Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2017, 13, 2521

Insights into human intrinsically disordered proteins from their gene expression profile†

Arup Panda, Debarun Acharya * and Tapash Chandra Ghosh*

Received 26th May 2017,
Accepted 9th October 2017

DOI: 10.1039/c7mb00311k

rsc.li/molecular-biosystems

Expression level provides important clues about gene function. Previously, various efforts have been undertaken to profile human genes according to their expression level. Intrinsically disordered proteins (IDPs) do not adopt any rigid conformation under physiological conditions, however, are considered as an important functional class in all domains of life. Based on a human tissue-averaged gene expression level, previous studies showed that IDPs are expressed at a lower level than ordered globular proteins. Here, we examined the gene expression pattern of human ordered and disordered proteins in 32 normal tissues. We noticed that in most of the tissues, ordered and disordered proteins are expressed at a similar level. Moreover, in a number of tissues IDPs were found to be expressed at a higher level than ordered proteins. Rigorous statistical analyses suggested that the lower tissue-averaged gene expression level of IDPs (reported earlier) may be the consequence of their biased gene expression in some specific tissues and higher protein length. When we considered the gene repertory of each tissue we noticed that a number of human tissues (brain, testes, *etc.*) selectively express a higher fraction of disordered proteins, which help them to maintain higher protein connectivity by forming disordered binding motifs and to sustain their functional specificities. Our results demonstrated that the disordered proteins are indispensable in these tissues for their functional advantages.

Introduction

Extensive research on intrinsically disordered proteins (IDPs) over the past few decades has led to a paradigm shift in our understanding of protein structural biology. These studies marked disordered proteins as a unique structural class, distinct from globular proteins in a number of structural and functional characteristics.[1–3] Differences between ordered and disordered proteins are manifested in multiple layers, starting from their sequence composition to functional consequences and evolutionary aspects. At the primary structure level, IDPs are devoid of hydrophobic and aromatic residues and highly enriched with polar and charged amino acids.[2,3] At the functional level, disordered proteins are enriched with processes complementary to the functions of globular proteins and are implicated in various regulatory and signaling cascades, such as control of cell division, apoptosis, post-translational modification, and transcription, *etc.*[4–7] Since IDPs are composed of low complexity regions and are enriched with highly mutable hydrophilic residues these proteins tend to evolve at a faster rate as compared to globular proteins.[8,9] Although IDPs lack three-dimensional structures under physiological conditions, these proteins can adopt well-defined conformations upon interaction with partner proteins (coupled folding and binding).[10] This unique feature of disordered proteins enables them to bind with a large number of partner molecules. Thus, disordered proteins often act as hubs in protein–protein interaction networks.[11] IDPs were initially regarded as a rare class of proteins. However, considering their abundance in different domains of life recent studies have suggested that IDPs constitute a very large class of proteins. Although there are controversies regarding the extent of the disorder, these studies suggested a general trend that IDPs are more common among complex genomes such as multi-cellular eukaryotes, however, are less abundant in unicellular bacterial and archaeal genomes.[12–17] Because of their functional advantages, recently it was proposed that IDPs play important roles in the evolution of complex organisms and their strategies to cope with environmental stresses.[18–21]

Although considerable progress has been achieved in our understanding of the characteristics of disordered proteins, many intriguing questions still remain elusive. One of the major goals of molecular biology is to profile transcripts in terms of their tissue distribution. Expression level provides a crucial indication of whether a gene is functional in a tissue or not. Moreover, gene expression profiles have major implications for understanding human disease etiology for the development of novel therapeutics.[22,23] Therefore, previously a number of initiatives have been undertaken to estimate the expression

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, West Bengal, India. E-mail: stararup@gmail.com, debarun@jcbose.ac.in, tapash@jcbose.ac.in; Fax: +91-33-2355-3886; Tel: +91-33-2355 6626

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c7mb00311k

levels of human genes at genomic scales.[22,24] However, until now little attention has been paid to investigating the gene expression signatures of disordered proteins at the tissue level. A few previous studies that have estimated their gene expression level considered the average gene expression values across all the tested tissues.[25,26] Thus, to date, we have no clear understanding of whether these proteins are expressed in all human tissues and to what extent. Therefore, in this study, we took an initiative to profile disordered proteins in terms of their gene expression level across various human tissues. Based on a mean gene expression level of several human tissues, previously it was ascertained that as compared to globular proteins, most of the disordered proteins tend to be expressed at a lower level.[25,26] However, tissue wise gene expression values, as we analyzed in this study, revealed a contrasting trend. Our study suggested that depending upon the nature of the tissue, disordered proteins may be expressed at a higher, lower or similar level as compared to ordered globular proteins. Moreover, here we found evidence that several human tissues selectively express a higher fraction of disordered proteins which help to sustain their functional specificities.

