Human long intrinsically disordered protein regions are frequent targets of positive selection

Arina Afanasyeva,[1,2,3,4,6] Mathias Bockwoldt,[5,6] Christopher R. Cooney,[1] Ines Heiland,[5] and Toni I. Gossmann[1]

[1]*Department of Animal and Plant Sciences, University of Sheffield, Sheffield S102TN, United Kingdom;* [2]*Institute of Nanobiotechnologies, Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg 195251, Russia;* [3]*Petersburg Nuclear Physics Institute, B.P. Konstantinov NRC Kurchatov Institute, Gatchina, Leningrad District 188300, Russia;* [4]*National Institutes of Biomedical Innovation, Health and Nutrition, Ibaraki City, Osaka 567-0085, Japan;* [5]*Department of Arctic and Marine Biology, UiT The Arctic University of Norway, 9037 Tromsø, Norway*

Intrinsically disordered regions occur frequently in proteins and are characterized by a lack of a well-defined three-dimensional structure. Although these regions do not show a higher order of structural organization, they are known to be functionally important. Disordered regions are rapidly evolving, largely attributed to relaxed purifying selection and an increased role of genetic drift. It has also been suggested that positive selection might contribute to their rapid diversification. However, for our own species, it is currently unknown whether positive selection has played a role during the evolution of these protein regions. Here, we address this question by investigating the evolutionary pattern of more than 6600 human proteins with intrinsically disordered regions and their ordered counterparts. Our comparative approach with data from more than 90 mammalian genomes uses a priori knowledge of disordered protein regions, and we show that this increases the power to detect positive selection by an order of magnitude. We can confirm that human intrinsically disordered regions evolve more rapidly, not only within humans but also across the entire mammalian phylogeny. They have, however, experienced substantial evolutionary constraint, hinting at their fundamental functional importance. We find compelling evidence that disordered protein regions are frequent targets of positive selection and estimate that the relative rate of adaptive substitutions differs fourfold between disordered and ordered protein regions in humans. Our results suggest that disordered protein regions are important targets of genetic innovation and that the contribution of positive selection in these regions is more pronounced than in other protein parts.

[Supplemental material is available for this article.]

There is substantial experimental evidence that proteins or protein regions may be deprived of a specific three-dimensional structure (Daughdrill et al. 2005; Oldfield and Dunker 2014) and instead exist as intrinsically disordered proteins or protein parts (IDPs). The lack of a particular conformation may be an advantage where structural flexibility is required, for example for multifunctional proteins or for functional flexibility and regulation (Tompa et al. 2005; Oldfield et al. 2008; Hsu et al. 2013). In some cases, IDP regions can adopt a specific three-dimensional structure when environmental conditions change, for example, as entropic bristles (Santner et al. 2012), entropic springs (Smagghe et al. 2010), and entropic clocks (Zandany et al. 2015), or when binding to protein partners (Daughdrill et al. 1997). Experimental evidence suggests that real time state transition between ordered and disordered states can occur (Mohan et al. 2006), illustrating that protein structures are not static arrangements, as they are generally perceived (Ahrens et al. 2017). Protein disorder may also provide a genome-wide mechanism of adaptation to environmental conditions and lifestyle; e.g., host-changing parasites have a higher level of predicted disorder compared to obligate intra-cellular parasites and endosymbionts (Pancsa and Tompa 2012). Other complex organismal roles for disordered regions have been suggested, such as tissue-specific alternative splicing of disordered regions that may alter protein functions and thus change protein–protein interaction networks by recruiting new interaction partners (Buljan et al. 2013). These examples clearly demonstrate that IDPs are a heterogeneous group of protein regions that increase the functional plasticity of proteins and the flexibility of intermolecular interactions in the cell (Buljan et al. 2012; Mosca et al. 2012).

Due to the functional complexity of these protein regions, much research has been conducted to characterize IDPs and explore the underlying evolutionary mechanisms (Brown et al. 2002; Chen et al. 2006a,b; Bellay et al. 2011; Szalkowski and Anisimova 2011; Zea et al. 2013). The molecular rate at which proteins evolve at the DNA level is an important quantity in evolutionary biology and population genetics to determine the selective forces that have shaped protein composition and is generally dominated by selective constraint across many taxa (Gossmann et al. 2014). This is most prominently illustrated by the fact that, in protein coding regions, amino acid changing substitutions occur much less frequently than synonymous changes, i.e., mutations that do not change the amino acid but the underlying codon. Consequently, the rate ratio of these two types of

[6]**These authors contributed equally to this work.**
Corresponding author: toni.gossmann@gmail.com
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.232645.117.

substitutions is often found to be ≪1, illustrating evolutionary constraint at the amino acid level between species. There is, however, compelling evidence that substantial rate heterogeneity in the protein molecular rate exists across the genome and within proteins (Echave et al. 2016). Variation in the rate at which proteins evolve is, but not exclusively, attributed to functional importance, genomic context (e.g., the recombination environment or local effective population size), and structural features. For example, functionally active parts of proteins are more susceptible to selective pressure, and active sites of enzymes evolve significantly slower than other parts of the protein (Dean et al. 2002). On the other hand, parts exposed to the solvent protein surface evolve much faster (Franzosa and Xia 2009). IDPs generally tend to evolve more rapidly, largely attributed to relaxed purifying selection due to the lack of structural constraint (Brown et al. 2011), although Ahrens et al. (2016) found that sites that were predicted to be disordered and to have secondary structure were evolving at a lower rate than sites that were predicted to be ordered and to have secondary structure. Other important evolutionary determinants of the rate at which disordered regions evolve are synonymous constraint elements (Macossay-Castillo et al. 2014) and gene age, as evolutionarily young proteins tend to be enriched in disordered regions (Wilson et al. 2017).

