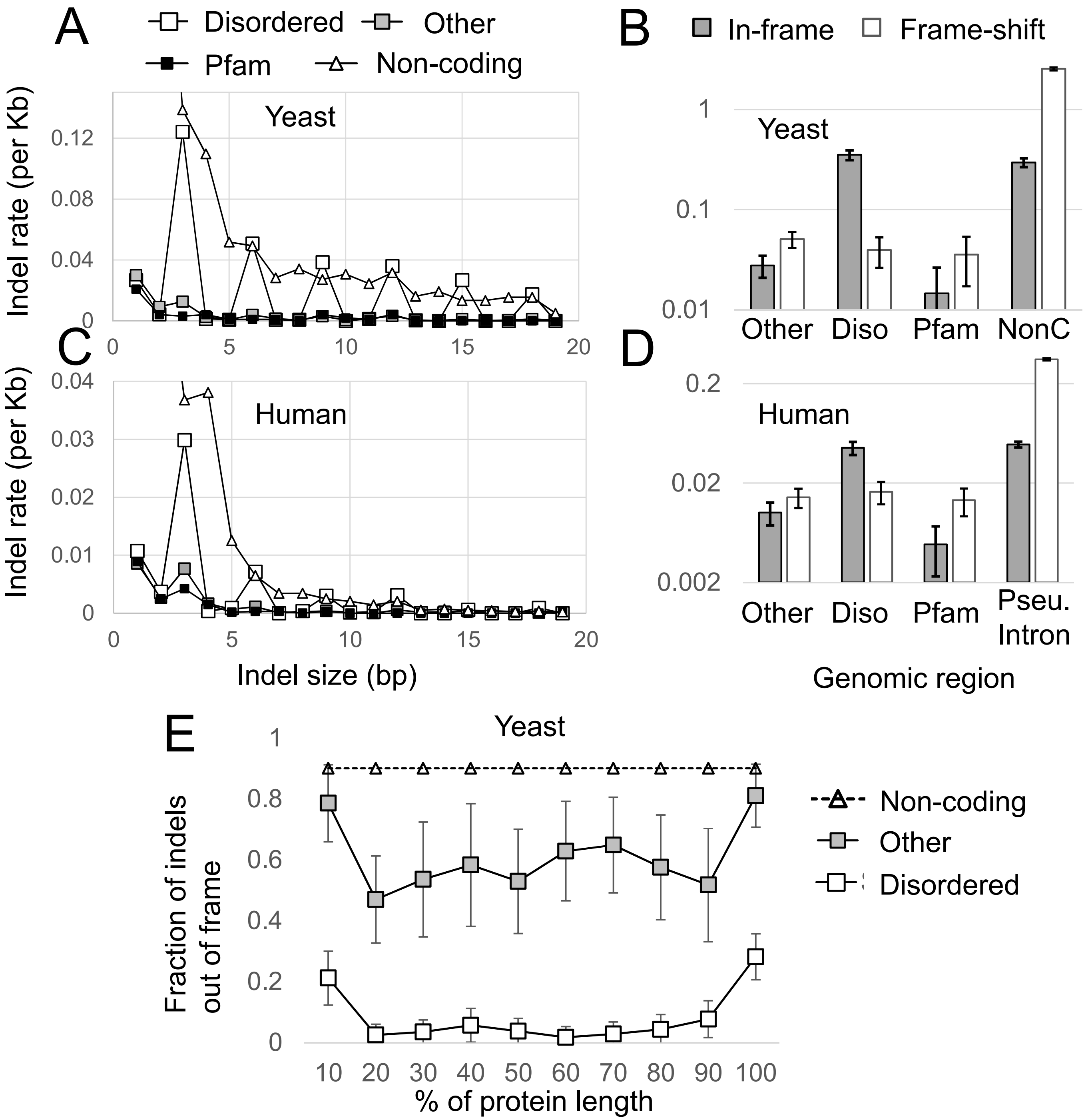

A
Pa/Ps rate ratio
0.5
0.4
0.3
0.2
0.1
0
Yeast
Human
Other
Disordered
Pfam
B
Other
Disordered
Pfam
Fraction of genes
0.25
0.2
0.15
0.1
0.05
0
Yeast
Human
0.01
0.1
1
0.01
0.1
1
Pa/Ps rate ratio

