Supplemental Material

Supplemental Methods

Protein annotations of disordered regions in human proteins

We obtained information of long intrinsically disordered regions for human proteins from **MobiDB v2.2** (Di Domenico et al. 2012), a database where such regions have been classified based on a consensus approach using ten different disordered protein region predictors. These include structural information from crystallographic data (**PDB**, Protein Data Bank) (Berman et al. 2000), experimental data from the disordered protein database (**DisProt** database (Sickmeier et al. 2007)), as well as bioinformatic approaches such as **ESpritz** (Walsh et al. 2012), **DisEMBL** (Linding et al. 2003a), **IUPred** (Dosztanyi et al. 2005), **GlobPlot** (Linding et al. 2003b), **VSL2** (Vucetic et al. 2003) and **RONN** (Yang et al. 2005). In brief, the detection of long disordered regions is optimized in MobiDB using an agreement factor $\geq 75\%$ across predictors and a regular expression on long regions with more than 20 consecutive amino acids (Di Domenico et al. 2012). Currently MobiDB contains 80,370,243 protein entries (release 2.2.2014.07 from 25/09/2014) and we restricted our analysis to human proteins (134,897 entries), focusing on the longest splicing variants. Ultimately, after filtering out entries containing only short disordered regions and shorter splicing variants, we obtained a dataset of 8,310 protein entries with disordered region annotation that was further processed to create high-quality alignments.

Multiple sequence alignments

To conduct a phylogenetically-based analysis we constructed multiple sequence alignments using a customized automated pipeline. First, we obtained orthologous genes for each human protein using the **NCBI RefSeq** database (O'Leary et al. 2016) that contains annotated genome information for more than 90 mammalian genomes by pairwise best BLAST hits with the human and mouse genomes - two of the best annotated mammalian genomes to date and HGNC identifiers (Gray et al. 2015) of all identified proteins as annotated in RefSeq. We then prepared each orthologous gene set separately by including sequence and annotation information from MobiDB to be able to identify disordered regions after alignment processing. In brief, we aligned protein sequences using **MSAProbs** (Liu et al. 2010), filtered out species with poor or little sequence information, too many insertions or deletions (indels) or that showed evidence of extremely high rates of evolution (as measured by d_N/d_S in a pairwise comparison with the human sequence) indicative of wrong orthologous assignment.

Phylogenetic framework

As a phylogenetic framework for PAML, we used the near-complete species-level mammalian consensus tree assembled by Bininda-Emonds et al. (2007) and updated by Rolland et al. (2014). To extract a phylogeny connecting the species in our study, we pruned the complete tree to leave only those species corresponding to samples in our genomic dataset.

Alignment pipeline details

To prepare input files for the phylogenetic analysis we developed an automated pipeline (Supplemental Fig S8) that includes multiple alignments, species filtering, re-alignment, and removing sequence positions of poor alignment quality. This pipeline steps are outlined below and customised scripts for the alignment processing pipeline are available as Supplemental Code and at (https://www.github.com/tonig-evo/3D_gaps).

Masking approach for site annotation information

To be able to restore the initial information of the disordered sites from MobiDB after all these filtering steps we used a custom method of site annotation. Based on the site types of the *Homo sapiens* protein sequence in MobiDB we constructed a corresponding artificial protein sequence with Phenylalanine (F) corresponding to the ordered sites and Lysin (K) corresponding to the disordered sites. The corresponding coding cDNA sequence was assigned accordingly (AAA indicates the codon for an ordered site and TTT for a disordered site). This annotation is mapped to the alignment based on the human sequence from MobiDB and contains positional information and is removed after alignment preparation.

Included proteins from MobiDB

First, we downloaded the MobiDB database data for all *Homo sapiens* proteins based on their UniProt identifiers. This data contains the protein sequence and general annotation information of the protein, such as name, sequence length, structural data availability (PDB codes) and location information of disordered regions. We prepared an initial set of files for which (I) the fasta formated file in MobiDB contained information of long disordered region(s). (II) Files with homologous proteins from mammalian species for each of MobiDB entry were available and constructed (III) a phylogenetic tree for all the mammalian species in newick format derived from the large mammalian phylogeny as described above.

Merging MobiDB entries with homologous sequences and filtering steps

The first step of the alignment preparation procedure was to merge the MobiDB database entry with the corresponding homologous sequences in RefSeq. For this we aligned the set of homologs together with the MobiDB sequence using MSAProbs v0.9.7 (Liu et al. 2010) with standard parameters and created a custom site type annotation for proteins and their corresponding cDNA sequences based on the annotation approach as described above. Some of the homologous protein sequences may affect alignment quality since they may contain large sequence insertions or deletions or show a low proportion of truly homologous positions to the human sequence. Furthermore, some of the sequences in other species may be lacking or contain little homologous positions of human long disordered regions, such sequences are not of interest for our analysis, so we conducted sequence filtration based on several statistics: We calculated the proportion of homologous positions relative to the human sequence for each species, the proportion of sites homologous to the human disordered sites and the proportion of the human sequence that will remain after removing gapped positions and stop codons. We also used similar statistics for disordered sites only. We defined an 80% threshold to filter sequences. For this we excluded sequences one by one starting from the sequence with the lowest number of homologous sites and recalculated the statistics until the sequences in the alignment covered more than 80% of sites for each of the three statistics. After this filtering procedure we performed a check for long insertions in homologous sequences: if more than 20% of sites in the sequence did not have homologous sites in the human sequence we excluded the sequence from further analysis. After applying these filering procedures, we performed a second alignment with MSAProbs (if some of the homologous sequences were excluded during filtering) and annotation procedure. We also checked for mismatches between the human Refseq sequence and the MobiDB sequence: if such mismatches occurred, we placed gaps in the mismatched sites. After these procedure we obtained aligned protein sequences and the corresponding (unaligned) cDNA sequences. We used PAL2NAL (Suyama et al. 2006) to retrieve the corresponding cDNA alignment from the protein alignment. Due to our customised annotation, we could easily retrieve the whole alignment or alignnments for ordered and disordered regions separately to conduct a separate analysis (Supplemental Fig S8).

Positional information through site masking and local realignment

After manual inspection we decided to additionally quality-check the resulting alignments using **ZORRO** (Wu et al. 2012) and **Gblocks v0.91b** (Talavera and Castresana 2007) to identify alignment columns of poor alignment quality. These poorly aligned columns were subsequently excluded from the analysis, i.e. sites with a ZORRO score of less than 9 or sites outside of the identified blocks in Gblocks using parameters -t=p -k=y -n=y -v=32000 -p=t. We then re-aligned the orthologous sets for the disordered regions with **MUSCLE** (Edgar 2004) and removed gene sets from the analysis for which the MUSCLE alignment disagreed with the second MSAProbs alignment. We also estimated pairwise substitution rates in a codon model for the disordered regions, and excluded species for which the median substitution rate exceeded two, a signature for saturation and hence potential misalignment. Due to these approaches, our method is more conservative regarding the alignment quality of the disordered regions in comparison to the ordered regions. This resulted in 6,663 human proteins with disordered regions and their corresponding orthologs in other mammalian species. These files were used to generate input files with **PAL2NAL** to conduct codon-based substitution rate analyses with **PAML** version 4.9a (Yang 2007).