Although our knowledge of the functional complexity and associated molecular pattern of IDPs is increasing, little is known about the role of positive selection in these important protein regions, in particular, in our own species. Rapid evolution of IDPs might suggest that positive selection contributed to the evolution of these protein regions. Disentangling the combined effects of relaxed purifying selection, genetic drift, and positive selection is crucial to identifying the underlying selective forces for the observed rapid evolution. Currently, evidence for positive selection in IDPs stems from comparative work in yeast species (Nilsson et al. 2011) and protein secondary structural elements in six *Drosophila* species (Ridout et al. 2010). However, the latter study finds evidence for an enrichment of positively selected residues in coiled-coil regions, but not in β-turns—regions that are also associated with protein disorder. Substitution-based tests of positive selection with few species are limited in power (Anisimova et al. 2001) and possibly prone to alignment quality issues (Markova-Raina and Petrov 2011); thus, it remains unclear if certain gene categories have experienced more positive selection than others. In humans, the selective effects of deleterious, but not advantageous, mutations (e.g., within a population genetic framework) in IDP regions have been investigated (Khan et al. 2015). Mutations in these regions experience less selective constraint than other structural protein elements, consistent with the observation that IDPs are rapidly evolving between species. The role of positive selection remained unexplored within this data set to circumvent potential alignment issues (Khan et al. 2015).

Here, we analyze human proteins with long intrinsically disordered regions to disentangle the evolutionary forces that have acted on these protein regions, with emphasis on the rate of positive selection.

Results

To analyze the evolutionary features of human proteins with long IDPs, we apply a comparative phylogenetic approach using publicly available genome data from human and other mammalian genomes. After rigorous alignment preprocessing, we focused on 6663 proteins with high-quality alignments (see Supplemental Methods).

Proteins with disordered regions are functionally associated with intermolecular binding

The vast majority of disordered regions in our protein data set correspond to small fractions of the entire protein (average disordered content is <15% and <100 amino acids) (Supplemental Fig. S1). Generally, these proteins tend to be involved in protein and DNA/RNA binding (Gene Ontology [GO] enrichment analysis) (Supplemental Table S1). Amino acid residues in disordered regions tend to occur predominantly at the surface of proteins (SASA scores) (Supplemental Table S2). Disordered regions are enriched in post-translational modification sites as well as regions and motifs (annotated sequence stretches of biological importance) in comparison to their ordered counterparts (Supplemental Table S3). Disease-associated SNPs tend to occur less frequently in disordered regions, with an exception for musculoskeletal disease-associated SNPs (Supplemental Table S4). This is in keeping with other research (Uversky et al. 2009; Marsh and Teichmann 2011; Gao and Xu 2012; Peng et al. 2015).

Ordered and disordered protein parts differ in their evolutionary rates due to genetic drift and differences in purifying selection

We subdivided our protein alignments into disordered and remaining parts of the protein (referred to hereafter as ordered protein regions) to separately estimate the molecular evolutionary rates in coding DNA ($\omega = d_N/d_S$) on a gene-by-gene basis. We found that ω ratios are significantly higher for disordered regions (Fig. 1) in comparison with their ordered counterparts (Wilcoxon signed-rank test, paired, $P < 2.2 \times 10^{-16}$). The difference in ω is largely driven by differences in the substitution rates at nonsynonymous sites (d_N, Wilcoxon signed-rank test, $P < 2.2 \times 10^{-16}$), as the difference in substitution rates at synonymous sites (d_S) between disordered and ordered regions is less prominent (Wilcoxon signed-rank test, $P < 4 \times 10^{-4}$). Furthermore, d_S, a proxy for local mutation rate, is significantly lower in disordered regions (median $\Delta d_S = d_S^{\text{disordered}} - d_S^{\text{ordered}} = -0.05$). This could indicate slight differences in the local mutation rate or differences in selection on synonymous sites between ordered and disordered sites, although these effects are currently difficult to disentangle (Smith et al. 2018). We estimated the proportion of substitutions fixed by genetic drift (i.e., $d_N/d_S = 1$) in ordered and disordered regions along with the intensity of purifying selection for nonneutral sites (Nearly Neutral Model sites model as implemented in PAML). The proportion of neutrally evolving sites is significantly different

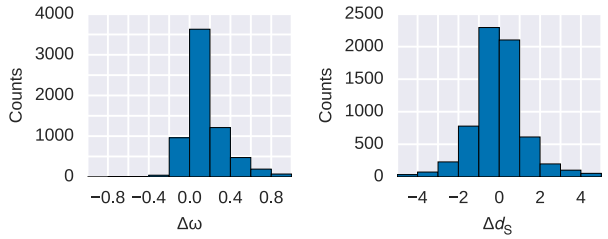

Figure 1. Histograms of paired differences in ω ($\omega = d_N/d_S$) and d_S values in proteins with disordered and ordered protein regions. Shown are the pairwise differences in disordered minus ordered protein regions ($\Delta\omega$ and Δd_S) of the same protein. Substitution rates for nonsynonymous (d_N) and synonymous (d_S) sites are obtained from a one-ratio model. ω values are significantly different for ordered and disordered regions (Wilcoxon signed-rank test, paired, $P < 2.2 \times 10^{-16}$); the difference d_S has a greater *P*-value ($P < 4 \times 10^{-4}$).

between ordered and disordered regions (Wilcoxon signed-rank test, $P < 2.2 \times 10^{-16}$, with genes with evidence for positive selection removed from the sample) (Fig. 2). We also found that ω for non-neutral sites is higher in disordered regions (median 0.078 compared to 0.041).