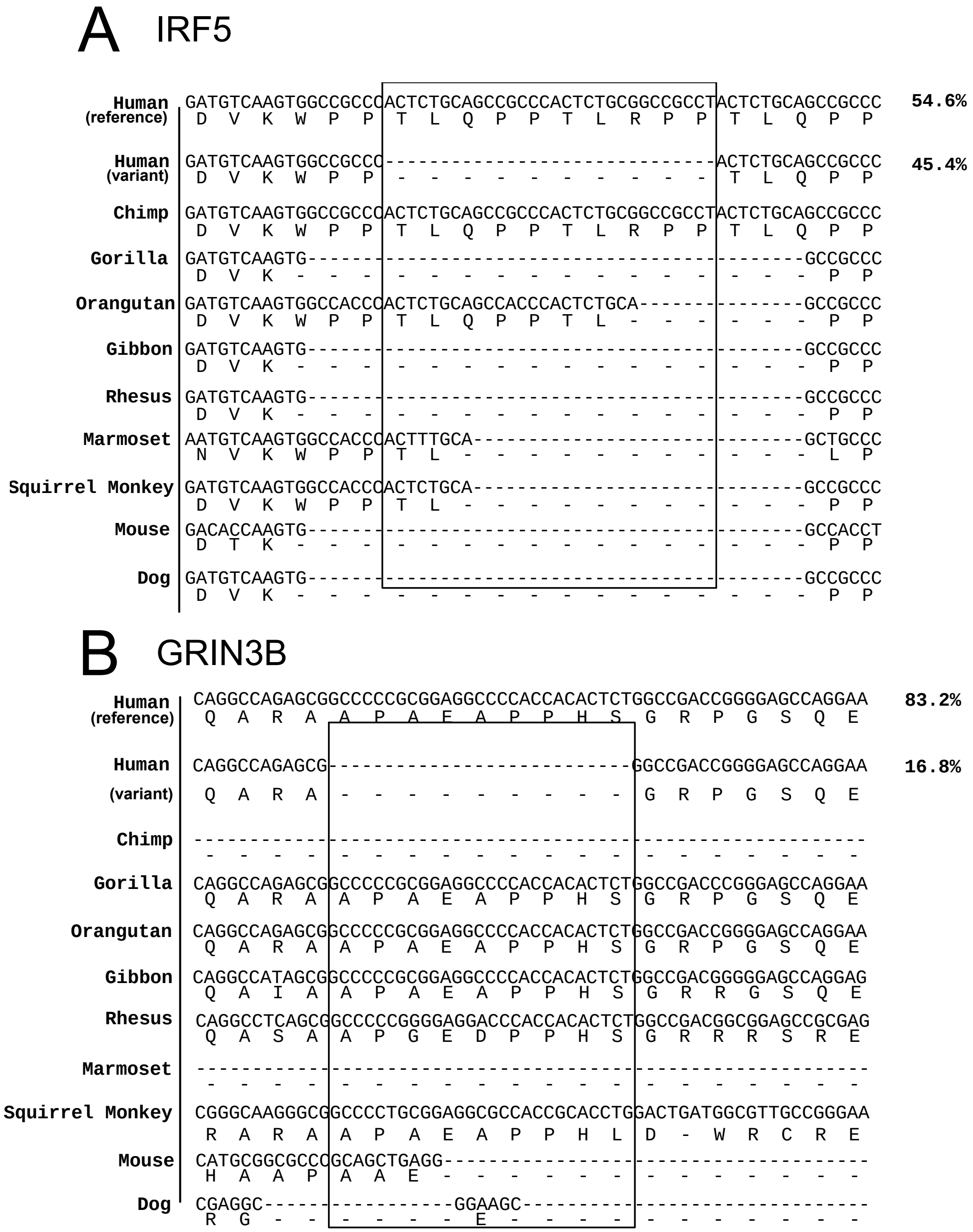

A
▨ Other
□ Disordered
■ Pfam
All SNPs pooled
0.3
0.25
0.2
0.15
0.1
0.05
0
B
Average of per gene estimates
Pa/Ps rate ratio
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
4
8
16
32
64
128
Expression level (ppm)