Phylogenetic models for site-specific analyses

Under the assumption that synonymous mutations evolve neutrally the evolutionary rate of a protein can be expressed as the ratio of non-synonymous to synonymous substitutions (i.e. $\omega = d_N/d_S$). This ratio can be interpreted as a measurement of selective pressure that has acted during the evolution of a protein, with ω values <1, =1, and >1 indicating purifying selection, neutral evolution, and diversifying selection, respectively. Hence this measure may be used to infer potential function(s) of proteins or protein domains.

In our analysis, we used site-specific d_N/d_S models (model M1a, nearly neutral model, and model M2a, direct test for positive selection) for which we assume that there is variation of selective pressures between different types of sites within a protein but not between species. Differences between models were assessed with a likelihood ratio test (LRT) assuming that twice the log likelihood difference is approximately χ^2 distributed with the respective degrees of freedom as indicated in the PAML manual. Although we cannot exclude the possibility of species-specific functions of disordered regions, we assume this is rather an exception. Additionally, as we excluded genes that show extreme rates of evolution for specific species, it is likely that most of the cases where this assumption is violated have been already excluded during file preparation. As for the codon-based analysis pipeline, we prepared three different alignment sets:

1. Joint analysis: Gene sets contain all sites (i.e. no prior information of site types was used)
2. Separate analysis: Gene sets contains only sites in disordered regions

3. Separate analysis: Gene sets contain only sites in ordered regions (i.e. non-disordered regions)

These models allow us to get more precise information about the specificity of evolutionary pressures that disordered protein regions have experienced and allow us to conduct comparative analyses between ordered and disordered parts of proteins, and by that controlling for the genomic context. Since these models are computationally very expensive, we had to compromise between computational time and the number of included species. We therefore randomly downsampled the number of species in cases when there were too many (threshold of 30 species). We found that downsampling is reasonable if the number of species is not too low. The large scale phylogenetic tree was pruned with the *nw_prune* module from the Newick Utilities tools for the processing of phylogenetic trees (http://cegg.unige.ch/newick_utils (Junier and Zdobnov 2010)) and then unrooted resulted trees with *unroot()* procedure from the Ape library in R package (http://ape-package.ird.fr/).

Sequence evolution simulation studies

We conducted sequence simulations using the Indelible package (Fletcher and Yang 2009), an extension of the evolver program package included in PAML. Indel-free simulations were conducted using parameter estimates obtained from the two separate codeml site analyses for ordered and disordered regions. Using these estimates we generated sets of 100 alignments for each protein by simultaneously simulating disordered and ordered protein regions in a partitioned model with Indelible. For computational reasons we focused on proteins with evidence for positive selection as well as 250 randomly chosen proteins from the remaining protein set. We determined the power and accuracy by conducting the same separate site analysis on the simulated sequences, split into ordered and disordered regions and counted how often LRTs were significant for each region. We expect that the proportion of significantly rejected site tests for genes initially identified to be under positive selection to be high, while the proportion of significant LRTs should be low when there was no positive selection inferred initially.

To determine how our alignment-processing pipeline performs, we applied it to simulated alignments with different indel rates for the disordered regions. For this, we constructed an artificial protein with a disordered region of 250 amino acids, flanked by ordered regions of 250 amino acids on each side. As insertions and deletions may produce shorter and less accurate alignments, and as PAML was run to ignore sites with gaps, a proportion of the remaining codons after gap removal might be incorrectly aligned and could generate false positive or negative results. We hence constructed 100 alignments with varying indel rates (equal rates for insertions and deletions with the indel rate being 1×, 5× and 10× relative to the ordered region) with and without positive selection. Except for the indel rate, we used the parameter estimates and tree topology from SIRT1 (a protein with similar length) and simulated either positive selection or neutral evolution for around 5% of the disordered sites that were initially estimated to be positively selected in SIRT1. We assumed a Lavalette distribution (Fletcher and Yang 2009) for indels with a maximum indel size of 50 and an a parameter of 2.0 which is within the range of generally observed values assuming this distribution (Gossmann and Schmid 2011). We applied the codeml site test separately for ordered and disordered regions and compared our processed alignments with the true alignments generated with Indelible.

Functional association from public databases (UniProt, NCBI SNP and PDB)

We collected information about functional sites annotation from the PDB database as well as from the UniProt database (regions and motifs - structural or binding, PTMs - sites of post-translational modifications and other functional or binding sites) to combine the results of our phylogenetic analysis with a potential functional classification of proteins containing disordered protein regions. UniProt contains data for almost all of the protein entries used in this analysis. We also obtained data of potential disease-related single-nucleotide polymorphism (SNP) positions in humans (http://www.uniprot.org/docs/humsavar.txt).

Structural data and site localization determination

As expected, PDB data was available for only 2,589 of the proteins in our dataset and only a small fraction of these had functional or binding site annotation. Although the vast majority of long disordered regions lack three-dimensional information, for a limited number of sequences there are structural information for a part or even for a whole disordered protein region available. For example, these data could be obtained by NMR or by a combination of several methods, or the disordered regions may be obtained in a structural conformation when bound to binding partners (Tan et al. 2009).

To determine the localization of long disordered regions in protein structures we combined our analysis with structural data from the PDB database. Based on the relative solvent-accessible surface area (SASA) it is possible to predict whether a protein site is buried or more likely to be positioned at the surface of the protein. Since amino acids considerably differ in size, an absolute measure of SASA would be difficult to compare in sense of the solvent accessibility, therefore should SASA values be normalized. The relative SASA represents the ratio of the surface area of a residue accessible to a solvent to its standard accessibilities in an unfolded state (calculated in the extended Gly-X-Gly tripeptide for all amino acid residue types) (Duarte et al. 2012). To calculate the relative SASA (rSASA) score we used ICM-Pro (http://www.molsoft.com/icm_pro.html), a program package for molecular modelling, assuming a standard water probe radius of 1.4 Åand the Shrake and Rupley algorithm (Shrake and Rupley 1973).

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 (hIL)-21 resolved by heteronuclear NMR spectroscopy (PDB code: 2OQP) was used (Bondensgaard et al. 2007) and periodic box conditions were set. Water was modelled explicitly with the TIP3P model. Calculations were performed with NVIDIA GPU acceleration on a workstation with a GeForce GTX 1080 graphic card. B-factors (atomic displacement parameter, (Yuan et al. 2005)) for each protein residue were calculated based on the whole time of the productive MD calculation (i.e. for 200 nano seconds with a time step of 0.002 pico seconds and recorded every pico second) using AmberTools 14 and atomfluct utility. The MD trajectory was pre-processed, i.e. box centred and superimposed by atoms of the protein main chain.