The role of positive selection is substantially more pronounced in disordered regions

To investigate whether elevated ω ratios are additionally caused by the effect of positive selection (e.g., a few sites for which $d_N/d_S > 1$), we conducted site-specific d_N/d_S analyses (Wong et al. 2004; Yang et al. 2005) using a gene-by-gene approach. First, we jointly analyzed ordered and disordered regions and found that 363 proteins (5.4%) showed evidence for positive selection based on a likelihood ratio test (FDR = 0.1) (Table 1). Second, we conducted the same analysis separately for ordered and disordered regions. We found evidence of positive selection in the disordered regions of 377 genes and in the ordered regions of 252 genes (significantly different, Fisher's exact test [2 df], $P < 0.001$, FDR = 0.1) (Supplemental Table S5). There were 240 genes (64%) with evidence for positive selection from the disordered set that were not identified as positively selected when the protein regions were jointly analyzed (in contrast to ≈11% for the ordered region). Moreover, considering that disordered regions are generally shorter than their ordered counterparts, this difference is likely to be an underestimate, as the power to detect positive selection decreases with shorter protein length (Anisimova et al. 2001). We therefore repeated our analysis by using the same number of sites for ordered and disordered regions by down-sampling to the number of sites of the shorter region. Using this approach, we found a roughly 10-fold (38 versus 363 genes) difference in the number of positively selected genes in ordered regions compared to disordered regions (Fisher's exact test [2 df], $P \ll 0.001$) (Table 1). We tested whether alignment scores before applying our alignment-processing pipeline are significantly different between the identified gene sets and find no evidence for that (Supplemental Fig. S2). Taken together, this illustrates that positive selection has played an intensified role in disordered regions and suggests that an a priori distinction of disordered sites substantially increases the power to detect adaptive processes. There is also no evidence that alignment quality issues have artificially created a signature of positive selection.

The detected level of positive selection is likely to be an underestimate

We investigated whether there is sufficient power and accuracy to detect positive selection for the estimated parameter range. For this, we conducted extensive sequence evolution simulations for proteins with a signature of positive selection, as well as 250 randomly chosen proteins from the remaining set. On average, more than 90% of the simulated data sets for which we simulated a signature of positive selection were identified as such and less than 10% were significant when no positive selection was simulated (Supplemental Fig. S3). We observed a slight increase in the detection of positive selection in the simulated data sets for disordered regions where only the corresponding ordered region was initially identified as positively selected. As we do not generally see this phenomenon for disordered regions, it suggests that, for these proteins, the disordered region might be contributing to the signal of positive selection. Possible reasons include a misclassification of ordered and disordered boundaries, heterogeneity of the lengths of ordered and disordered regions across species, and heterogeneity of positively selected sites within the disordered regions. As PAML does not consider gapped positions, we investigated the impact of indels on our alignment-processing pipeline. We additionally simulated protein sequences composed of disordered and ordered regions under varying indel rates, to simulate the alignment difficulties usually observed for disordered regions. Applying the site test to the processed alignments in comparison to the true simulated alignment, we find that, with an increasing indel rate, the power to detect positive selection decreases (Supplemental Fig. S4). The false positive rate stays low and remains on a comparable level when the test is applied on the true alignment. Taken together, these simulations suggest that our alignment pipeline does not artificially create signatures of positive selection but, on the contrary, is rather conservative, and that the true rate of positive selection in disordered regions is likely to be even more pronounced than reported here.

Table 1. Number of genes with evidence of positive selection in ordered, disordered, and joint estimates

Protein region	Evidence for positive selection	No evidence for positive selection
Ordered	252 (38)	6411 (6625)
Disordered	377 (363)	6286 (6300)
Joint	363 (363)	6300 (6300)

Number of genes with significant test of positive selection when down-sampled to the same length is given in brackets. Disordered regions show disproportionally more evidence of positive selection compared to their ordered counterparts (FDR = 0.1, Fisher's exact test [2 df], $P < 0.001$).

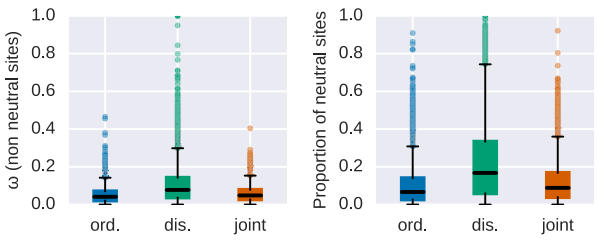

Figure 2. Estimates of sequence evolution in a nearly neutral model. Distributions of $\omega = d_N/d_S$ values (*left* panel) and the proportion of the neutrally evolved sites (*right* panel) for ordered (blue), disordered (green), and ordered and disordered protein regions jointly analyzed (orange).

Biological implications of disordered regions with positively selected target sites

We were interested in the functional associations and interactions of the proteins for which the corresponding genes showed evidence of positive selection. We performed a protein network analysis using STRING (Szklarczyk et al. 2017) for protein products of the genes that were uniquely identified to have evolved under positive selection in disordered and ordered regions (322 and 197 genes, respectively). We find that protein interaction networks for the two sets are significantly enriched in interactions but differ drastically in their layout (Supplemental Fig. S5). While the proteins from the ordered set show small-sized clusters with less than eight interaction partners, we find one large-size cluster connecting more than 50 proteins from the disordered set that includes important proteins such as breast cancer type 1

susceptibility protein BRCA1, the telomere binding protein TERF1, involved in telomere length homeostasis, and the NAD-dependent de-acetylase SIRT1 that is an important drug target in aging research (Hubbard et al. 2013). Differences in maximum cluster size are also observed when differences in the protein numbers are accounted for (Supplemental Fig. S6). The interaction network from the disordered gene set showed significant functional enrichment in RNA binding and nucleic acid binding, while the ordered gene set was associated with immune response and T-cell activation (Supplemental Table S6). This suggests an important, and to our knowledge hitherto unattributed, role of positive selection of disordered regions in transcriptional and/or translational regulation and may be driven by co-evolutionary mechanisms in regulatory arms races. Rapid evolution of immune defense genes has been reported in many taxa (Mondragón-Palomino et al. 2002; Viljakainen et al. 2009; Bonneaud et al. 2011; McTaggart et al. 2012) and is often associated with adaptive immune response.