A
Yeast
■ Non-synonymous
□ Synonymous
Fraction of SNPs
0
0.1
0.2
0.3
0.4
0.5
Other
Disordered
Pfam
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Derived Allele Frequency
B
Human
■ Non-synonymous
□ Synonymous
Fraction of SNPs
0
0.1
0.2
0.3
0.4
0.5
Other
Disordered
Pfam
1 2 3 4 5 6 7 8 9 10
Number of alleles

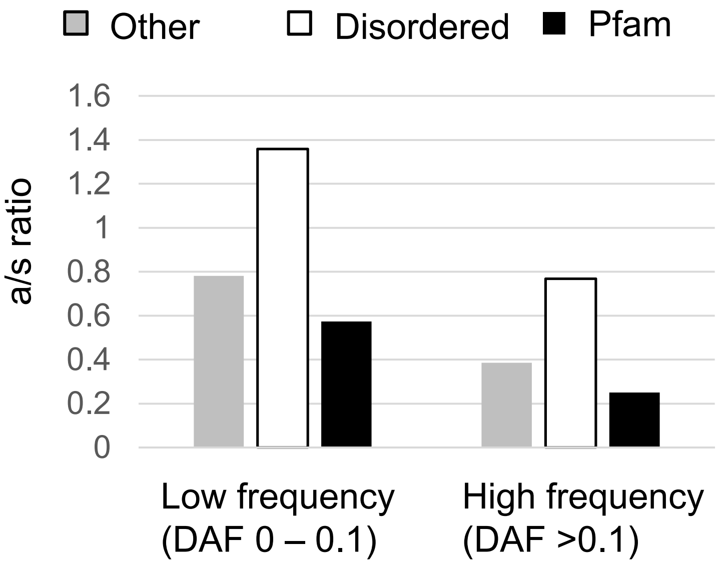

Proportion of nonsynonymous sites
0.0
0.2
0.4
0.6
0.8
□ Disordered regions
▨ Other ordered regions
■ Pfam domains
**
*
**
< 1
1 − 10
10 − 100
> 100
$N_e s$