Methods

Data collection

Tissue wise gene expression values of human protein-coding sequences were obtained from Uhlén *et al.*,[27] In this dataset, average FPKM (fragments per kilobase of exon model per million mapped reads) values were provided for a total of 20 344 genes across 32 human tissues (adipose tissue, adrenal gland, appendix, bone marrow, brain, colon, duodenum, endometrium, esophagus, fallopian tube, gallbladder, heart muscle, kidney, liver, lung, lymph node, ovary, pancreas, placenta, prostate, rectum, salivary gland, skeletal muscle, skin, small intestine, smooth muscle, spleen, stomach, testis, thyroid gland, tonsil, urinary bladder). All these genes were tested for evidence at a protein level through various biochemical approaches; for details see Uhlen *et al*.[27] Here, we discarded 3174 genes either with no evidence or with evidence only at the transcript (RNA) level and further removed 535 genes with undetectable gene expression values (FPKM < 1 in all tissues). Following the gene annotation of Uhlén *et al.*, protein coding sequences of these genes were retrieved from Ensembl release 75.[28] For genes with more than one transcript, we considered the longest transcript. Sequences containing internal stop codons and partial codons were detected by CodonW (J Peden, http://codonw.sourceforge.net) and removed.

Prediction of protein intrinsic disorder content

Disorder predictions for human proteins were retrieved from the Database of Disordered Protein Predictions (D2P2) database.[29] Currently, D2P2 houses disorder predictions for more than 10 429 760 unique proteins from 1765 individual genomes. Each protein in this database was checked with nine disorder prediction algorithms, namely VL-XT, VSL2b, PrDOs, PV2, IUPred-S, IUPred-L, Espritz-X, Espritz-N and Espritz-D, and searched for several other biologically relevant information such as the number of phosphorylation sites, domain annotations, *etc.* D2P2 allows users to retrieve disorder predictions in several useful formats. To calculate the disorder content of proteins in our dataset we retrieved a prediction for all human proteins currently available in this database. However, we considered disorder predictions only when we found an exact match between the sequences in the D2P2 database and the sequences in our dataset. We found disorder predictions for 15 472 proteins in our dataset all of which were considered for this analysis. To estimate the fraction of disordered residues in each protein, we considered the residues predicted as disordered residues by at least five of the nine algorithms. The disorder content was calculated as the fraction of the total number of such disordered residues in a protein to the length of that protein. We also checked the consistency of the results by calculating protein disordered content considering residues predicted as disordered by at least 6 and 7 algorithms.

Calculations of tissue selectively

To determine the genes that are selectively expressed in different tissues we considered two approaches. At first, we followed the tissue annotation of Uhlén *et al.*, from where we retrieved gene expression values.[27] Based on the expression profile they classified human genes into six general categories (i) tissue enriched genes, (ii) group enriched genes, (iii) tissue enhanced genes, (iv) mixed genes, (v) genes which are expressed in all tissues and (vi) genes which are not expressed in any tissue. Among these categories, tissue enriched genes were defined with most stringent criteria, 5-fold higher FPKM in one tissue as compared to all the remaining tissues. To compile the list of genes that are selectively expressed in each tissue, we considered the genes that were annotated as 'tissue enriched genes' and associated with only one tissue. However, the genes that were identified as tissue-selective genes by this approach lack any statistical validation. Therefore, we considered another approach that defines tissue-selective genes through rigorous statistical analysis.[30,31] Following Chang *et al.* and Greco *et al.* for each tissue–gene pair we calculated a tissue-selectivity score S_{ij} from the gene expression matrix as:

$$S_{ij} = W_i \times X_{ij}$$

Here, X_{ij} is the normalized gene expression (FPKM) value of gene 'i' in tissue 'j' and W_i is a gene-specific weight. The gene-specific weight W_i was calculated as follows:

$$W_i = \frac{1}{(N-1)} \sum_{k=1}^{N} (1 - X_{ik})$$

Here, X_{ik} is the FPKM value of gene 'i' in tissue 'k' and N is the total number of tested tissues.