Molecular dynamics of positively selected target sites in IL21

As we applied a site-specific test of positive selection, it is possible to specifically pinpoint amino acid residues that potentially have evolved under positive selection. Unfortunately, three-dimensional structural information for IDPs is difficult to obtain and, hence, substantially underrepresented in respective databases. A systematic study of important residues in these regions on the three-dimensional features is therefore very limited. Out of 7652 residues in IDPs that we classified to be on the surface and, hence, accessible to interaction partners (Supplemental Table S2), 83 residues (1.08%) show evidence for positive selection. In stark contrast, none of the residues identified to be on the surface in the ordered regions show evidence for positive selection, suggesting that the disordered state, but not the ligand accessibility, lead to increased molecular rates. An exception suitable for a more detailed study of three-dimensional properties of positively selected sites in IDPs is one of our candidate genes under positive selection, interleukin 21, a protein that plays a fundamental role in the innate and adaptive immune responses (Yi et al. 2010; Ju et al. 2016). Based on a partly resolved NMR structure of the disordered region, we pinpointed three residues under positive selection (Fig. 3A). This region has been shown to be a part of a helix C motif, which exists in both ordered and disordered states in different conformers and is involved in receptor binding (Bondensgaard et al. 2007). As disordered regions exist as ensembles of structures rather than as a static snapshot, we conducted a molecular dynamic (MD) analysis by simulating molecular movements of the disordered region based on the resolved NMR structure (PDB: 2OQP) for 200 nanoseconds (nsec) in explicit water solvent (Fig. 3B). This allows us to investigate the effect of the three identified residues on the structural flexibility within this protein region (Rajasekaran et al. 2011; Papaioannou et al. 2015). All three residues (S81, G85, R91) are located in a highly flexible region according to B-factor values, suggesting that these sites contribute to the structural flexibility in the disordered state (Radivojac et al. 2004). Comparing these positions across species (Supplemental Fig. S7), we find at least three different variant types at each site and with the exception of *Tarsius syrichta* (I91), none of these variants is a protein order-promoting amino acid (Oldfield and Dunker 2014). The MD simulation also reveals three neighboring residues of high flexibility at positions 61–64 outside of the annotated disordered region (Fig. 3B). The initial study by Bondensgaard et al. (2007) reported a segment from position 57 to 84 as regions of two interchangeable conformers (i.e., disordered region), adding to the notion that precise boundaries of disordered regions are difficult to capture with different methods.

The distribution of fitness effects differ between ordered and disordered protein regions in humans

To get a better understanding of the role of genetic drift and the purifying selective forces currently acting in IDP regions, we used polymorphism data from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015). The rate ratio of nonsynonymous to synonymous diversity (π_N/π_S) in coding regions can be regarded as a rough indicator of the effectiveness of negative selection on amino acid-changing mutations, with larger values indicating less effective purifying selection (Chen et al. 2017). Mean π_N/π_S is increased for disordered regions compared to their ordered counterparts (0.277 versus 0.17, Wilcoxon signed-rank test, $P < 3.9 \times 10^{-18}$) (Fig. 4A). We inferred the distribution of fitness effects (DFE) of new amino acid-changing mutations by applying a method that uses the frequency distribution of mutations at nonsynonymous sites relative to synonymous sites as neutral reference (Keightley and Eyre-Walker 2007). The DFE in ordered and disordered regions differs significantly, with roughly 23% of nonsynonymous mutations being effectively neutral in disordered regions compared to 12% for ordered regions. In contrast, the proportion of mutations with strong selective effects (i.e., $N_e s > 100$) is reduced for disordered regions relative to their ordered counterparts (48% versus 63%). Taken together, these results are in full agreement with the substitution rate analyses, hinting at a prominent role of genetic drift and reduced purifying selection in disordered regions. When conducting a joint analysis of ordered and disordered regions, results are

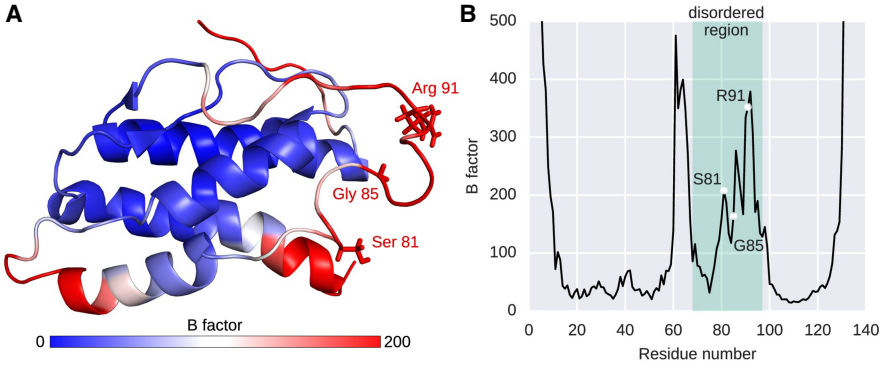

Figure 3. Three-dimensional features of positively selected sites in the disordered region of human interleukin 21. (*A*) Cartoon of the NMR structure of human interleukin 21 (PDB Code: 2OQP) including its disordered region indicating B-factor scores from the molecular dynamics analysis. Three residues have been identified as positively selected in a PAML branch-site test (Ser81, Gly85, and Arg91) in the disordered region. (*B*) Molecular dynamics analysis; shown are the B-factors of all residues. Here, B-factors reflect the fluctuation of single amino acids (C_α atom) about their average positions during the MD simulation. The predicted disordered region by MobiDB is indicated in green as well as the three identified residues under positive selection (S81, G85, and R91).