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

We obtained whole genome information for 46 unrelated Yorubian individuals (i.e. 92 haplotypes) from the human 1000K genome project (The 1000 Genomes Project Consortium 2015) and extracted their genic variation (genome annotation file http://ftp.ensembl.org/pub/grch37/release-87/gtf/homo_sapiens/). We excluded genes on the X Chromosome as well as genes that could not clearly be assigned to the respective MobiDB database entry. We focused on bi-allelic SNP variation and created site frequency spectra for synonymous and nonsynonymous sites on a gene-by-gene basis using the Python egglib package (De Mita and Siol 2012). 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 dataset. We then split the information into ordered and disordered regions and summed data across genes, because some genes are very short or contain little polymorphisms. Statistics presented here, unless otherwise stated, are obtained from the summed data. 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. The McDonald Kreitman test (McDonald and Kreitman 1991) is a classic test of positive selection that uses the contrast of divergence and diversity at selected (e.g. nonsynonymous) relative to neutral (e.g. synonymous) sites to infer the rate of positive selection. In its classic form, it neglects the effect of slightly deleterious mutations and the rate of adaptive evolution can be denoted as α, where $\alpha = 1 - (D_S P_N)/(D_N P_S)$ (where D and P denote the relative rate of (non)synonymous substitutions and polymorphisms, respectively (Eyre-Walker 2006)). Using this notation $\omega_a = D_{nA}/D_S = \alpha D_N/D_S$, where D_{nA} is the rate of adaptive nonsynonymous substitutions. Since we expect the number of non-adaptive nonsynonymous substitutions to vary between disordered and ordered region because of the differences in the DFE, α would vary between these protein regions even if D_{nA} would be the same. Hence ω_a is a better measure to compare the role of adaptation between these two protein regions.

Statistical and GO enrichment analysis

For statistical analysis, we used the scipy Python package. Graphs were generated with matplotlib and seaborn in Python3. In a box plot, the box represents the range between upper and lower quartiles, the horizontal line within the box shows the median, and the whiskers show the most extreme data point, which is no more than 1.5 times the length of the box away from the box. In a barplot the error bars denote the standard error. To test for enrichment of genes with different gene ontology (GO) classifications, we used PANTHER (Mi et al. 2016) as well as STRING database (Szklarczyk et al. 2017). PyMOL 1.7.2.1 was used for protein structure visualisation.

Supplemental Figures

Supplemental Figure S1

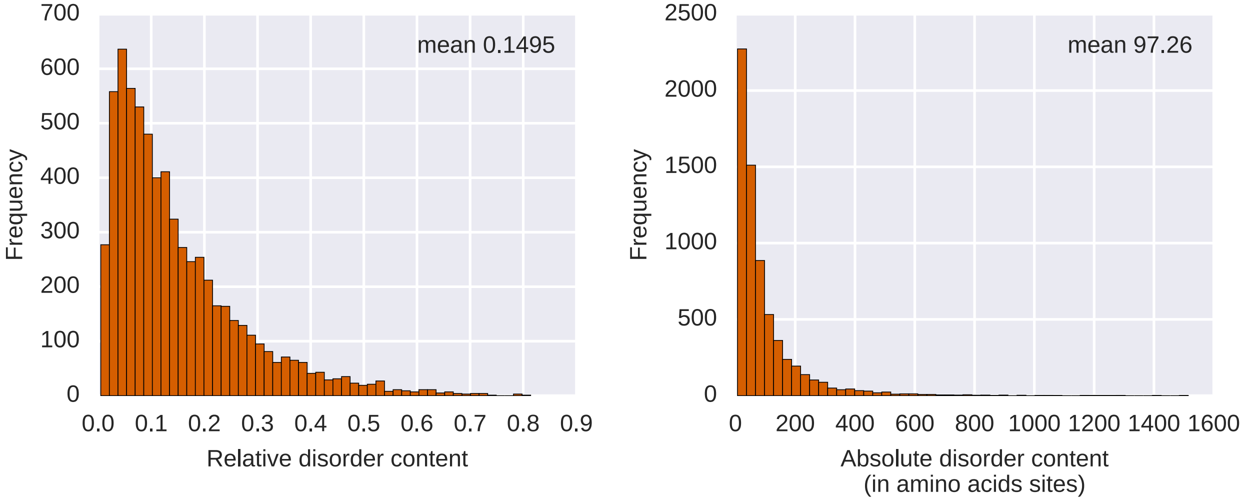

Figure S1: **Disordered protein content in the analysed dataset.** Histograms of relative (left panel) and absolute (right panel) number of sites in proteins that are predicted to be part of a long intrinsically disordered protein region in our dataset, based on human protein sequences from MobiDB (Di Domenico et al. 2012) after filtering and alignment processing (6,663 entries).

Supplemental Figure S2

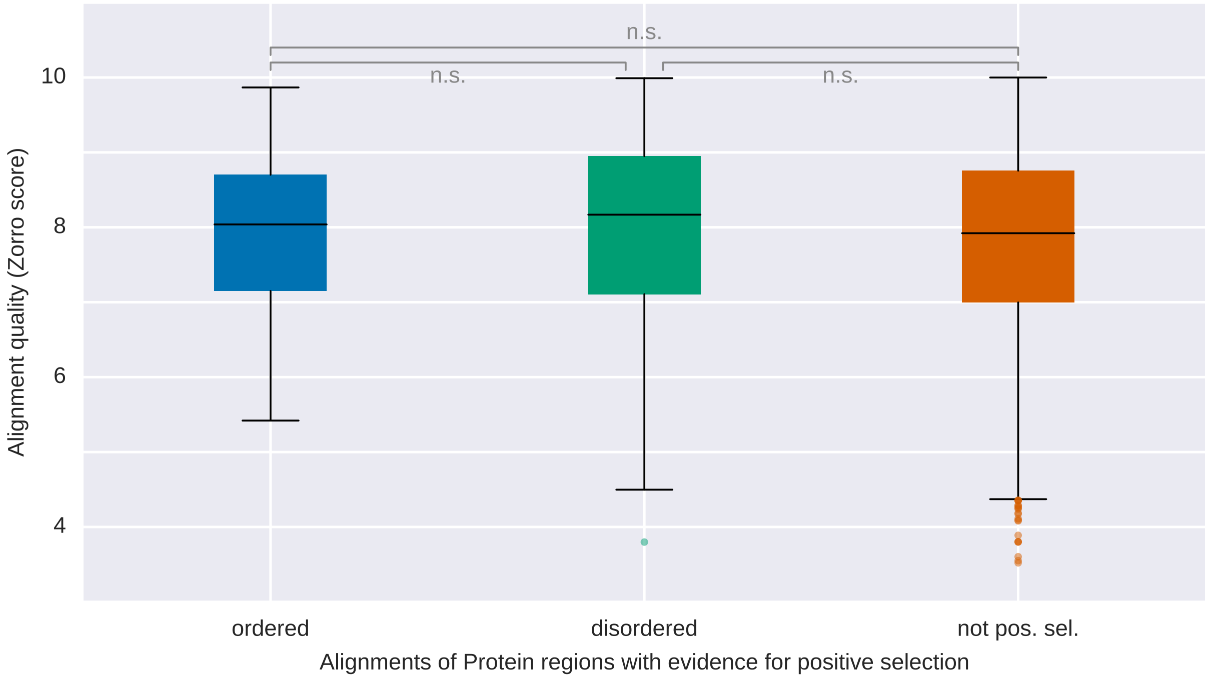

Figure S2: **Alignment quality scores before alignment pipeline was applied.** Average per site zorro scores for alignments for which positive selection was inferred in the disordered and ordered regions and for the remaining gene sets are shown. No significant difference was observed for alignment quality between the groups ($P > 0.05$, Mann-Whitney-U test).