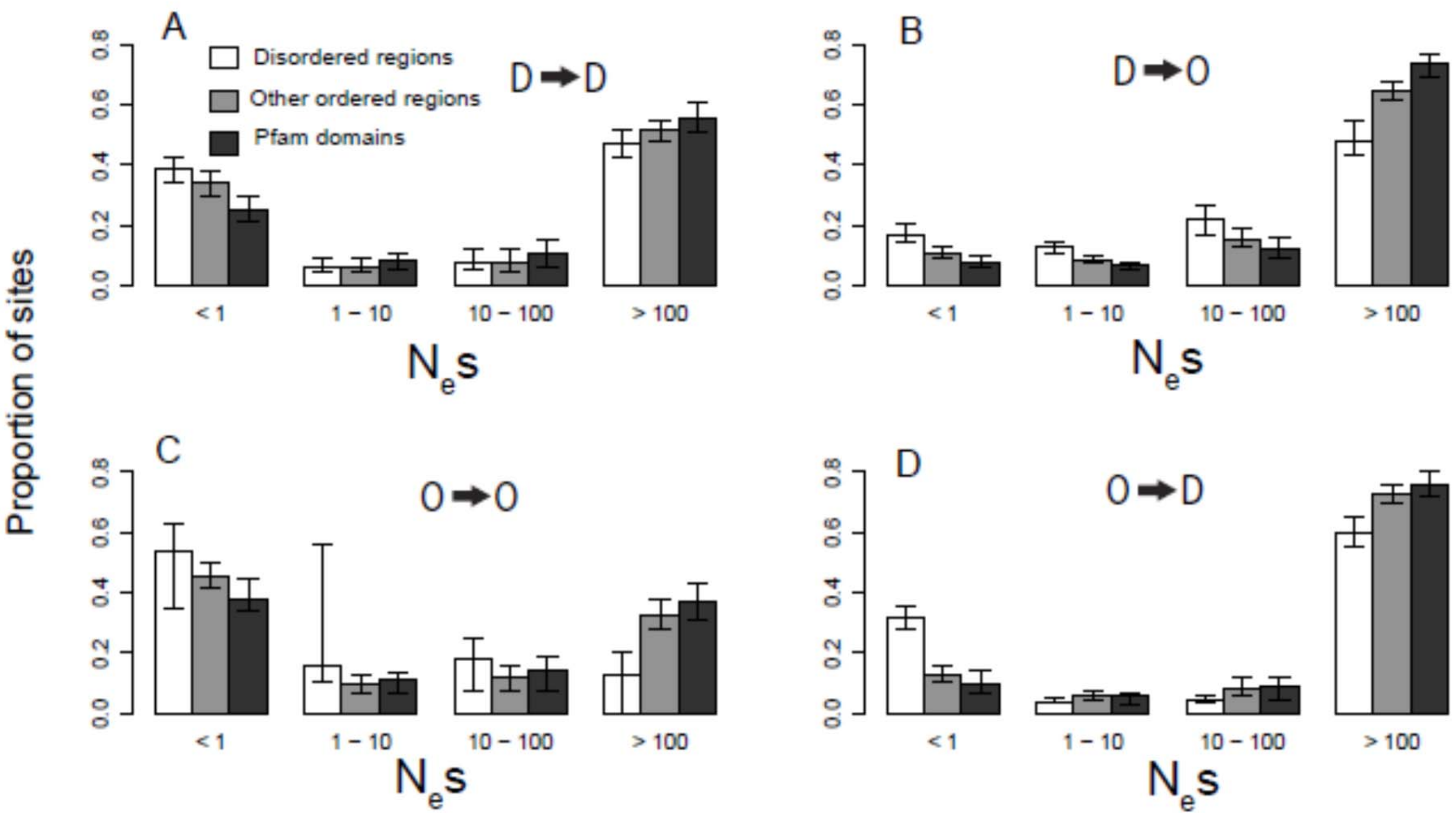

A
Fraction neutral D→O sites
0.1
0.2
0.3
0.4
Disordered regions
Other ordered regions
Pfam domains
B
Fraction neutral O→D sites
0.1
0.2
0.3
0.4
Disordered regions
Other ordered regions
Pfam domains