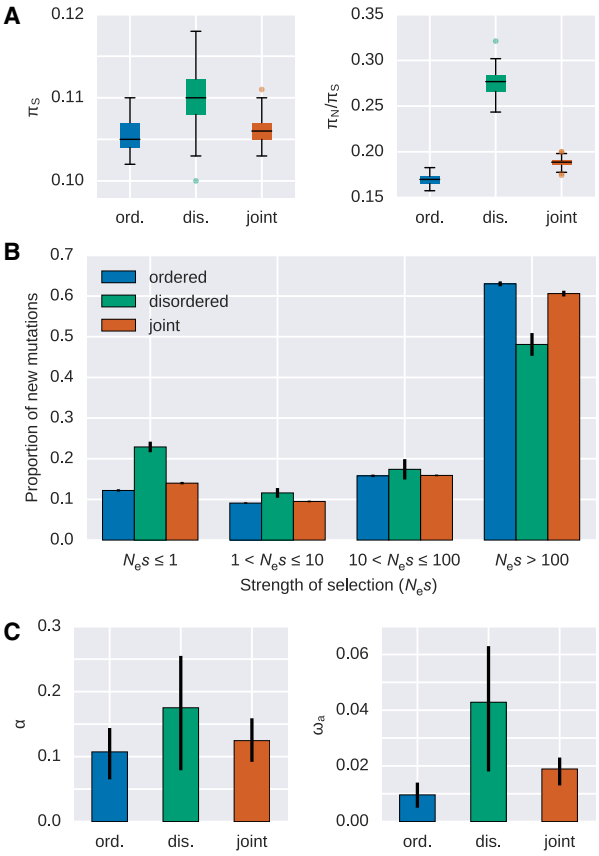

Figure 4. Evidence for differences in the selective effects in disordered and ordered protein regions in humans. (*A*) Nucleotide diversity at synonymous sites (π_S, *left* panel) and the ratio of nucleotide diversity at nonsynonymous sites over synonymous sites (π_N/π_S, *right* panel) for ordered, disordered, and jointly obtained protein regions in humans. (*B*) The distribution of fitness effects of nonsynonymous mutations estimated separately for ordered and disordered protein regions, as well as when jointly estimated. Error bars represent the standard error. $N_e s$ denotes the effective population size (N_e) scaled strength of selection (s). (*C*) Estimates of the role of positive selection for the analyzed protein set. The proportion of nonsynonymous substitutions that can be attributed to positive selection (α) and the adaptive divergence relative to the synonymous divergence (ω_a) estimated separately for ordered and disordered protein regions, as well as when jointly estimated. All pairwise comparisons are significantly different (Wilcoxon signed-rank test, paired, $P < 3 \times 10^{-12}$).

very similar to ordered regions. Taking into account that the majority of sites lie in ordered protein regions, indicating that the selective effects in disordered regions are somewhat obscured by the vast number of mutations in the ordered protein parts.

Adaptive amino acid changes have substantially contributed to the evolution of disordered protein regions in recent human evolution

To infer the role of positive selection in IDP regions in recent human and ape evolution, we compared the ratio of nonsynonymous to synonymous sites between intra-species polymorphisms and inter-species divergence (McDonald-Kreitman test [MK test]) (McDonald and Kreitman 1991). This mathematical contrast can be used to obtain the proportion of fixed substitutions that were driven by positive selection (α) and the rate of adaptive substitutions relative to the synonymous divergence (ω_a) (Gossmann et al. 2010). Since the DFEs for ordered and disordered regions differ (Fig. 4B), we applied a derivate of the MK test that corrects for the effect of slightly deleterious mutations based on the DFE (Eyre-Walker and Keightley 2009). We find that disordered regions show a higher proportion of adaptive mutations ($\alpha = 17\%$ versus 11% for ordered regions) (Fig. 4C). Since there is strong evidence that the proportion of effectively neutral nonsynonymous mutations and substitutions appear to be different (Figs. 2 and 4B) between ordered and disordered regions, a contrast of ω_a between the two regions will reveal the contribution of positive selection in absolute terms independent of fixation events of neutral and slightly deleterious mutations contributing to the nonsynonymous divergence. We observe a significant difference in ω_a values (≈fourfold, $\omega_a^{\text{disordered}} = 0.043$ versus $\omega_a^{\text{ordered}} = 0.0096$) (Fig. 4C), suggesting that the rate of adaptive substitutions relative to the synonymous rate is elevated in disordered regions. As expected, we also find reduced rates of adaptive evolution when the MK test is conducted jointly on ordered and disordered regions. This illustrates that the role of positive selection in IDP regions is difficult to capture when not taken into account a priori.

Discussion

Here, we investigated the evolutionary pressures that may have acted on proteins containing long IDP regions in humans. Our approach differs from previous analyses of the evolutionary rate analysis of these regions. First, in our comparative analysis, we used sequence data from more than 90 species, many more species than in any other previous study on this topic. In doing so, we were able to exclude low-quality sequences by excluding sequences of questionable alignability but still were able to conduct our test statistics with a sufficient number of orthologs in high-quality alignments. Second, we accounted for the genomic context and differences in the amino acid composition of disordered regions by conducting our test statistics separately for ordered and disordered regions of the same protein in a pairwise manner. Third, unlike previous approaches, we have combined inter-specific and intra-specific data for humans and other ape species to conduct a derivative of the McDonald-Kreitman test to investigate whether positive selection has played a role in the evolution of disordered regions in the human lineage.

