RESEARCH ARTICLE

REVISED Disease association and comparative genomics of compositional bias in human proteins [version 2; peer review: 2 approved]

Christos E. Kouros[1], Vasiliki Makri[1], Christos A. Ouzounis[1,2], Anastasia Chasapi[2]

[1]BCCB-AIIA, School of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
[2]BCPL, Chemical Process & Energy Resources Institute, Centre for Research & Technology Hellas (CERTH), Thessaloniki, Greece

v2 **First published:** 20 Feb 2023, **12**:198
https://doi.org/10.12688/f1000research.129929.1
Latest published: 14 Apr 2023, **12**:198
https://doi.org/10.12688/f1000research.129929.2

Abstract

Background: The evolutionary rate of disordered protein regions varies greatly due to the lack of structural constraints. So far, few studies have investigated the presence/absence patterns of compositional bias, indicative of disorder, across phylogenies in conjunction with human disease. In this study, we report a genome-wide analysis of compositional bias association with disease in human proteins and their taxonomic distribution.
Methods: The human genome protein set provided by the Ensembl database was annotated and analysed with respect to both disease associations and the detection of compositional bias. The Uniprot Reference Proteome dataset, containing 11297 proteomes was used as target dataset for the comparative genomics of a well-defined subset of the Human Genome, including 100 characteristic, compositionally biased proteins, some linked to disease.
Results: Cross-evaluation of compositional bias and disease-association in the human genome reveals a significant bias towards biased regions in disease-associated genes, with charged, hydrophilic amino acids appearing as over-represented. The phylogenetic profiling of 17 disease-associated, proteins with compositional bias across 11297 proteomes captures characteristic taxonomic distribution patterns.
Conclusions: This is the first time that a combined genome-wide analysis of compositional bias, disease-association and taxonomic distribution of human proteins is reported, covering structural, functional, and evolutionary properties. The reported framework can form the basis for large-scale, follow-up projects, encompassing the entire human genome and all known gene-disease associations.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 14 Apr 2023	✓ view	✓ view
version 1 20 Feb 2023	✓ view	? view

1. **Stella Tamana**, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus

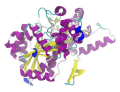
2. **Gregory Amoutzias**, University of Thessaly, Larissa, Greece

Any reports and responses or comments on the article can be found at the end of the article.

Keywords
disease-associated gene, low complexity, compositional bias, intrinsically disordered protein (IDP), intrinsically disordered region (IDR), phylogenetic profile, human genome, human disease

This article is included in the HEAL1000 gateway.

This article is included in the Genomics and Genetics gateway.



This article is included in the Structural & Comparative Genomics collection.

Corresponding author: Anastasia Chasapi (chasapia@gmail.com)

Author roles: Kouros CE: Data Curation, Formal Analysis, Investigation, Methodology, Software; **Makri V**: Data Curation, Investigation; **Ouzounis CA**: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – Review & Editing; **Chasapi A**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: This research was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the project "Reinforcement of Postdoctoral Researchers - 2nd Cycle" (MIS-5033021), implemented by the State Scholarships Foundation (IKY). The work was also supported by Elixir-GR (grant # MIS 5002780), implemented under the Action "Reinforcement of the Research & Innovation Infrastructure," funded by the Operational Program Competitiveness, Entrepreneurship, & Innovation (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

How to cite this article: Kouros CE, Makri V, Ouzounis CA and Chasapi A. **Disease association and comparative genomics of compositional bias in human proteins [version 2; peer review: 2 approved]** F1000Research 2023, **12**:198 https://doi.org/10.12688/f1000research.129929.2

First published: 20 Feb 2023, **12**:198 https://doi.org/10.12688/f1000research.129929.1

REVISED **Amendments from Version 1**

Version 2 attempts to address all comments from the reviewers for version 1. Specifically, we have adjusted the abstract to include only the term of compositional bias, when referring to protein disorder, instead of several terms that were used before. The introduction has been enriched to accommodate the reviewers' suggestions, including several references and a few sentences regarding the correlation between CB and disease-association. Moreover, a new paragraph has been added, providing a concrete example of intrinsic disorder association to human disease. Some explanatory details have been added to the methods and results. Table 1 legend has been adjusted for precision, Figure 1 legend was enriched, Figure 3 legend was significantly changed for clarity. One of the keywords was adjusted and a new, enriched dataset was deposited in Zenodo accompanied by an updated DOI which is provided in Version 2. The extended data file has also been updated and a new DOI is provided.

Any further responses from the reviewers can be found at the end of the article

Introduction

Disordered proteins exhibit specific patterns at the sequence level

The classical view that protein function requires a defined three-dimensional (3D) structure has been challenged by recent developments where many proteins and protein regions are shown to perform distinct biological functions, despite their propensity for disordered conformations (Bürgi *et al.*, 2016; Ntountoumi *et al.*, 2019; Tantos *et al.*, 2012). These intrinsically disordered proteins (IDPs) and intrinsically disordered protein regions (IDRs) are defined as lacking a precise 3D folding pattern. The difference between ordered and disordered proteins is already reflected at the primary structure level, with IDPs being characterized by regions typically enriched in specific amino acids, resulting in an overall low sequence complexity. Specifically, IDPs and IDRs contain substantially fewer residues that promote order (typically C, F, I, L, N, V, W and Y) and are substantially enriched in residues that promote disorder (typically A, E, G, K, P, Q, R, and S) (Williams *et al.*, 2000; Dunker *et al.*, 2001; Harbi *et al.*, 2011). These regions are characterized as low complexity of compositionally biased regions, both terms indicating disorder (Dunker *et al.*, 2001). Residues that promote order are typically found within the hydrophobic cores of foldable proteins, as opposed to residues that promote disorder which are mostly located at solvent-exposed surfaces (Theillet *et al.*, 2013; Uversky, 2013). When exposed, hydrophobic-enriched regions can induce self-aggregation and/or intermolecular interactions with surrounding proteins, triggering aggregate formation (Grignaschi *et al.*, 2018).

IDP/IDR identification

With the increasing number of predicted and experimentally validated IDPs and proteins containing IDRs, disordered proteins and regions are no longer considered as exceptions, but rather the object of extensive study with regard to their structure and function. A wide range of disorder predictors has been successfully developed over the past years, adopting different approaches such as Compositional Bias Detection (Harrison, 2021; Promponas *et al.*, 2000; Wootton & Federhen, 1993), residual energy-based disorder prediction (Dosztanyi *et al.*, 2009, 2005) and others (Linding *et al.*, 2003; Tang *et al.*, 2021; Walsh *et al.*, 2012; Wang *et al.*, 2016; Zhang *et al.*, 2012). Integrative tools have made their appearance, such as MobiDB-lite (Necci *et al.*, 2017), a data fusion tool making use of eight distinct predictors. The prediction accuracy of such tools varies greatly, with deep learning-based methods typically outperforming methods based on physicochemical characteristics (CAID Predictors *et al.*, 2021). DisProt, a manually curated, dedicated database for IDPs (Sickmeier *et al.*, 2007) has developed into the main resource for IDP/IDR information (Hatos *et al.*, 2019; Quaglia *et al.*, 2022).

IDPs, phylogeny and disease

Multiple computational and experimental analyses of a wide range of species at the genome level have established widespread presence of intrinsic disorder across the tree of life (Hatos *et al.*, 2019; Ntountoumi *et al.*, 2019; Peng *et al.*, 2015; Ward *et al.*, 2004). In fact, proteins at all taxonomic levels, including viruses, exhibit noticeable intrinsic disorder that apparently increases with organism complexity. Disorder presence is particularly prominent in eukaryotes, in which at least half of their genome-encoded proteins possess long IDRs (Ahrens *et al.*, 2017; Basile *et al.*, 2019; Peng *et al.*, 2015; Ward *et al.*, 2004; Xue *et al.*, 2012). This high prevalence of IDPs and IDRs in eukaryotes indicates that key functions, such as cell signalling and regulation, are dynamically associated with intrinsic disorder in nucleated cells (Bürgi *et al.*, 2016; Tantos *et al.*, 2012).

The same trend holds for an ever-increasing emergence of disease-associated genes in more recent speciation events (Dickerson & Robertson, 2012; Lopez-Bigas & Ouzounis, 2004), raising the question whether specific residues can be directly implicated in particular diseases. A correlation between intrinsic disorder and various human diseases such as cancer, diabetes, amyloidosis, and neurodegenerative diseases has already been established in specific cases (Choudhary *et al.*, 2022; Monti *et al.*, 2021, 2022), and is emerging as a significant biomedical research endeavour.