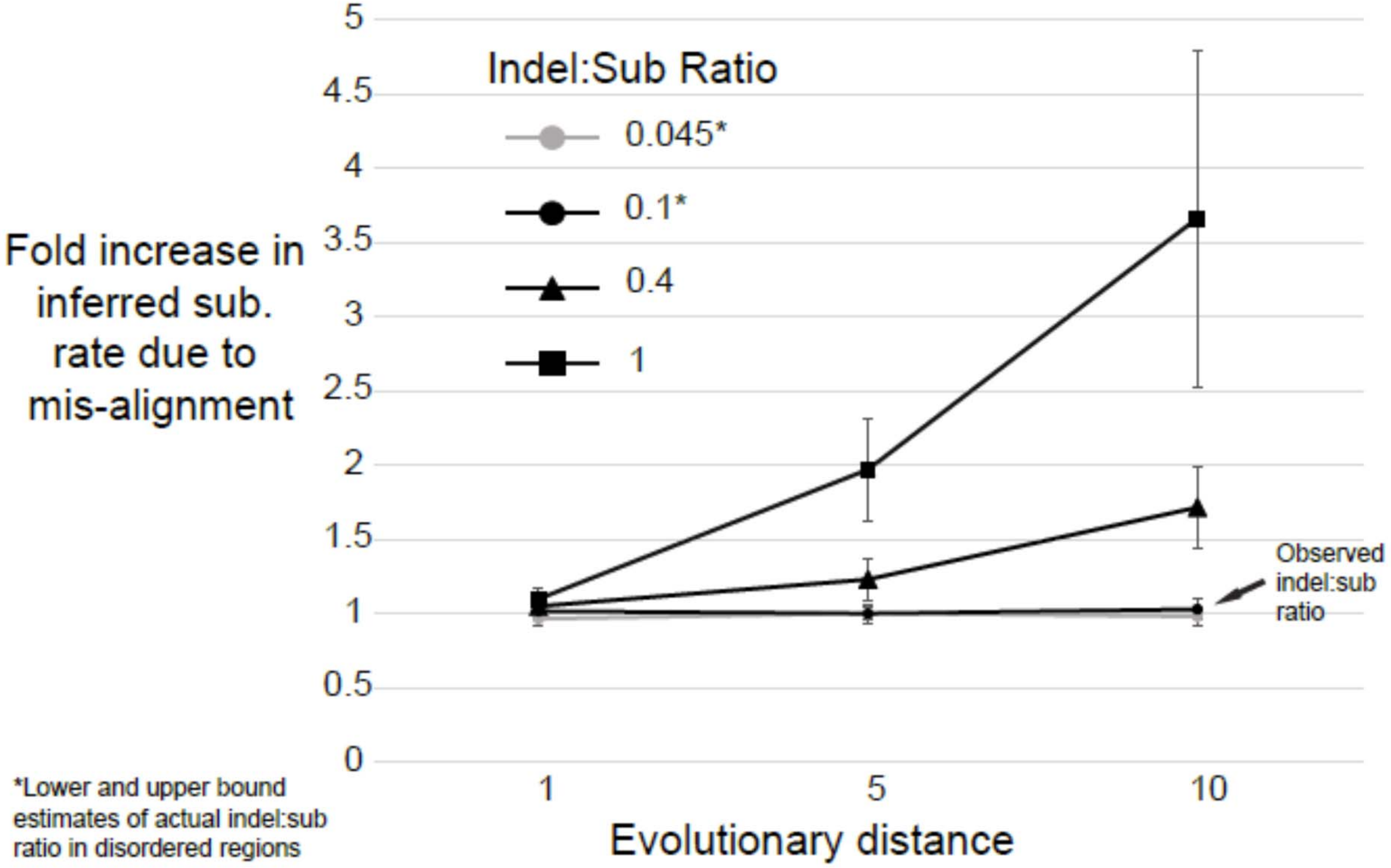

A
Disordered
Other
Pfam
Non-coding
Yeast
Indel rate (per Kb)
0.12
0.08
0.04
0
0
5
10
15
20
B
In-frame
Frame-shift
Yeast
1
0.1
0.01
Other
Diso
Pfam
NonC
C
Human
Indel rate (per Kb)
0.04
0.03
0.02
0.01
0
0
5
10
15
20
Indel size (bp)
D
Human
0.2
0.02
0.002
Other
Diso
Pfam
Pseu.
Intron
Genomic region
E
Yeast
Fraction of indels out of frame
1
0.8
0.6
0.4
0.2
0
10
20
30
40
50
60
70
80
90
100
% of protein length
Non-coding
Other
Disordered