We show that the molecular evolutionary rates, as measured by the nonsynonymous to synonymous rate ratio in coding regions, are elevated in disordered regions compared to their ordered counterparts and identify three main contributors for these elevated rates: (1) relaxed purifying selection; (2) intensified genetic drift; and (3) positive selection. This is in agreement with studies from other taxa such as yeast species and *Drosophila* and suggests general features of the evolvability of IDP regions. The lack of three-dimensional constraint may explain the fast evolution of these protein regions. However, it has been shown that there are amino acid residues that maintain the intrinsic disorder and that these residues are under stronger evolutionary constraint than in ordered regions (Ahrens et al. 2016). It is therefore important to note that the vast majority of IDP regions experienced substantial purifying selection during their evolution (Fig. 2), hinting at their important functional role despite their relaxed structural features for certain residues. The difference in the selective effects between ordered and disordered regions may be attributed to a net shift of strongly deleterious to slightly deleterious and effectively neutral mutations (Fig. 4B). Hence, the selective effects on mutations in

IDP regions are more dependent on fluctuations of the effective population size (Charlesworth 2009)—which may explain rapid evolution and higher molecular diversification due to periodic episodes of random genetic drift (Nabholz et al. 2013).

A major outcome of our study is that positive selection has played a pronounced role in certain proteins with disordered regions across the entire mammalian tree, as we find evidence for positive selection in 377 IDP regions. To our knowledge, this is the first time that the increased role of positive selection in these regions has been attributed in mammals. We show that there is limited power to detect adaptive processes if intrinsic features of disorder are not taken into account a priori. This may explain why the role of adaptive substitutions has not been emphasized thus far. We also identify, based on an extended version of the McDonald-Kreitman test that takes into account the distribution of fitness effects of new mutations, that there is a fourfold increase in the rate of adaptive substitutions relative to the rate of synonymous substitutions (ω_a) in intrinsically disordered regions. It is important to note that ω_a, but not α, should be used to compare the adaptive rates in these regions. If ordered and disordered regions would experience the same amount of adaptive substitutions, we would expect α to be lower in disordered regions; as there are more neutral fixations, this reduces the relative proportion of adaptive substitutions.

It has recently been suggested that younger proteins tend to be enriched in disordered regions and over evolutionary time become more ordered (Wilson et al. 2017). If the loss of intrinsic disorder is potentially associated with a selective advantage (e.g., gain of protein domain function), then positive selection could be a general mechanism explaining the evolvability of these IDPs. However, it is difficult to determine the turnover of intrinsic disordered regions across various species from our data set and consequently whether our set of positively selected genes between species supports such a model of adaptive losses. Evidence for repeated events of positive selection across the entire mammalian tree may suggest repeated functional diversification within the mammalian lineage due to multiple independent losses of intrinsic disorder. It is well possible that we underestimate the true role of positive selection in IDP regions. Our estimates suggest that there might be up to a 10-fold difference in the amount of adaptive changes between disordered and ordered protein parts when sequence lengths are accounted for. This is due to technical limitations of the applied test statistic, with short sequence length and high quality alignments that are difficult to obtain for some IDPs due to increased rate of fixed indels (Khan et al. 2015). Hence, disordered residues in regions of ambiguous alignability are not included in this study and their evolutionary patterns remain unexplored. However, we were able to elucidate the role of positive selection in IDPs because we included structural information a priori, and it is likely that our estimates of the amount of positive selection in these regions are conservative. Taken together, our results suggest that IDP regions are important targets of genetic innovation and that the contribution of positive selection in these regions is more pronounced than in other parts of proteins.

Methods

Protein annotations of disordered regions in human proteins and multiple sequence alignments

We obtained information of long intrinsically disordered regions for human proteins from MobiDB v2.2 (Di Domenico et al. 2012). To conduct a phylogenetically based analysis, we performed multiple sequence alignments using a customized automated pipeline (Supplemental Methods; Supplemental Fig. S8). As a phylogenetic framework, we used the near-complete species-level mammalian consensus tree assembled by Bininda-Emonds et al. (2007) and updated by Rolland et al. (2014) and pruned the complete tree to leave only those species corresponding to samples in our genomic data set (Supplemental Fig. S9). This resulted in 6663 human proteins with disordered regions and their corresponding orthologs in other mammalian species. Details are described in the Supplemental Methods.

Phylogenetic models for site-specific analyses and site annotation

The ratio of nonsynonymous to synonymous substitutions (i.e., $\omega = d_N/d_S$) can be interpreted as a measurement of selective pressure that has acted during the evolution of a protein. Here, we use site-specific d_N/d_S models for which we assume that there is variation of selective pressures between different types of sites within a protein but not between species. Since these models are computationally very expensive, we randomly down-sampled the number of species in cases when there were too many (Supplemental Fig. S10). We conducted sequence simulations using the INDELible package (Fletcher and Yang 2009). Functional associations and structural data were obtained from UniProt and PDB Details, and based on the relative solvent-accessible surface area (SASA), we predicted whether a protein site is buried or more likely to be positioned at the surface of the protein. Details are described in the Supplemental Methods.

Molecular dynamics analysis

Molecular dynamics simulations were performed using a standard protocol for pmemd simulations included in the AMBER 14 software package (Salomon-Ferrer et al. 2013). A high-resolution three-dimensional structure of human interleukin 21 (IL21) resolved by heteronuclear NMR spectroscopy (PDB code: 2OQP) was used (Bondensgaard et al. 2007).