A typical example of intrinsic disorder association with human disease is the role of α-synuclein protein in Synucleinopathies such as Parkinson's disease. A-synuclein is a small IDP protein that plays a role in the regulation of neurotransmitters, and has tremendous conformational plasticity (Uversky *et al.*, 2008). In Parkinson's disease, alpha-synuclein misfolds and aggregates to form Lewy bodies, which are pathological hallmarks of the disease. The intrinsic disorder of alpha-synuclein is thought to play a key role in its misfolding and aggregation, which leads to the death of dopaminergic neurons and the onset of Parkinson's disease symptoms (Breydo *et al.*, 2012).

Due to a lack of structural constraints, the evolutionary rate of disordered proteins varies, with some IDPs/IDRs being highly conserved while others appearing rapidly evolving (Brown *et al.*, 2011; Khan *et al.*, 2015; Xue *et al.*, 2013). So far, few studies have investigated the phylogenetic profiling of IDRs in conjunction with human disease (Pajkos *et al.*, 2020). To assess this hypothesis, we use a curated list of 100 annotated proteins from the human genome with well-characterised compositionally biased regions (CBRs) (Mier *et al.*, 2020), as a first step for the comparative genomics of compositionally biased genes, some of which are in fact disease associated. We identify those instances known to be linked with human disease and assess their phylogenetic depth. This framework, with human queries against multiple species, forms the basis for follow-up, large-scale studies that would encompass the entire human genome and all known gene-disease associations.

Methods

Data compilation

The Human Genome protein set recorded in the Ensembl database (GRCh38.p13) was retrieved, containing ~119K gene transcripts to be used as reference (Yates *et al.*, 2016).

For the disease mapping on gene transcripts, the DISEASES database was chosen, which integrates disease-gene associations derived from text mining, as well as manually curated disease–gene associations, cancer mutation data, and genome-wide association studies from existing databases (Pletscher-Frankild *et al.*, 2015). Specifically, the "Knowledge channel" was selected, containing manually curated associations from the Genetics Home Reference (GHR) (Koos & Bassett, 2018) and UniProtKB (The UniProt Consortium *et al.*, 2022), a total of 7269 disease-gene, high-confidence associations, regarding 3837 genes and 1097 unique disease identifiers.

Disease associations are provided with the use of Disease Ontology identifiers (DOID) (Schriml *et al.*, 2019). For each entry of the Ensembl dataset, DOIDs were mapped from the DISEASES knowledge channel dataset and added to the header description of the corresponding gene transcript.

For phylogenetic analysis, the Uniprot Reference Proteome (URP) dataset was selected, containing a total of 11297 proteomes, excluding viruses. The URP set has been selected manually and algorithmically among all proteomes, to provide broad coverage of the tree of life, representing the taxonomic diversity found within UniProtKB and including the proteomes of well-studied model organisms and other proteomes of interest for biomedical research (Chen *et al.*, 2011). Specifically, the URP (version: Reference_Proteomes_2022_04) contains 349 Archaeal, 8763 Bacterial and 2185 Eukaryotic proteomes.

The low complexity query set investigated both for its disease association and phylogenetic depth was previously recorded (Mier *et al.*, 2020) and contains 100 human proteins with characteristic compositional bias.

Data transformation

The computational pipeline **cogent_utils,** part of CGG toolkit v1.0.1. (Vasileiou et al, submitted), was used to create a CoGenT-style sequence collection (Janssen *et al.*, 2003) from Ensembl GRCh38.p13 as well as the URP, selected as a robust and convenient identifier encoding scheme both for human interpretation and programming convenience. Specifically, cogent_utils enables header modification for all entries of FASTA sequence files, based on user-defined criteria. Below we present the example of the oleosine protein of *Camellia sinensis*, as it appears originally in URP and after cogent_utils transformation:

URP original header

>tr|A0A7J7IAQ7|A0A7J7IAQ7_CAMSI Oleosin OS=Camellia sinensis OX=4442 GN=HYC85_002860 PE=3 SV=1

Modified header

>UP000593564-00004442-Came_sine-22-000001-E-000699 tr|A0A7J7IAQ7|A0A7J7IAQ7_CAMSI Oleosin OS=Camellia sinensis OX=4442 GN=HYC85_002860 PE=3 SV=1

The first part of the header has been added, and corresponds to the following format: [URP identifier]-[NCBI Taxonomy ID]-[organism name]-[URP year release]-[proteome counter]-[taxonomic domain]-[protein counter].

MagicMatch v1.0.1 (Smith *et al.*, 2005) was used for sequence matching across databases to verify the identity of the reference proteome collection against the modified identifier space. MagicMatch enables sequence-identity check at 100%, based on sequence only, with no comparison of identifiers.

Masking, searching, phylogenetic profiling

For the detection of compositional bias as a proxy for low-complexity sequence tracts, we deployed **CAST v1.0.1** (Promponas *et al.*, 2000), for all protein sequences of the human genome. The CAST algorithm was applied on the DOID annotated Ensembl FASTA format dataset using default parameters, i.e. threshold score 40 for reported regions. The outcome of the analysis were 2 files dividing the original dataset; one containing all entries where low complexity regions were detected and one containing all remaining entries.

Searching with query datasets against Proteomes for the creation of phylogenetic profile patterns was performed with **DIAMOND** blastp (Buchfink *et al.*, 2021), using the URP dataset as target database and adjusting the alignment algorithm to enable compositional bias statistics (option: --comp-based-stats 3), conditioned on sequence properties (Yu & Altschul, 2005). All hits considered as significant recorded an E-value<0.001 and exhibit sequence similarities of 21% and above, without consideration to alignment geometry, e.g. coverage.

For the calculation of amino acid frequencies across the Ensembl protein set, the BioPython Bio.SeqUtils.ProtParam module (Cock *et al.*, 2009) was used, which takes input files of sequences (typically FASTA or FASTQ), counts all the letters in each sequence, and returns a summary table of their counts and percentages. The output was used for data normalisation as explained in Figure 1.

The phylogenetic profile heatmap (Figure 4) was produced using the heatmap3 R library (Zhao *et al.*, 2014) with default dissimilarity matrix calculation parameters. The plotted values have been normalized using the scale = row parameter, where each row is scaled to have a mean of 0 and standard deviation of 1.

The 2×2 chi-square test, comparing low complexity presence in protein transcripts and disease-association (Table 1) was performed with 0.01 significance threshold and no Yates continuity correction.

Lifemap, an interactive cartography-type tool to explore the NCBI taxonomy was chosen for the visualisation of the taxonomic distribution of data subsets (de Vienne, 2016). For each visualisation, a list of the NCBI Taxonomy IDs of interest were used as input for the tool, which were retrieved, in each case, from the phylogenetic profiling hit list. NCBI taxonomy ID visualisations are provided for all UPR hits of ALG13, SIX3 and RP9 (Figure 5).

Results

Disease association across the human genome

The dataset upon which all transformations and analyses were performed was the Ensembl Human Genome export (GRCh38.p13), containing 119068 gene transcripts, corresponding to 23506 genes. The dataset was annotated with regard to disease association, using curated associations from GHR and UniProtKB, which are indexed in the DISEASES database (Pletscher-Frankild *et al.*, 2015). Of these, 3625 transcripts are confidently associated with disease, whereas the remaining 115443 are not verified for any strong disease association in the "knowledge channel" of DISEASES.

To remove noise, e.g. putative or alternative mini-transcripts (some with multiple stop codons), the Ensembl dataset was filtered and all transcripts with length <80 amino acid residues were removed, with the exception of short transcripts with at least one disease (i.e. DOID) association. The filtered set contains 102702 transcripts, which include all 3625 instances associated with disease (Table 1).

Table 1. Contingency table between disease association in gene transcripts presenting compositionally biased regions versus gene transcripts without any detectable compositional bias.

	Low complexity	High complexity	
Non-disease	36250	62827	99077
Disease	1845	1780	3625
	38095	64607	102702

Compositional bias and human disease

For the evaluation of low complexity presence in the transcripts of the human genome we performed compositional bias detection using CAST (Promponas *et al.*, 2000). Out of the 102702 transcripts of the filtered Ensembl human genome dataset, compositional bias was detected in 38095 instances, with at least one compositionally biased sequence tract. Cross-evaluation of compositional bias and disease-association presence in the dataset using chi-square test of independence, revealed a significant bias towards low complexity regions in disease-associated, X^2 (1, N = 102702) = 306.8467, p-value < 1e-5, with an enrichment fold of 1.8 ((1845/1780) / (36250/62827) (Table 1). This significant pattern alone provides a strong indication for the involvement of low complexity in human disease on genome scale, seen here for the first time, complementing previous, well-established classifications of protein structure and function (Ouzounis *et al.*, 2003).