A
IRF5
Human (reference) GATGTCAAGTGGCCGCCCACTCTGCAGCCGCCCACTCTGCGGCCGCCTACTCTGCAGCCGCCC 54.6%
D V K W P P T L Q P P T L R P P T L Q P P
Human (variant) GATGTCAAGTGGCCGCCC------------------------------ACTCTGCAGCCGCCC 45.4%
D V K W P P - - - - - - - - - - T L Q P P
Chimp GATGTCAAGTGGCCGCCCACTCTGCAGCCGCCCACTCTGCGGCCGCCTACTCTGCAGCCGCCC
D V K W P P T L Q P P T L R P P T L Q P P
Gorilla GATGTCAAGTG---GCCGCCC
D V K - - - - - - - - - - - - - - - - P P
Orangutan GATGTCAAGTGGCCACCCACTCTGCAGCCACCCACTCTGCA---------------GCCGCCC
D V K W P P T L Q P P T L - - - - - - P P
Gibbon GATGTCAAGTG---GCCGCCC
D V K - - - - - - - - - - - - - - - - P P
Rhesus GATGTCAAGTG---GCCGCCC
D V K - - - - - - - - - - - - - - - - P P
Marmoset AATGTCAAGTGGCCACCCACTTTGCA------------------------------GCTGCCC
N V K W P P T L - - - - - - - - - - - L P
Squirrel Monkey GATGTCAAGTGGCCACCCACTCTGCA------------------------------GCCGCCC
D V K W P P T L - - - - - - - - - - - P P
Mouse GACACCAAGTG---GCCACCT
D T K - - - - - - - - - - - - - - - - P P
Dog GATGTCAAGTG---GCCGCCC
D V K - - - - - - - - - - - - - - - - P P
B
GRIN3B
Human (reference) CAGGCCAGAGCGGCCCCCGCGGAGGCCCCACCACACTCTGGCCGACCGGGGAGCCAGGAA 83.2%
Q A R A A P A E A P P H S G R P G S Q E
Human (variant) CAGGCCAGAGCG---------------------------GGCCGACCGGGGAGCCAGGAA 16.8%
Q A R A - - - - - - - - - G R P G S Q E
Chimp --
- - - - - - - - - - - - - - - - - - - -
Gorilla CAGGCCAGAGCGGCCCCCGCGGAGGCCCCACCACACTCTGGCCGACCCGGGAGCCAGGAA
Q A R A A P A E A P P H S G R P G S Q E
Orangutan CAGGCCAGAGCGGCCCCCGCGGAGGCCCCACCACACTCTGGCCGACCGGGGAGCCAGGAA
Q A R A A P A E A P P H S G R P G S Q E
Gibbon CAGGCCATAGCGGCCCCCGCGGAGGCCCCACCACACTCTGGCCGACGGGGGAGCCAGGAG
Q A I A A P A E A P P H S G R R G S Q E
Rhesus CAGGCCTCAGCGGCCCCCGGGGAGGACCCACCACACTCTGGCCGACGGCGGAGCCGCGAG
Q A S A A P G E D P P H S G R R R S R E
Marmoset --
- - - - - - - - - - - - - - - - - - - -
Squirrel Monkey CGGGCAAGGGCGGCCCCTGCGGAGGCGCCACCGCACCTGGACTGATGGCGTTGCCGGGAA
R A R A A P A E A P P H L D - W R C R E
Mouse CATGCGGCGCCCGCAGCTGAGG--------------------------------------
H A A P A A E - - - - - - - - - - - - -
Dog CGAGGC----------------GGAAGC--------------------------------
R G - - - - - - E - - - - - - - - - - -

Supplementary figures


Other
Disordered
Pfam
a/s ratio
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Low frequency (DAF 0 – 0.1)
High frequency (DAF >0.1)


Proportion of sites
A
Disordered regions
Other ordered regions
Pfam domains
D ➡ D
B
D ➡ O
C
O ➡ O
D
O ➡ D
0.0
0.2
0.4
0.6
0.8
< 1
1 – 10
10 – 100
> 100
$N_e s$


Indel:Sub Ratio
0.045*
0.1*
0.4
1
Fold increase in inferred sub. rate due to mis-alignment
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1
5
10
Evolutionary distance
Observed indel:sub ratio
*Lower and upper bound estimates of actual indel:sub ratio in disordered regions