Supplemental Figure S3

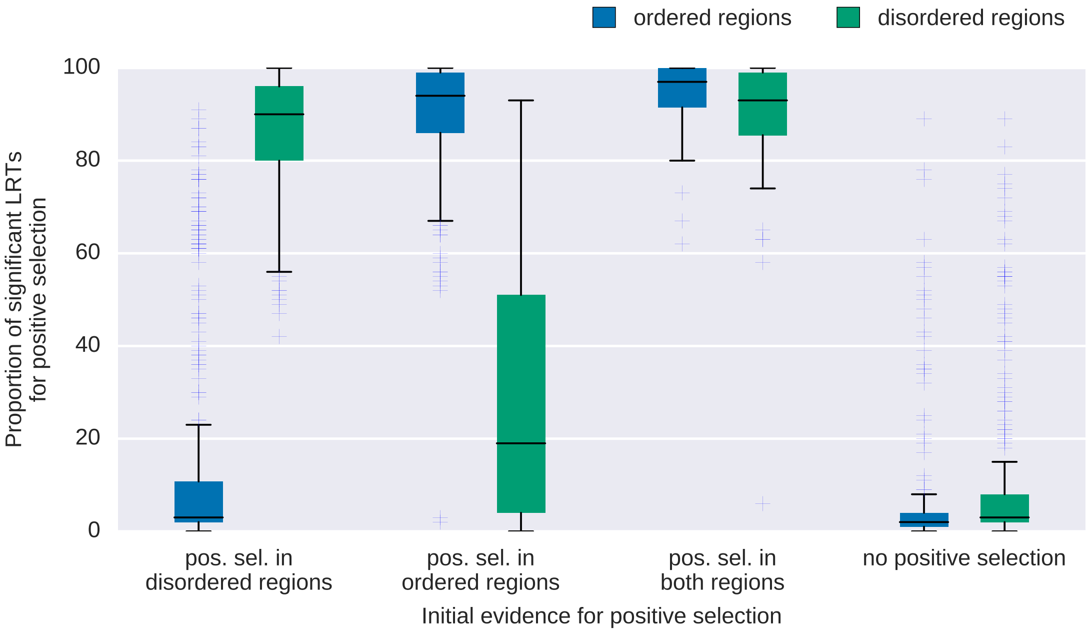

Figure S3: **Simulation studies of proteins with evidence of positive selection** and a random subset of 250 proteins without a signature of positive selection. Sets of 100 alignments per protein were simulated with INDELIBLE in a partioned model using parameters estimates from the separate codeml analysis for ordered and disordered regions**.**

Supplemental Figure S4

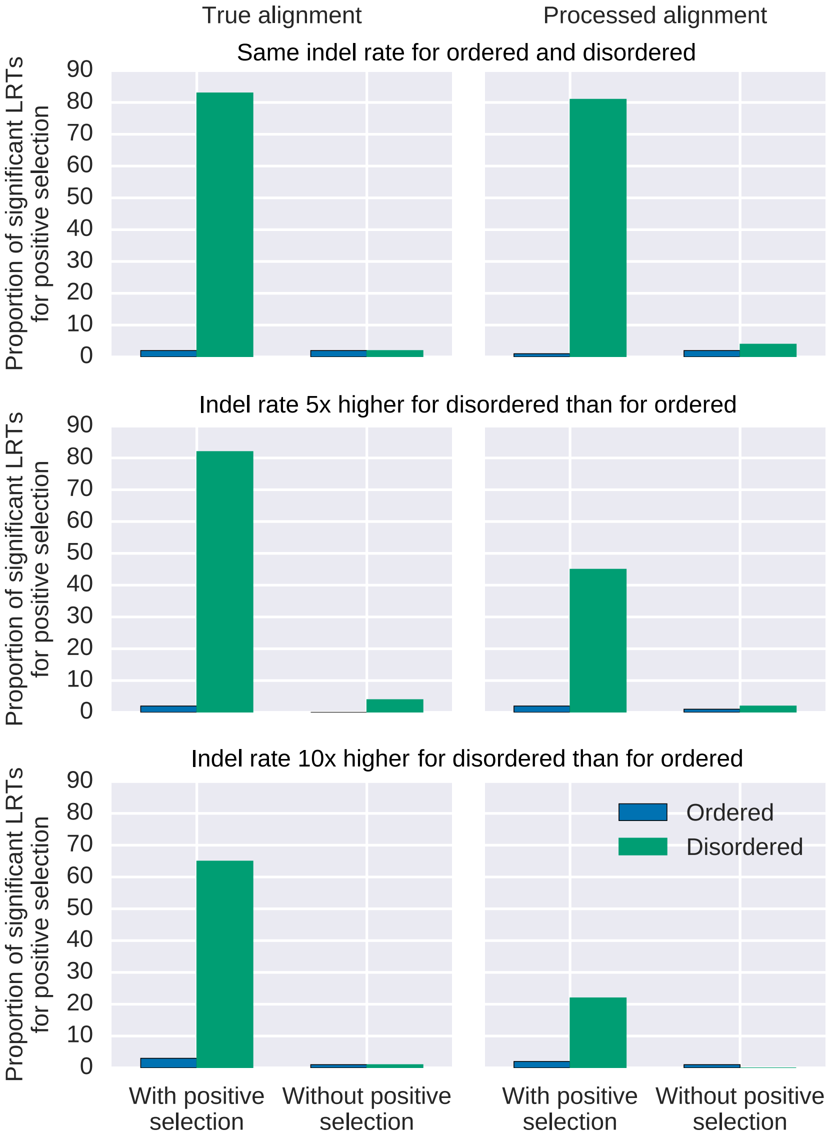

Figure S4: **Simulation studies for an artifical protein** with a disordered protein region of 250 amino acids flanked by ordered regions of 250 amino acids on each site under varying indel rates in the disordered region (1×, 5× and 10× relative to the ordered region). 100 alignments were simulated and processed for each group and positive selection was assumed for $\approx 5\%$ of the disordered sites with $\omega = 3.28$, otherwise this fraction was set no evolve neutral ($\omega = 1$).