Examination of the low complexity gene dataset features highlighted the significant difference among amino acid-related, low complexity frequencies. Figure 1 shows the amino acid-specific rich regions, expressed by the sum of CAST scores for each compositionally biased region and normalised with respect to general amino acid frequency in the human genome as calculated from the filtered Ensembl dataset using Biopython protein analysis modules (Cock *et al.*, 2009). Charged, hydrophilic residues (E, K, Q, S, T) appear over-represented, while hydrophobic, order-promoting amino acids are less frequent, in agreement with what is known about IDP/IDR composition (Williams *et al.*, 2000; Dunker *et al.*, 2001; Harbi *et al.*, 2011). The striking over-representation of serine/threonine (S/T) tracts, along with glutamate/glutamine (E/Q) and proline (P) followed by lysine (K) is indicative of the main residue types that might affect functional properties of human proteins, including their potential association with known phenotypes, such as polyglutamine tracts with neurodegenerative diseases (Bunting *et al.*, 2022). Indeed, it has been shown that if certain pH and temperature conditions are met, tandem repeats of short oligopeptides containing glycine, proline, serine and threonine can form flexible structures that bind ligands (Matsushima *et al.*, 2008; Williamson *et al.*, 1994).

For the assessment of the relationship among disease association and compositional bias across the human proteome, the associated DOID vector for each amino acid enriched region was used as a multidimensional clustering parameter for Principal Component Analysis (PCA) (Figure 2). Consistent with the above, the presence of amino acid types in low complexity regions (e.g. S, E, P, Q) exhibit the highest contribution to the main principal components with regard to disease association, thus amplifying the link between low complexity and disease and establishing a direction for further study.

Phylogenetic profiling of disease-associated LC proteins

To further our investigation into the phylogenetic depth of low complexity proteins with or without known disease associations, we selected a published list of 100 human proteins with well-characterised compositionally biased regions (Mier *et al.*, 2020). The proteins were mapped to the enriched human genome datasets derived from Ensembl. Out of the 100 proteins in this curated dataset, 17 are confidently associated with disease, with one or more associated DOIDs, covering a wide range of disorders from metabolic and cardiovascular diseases to autoimmune conditions and cancer (Figure 3).

To examine in more detail the emergence of compositional bias for the curated dataset of 100 human proteins as an exemplary case, protein sequence alignment using DIAMOND (see Methods) was performed against the URP dataset. Homologues were detected in >11000 species, with just 269 cases not containing any of these regions, largely

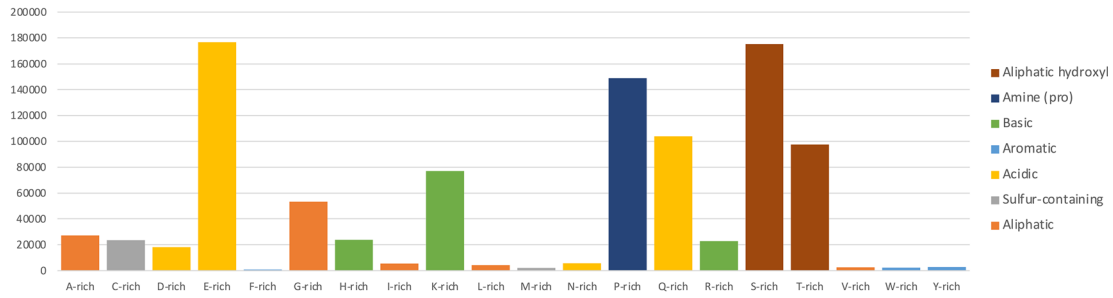


Figure 1. Compositionally biased protein regions identified using CAST on the Ensembl human genome. Sum of CAST scores for all compositionally biased occurrences by amino acid, normalised with respect to general amino acid frequency in the human genome (i.e. total score/frequency). Higher CAST scores reflect strong compositional bias for the specific region. The column colour corresponds to an amino acid classification according to the chemical nature of their side chains (Katchalski-Katzir *et al.*, 2006).

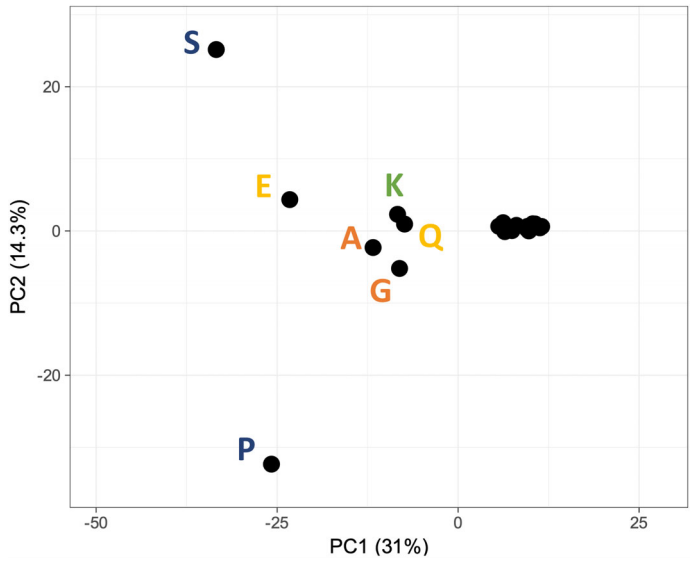


Figure 2. PCA analysis of DOID correlation to proteins with low complexity regions, across the human genome. The colour coding of each amino acid is the same as Figure 1 and reflects the chemical nature of their side chains. DOID=Disease Ontology identifier.

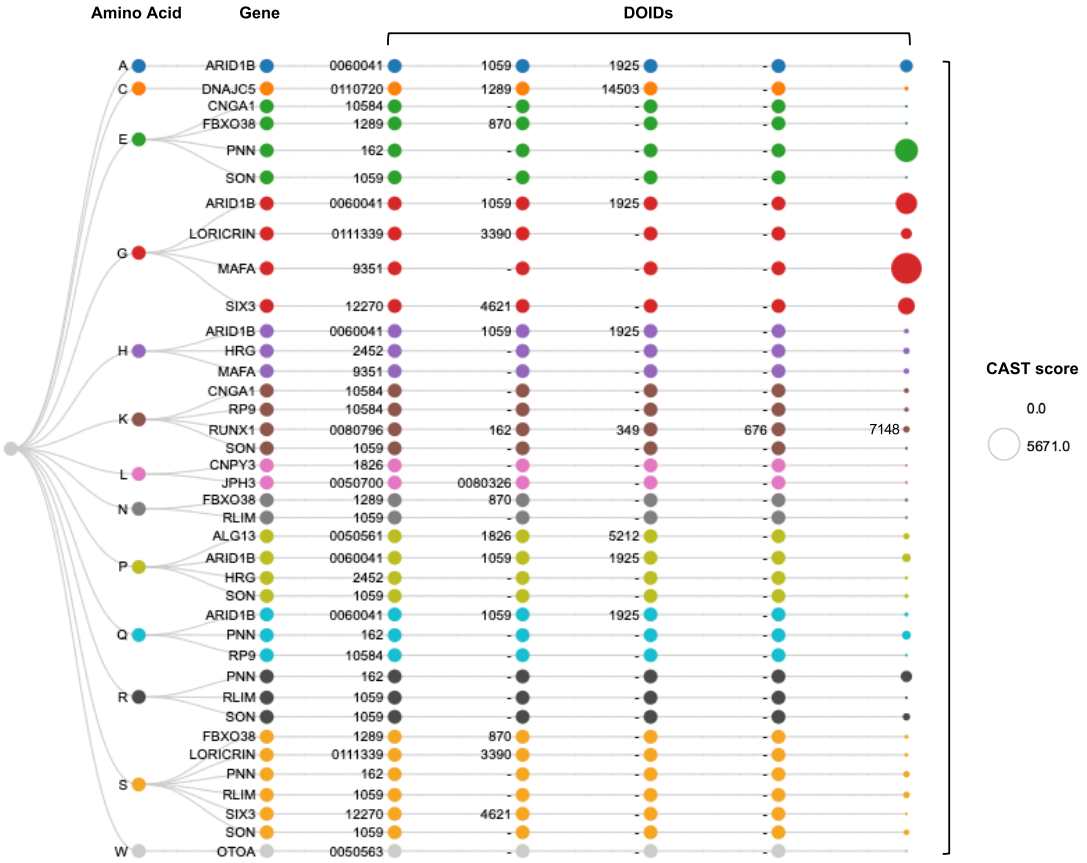


Figure 3. The 17 disease-associated genes from the set of 100 human proteins with well-characterised compositionally biased regions, linked to associated DOIDs. The first column corresponds to the amino acid that appears to have enriched presence in each case. The last column's size is proportional to the sum of CAST scores for amino acid rich regions in each gene. Although specific genes have originally been listed as exemplary for one compositional bias type, in this analysis they can be observed more than once along their DOID associations, as the result of CAST analysis for the derivation of total scores. DOID=Disease Ontology identifier.

corresponding to Archaeal and Bacterial taxa. This preliminary, targeted comparative analysis using a limited query of 100 human proteins is a first glimpse into the dynamics of compositional bias across phylogenies. Our ongoing effort to investigate the presence of compositional bias and the connection to human disease will assess these discovered phylogenetic patterns across the entire human genome in the near future. The complete phylogenetic profiling matrix is provided as *Extended data* (Chasapi, 2022).