Polymorphism statistics, DFE, and McDonald-Kreitman type test of positive selection

We obtained coding gene information for 46 unrelated Yorubian individuals from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2015) and excluded genes on the X Chromosome as well as genes that could not clearly be assigned to the respective MobiDB database entry. Divergence data for the respective gene was obtained by randomly obtaining the ortholog from a closest related non-ape species we had in our between-species data set. We used DFE-alpha (Keightley and Eyre-Walker 2007; Eyre-Walker and Keightley 2009) to estimate the distribution of fitness effects of new nonsynonymous mutations (Eyre-Walker and Keightley 2007) along with the proportion of substitutions attributed to positive selection and the relative rate of adaptive substitutions to synonymous divergence (ω_a) (Gossmann et al. 2010) for ordered and disordered regions as well as jointly for both together. Details are described in the Supplemental Methods.

Software availability

Customized scripts for the alignment processing pipeline are available in Supplemental Material and at https://www.github.com/tonig-evo/3D_gaps.

Acknowledgments

We thank Tobias Warnecke for helpful comments on an earlier version of this manuscript, and we also thank three anonymous reviewers for comments that have helped to improve the quality of this manuscript. The computations were partially performed on resources provided by UNINETT Sigma2–the National Infrastructure for High Performance Computing and Data Storage in Norway. Financial support for the work came from a FEBS Short-Term Fellowship to A.A., a UiT BFE mobility grant to M.B., and a Leverhulme Early Career Fellowship Grant (ECF-2015-453) and Natural Environment Research Council grant (NE/N013832/1) to T.I.G.

Author contributions: T.I.G. designed the study; A.A., M.B., and T.I.G. conducted the research; C.R.C. contributed to the research; A.A. and T.I.G. wrote the draft of the manuscript; and all authors contributed to editing of the manuscript.

References

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* **526:** 68–74.

Ahrens J, Dos Santos HG, Siltberg-Liberles J. 2016. The nuanced interplay of intrinsic disorder and other structural properties driving protein evolution. *Mol Biol Evol* **33:** 2248–2256.

Ahrens JB, Nunez-Castilla J, Siltberg-Liberles J. 2017. Evolution of intrinsic disorder in eukaryotic proteins. *Cell Mol Life Sci* **74:** 3163–3174.

Anisimova M, Bielawski JP, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* **18:** 1585–1592.

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:** 1.

Bininda-Emonds ORP, Cardillo M, Jones KE, MacPhee RDE, Beck RMD, Grenyer R, Price SA, Vos RA, Gittleman JL, Purvis A. 2007. The delayed rise of present-day mammals. *Nature* **446:** 507–512.

Bondensgaard K, Breinholt J, Madsen D, Omkvist DH, Kang L, Worsaae A, Becker P, Schiødt CB, Hjorth SA. 2007. The existence of multiple conformers of interleukin-21 directs engineering of a superpotent analogue. *J Biol Chem* **282:** 23326–23336.

Bonneaud C, Balenger SL, Russell AF, Zhang J, Hill GE, Edwards SV. 2011. Rapid evolution of disease resistance is accompanied by functional changes in gene expression in a wild bird. *Proc Natl Acad Sci* **108:** 7866–7871.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Keith Dunker A. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol* **55:** 104–110.

Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. 2011. Evolution and disorder. *Curr Opin Struct Biol* **21:** 441–446.

Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM. 2012. Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* **46:** 871–883.

Buljan M, Chalancon G, Dunker AK, Bateman A, Balaji S, Fuxreiter M, Babu MM. 2013. Alternative splicing of intrinsically disordered regions and rewiring of protein interactions. *Curr Opin Struct Biol* **23:** 443–450.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* **10:** 195–205.

Chen JW, Romero P, Uversky VN, Dunker AK. 2006a. Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J Proteome Res* **5:** 879–887.

Chen JW, Romero P, Uversky VN, Dunker AK. 2006b. Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder. *J Proteome Res* **5:** 888–898.

Chen J, Glémin S, Lascoux M. 2017. Genetic diversity and the efficacy of purifying selection across plant and animal species. *Mol Biol Evol* **34:** 1417–1428.

Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW. 1997. The C-terminal half of the anti-σ factor, Flgm, becomes structured when bound to its target, σ 28. *Nat Struct Biol* **4:** 285–291.

Daughdrill GW, Pielak GJ, Uversky VN, Cortese MS, Dunker AK. 2005. Natively disordered proteins. In *Protein folding handbook* (ed. Buchner J, Kiefhaber T), pp. 275–357. Wiley, Hoboken, NJ.

Dean AM, Neuhauser C, Grenier E, Golding GB. 2002. The pattern of amino acid replacements in α/β-barrels. *Mol Biol Evol* **19:** 1846–1864.

Di Domenico T, Walsh I, Martin AJ, Tosatto SC. 2012. MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* **28:** 2080–2081.

Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet* **17:** 109–121.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet* **8:** 610–618.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol* **26:** 2097–2108.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol* **26:** 1879–1888.

Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol* **26:** 2387–2395.

Gao J, Xu D. 2012. Correlation between posttranslational modification and intrinsic disorder in protein. *Pac Symp Biocomput* **2012:** 94–103.

Gossmann TI, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol Biol Evol* **27:** 1822–1832.

Gossmann TI, Santure AW, Sheldon BC, Slate J, Zeng K. 2014. Highly variable recombinational landscape modulates efficacy of natural selection in birds. *Genome Biol Evol* **6:** 2061–2075.

Hsu W-L, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, Uversky VN, Dunker AK. 2013. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci* **22:** 258–273.

Hubbard BP, Gomes AP, Dai H, Li J, Case AW, Considine T, Riera TV, Lee JE, E SY, Lamming DW, et al. 2013. Evidence for a common mechanism of SIRT1 regulation by allosteric activators. *Science* **339:** 1216–1219.

Ju B, Li D, Ji X, Liu J, Peng H, Wang S, Liu Y, Hao Y, Yee C, Liang H, et al. 2016. Interleukin-21 administration leads to enhanced antigen-specific T cell responses and natural killer cells in HIV-1 vaccinated mice. *Cell Immunol* **303:** 55–65.