Supplemental Figure S5

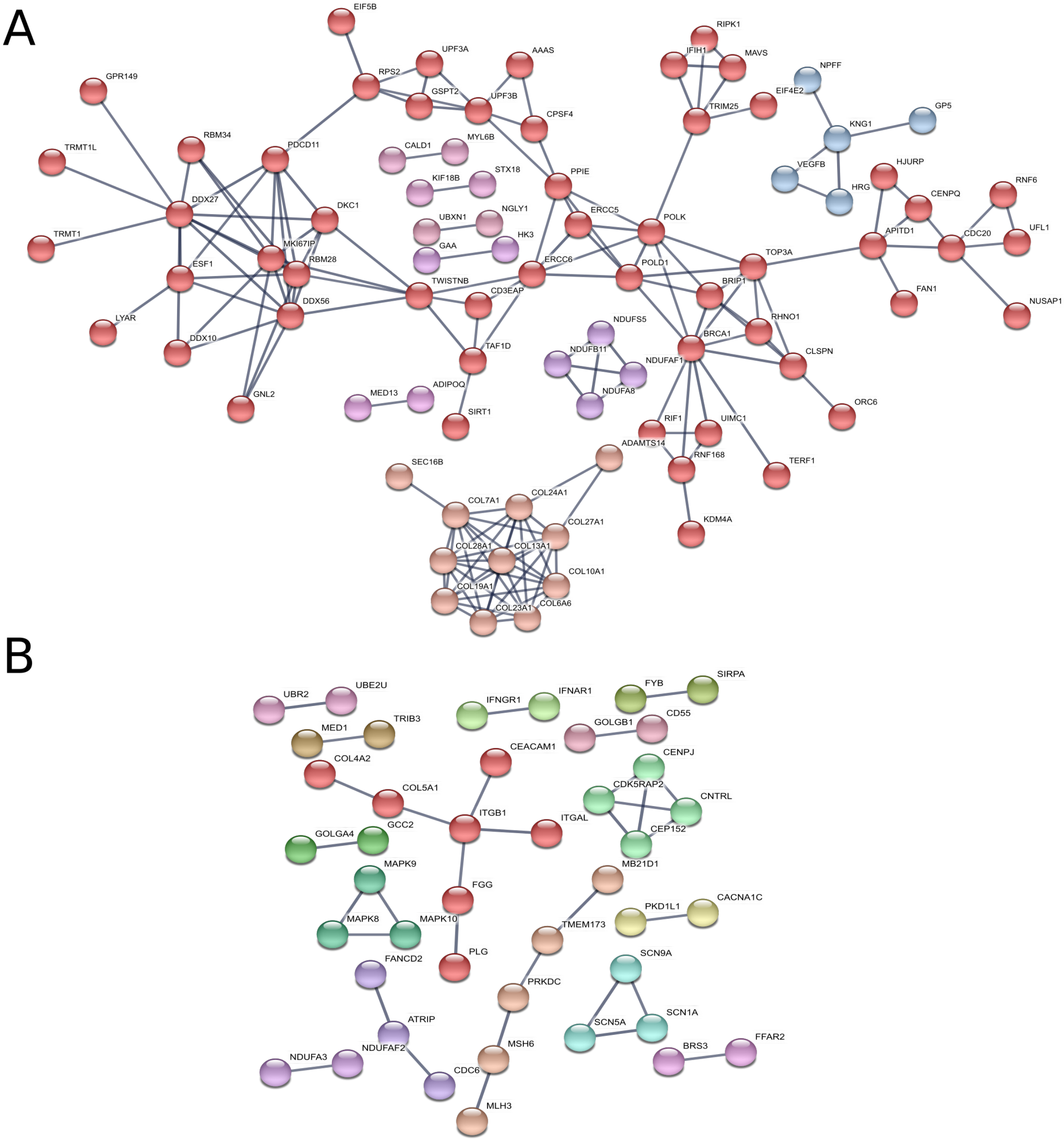

Figure S5: **Protein interaction networks of proteins for which their coding genes showed evidence for positive selection** (A) Proteins with unique evidence for positive selection in disordered regions (B) Proteins with unique evidence of positive selection in ordered regions. Colors are arbitrarily assigned based on an MCL clustering algorithm.

Supplemental Figure S6

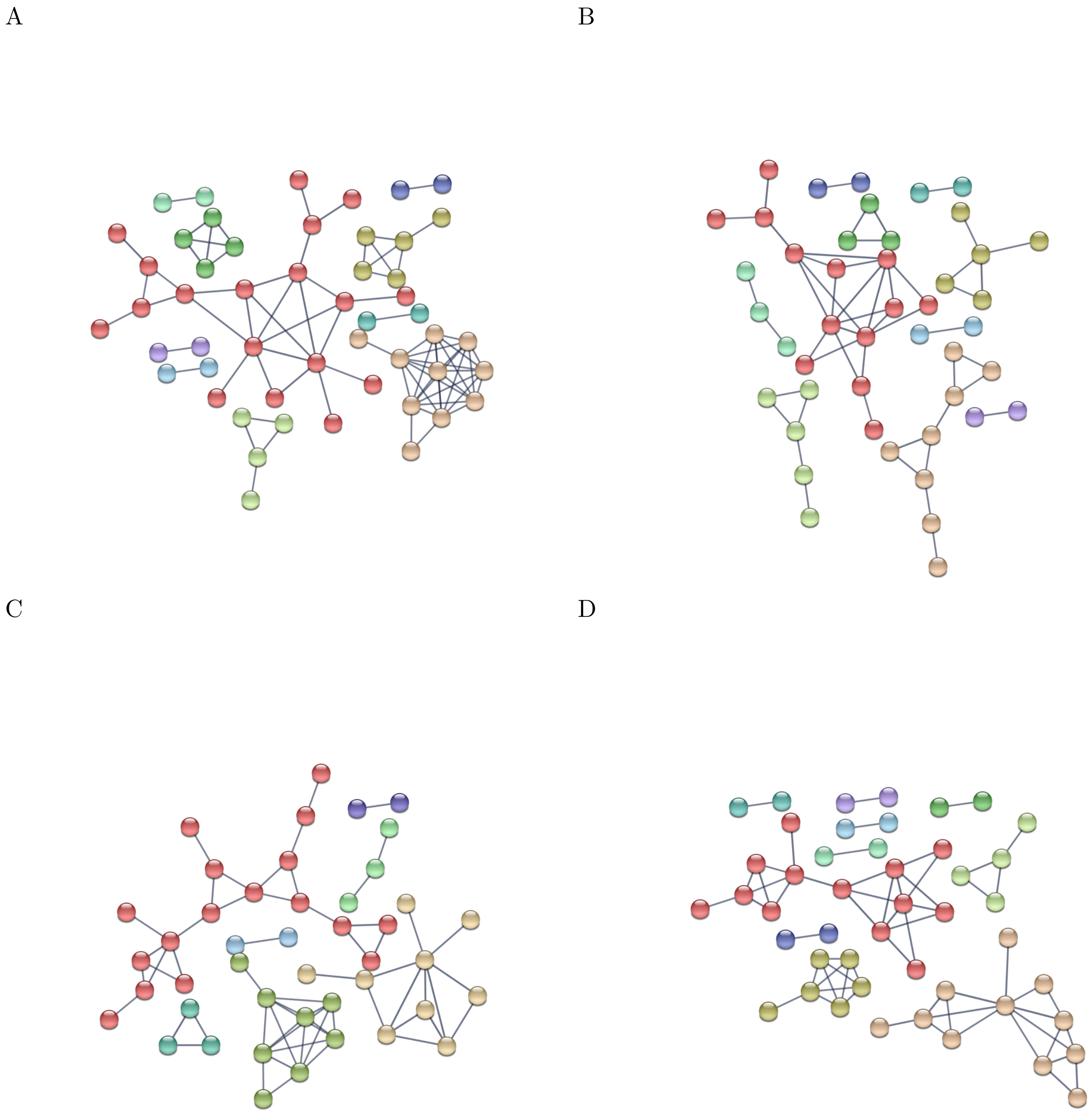

Figure S6: **Protein interaction networks of proteins for which their coding genes showed evidence for positive selection in disordered regions** (A-D) Proteins with unique evidence for positive selection in disordered regions randomly downsampled to 197 proteins to account for difference in protein numbers between the ordered and disordered sets. Maximum cluster size varies between 13 and 18