Focusing on the 100-gene subset with confident disease associations (i.e. the 17 proteins), most disease-associated genes had detectable homologues across Eukaryotic organisms, with only a few, scarce Bacterial hits (Figure 4). An exception is the DnaJ heat shock protein family (Hsp40) member C5 (DNAJC5) which exhibits an extended phylogenetic depth, covering 86% of the URP (i.e. 9751 proteomes), verifying the observation as a well-known, abundant domain (Stetler *et al.*, 2010; Qiu *et al.*, 2006).

The remaining, 16 disease-associated genes were detected in 1350 proteomes with one or several hits. Figure 4 shows the phylogenetic profile map of these genes across the URP target proteome set. Most genes display homologues in higher eukaryotic organisms, whereas, with the exception of the E3 ubiquitin-protein ligase RLIM, almost no homologous genes are detected in plant genomes. Similarly, the subunit of the rod cyclic GMP-gated cation channel (CNGA1) is the only query gene with ion channel homologues in ciliates and fungi, with the exception of the Ascomycota. In the case of genes

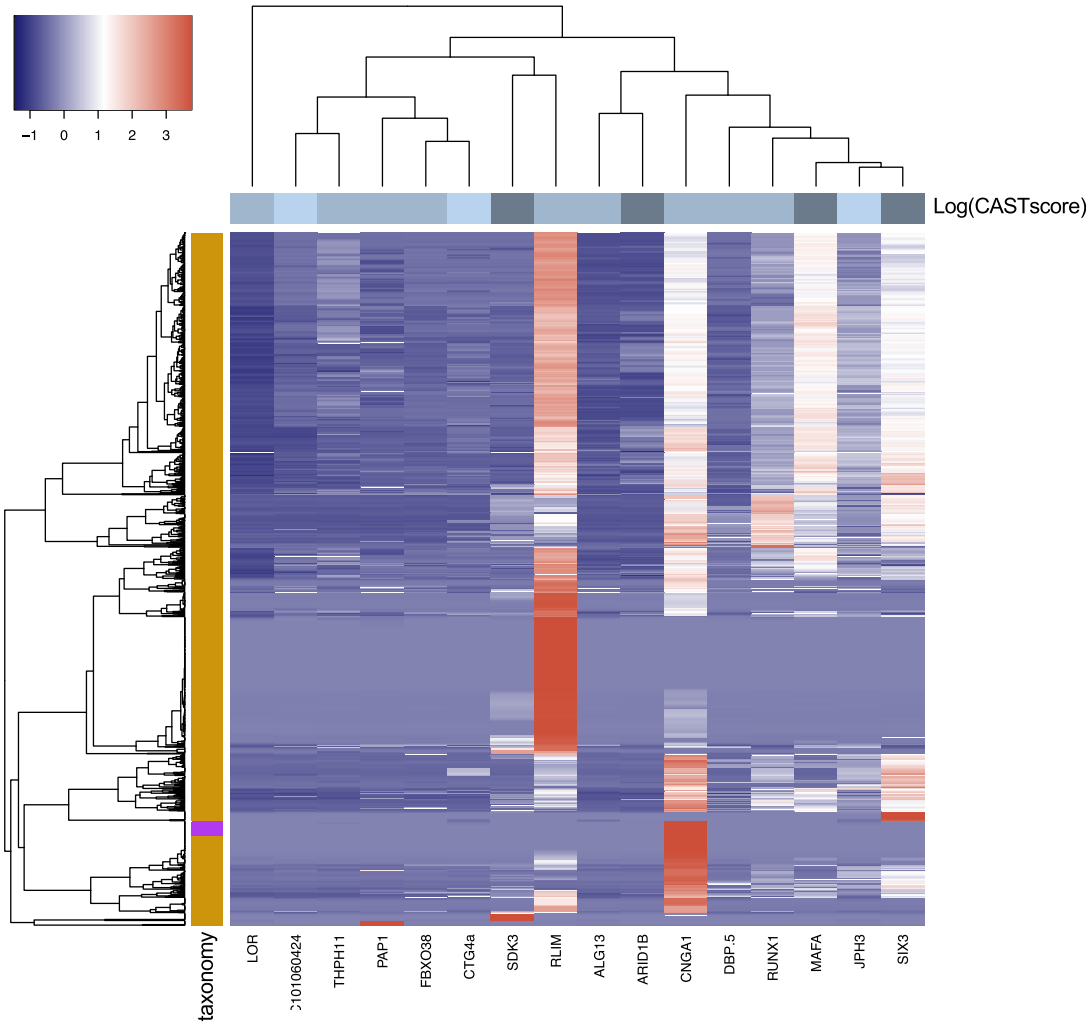


Figure 4. Phylogenetic profile of the 100-gene subset with confident disease associations. The heatmap represents the number of homologues found across species for each gene. The plotted values correspond to the normalized number of homologues (see Methods). The row side colours correspond to the log of the sum of CAST scores for all detected compositionally biased regions for each protein (darker colour indicates higher sum of CAST scores). The column side colours indicate the taxonomic level of each target proteome (orange = eukaryotes, purple = bacteria). DNAJC5 is not displayed.

with an overall high CAST score, there seem to be more sequence hits, both in number and in taxonomic distribution. This can be, in part, due to the sequence alignment analysis which was tailored to compositionally biased sequences, thus increasing hit sensitivity.

This comparative genomics framework is a useful tool both for the investigation of tendencies among gene sets confidently associated with diseases, containing compositionally biased regions, as well as for the identification of specific taxonomic signatures for each gene. A selected number of specific cases are reviewed below.

ALG13 has a restricted phylogenetic depth

The protein encoded by ALG13 is a subunit of a bipartite UDP-N-acetylglucosamine transferase, which heterodimerizes with asparagine-linked glycosylation 14 homolog to form a functional UDP-GlcNAc glycosyltransferase that catalyses the second sugar addition of the highly conserved oligosaccharide precursor in endoplasmic reticulum N-linked glycosylation. ALG13 has been associated with several disease conditions including developmental and epileptic encephalopathy as well as genetic intellectual disability (Epi4K Consortium & Epilepsy Phenome/Genome Project, 2013; Bissar-Tadmouri *et al.*, 2014; Ng *et al.*, 2020). ALG13 homologs are detected in 248 proteomes. Moreover, all hits correspond to higher Eukaryotes, specifically to the infraphylum Gnathostomata, including mostly Euteleostomi representatives. Figure 5A shows a general view of the tree of life, highlighted for species where ALG13 homologue hits were retrieved, whereas Figure 5B provides a closer look of the same result. The restricted phylogenetic depth of