The normalized gene expression value X_{ij} was calculated by dividing the FPKM value of gene 'i' in tissue 'j' with its highest FPKM across all the tested tissues.

$$X_{ij} = \frac{Y_{ij}}{\max\{Y_{ik}\}_{K=1}^{N}}$$

Tissue selectivity score S_{ij} ranges between zero and one, where one denotes a higher propensity of tissue-selective expression. The significance threshold for the tissue-selectivity score was computed through a permutation test. Briefly, we generated 1000 arbitrary gene expression datasets by sampling tissue–gene pairs randomly and calculated tissue-selectivity scores for each such dataset. For each tissue gene pair, we calculated the threshold value as the number of times the random tissue selective scores are greater than the real tissue selective score divided by the number of randomized datasets (1000). For a gene, if we found a tissue with FPKM $>$ 100 with threshold value $<10^{-2}$ then the gene was considered to be selectivity expressed in that tissue.[30,31]

Prediction of molecular recognition features (MoRFs)

Protein binding sites embedded within disordered regions were predicted by the ANCHOR algorithm[32,33] and fMoRFpred algorithm.[34] ANCHOR predicts protein–protein interaction sites that undergo disorder to order transition upon binding on the basis of pairwise inter-residue interaction energies irrespective of its amino acid composition and its secondary structure.[32] This method was proposed to give an unbiased estimate of protein binding capacity.[12] fMoRFpred predicts MoRF regions with the help of support vector machine based on 20 features related to the structural and biochemical characteristics of the input protein sequence. This algorithm was tested with several benchmarking datasets and validated against experimentally supported results in small scales.[34] For each residue in the input sequence, fMorfPred provides a binary classification where '1' denotes an MoRF residue and '0' a non-MoRF residue. Currently, fMoRFpred supports prediction for proteins less than 1000 residues in length. Here, we could predict MoRF regions for 13 281 of 15 472 proteins in our dataset and then we calculated the percentage of MoRF residues in those proteins.

Protein–protein interaction data

Human protein–protein interaction data were retrieved from BioGRID protein interaction repository (v-3.4.144).[35] Currently, BioGRID houses the largest number of interaction pools as compared to the other human interaction databases like HPRD,[36] MIPS,[37] FlyBase,[38] *etc.* Therefore, for systematic analysis of the interaction network, we chose the BioGRID database. Currently, there are interaction data for 21 270 unique human proteins collectively annotated with 279 852 non-redundant interactions. To compute protein connectivity, we considered human binary protein interactions with experimental evidence of physical connections. We removed self-interactions and counted the number of unique interaction partners that a protein connects with (protein connectivity).

Functional enrichment analysis

To determine the functional categories that are significantly over-represented among the genes that are selectively expressed in different human tissues we used the GOrilla Gene Ontology (GO) enrichment analysis tool.[39,40] GOrilla automatically retrieves GO terms (biological process, molecular functions, and cellular components) from gene names or identifiers and compares their distribution either in a ranked gene list or between a target and a background list of genes through rigorous statistical analysis. Along with the details (identifier, description, *P* values, *etc.*) of the terms that are significantly overrepresented in the target list, GOrilla provides a graphical overview of their hierarchical relationships. Here, we compared the distribution of GO terms in tissue-selective genes with respect to their distribution in the total of 15 472 human genes considered for this study.

Statistical analyses

All statistical tests were performed using the SPSS package. Following their non-parametric distribution, we compared the measures of different variables (protein disorder content, gene expression level, and protein length) by the Kruskal–Wallis *H* test, an extended version of the Mann–Whitney *U* test, applicable for comparing distributions between multiple independent groups. To determine significant differences we considered adjusted *P* values corrected for multiple comparisons. For correlation analysis, we calculated non-parametric Spearman's Rank correlation coefficient ρ, where significant correlations were denoted by $P < 0.05$.