Supplemental Figure S7

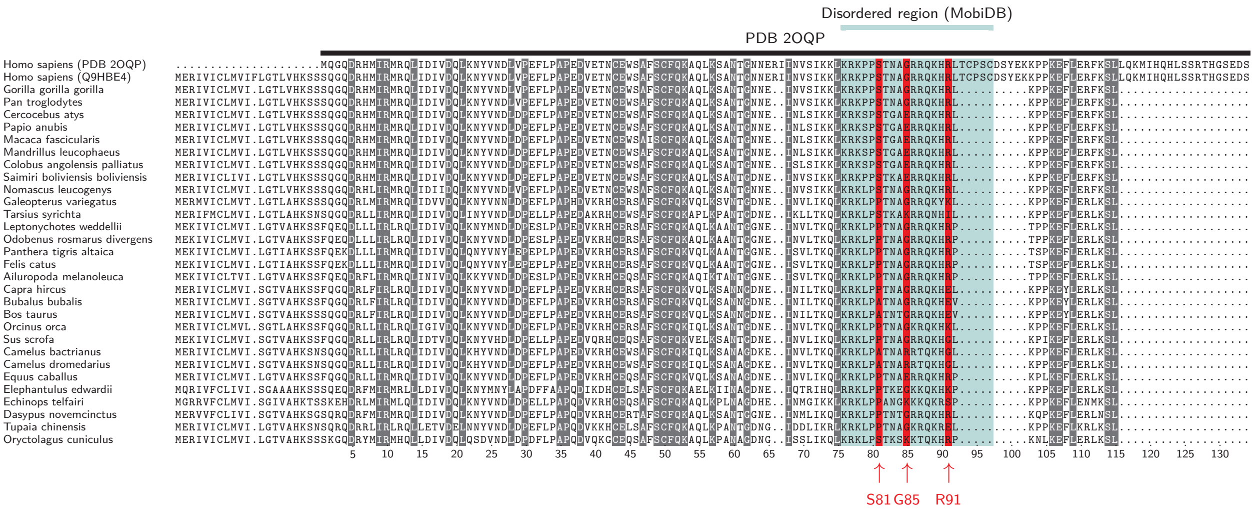

Figure S7: **Protein alignment of 30 species including human interleukin-21 used to identify sites under positive selection** The protein sequence of the NMR structure of human interleukin-21 (PDB Code: 2OQP) is shown on top and the corresponding human disordered region from MobiDB is indicated in green as well as three residues that have been identified as positively selected in a PAML branch-site test (S81, G85, and R91) in red. Note that sites have been counted relative to the PDB protein sequence and that sites in sequences from non-human species that are indicated with gaps have been excluded from the paml analysis as part of the alignment processing pipeline.

Supplemental Figure S8

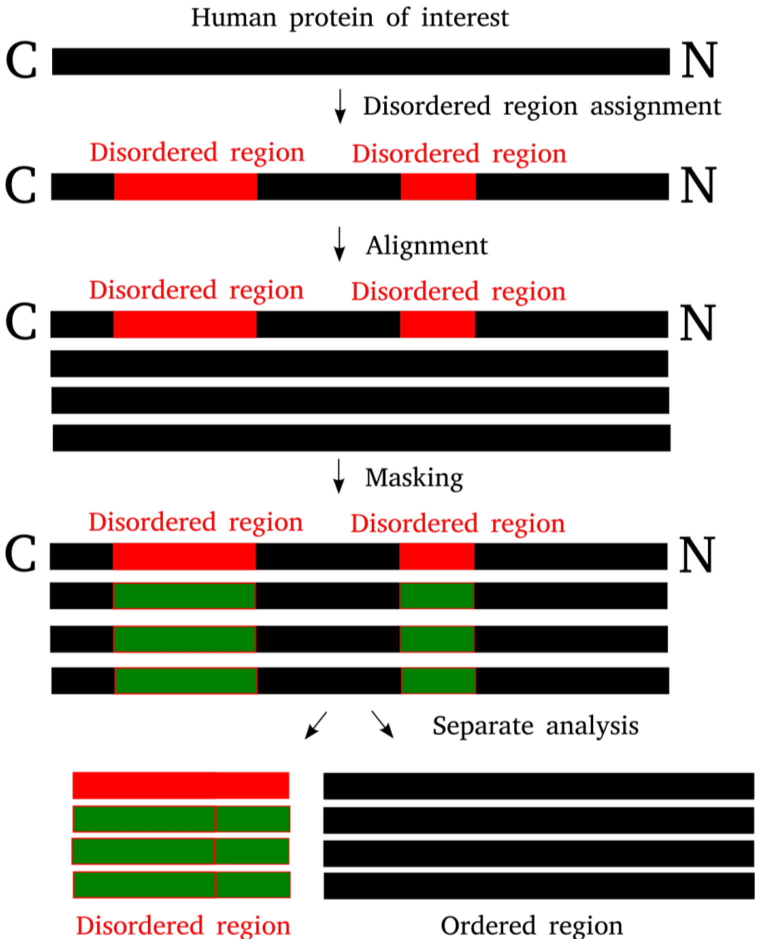

Figure S8: **Graphical outline of the analysis pipeline to obtain evolutionary rates of human disordered and ordered regions in a comparative framework.** We extracted human proteins with intrinsically disordered regions (C and N denote the C and N terminal protein regions, respectively). These were aligned with sequence orthologs from other species and human disordered residues were masked in the other sequences accordingly. To take the genomic context into account we only considered paired regions (disordered and ordered region of the same protein).