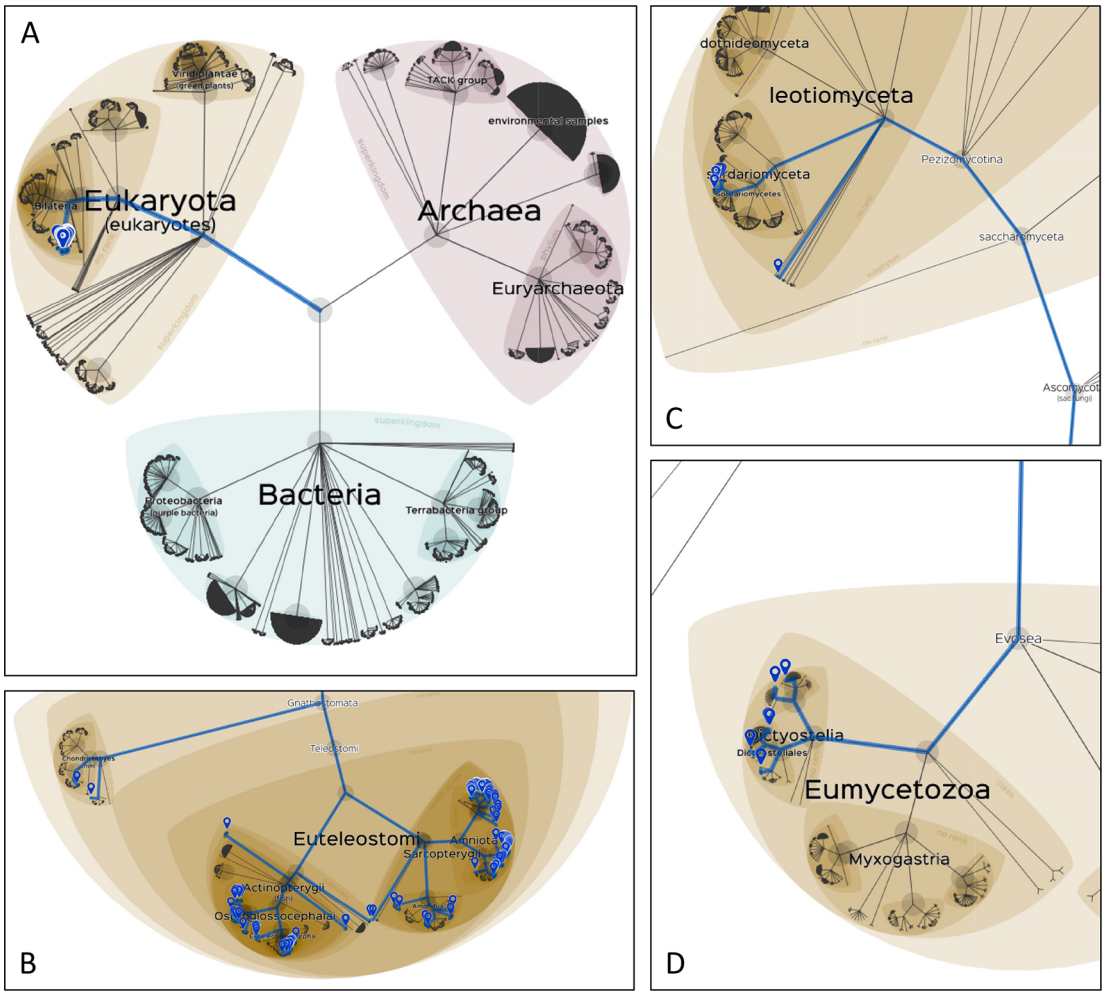


Figure 5. Taxonomic distribution of sequence alignment hits for selected genes, across the URP proteome dataset. A) a broad view of the tree of life, visualised by Lifemap (de Vienne, 2016). ALG13 hits are highlighted in blue. B) A zoomed in view of all ALG13 hits. C) SIX3 hits that belong to the Ascomycota phylum and are not found for any other gene of the dataset. D) RP9 hits that correspond to the Dictyostelia clade and are not found for any other gene of the dataset.

ALG13 may indicate that the interaction pathways including ALG13 are restricted to functions specific to bony vertebrates, a hypothesis that can be assessed by jointly analysing all participating proteins for their evolutionary emergence.

SIX3 has a unique conservation signature

SIX Homeobox 3 (SIX3) encodes a member of the sine oculis homeobox transcription factor family. The expressed protein plays a role in brain and eye development, and its mutations are associated with Holoprosencephaly and Schizencephaly abnormalities (Wallis *et al.*, 1999, 3; Hehr *et al.*, 2010, 3). SIX3 homologues were detected in 869 reference proteomes, including filamentous ascomycetes proteome sequences in which SIX3 is the only disease-associated gene with significant hits (Figure 5C). A follow-up study could further investigate this distinct conservation pattern.

RP9 is uniquely matched in Dictyostelia

Retinitis Pigmentosa 9 (RP9 or PAP1) is thought to be a target protein for the PIM1 serine/threonine protein kinase. The protein localises in nuclear speckles and has a role in pre-mRNA splicing. Mutations in the RP9 gene result in autosomal dominant retinitis pigmentosa (Maita *et al.*, 2004; Keen *et al.*, 2002). The comparative genomics analysis of RP9 presence detects homologues in 507 species, including all representatives of the Dictyostelia clade, that were uniquely matched to RP9 among all disease genes (Figure 5D). *Dictyostelium discoideum*, the most studied representative of Dictyostelia (i.e. dictyostelid cellular slime molds), has been used extensively as model organism for cell communication, differentiation, and programmed cell death studies (Kawabe *et al.*, 2019; Strassmann *et al.*, 2000). The specific presence of RP9 homologues in Dictyostelia including D. *discoideum*, raises questions about their specific roles in this taxon and the possibility that functional analysis can shed further light into the human disease.

Discussion

A major research objective for biomedical research is the detection of genetic factors involved in human disease at multiple levels including variation, gene expression and cellular roles. The evolutionary perspective of human disease is less appreciated, compared to the functional genomics of human genes and proteins, by either computational or experimental means. Combining evolutionary characters to structural features such as IDR presence which has yet to be systematically studied in conjunction with specific disease classes, can provide a novel analysis framework of the human genome with respect to disease.

In this study, we report a genome-wide analysis of the compositional bias association with disease in human proteins and their taxonomic distribution. It is the first time that a combined genome-wide analysis of these aspects is reported, from various structural, functional and evolutionary angles. Our analysis includes novel views on the relation between compositional bias and disease-association, demonstrating a strong correlation between the two features. Delving deeper into the contribution of specific amino acids to compositionally biased regions of disease-associated genes across the human genome, we demonstrate that charged, hydrophilic residues are over-represented in genes with confident disease associations.

We adopt a comparative genomics perspective for the evaluation of disease association of compositional bias in human proteins, using a curated list of 100 human proteins, as a first step towards this direction in a controlled manner. We delineate conservation patterns of the annotated gene set across taxonomic categories, taking advantage of the great plethora of sequenced genomes across the tree of life, using a total of 11297 representative proteomes.

The described framework of structurally and functionally annotated gene queries against multiple species has been developed with the view of future directions, encompassing the entire human genome and all known gene-disease associations. This will potentially allow us to elucidate specific evolutionary patterns of groups of genes involved in the same disease, serving as a tool to better understand the underlying mechanisms and identify appropriate model organisms for experimental investigation.

Data availability

Underlying data

All data underlying the analyses are available as part of the article or as referenced external data sources and no additional source data are required.

Extended data

Zenodo: Phylogenetic profile of 100 annotated low complexity proteins against the Uniprot Reference Proteome dataset. https://doi.org/10.5281/zenodo.7486339 (Chasapi, 2022).

This project contains the following extended data:

- cb100-query-20221223.csv (The phylogenetic profile of the 100 selected annotated low complexity proteins against the Uniprot Reference Proteome dataset)

Acknowledgements

The authors would like to thank the IDP implementation study community participating in the Elixir Commissioned Service entitled "Standardising Intrinsically Disordered Proteins (IDPs) Data" for the useful knowledge exchange and excellent collaboration on the topic of IDP standardisation.

References

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

Basile W, Salvatore M, Bassot C, *et al.*: **Why do eukaryotic proteins contain more intrinsically disordered regions?** *PLoS Comput. Biol.* 2019; **15**: e1007186.
PubMed Abstract | Publisher Full Text | Free Full Text

Bissar-Tadmouri N, Donahue WL, Al-Gazali L, *et al.*: **X chromosome exome sequencing reveals a novel *ALG 13* mutation in a nonsyndromic intellectual disability family with multiple affected male siblings.** *Am. J. Med. Genet. A.* 2014; **164**: 164–169.
Publisher Full Text

Breido L, Wu JW, Uversky VN: **α-Synuclein misfolding and Parkinson's disease**. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease.* 2012; Feb 1; **1822**(2):261–85.
PubMed Abstract | Publisher Full Text

Brown CJ, Johnson AK, Dunker AK, *et al.*: **Evolution and disorder.** *Curr. Opin. Struct. Biol.* 2011; **21**: 441–446.
PubMed Abstract | Publisher Full Text | Free Full Text

Buchfink B, Reuter K, Drost H-G: **Sensitive protein alignments at tree-of-life scale using DIAMOND.** *Nat. Methods.* 2021; **18**: 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Bunting EL, Hamilton J, Tabrizi SJ: **Polyglutamine diseases.** *Curr. Opin. Neurobiol.* 2022; **72**: 39–47.
Publisher Full Text