Results

Gene expression level of disordered proteins across 32 human tissues

To analyze the gene expression pattern of human disordered proteins at the tissue level we considered the dataset provided by Uhlén *et al.*,[27] with two restrictions that (i) only genes with detectable expression (FPKM value $\geq$ 1) at least in one tissue and (ii) only genes with evidence at the protein level were selected. Genes with no evidence at the protein level were regarded as missing genes or non-coding genes and were suggested to be removed from the list of human protein-coding sequences.[27] Disorder predictions were retrieved from the D2P2[29] database and disorder content was estimated based on the consensus of 5 of 9 prediction algorithms (see materials and methods). Following Edwards *et al.*,[25] we categorized our dataset into five bins in ranges of 0–20% (ordered), 20–40% (moderately disordered), 40–60% (disordered), 60–80% (highly disordered) and 80–100% (extremely disordered) predicted disorder content. As has been suggested earlier,[25] here we noticed that both highly disordered (predicted disorder content $>$60%) and extremely disordered proteins (predicted disorder content $>$80%) are relatively rare in the human proteome (Fig. S1 in Supplementary file 1, ESI†). Following previous studies,[22,27] an FPKM value of 1 was taken as a threshold to estimate the genes expressed in different tissues. Interestingly, among the genes expressed in different tissues (with FPKM $>$ 1), $\sim$3% of genes were found to be extremely disordered (predicted disorder content $>$80%) and $\sim$11–12% of genes were predicted as highly disordered (predicted disorder content $>$60%). Next, we calculated the mean gene expression intensities of ordered and disordered proteins in each individual tissue (Fig. S2 and Table S1 in Supplementary file 1, ESI†).

In 21 of 32 tested tissues (adipose tissue, adrenal gland, appendix, colon, duodenum, esophagus, fallopian tube, gallbladder, heart muscle, lung, pancreas, placenta, prostate, rectum, salivary gland, small intestine, smooth muscle, stomach, thyroid gland, tonsil, and urinary bladder), we did not find a significant difference in gene expression level between any order and disorder categories ($P > 0.05$ for all the pairwise comparison among five disorder categories by Kruskal–Wallis H Test). However, in tissues like the brain, endometrium, lymph node, ovary, skeletal muscle, skin, spleen, and testes, disordered (bin3) and/or highly disordered (bin4) proteins were found to be expressed at a significantly higher level as compared to ordered proteins (bin1) ($P < 0.05$). In contrast, an opposite trend was noticed in three tissues – liver, kidney and bone marrow, where ordered proteins were found to be expressed at a relatively higher level than disordered proteins (Fig. S2 and Table S1 in Supplementary file 1, ESI†). We also tested whether the observed variations in mean gene expression levels between proteins in different disorder bins are due to random chance. For this analysis we generated 100 arbitrary gene expression matrices from our real gene expression dataset by random permutation of tissue gene pairs. Next, in each random dataset we found out the tissues where ordered and disordered proteins differ significantly in their mean gene expression level. Considering all those random datasets (32×100 tissue wise comparisons) we found significant differences in 142 tissues ($\sim$1.5 tissues per random dataset) (Supplementary results S2 in Supplementary file 3, ESI†). In $\sim$50% of tissues where we found significant differences, disordered proteins were found to be expressed at a higher level, while in the remaining $\sim$50% ordered proteins were found to be expressed at a higher level. Moreover, here we did not find any general trend in these tissues. Altogether this suggested that the observed variations are not due to random chance. In addition, the mean gene expression intensity values may have been biased by the very high expression of a few genes in some tissues. To check this possibility we calculated average gene expression intensities after removing the genes with expression intensity >1000 (Fig. S3 in Supplementary file 1, ESI†) and >5000 (Fig. S4 in Supplementary file 1, ESI†) in any of the tested tissues and considered median values instead of mean values (Supplementary file 2, ESI†). Further, we re-annotated proteins into five disordered bins based on the disorder content predicted by the consensus of 6 (Fig. S5 in Supplementary file 1, ESI†) and 7 algorithms (Fig. S6 in Supplementary file 1, ESI†) and compared their mean gene expression levels. When we compared among these datasets (Fig. S2–S6 in Supplementary file 1, ESI†) we noticed a similar trend that in most of the human tissues there is no significant difference in gene expression between any disorder bins (Table S1 in Supplementary file 1, ESI†). For most of the tissues in which we found a significant difference we didn't find any consistent trend, however disordered proteins were found to be expressed at a lower level in the liver and kidneys across all these datasets, while at a higher level in the testes, ovaries and to some extent the brain. Previously, it was ascertained that disordered proteins tend to be expressed at a lower level than ordered globular proteins.[25,26] However, these results imply that in most of the human tissues proteins are expressed at a similar level irrespective of their order and disorder tendencies. Moreover, here we found evidence that disordered proteins may be expressed at a higher level than ordered proteins depending upon the tissue physiology.