Supplemental Figure S9

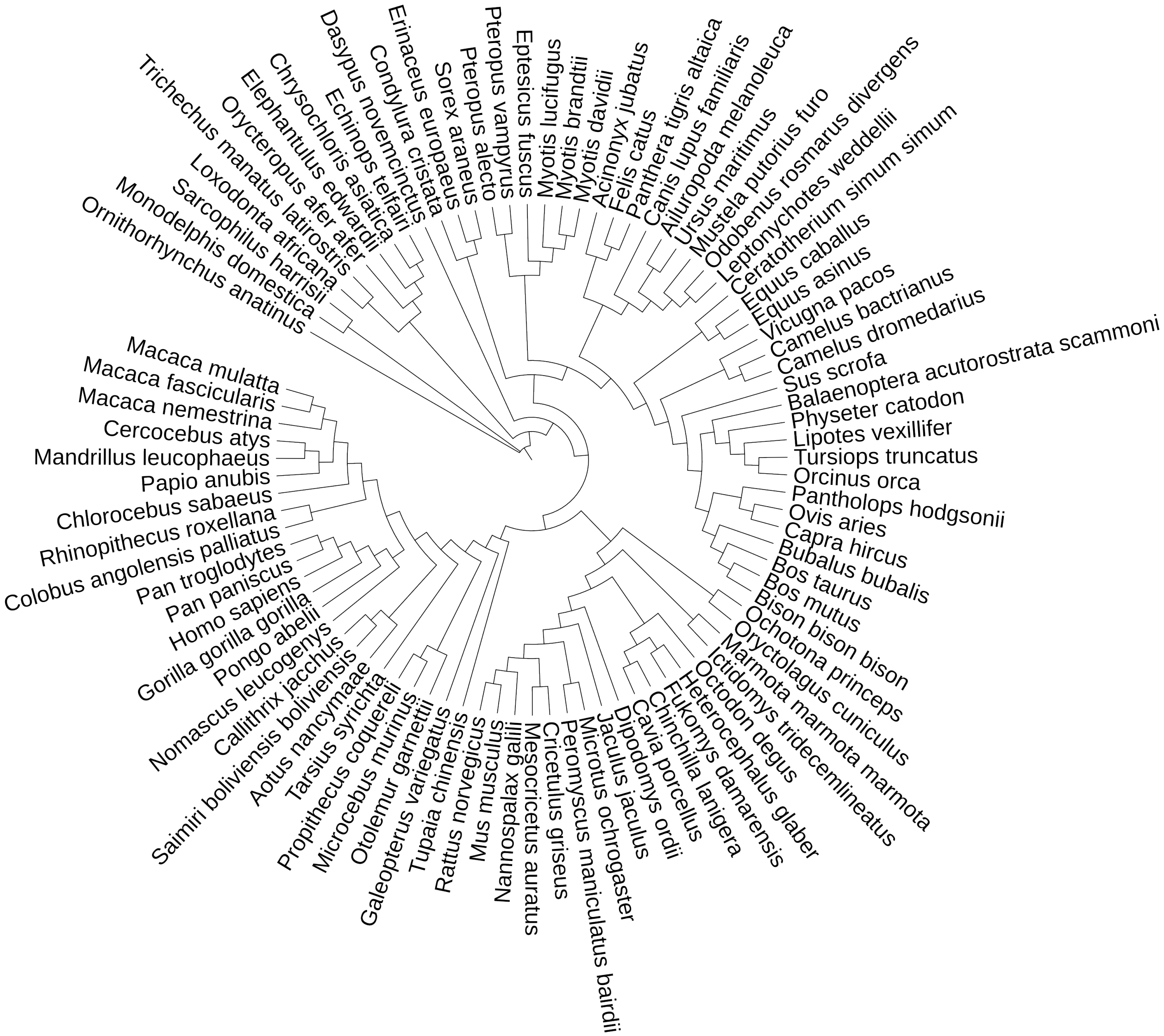

Figure S9: **Phylogenetic tree of the mammalian species used in this analysis.**

Supplemental Figure S10

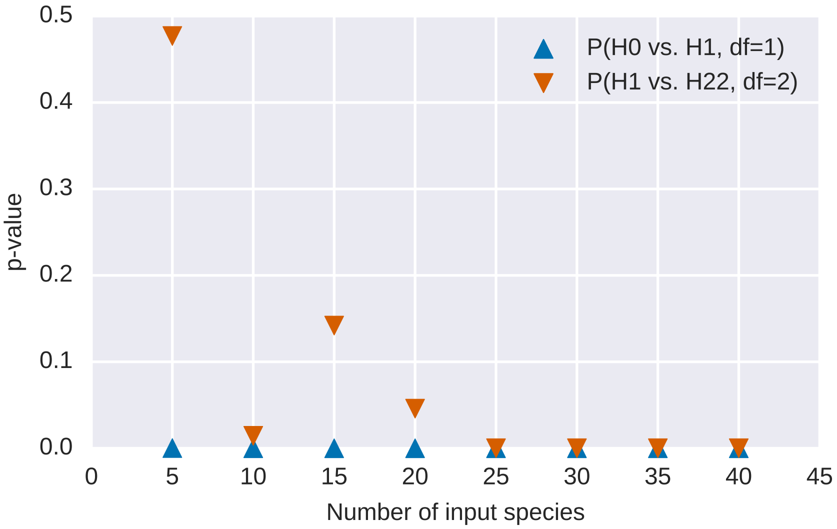

Figure S10: **Likelihood tests for phylogenetic calculations** with different numbers of randomly chosen species from the list of homologous sequences, where H0, H1 and H22 refer to different PAML site models (hypothesis) for the phylogenetic analysis (one-ratio, neutral, positive selection, respectively, **df** denotes the degrees of freedom). Starting from 20 species, *p*-values for both tests are less than 0.05, hypothesis difference becomes significant.

Supporting Tables

Supplementary Table S1

Table S1: GO enrichment analysis with PANTHER of the gene set used in phylogenetic analysis (6,663 human genes annotated in MobiDB database) in comparison to the entire human proteome. Enrichment was tested with an exact Fisher's exact test with df=2.

GO name	GO id	in background*	in analysed**	expected	Fold Enrichment	p-value
mRNA binding	GO:0003729	96	64	30.85	2.07	1.95×10^{-05}
sequence-specific DNA binding RNA polymerase II transcription factor activity	GO:0000981	190	122	61.05	2.00	5.77×10^{-10}
chromatin binding	GO:0003682	166	106	53.34	1.99	1.95×10^{-08}
small GTPase regulator activity	GO:0005083	287	166	92.21	1.80	3.66×10^{-10}
guanyl-nucleotide exchange factor activity	GO:0005085	146	81	46.91	1.73	6.34×10^{-04}
RNA binding	GO:0003723	359	195	115.35	1.69	9.96×10^{-10}
sequence-specific DNA binding transcription factor activity	GO:0003700	1167	610	374.96	1.63	1.29×10^{-28}
enzyme regulator activity	GO:0030234	678	354	217.84	1.63	6.99×10^{-16}
protein kinase activity	GO:0004672	406	207	130.45	1.59	4.47×10^{-08}
DNA binding	GO:0003677	1392	704	447.25	1.57	2.23×10^{-29}
nucleic acid binding	GO:0003676	2080	1042	668.31	1.56	9.48×10^{-44}
transcription factor binding transcription factor activity	GO:0000989	231	114	74.22	1.54	1.70×10^{-03}
protein binding transcription factor activity	GO:0000988	232	114	74.54	1.53	2.04×10^{-03}
transcription cofactor activity	GO:0003712	222	109	71.33	1.53	3.21×10^{-03}
kinase activity	GO:0016301	573	251	184.11	1.36	2.07×10^{-04}
binding	GO:0005488	5024	2064	1614.22	1.28	1.00×10^{-33}
protein binding	GO:0005515	2607	1009	837.64	1.20	7.07×10^{-08}

* in background is the number of human proteins in the PANTHER database with this GO term (total 21002), ** in analysed is the number of analysed proteins with this GO term (total 6663)

Supplementary Table S2

Table S2: Number of amino acids with structural features based on SASA scores in ordered and disordered regions of the analysed proteins. Structural information was available for 2115 of the 6663 proteins. The *p*-value is based on a χ^2 test of the two-by-two matrix.