Bürgi J, Xue B, Uversky VN, *et al.*: **Intrinsic Disorder in Transmembrane Proteins: Roles in Signaling and Topology Prediction.** *PLoS One.* 2016; **11**: e0158594.
PubMed Abstract | Publisher Full Text | Free Full Text

Chasapi A: Phylogenetic profile of 100 annotated low complexity proteins against the Uniprot Reference Proteome dataset. [Dataset]. *Zenodo.* 2022.
Publisher Full Text

Chen C, Natale DA, Finn RD, *et al.*: **Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation.** *PLoS One.* 2011; **6**: e18910.
PubMed Abstract | Publisher Full Text | Free Full Text

Choudhary S, Lopus M, Hosur RV: **Targeting disorders in unstructured and structured proteins in various diseases.** *Biophys. Chem.* 2022; **281**: 106742.
PubMed Abstract | Publisher Full Text

Cock PJA, Antao T, Chang JT, *et al.*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics.* 2009; **25**: 1422–1423.
PubMed Abstract | Publisher Full Text | Free Full Text

Dickerson JE, Robertson DL: **On the Origins of Mendelian Disease Genes in Man: The Impact of Gene Duplication.** *Mol. Biol. Evol.* 2012; **29**: 61–69.
PubMed Abstract | Publisher Full Text | Free Full Text

Dosztanyi Z, Csizmok V, Tompa P, *et al.*: **IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.** *Bioinformatics.* 2005; **21**: 3433–3434.
Publisher Full Text

Dosztanyi Z, Meszaros B, Simon I: **ANCHOR: web server for predicting protein binding regions in disordered proteins.** *Bioinformatics.* 2009; **25**: 2745–2746.
PubMed Abstract | Publisher Full Text | Free Full Text

Dunker AK, Lawson JD, Brown CJ, *et al.*: **Intrinsically disordered protein.** *J. Mol. Graph. Model.* 2001; **19**: 26–59.
Publisher Full Text

Epi4K Consortium & Epilepsy Phenome/Genome Project: **De novo mutations in epileptic encephalopathies.** *Nature.* 2013; **501**: 217–221.
PubMed Abstract | Publisher Full Text | Free Full Text

Grignaschi E, Cereghetti G, Grigolato F, *et al.*: **A hydrophobic low-complexity region regulates aggregation of the yeast pyruvate kinase Cdc19 into amyloid-like aggregates in vitro.** *J. Biol. Chem.* 2018; **293**(29): 11424–32.
PubMed Abstract | Publisher Full Text | Free Full Text

Harbi D, Kumar M, Harrison PM: **LPS-annotate: complete annotation of compositionally biased regions in the protein knowledgebase.** *Database.* 2011; **2011**: baq031–baq031.
Publisher Full Text

Harrison PM: **fLPS 2.0: rapid annotation of compositionally-biased regions in biological sequences.** *PeerJ.* 2021; **9**: e12363.
PubMed Abstract | Publisher Full Text | Free Full Text

Hatos A, Hajdu-Soltész B, Monzon AM, *et al.*: **DisProt: intrinsic protein disorder annotation in 2020.** *Nucleic Acids Res.* 2019; **48**: D269–D276.
PubMed Abstract | Publisher Full Text | Free Full Text

Hehr U, Pineda-Alvarez DE, Uyanik G, *et al.*: **Heterozygous mutations in SIX3 and SHH are associated with schizencephaly and further expand the clinical spectrum of holoprosencephaly.** *Hum. Genet.* 2010; **127**: 555–561.
PubMed Abstract | Publisher Full Text | Free Full Text

Janssen P, Enright AJ, Audit B, *et al.*: **COmplete GENome Tracking (COGENT): a flexible data environment for computational genomics.** *Bioinformatics.* 2003; **19**: 1451–1452.
PubMed Abstract | Publisher Full Text

Katchalski-Katzir E, Kasher R, Fridkin M: **Amino Acids: Physicochemical Properties.** *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine.* Berlin Heidelberg: Springer; 2006; pp 55–68.

Kawabe Y, Du Q, Schilde C, *et al.*: **Evolution of multicellularity in Dictyostelia.** *Int. J. Dev. Biol.* 2019; **63**: 359–369.
PubMed Abstract | Publisher Full Text | Free Full Text

Keen TJ, Hims MM, McKie AB, *et al.*: **Mutations in a protein target of the Pim-1 kinase associated with the RP9 form of autosomal dominant retinitis pigmentosa.** *Eur. J. Hum. Genet.* 2002; **10**: 245–249.
PubMed Abstract | Publisher Full Text

Khan T, Douglas GM, Patel P, *et al.*: **Polymorphism Analysis Reveals Reduced Negative Selection and Elevated Rate of Insertions and Deletions in Intrinsically Disordered Protein Regions.** *Genome Biol. Evol.* 2015; **7**: 1815–1826.
PubMed Abstract | Publisher Full Text | Free Full Text

Koos JA, Bassett A: **Genetics Home Reference: A Review.** *Med. Ref. Serv. Q.* 2018; **37**: 292–299.
Publisher Full Text

Linding R, Jensen LJ, Diella F, *et al.*: **Protein Disorder Prediction.** *Structure.* 2003; **11**: 1453–1459.
Publisher Full Text

Lopez-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease.** *Nucleic Acids Res.* 2004; **32**: 3108–3114.
PubMed Abstract | Publisher Full Text | Free Full Text

Maita H, Kitaura H, Keen TJ, *et al.*: **PAP-1, the mutated gene underlying the RP9 form of dominant retinitis pigmentosa, is a splicing factor.** *Exp. Cell Res.* 2004; **300**: 283–296.
PubMed Abstract | Publisher Full Text

Matsushima N, Yoshida H, Kumaki Y, *et al.*: **Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins**. *Curr. Protein. Pept. Sci.* 2008; **9**(6): 591–610.
PubMed Abstract | Publisher Full Text

Mier P, Paladin L, Tamana S, *et al.*: **Disentangling the complexity of low complexity proteins.** *Brief. Bioinform.* 2020; **21**: 458–472.
PubMed Abstract | Publisher Full Text | Free Full Text

Monti SM, De Simone G, Langella E: **The Amazing World of IDPs in Human Diseases.** *Biomolecules.* 2021; **11**: 333.
PubMed Abstract | Publisher Full Text | Free Full Text

Monti SM, De Simone G, Langella E: **The Amazing World of IDPs in Human Diseases II.** *Biomolecules.* 2022; **12**: 369.
PubMed Abstract | Publisher Full Text | Free Full Text

Necci M, Piovesan M, CAID Predictors, *et al.*: **Critical assessment of protein intrinsic disorder prediction.** *Nat. Methods.* 2021; **18**: 472–81.

Necci M, Piovesan D, Dosztányi Z, *et al.*: **MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins.** *Bioinformatics.* 2017; **33**: 1402–1404.
Publisher Full Text

Ng BG, Eklund EA, Shiryaev SA, *et al.*: **Predominant and novel de novo variants in 29 individuals with *ALG13* deficiency: Clinical description, biomarker status, biochemical analysis, and treatment suggestions.** *J. Inherit. Metab. Dis.* 2020; **43**: 1333–1348.
PubMed Abstract | Publisher Full Text | Free Full Text

Ntountoumi C, Vlastaridis P, Mossialos D, *et al.*: **Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved.** *Nucleic Acids Res.* 2019; **47**: 9998–10009.
PubMed Abstract | Publisher Full Text | Free Full Text

Ouzounis CA, Coulson RMR, Enright AJ, *et al.*: **Classification schemes for protein structure and function.** *Nat. Rev. Genet.* 2003; **4**: 508–519.
Publisher Full Text

Pajkos M, Zeke A, Dosztányi Z: **Ancient Evolutionary Origin of Intrinsically Disordered Cancer Risk Regions.** *Biomolecules.* 2020; **10**: 1115.
PubMed Abstract | Publisher Full Text | Free Full Text

Peng Z, Yan J, Fan X, *et al.*: **Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.** *Cell. Mol. Life Sci.* 2015; **72**: 137–151.
PubMed Abstract | Publisher Full Text

Pletscher-Frankild S, Pallejà A, Tsafou K, *et al.*: **DISEASES: Text mining and data integration of disease–gene associations.** *Methods.* 2015; **74**: 83–89.
PubMed Abstract | Publisher Full Text

Promponas VJ, Enright AJ, Tsoka S, *et al.*: **CAST: an iterative algorithm for the complexity analysis of sequence tracts.** *Bioinformatics.* 2000; **16**: 915–922.
Publisher Full Text