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:** 2251–2261.

Khan T, Douglas GM, Patel P, Nguyen Ba AN, Moses AM. 2015. Polymorphism analysis reveals reduced negative selection and elevated rate of insertions and deletions in intrinsically disordered protein regions. *Genome Biol Evol* **7:** 1815–1826.

Macossay-Castillo M, Kosol S, Tompa P, Pancsa R. 2014. Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments. *PLoS Comput Biol* **10:** e1003607.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res* **21:** 863–874.

Marsh JA, Teichmann SA. 2011. Relative solvent accessible surface area predicts protein conformational changes upon binding. *Structure* **19:** 859–867.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.

McTaggart SJ, Obbard DJ, Conlon C, Little TJ. 2012. Immune genes undergo more adaptive evolution than non-immune system genes in *Daphnia pulex*. *BMC Evol Biol* **12:** 63.

Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. 2006. Analysis of molecular recognition features (MoRFs). *J Mol Biol* **362:** 1043–1059.

Mondragón-Palomino M, Meyers BC, Michelmore RW, Gaut BS. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. *Genome Res* **12:** 1305–1315.

Mosca R, Pache RA, Aloy P. 2012. The role of structural disorder in the rewiring of protein interactions through evolution. *Mol Cell Proteomics* **11:** M111.014969.

Nabholz B, Uwimana N, Lartillot N. 2013. Reconstructing the phylogenetic history of long-term effective population size and life-history traits using patterns of amino acid replacement in mitochondrial genomes of mammals and birds. *Genome Biol Evol* **5:** 1273–1290.

Nilsson J, Grahn M, Wright APH. 2011. Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome Biol* **12:** R65.

Oldfield CJ, Dunker AK. 2014. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83:** 553–584.

Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. 2008. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics* **9:** S1.

Pancsa R, Tompa P. 2012. Structural disorder in eukaryotes. *PLoS One* **7:** e34687.

Papaioannou A, Kuyucak S, Kuncic Z. 2015. Molecular dynamics simulations of insulin: elucidating the conformational changes that enable its binding. *PLoS One* **10:** e0144058.

Peng Z, Yan J, Fan X, Mizianty MJ, Xue B, Wang K, Hu G, Uversky VN, Kurgan L. 2015. Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci* **72:** 137–151.

Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. 2004. Protein flexibility and intrinsic disorder. *Protein Sci* **13:** 71–80.

Rajasekaran M, Abirami S, Chen C. 2011. Effects of single nucleotide polymorphisms on human N-acetyltransferase 2 structure and dynamics by molecular dynamics simulation. *PLoS One* **6:** e25801.

Ridout KE, Dixon CJ, Filatov DA. 2010. Positive selection differs between protein secondary structure elements in *Drosophila*. *Genome Biol Evol* **2:** 166–179.

Rolland J, Condamine FL, Jiguet F, Morlon H. 2014. Faster speciation and reduced extinction in the tropics contribute to the mammalian latitudinal diversity gradient. *PLoS Biol* **12:** e1001775.

Salomon-Ferrer R, Case DA, Walker RC. 2013. An overview of the amber biomolecular simulation package. *Wiley Interdiscip Rev Comput Mol Sci* **3:** 198–210.

Santner AA, Croy CH, Vasanwala FH, Uversky VN, Van Y-YJ, Dunker AK. 2012. Sweeping away protein aggregation with entropic bristles: Intrinsically disordered protein fusions enhance soluble expression. *Biochemistry* **51:** 7250–7262.

Smagghe BJ, Huang P-S, Ban Y-EA, Baker D, Springer TA. 2010. Modulation of integrin activation by an entropic spring in the β-knee. *J Biol Chem* **285:** 32954–32966.

Smith TCA, Arndt PF, Eyre-Walker A. 2018. Large scale variation in the rate of germ-line *de novo* mutation, base composition, divergence and diversity in humans. *PLoS Genet* **14:** e1007254.

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

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, et al. 2017. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res* **45:** D362–D368.

Tompa P, Szász C, Buday L. 2005. Structural disorder throws new light on moonlighting. *Trends Biochem Sci* **30:** 484–489.

Uversky VN, Oldfield CJ, Midic U, Xie H, Xue B, Vucetic S, Iakoucheva LM, Obradovic Z, Dunker AK. 2009. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics* **10:** S7.

Viljakainen L, Evans JD, Hasselmann M, Rueppell O, Tingek S, Pamilo P. 2009. Rapid evolution of immune proteins in social insects. *Mol Biol Evol* **26:** 1791–1801.

Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted by the preadaptation hypothesis of *de novo* gene birth. *Nat Ecol Evol* **1:** 0146-146.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168:** 1041–1051.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22:** 1107–1118.

Yi JS, Cox MA, Zajac AJ. 2010. Interleukin-21: a multifunctional regulator of immunity to infections. *Microbes Infect* **12:** 1111–1119.

Zandany N, Lewin L, Nirenberg V, Orr I, Yifrach O. 2015. Entropic clocks in the service of electrical signaling: 'ball and chain' mechanisms for ion channel inactivation and clustering. *FEBS Lett* **589:** 2441–2447.

Zea DJ, Monzon AM, Fornasari MS, Marino-Buslje C, Parisi G. 2013. Protein conformational diversity correlates with evolutionary rate. *Mol Biol Evol* **30:** 1500–1503.

Received November 21, 2017; accepted in revised form June 1, 2018.

Human long intrinsically disordered protein regions are frequent targets of positive selection

Arina Afanasyeva, Mathias Bockwoldt, Christopher R. Cooney, et al.

Genome Res. 2018 28: 975-982 originally published online June 1, 2018
Access the most recent version at doi:10.1101/gr.232645.117