Tissue averaged gene expression level of disordered proteins

Based on the tissue averaged gene expression values, previously it was shown that human disordered proteins (predicted disorder content 40–80%), tend to be expressed at a comparatively lower level than ordered globular proteins.[25,26] Thus our results are in apparent conflict with the results shown based on tissue averaged values. To check whether tissue averaged values would reflect a different scenario than what we found in individual tissues, we considered the average gene expression level of all the 32 tissues. As has been reported earlier, here we noticed that disordered and highly disordered proteins (predicted disorder content 40–80%) indeed have significantly lower tissue averaged gene expression levels than ordered proteins (Fig. 1A). However, all significant differences disappeared when we calculated the mean values without considering the tissues (liver, pancreas, and salivary gland) where we found large variation in gene expression between ordered and disordered proteins (Fig. 1B). Thus, these results suggest that the lower tissue averaged gene expression level of disordered proteins, as reported earlier, may have been caused by biased gene expression of these proteins in some specific tissues. To further evaluate the effect of other

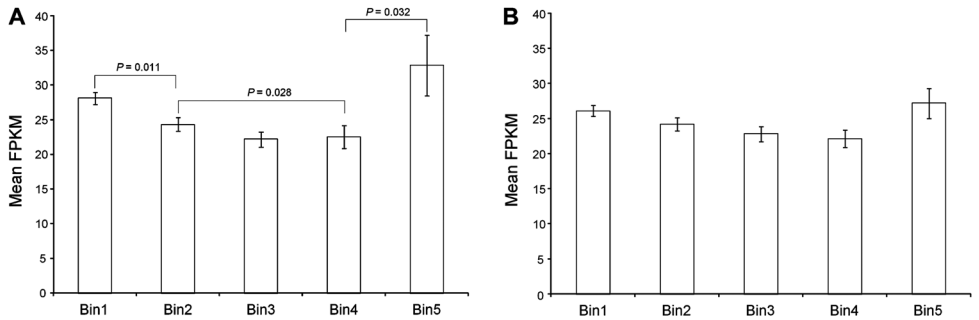

Fig. 1 Tissue-averaged mean gene expression levels of proteins in different disorder bins. (A) Average values calculated considering all 32 tissues, (B) considering 29 tissues (without considering liver, pancreas, and salivary gland where we found large variation in gene expression between ordered and disordered proteins).

factors we considered the impact of protein length. Gene length was regarded as a major determinant of gene expression level[41,42] and was also shown to be correlated with protein disorder content.[8,18] In accordance, here we noticed a significant negative correlation between protein length and tissue averaged gene expression level (Spearman $\rho = -0.132$, $P = 1 \times 10^{-6}$). Consequently, proteins in moderately and highly disorder bins were found to have a significantly higher length as compared to ordered proteins (Fig. 2), suggesting that disordered proteins may have lower gene expression due to their higher protein length. To analyze how protein length influences the correlation between gene expression level and protein disorder content we controlled this effect using partial correlation analysis. The weak correlation between gene expression and protein disorder content was found to disappear (Spearman $\rho = -0.018$, $P = 2.3 \times 10^{-2}$ for correlation between protein disorder content and average gene expression) controlling protein length. To evaluate whether the observed distribution of tissue averaged gene expression intensities has really been influenced by protein length we compared gene expression levels between ordered and disordered proteins of comparable length (in protein length bins). When we controlled the effect of protein length in this way we found no significant difference in mean gene expression intensity between ordered and disordered proteins in most of these bins (Table S2 in Supplementary file 1, ESI†). However, one probable reason for not finding a significant difference may be the lower sample size *i.e.* the number of ordered and disordered proteins to compare in each length bin. To consider this possibility, we randomly sampled 500 proteins from each of the ordered, moderately disordered, disordered and highly disordered protein groups such that the average gene lengths of these groups do not differ significantly. We then checked whether the tissue averaged gene expression level varies significantly between these groups. The extremely disordered group of proteins was not considered for this analysis due to the insufficiency of the dataset required for the randomization procedure. We repeated the procedure 1000 times, however, in more than 95% of cases we did not find any significant difference (at 95% confidence level) in the tissue averaged gene expression level between any disorder bins. Thus, these results suggest that a lower tissue averaged gene expression level among the moderately and highly disordered proteins as has been reported earlier may be the consequence of their higher gene length.

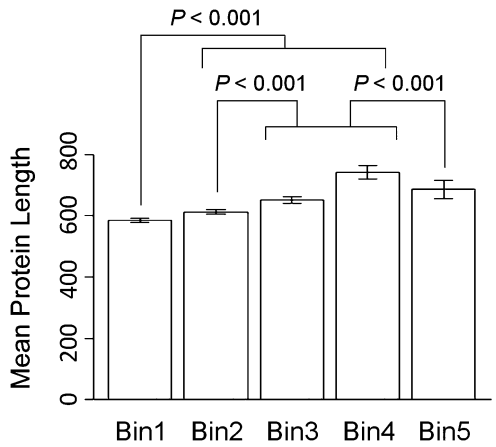

Fig. 2 Average length of proteins in different disorder bins. Significant differences for pair-wise comparison between different groups were evaluated through Kruskal Wallis *H* test and shown with *P*-values.