Qiu X-B, Shao Y-M, Miao S, *et al.*: **The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones.** *Cell. Mol. Life Sci.* 2006; **63**: 2560–2570.
PubMed Abstract | Publisher Full Text

Quaglia F, Mészáros B, Salladini E, *et al.*: **DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation.** *Nucleic Acids Res.* 2022; **50**: D480–D487.
Publisher Full Text

Schriml LM, Mitraka E, Munro J, *et al.*: **Human Disease Ontology 2018 update: classification, content and workflow expansion.** *Nucleic Acids Res.* 2019; **47**: D955–D962.
PubMed Abstract | Publisher Full Text | Free Full Text

Sickmeier M, Hamilton JA, LeGall T, *et al.*: **DisProt: the Database of Disordered Proteins.** *Nucleic Acids Res.* 2007; **35**: D786–D793.
PubMed Abstract | Publisher Full Text | Free Full Text

Smith M, Kunin V, Goldovsky L, *et al.*: **MagicMatch--cross-referencing sequence identifiers across databases.** *Bioinformatics.* 2005; **21**: 3429–3430.
PubMed Abstract | Publisher Full Text

Stetler RA, Gan Y, Zhang W, *et al.*: **Heat shock proteins: Cellular and molecular mechanisms in the central nervous system.** *Prog. Neurobiol.* 2010; **92**: 184–211.
PubMed Abstract | Publisher Full Text | Free Full Text

Strassmann JE, Zhu Y, Queller DC: **Altruism and social cheating in the social amoeba Dictyostelium discoideum.** *Nature.* 2000; **408**: 965–967.
PubMed Abstract | Publisher Full Text

Tang Y-J, Pang Y-H, Liu B: **IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning.** *Bioinformatics.* 2021; **36**: 5177–5186.
PubMed Abstract | Publisher Full Text

Tantos A, Han K-H, Tompa P: **Intrinsic disorder in cell signaling and gene transcription.** *Mol. Cell. Endocrinol.* 2012; **348**: 457–465.
Publisher Full Text

The UniProt Consortium, Bateman A, Martin M-J, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2022; gkac1052.

Theillet FX, Kalmar L, Tompa P, *et al.*: **The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins.** *Intrinsically Disord Proteins.* 2013; **1**(1): e24360.
PubMed Abstract | Publisher Full Text | Free Full Text

Uversky VN: **The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins**. *Intrinsically Disord Proteins.* 2013; **1**(1): e24684.
PubMed Abstract | Publisher Full Text | Free Full Text

Uversky VN, Oldfield CJ, Dunker AK.: **Intrinsically disordered proteins in human diseases: introducing the D2 concept**. *Annu. Rev. Biophys.* 2008; **37**: 215–46.
Publisher Full Text

de Vienne DM: **Lifemap: Exploring the Entire Tree of Life.** *PLoS Biol.* 2016; **14**: e2001624.
PubMed Abstract | Publisher Full Text | Free Full Text

Wallis DE, Roessler E, Hehr U, *et al.*: **Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly.** *Nat. Genet.* 1999; **22**: 196–198.
Publisher Full Text

Walsh I, Martin AJM, Di Domenico T, *et al.*: **ESpritz: accurate and fast prediction of protein disorder.** *Bioinformatics.* 2012; **28**: 503–509.
PubMed Abstract | Publisher Full Text

Wang S, Ma J, Xu J: **AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields.** *Bioinformatics.* 2016; **32**: i672–i679.
PubMed Abstract | Publisher Full Text | Free Full Text

Ward JJ, Sodhi JS, McGuffin LJ, *et al.*: **Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life.** *J. Mol. Biol.* 2004; **337**: 635–645.
PubMed Abstract | Publisher Full Text

Williams RM, Obradovic Z, Mathura V, *et al.*: **The Protein Non-folding Problem: Amino Acid Determinants of Intrinsic Order and Disorder**. *Biocomputing 2001*. Mauna Lani, Hawaii: World Scientific; 2000; pp. 89–100.

Williamson MP: **The structure and function of proline-rich regions in proteins**. *Biochem. J.* 1994; **297**(Pt 2) (Pt 2): 249–60.
PubMed Abstract | Publisher Full Text | Free Full Text

Wootton JC, Federhen S: **Statistics of local complexity in amino acid sequences and sequence databases.** *Comput. Chem.* 1993; **17**: 149–163.
Publisher Full Text

Xue B, Brown CJ, Dunker AK, *et al.*: **Intrinsically disordered regions of p53 family are highly diversified in evolution.** *Biochim Biophys Acta BBA - Proteins Proteomics.* 2013; **1834**: 725–738.
PubMed Abstract | Publisher Full Text | Free Full Text

Xue B, Dunker AK, Uversky VN: **Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life.** *J. Biomol. Struct. Dyn.* 2012; **30**: 137–149.
PubMed Abstract | Publisher Full Text

Yates A, Akanni W, Amode MR, *et al.*: **Ensembl 2016.** *Nucleic Acids Res.* 2016; **44**: D710–D716.
PubMed Abstract | Publisher Full Text | Free Full Text

Yu Y-K, Altschul SF: **The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions.** *Bioinformatics.* 2005; **21**: 902–911.
PubMed Abstract | Publisher Full Text

Zhang T, Faraggi E, Xue B, *et al.*: **SPINE-D: Accurate Prediction of Short and Long Disordered Regions by a Single Neural-Network Based Method.** *J. Biomol. Struct. Dyn.* 2012; **29**: 799–813.
PubMed Abstract | Publisher Full Text | Free Full Text

Zhao S, Guo Y, Sheng Q, *et al.*: **Heatmap3: an improved heatmap package with more powerful and convenient features.** *BMC Bioinformatics.* 2014; **15**(10): 1–2.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 17 April 2023

https://doi.org/10.5256/f1000research.146550.r169726

✔ **Stella Tamana**
Molecular Genetics Thalassaemia Department, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus

The authors have addressed my comments/recommendations and I have no further comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Compositionally Biased Regions, Comparative genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 17 April 2023

https://doi.org/10.5256/f1000research.146550.r169727

✔ **Gregory Amoutzias**
Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, Larissa, Greece

The authors have addressed my suggestions/corrections and I have no further comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, evolution, sequence analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 13 March 2023

https://doi.org/10.5256/f1000research.142650.r164958

?

Gregory Amoutzias
Bioinformatics Laboratory, Department of Biochemistry and Biotechnology, University of Thessaly, Larissa, Greece

The authors identified disordered regions in human proteins, they investigated their annotation and their association with diseases and then investigated the presence/absence pattern of these disordered proteins in homologs from various taxa. Interestingly, the authors found that disease-related proteins tend to have LCRs that are enriched in charged, hydrophilic amino acids. In addition, they investigated the phylogenetic distribution of these proteins. The authors discuss in detail some very interesting disease-associated genes with LCRs and their distinct phylogenetic depth. This is an interesting and very useful study that sheds more light (also from an evolutionary point of view) on the intriguing link between LCRs and disease, and importantly, it functions as a pilot study for a larger-scale planned analysis by the authors.

Comments/suggestions:

In introduction, I would also add a paragraph and explain in more detail the terms 'compositional bias' and 'low complexity' and explain that they are strongly linked to IDRs. Concerning the LCRs part, I would include the references of Wooton *et al.* (1994[1]), of Karlin *et al.* (2002[2]), of Schaper *et al.* (2014[3]) and explain that they were originally thought of as junk or linker regions, but not anymore (Haerty *et al.*, 2010[4]).

In Methods: “Searching with query datasets against Proteomes for the creation of phylogenetic profile patterns was performed with DIAMOND blastp”. I understand that the authors search for homologs and not only orthologs, is that correct? Also, concerning the cutoff of 21% sequence similarity, is it for any local alignment that DIAMOND reports, or is it for a certain query coverage too?

Maybe Figure 1 and Table 1 could be moved to Results.

In Results, third paragraph. The authors clearly identify a strong compositional bias for disease-associated transcripts. It would be very informative to mention the enrichment fold of 1.8 ((1845/1780) / (36250/62827)). Also, better mention the p-value as 1e-5. I wonder if disease-associated genes tend to have more transcripts on average than non-disease genes? If they do, are the transcripts-isoforms more frequently retaining the compositionally biased region? Maybe the authors could repeat this chi-square test, where they use the longest transcript per gene.

Some thoughts for the future, when the authors perform their planned large-scale analysis: Do disease-associated genes and transcripts with compositional bias have more wide or more restricted gene expression profiles?