Site category	ordered	disordered	p-value
Surface	189163	7652	
Core	154962	820	
—————-	———	————-	———-
Surface to Core ratio	1.22	9.33	0

Supplementary Table S3

Table S3: Proportion of number of amino acids with Uniprot regional features in ordered and disordered regions of analysed proteins. Uniprot feature information was available for 6649 of the 6663 proteins. The *p*-value is based on the χ^2-test of the two-by-two matrix of the raw counts. The enriched pair is marked with an asterisk. Regions and motifs are protein regions (longer or shorter than 20 amino acids, respectively) with a biological significance. PTMs are single amino acids that may undergo post-translational modification. Site are single amino acids with biological significance that are not PTMs.

Annotation type	ordered	disordered	p-value
Region	0.10640	0.10907*	7×10^{-08}
Motif	0.00245	0.00505*	2×10^{-273}
PTM	0.00027	0.00067*	4×10^{-58}
Site	0.00058*	0.00003	2×10^{-73}

Supplementary Table S4

Table S4: Number of amino acids in ordered and disordered regions with an annotated disease association. The *p*-value is based on a χ^2 test, given that there is a total number of 641061 ordered and 91845 disordered sites. The enriched pair is marked with an asterisk. † Not enough counts for χ^2 test. ICD 10 - International Statistical Classification of Diseases and Related Health Problems

Disease type	ordered	disordered	*p*-value
Blood	410*	11	1×10^{-09}
Chromosomal	29	2	$-\dagger$
Circulatory	472*	58	3×10^{-01}
Digestive	134*	6	5×10^{-03}
Endocrine	1154*	35	3×10^{-23}
EyeEar	690*	20	8×10^{-15}
Genitourinary	214*	15	8×10^{-03}
Infection	4	0	$-\dagger$
Mental	155*	19	6×10^{-01}
Musculoskeletal	594	131*	9×10^{-06}
Neoplasm	367*	22	6×10^{-05}
Nervous	1807*	83	1×10^{-26}
OtherCongenital	1188*	100	3×10^{-07}
Pregnancy	5	1	$-\dagger$
Respiratory	21	0	$-\dagger$
Skin	23	0	$-\dagger$
Surgery	80	4	$-\dagger$
No ICD10	294*	20	1×10^{-03}

Supplementary Table S5

Table S5: Protein identifiers (HGNC), Uniprot Ids as well ω values for proteins with a signature of positive selection in ordered and disordered region

Table provided as Supplementary Excel file (Supplementary Table S5)

Supplementary Table S6

Table S6: Molecular functional enrichment categories (FDR<0.01) in the STRING network analyses (Supplemental Fig. S5) of proteins identified to be evolving under positive selection for disordered and ordered protein regions.

pathway ID	pathway description	count in network	false discovery rate
Disordered			
GO:0003723	RNA binding	48	0.00393
GO:0003676	nucleic acid binding	87	0.00927
Ordered			
GO:0002682	regulation of immune system process	33	0.00104
GO:0050776	regulation of immune response	24	0.00133
GO:0022407	regulation of cell-cell adhesion	15	0.00149
GO:0050863	regulation of T cell activation	13	0.00149
GO:0031347	regulation of defense response	22	0.0015
GO:0022610	biological adhesion	25	0.00254
GO:0051607	defense response to virus	10	0.00295
GO:0002697	regulation of immune effector process	15	0.00425
GO:0007155	cell adhesion	24	0.00468
GO:0050670	regulation of lymphocyte proliferation	10	0.00468
GO:0045088	regulation of innate immune response	14	0.00504
GO:0002252	immune effector process	15	0.00683
GO:0009615	response to virus	11	0.00686

Supplemental References

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic acids research* **28**: 235–242.

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. *The Journal of biological chemistry* **282**: 23326–23336.

De Mita S, Siol M. 2012. EggLib: Processing, analysis and simulation tools for population genetics and genomics. *BMC genetics* **13**: 27.

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

Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of molecular biology* **347**: 827–839.

Duarte JM, Srebniak A, Schärer MA, Capitani G. 2012. Protein interface classification by evolutionary analysis. *BMC bioinformatics* **13**: 1.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**: 1792–1797.

Eyre-Walker A. 2006. The genomic rate of adaptive evolution. *Trends in ecology & evolution* **21**: 569–575.

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

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nature reviews Genetics* **8**: 610–618.

Fletcher W, Yang Z. 2009. INDELible: A flexible simulator of biological sequence evolution. *Molecular biology and evolution* **26**: 1879–1888.

Gossmann TI, Schmid KJ. 2011. Selection-driven divergence after gene duplication in arabidopsis thaliana. *Journal of molecular evolution* **73**: 153–165.

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. *Molecular biology and evolution* **27**: 1822–1832.

Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. 2015. Genenames.org: The hgnc resources in 2015. *Nucleic acids research* **43**: D1079–D1085.

Junier T, Zdobnov EM. 2010. The Newick utilities: High-throughput phylogenetic tree processing in the UNIX shell. *Bioinformatics* **26**: 1669–1670.

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.

Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003a. Protein disorder prediction: Implications for structural proteomics. *Structure* **11**: 1453–1459.

Linding R, Russell RB, Neduva V, Gibson TJ. 2003b. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic acids research* **31**: 3701–3708.

Liu Y, Schmidt B, Maskell DL. 2010. MSAProbs: Multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* **26**: 1958–1964.

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

Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD. 2016. PANTHER version 10: Expanded protein families and functions, and analysis tools. *Nucleic acids research* **44**: D336–D342.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. 2016. Reference sequence (refseq) database at ncbi: Current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**: D733–D745.

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 biology* **12**: e1001775.

Salomon-Ferrer R, Case DA, Walker RC. 2013. An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**: 198–210.

Shrake A, Rupley J. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of molecular biology* **79**: 351–371.

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, et al. 2007. DisProt: The database of disordered proteins. *Nucleic acids research* **35**: D786–D793.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34**: W609–W612.

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 research* **45**: D362–D368.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* **56**: 564–577.

Tan K, Duquette M, Joachimiak A, Lawler J. 2009. The crystal structure of the signature domain of cartilage oligomeric matrix protein: Implications for collagen, glycosaminoglycan and integrin binding. *The FASEB Journal* **23**: 2490–2501.

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

Vucetic S, Brown CJ, Dunker AK, Obradovic Z. 2003. Flavors of protein disorder. *Proteins: Structure, Function, and Bioinformatics* **52**: 573–584.

Walsh I, Martin AJ, Di Domenico T, Tosatto SC. 2012. ESpritz: Accurate and fast prediction of protein disorder. *Bioinformatics* **28**: 503–509.

Wu M, Chatterji S, Eisen JA. 2012. Accounting for alignment uncertainty in phylogenomics. *PloS one* **7**: e30288.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**: 1586–1591.

Yang ZR, Thomson R, Mcneil P, Esnouf RM. 2005. RONN: The bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* **21**: 3369–3376.

Yuan Z, Bailey TL, Teasdale RD. 2005. Prediction of protein B-factor profiles. *Proteins* **58**: 905–912.