Results: "The striking over-representation of serine/threonine (S/T) tracts, along with glutamate/glutamine (E/Q) and proline (P) followed by lysine (K) is indicative of the main residue types that might affect functional properties of human proteins, including their potential association with known phenotypes, such as polyglutamine tracts with neurodegenerative diseases (Bunting et al., 2022)"
It would also be informative to add that tandem repeats of short oligopeptides that are rich in glycine, proline, serine or threonine are capable of forming flexible structures that bind ligands under certain pH and temperature conditions (Matsushima *et al.*, 2008[5]). Maybe also add Williamson *et al.* (1994[6]), concerning proline rich regions.

I would recommend that the authors explain somewhere in detail what the CAST score means.

Concerning the section of phylogenetic profiling, first paragraph: It would be informative to show what is the enrichment of disease in the well characterized CB proteins, compared to the background, and the statistical significance (Hypergeometric test probably). For example, the background is X disease-proteins in Y total proteins (or genes) of the genome. Enrichment: (17/100)/(X/Y).
Same section, second paragraph: I guess its pairwise local sequence alignment with DIAMOND?
Same section, third paragraph: "Focusing on the 100-gene subset with confident disease associations". I guess these are the 17 proteins of the 100 gene subset with known IDs?

Within the abstract, the authors use all three related terms, intrinsically disordered regions (IDRs), compositional bias, low complexity regions (LCRs). Maybe they could use only one (in the abstract). Also, they should explain in the Introduction their inter-changeability.

Abstract: "The evolutionary rate of disordered proteins". I would rephrase that as disordered protein regions.

Abstract: Low complexity proteins or proteins with LCRs?

In keywords, I would rephrase towards: low complexity region (LCR).

In Introduction, please correct the CAID Predictors et al reference.

In Introduction: "are transiently associated with intrinsic disorder in nucleated cells". Could the authors please explain in more detail what they mean with the term "transiently"?

Introduction: the paragraph that discusses the link between intrinsic disorder and disease: It would be nice to briefly mention one specific example of how intrinsic disorder is associated with human disease.

Introduction: “while others appearing particularly diversified”. Maybe the term “rapidly evolving” would be better.

In Methods, when first mentioning GHR, I would include the entire name “Genetics Home Reference”.

In Methods: “a total of 7269 disease-gene, high-confidence associations”. It would be even more informative if the number of unique genes and the number of unique diseases in these associations were included too.

In Methods, could the authors please elaborate more on what MagicMatch does? I did not exactly understand this part: “to verify the identity of the reference proteome collection against the modified identifier space”.

Concerning the calculation of amino acid frequencies across the Ensembl protein set, was that done for all protein isoforms of a certain gene, or only for one representative isoform (i.e. the longest one) from each gene?

In table 1, the numbers are for human genes or transcripts/protein isoforms?

In Methods: “For each visualisation, a list of the NCBI IDs”. Which types of IDs?

Results: “(GRCh38.p13), containing 119068 gene transcripts”. Could the authors also mention the number of genes?

Results: “Examination of the low complexity gene dataset features highlighted the significant divergence among amino acid related, low complexity frequencies.” Maybe better use the term difference because divergence relates to conservation.

Results: “Charged, hydrophilic residues appear over-represented”. I would also mention them in parenthesis.

In figure 3, this must be the subset of 17 disease-associated genes from the set of 100 human proteins with well-characterised compositionally biased regions? . The title of Figure 3 should be changed accordingly. Is it possible to also include a key of DOIDS-disease next to the figure, or is it too many of them?

Figure 4 legend: could the authors explain this more?: “The heatmap range reflects the dissimilarity matrix of the plotted values.”. I am not sure I understood figure 4. Does the figure only show if a homologue is present in a certain species, or does it show as well if the homologue also contains an LCR or CB region as well? Does this correspond to the red/blue colour of the matrix?
Concerning figure 4, as a thought/suggestion for future studies, when the authors move to larger-scale analyses, maybe they could also include an analogous analysis, where they show the

presence of orthologs, not homologs. For that, they would have to use best reciprocal blast and one representative protein (the longest) from each gene of a genome. Or, maybe the authors could do that using the orthology presence from the OMA database (https://omabrowser.org/oma/home/).

Figure 5: visualized by Lifemap.

Concerning supplementary .map file, I would simply convert it to a csv file for import in excel. I would also add two columns, one with the species name of the proteome and another one with the wider taxonomic group that the species belongs to. Also, could the authors explain what the numbers in the cells correspond to, I guess it's the number of homologs in that species?

References

1. Wootton JC: Non-globular domains in protein sequences: automated segmentation using complexity measures.*Comput Chem*. 1994; **18** (3): 269-85 PubMed Abstract | Publisher Full Text
2. Karlin S, Brocchieri L, Bergman A, Mrazek J, et al.: Amino acid runs in eukaryotic proteomes and disease associations.*Proc Natl Acad Sci U S A*. 2002; **99** (1): 333-8 PubMed Abstract | Publisher Full Text
3. Schaper E, Gascuel O, Anisimova M: Deep conservation of human protein tandem repeats within the eukaryotes.*Mol Biol Evol*. 2014; **31** (5): 1132-48 PubMed Abstract | Publisher Full Text
4. Haerty W, Golding GB: Low-complexity sequences and single amino acid repeats: not just "junk" peptide sequences.*Genome*. 2010; **53** (10): 753-62 PubMed Abstract | Publisher Full Text
5. Matsushima N, Yoshida H, Kumaki Y, Kamiya M, et al.: Flexible structures and ligand interactions of tandem repeats consisting of proline, glycine, asparagine, serine, and/or threonine rich oligopeptides in proteins.*Curr Protein Pept Sci*. 2008; **9** (6): 591-610 PubMed Abstract | Publisher Full Text
6. Williamson MP: The structure and function of proline-rich regions in proteins.*Biochem J*. 1994; **297 (Pt 2)** (Pt 2): 249-60 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, evolution, sequence analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 March 2023

https://doi.org/10.5256/f1000research.142650.r164957

Stella Tamana
Molecular Genetics Thalassaemia Department, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus

The presented manuscript "Disease association and comparative genomics of compositional bias in human proteins" focuses on the very interesting, but challenging to analyze and interpret, study of the structural, functional and evolutionary properties of compositional bias in human proteins. The authors report novel findings and a strong correlation between compositional bias and disease association. Studying compositional bias and disease association, both computationally and experimentally, has been quite challenging over the years, mainly because proteins with compositional bias tend to conform into non-globular structures or be in a disordered state and thus, requiring special treatment in fundamental steps of comparative genomic analyses and experimental procedures to determine their three-dimensional (3D) structures. This is the first time that, successfully, a computational framework, combined genome-wide disease-association analysis of compositional bias with regards to their functional, structural, and evolutionary properties. The presentation is very clear, well-structured, and easy to comprehend. The reported computational framework and methodology take advantage of up-to-date computational tools and statistical packages. This paper is of interest to scientists within the field of human diseases and variant interpretation as elucidating specific functional and structural patterns of groups of genes (especially those presenting extreme compositional bias) involved in the same disease will ultimately help discover the underlying mechanisms and develop new experimental procedures.

Minor comments:

- **Introduction – Disordered proteins exhibit specific patterns at the sequence level**
  - This section will be greatly enhanced if the authors consider adding a paragraph (or 2-3 lines) of the structural characteristics and underlying mechanisms of IDPs/IDRs and thus, giving the reader the opportunity right from the start of the paper to understand the correlation between compositional bias and disease-association. For example, providing in a bit more detail that order-promoting residue types are commonly found within the hydrophobic cores of foldable proteins as opposed to

disorder-promoting residues typically located at the surface of foldable proteins (for references see Theillet *et al.*, 2013[1]; Uversky, 2013[2]). Also, another example could be that hydrophobic enriched regions are prone to induce either self-aggregation or/and intermolecular interactions with surrounding proteins when exposed and thus, trigger aggregation (for reference see Grignaschi *et al.*, 2018[3]).

Line 1-3: please consider adding references.

References

1. Theillet FX, Kalmar L, Tompa P, Han KH, et al.: The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins.*Intrinsically Disord Proteins*. 2013; **1** (1): e24360 PubMed Abstract | Publisher Full Text

2. Uversky VN: The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins.*Intrinsically Disord Proteins*. 2013; **1** (1): e24684 PubMed Abstract | Publisher Full Text

3. Grignaschi E, Cereghetti G, Grigolato F, Kopp MRG, et al.: A hydrophobic low-complexity region regulates aggregation of the yeast pyruvate kinase Cdc19 into amyloid-like aggregates in vitro.*J Biol Chem*. 2018; **293** (29): 11424-11432 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, Compositionally Biased Regions, Comparative genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.