Disordered proteins and tissue functionality

Our results suggested that although in most of the human tissues ordered and disordered proteins are expressed at a similar level, in some specific tissues disordered proteins tend to be expressed at a higher level than ordered proteins. To delve into this issue further, we analyzed gene expression specificities of ordered and disordered proteins in each individual tissue. Genes that are expressed predominantly in a particular tissue were considered to be important for functional specificities of that tissue.[30,31,43] Therefore, here we considered the genes that are selectively expressed in each of the 32 tissues identified by two approaches (see Materials and methods). From Uhlén's *et al.*,[27] we retrieved 1707 tissue enriched genes, as compared to 1086 tissue-selective genes identified by the second method.[30,31] For most of the tissues, we noticed a high degree of overlap between the lists of tissue-selective genes identified by these two methods (Fig. S7 in Supplementary file 1, ESI†). Moreover, genes which were identified as tissue-selective genes by both these methods (602 genes) were found to have the same tissue specificity. When we compared their protein disorder content, we found that genes that are selectively expressed in tissues like the testes, brain, *etc.* have a higher protein disorder content as compared to the genes that are expressed selectively in the liver, pancreas, kidney *etc.* tissues (Fig. 3). The higher protein disordered content of tissue-selective genes may suggest that disordered residues are indispensable for the proper functioning of the former group of tissues. In this context, we found it interesting to analyze why the genes that are selectively expressed in the former group of tissues encode more disordered residues as compared to the other groups of tissue-selective genes. Previously, it was ascertained that proteins that connect with a large number of partners in their interaction network (hub proteins) are more disordered as compared to the proteins that interact with few partners.[11,44] Consequently, we compared different groups of tissue-selective genes in terms of their protein connectivity. In favor of their higher disorder content, here we noticed that genes that are selectively expressed in tissues like the testes and brain, *etc.* share higher protein connectivity than the genes that are selectively expressed in the liver or kidneys (Fig. 4). Proteins with higher connectivity were shown to encode a large number of disordered binding regions (protein binding sites within disordered regions) for their binding promiscuity.[32] Therefore, we predicted disordered binding sites using two algorithms – (1) ANCHOR[33] and (2) fMoRFpred,[34] both of which suggested that the genes that are selectively expressed in the former group of tissues (testes, brain, *etc.*) encode a greater fraction of such motifs than the liver and kidney *etc.* tissue-selective genes (Fig. 5). This may suggest that a higher fraction of disordered residues among the

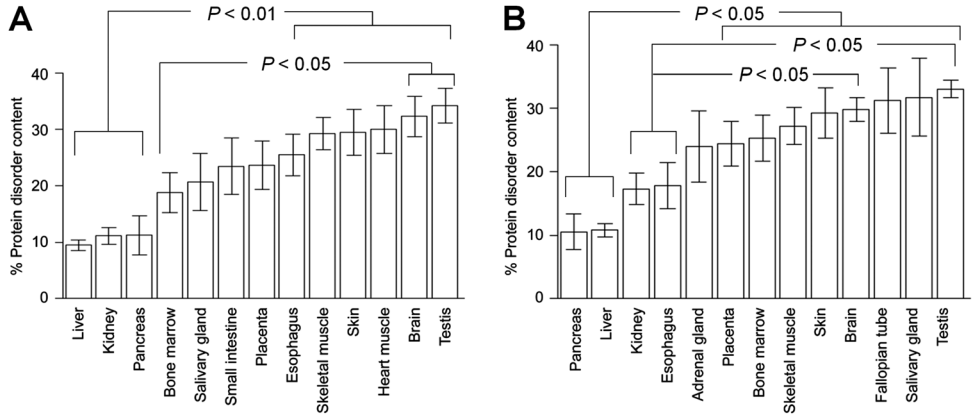

Fig. 3 Average protein disorder content of different groups of tissue-enriched genes. A protein disorder content was retrieved from the D2P2 database and tissue-selective genes were identified using two methods: (A) Uhlen *et al.* and (B) Chang *et al.*, and Greco *et al.* Here, tissues with 30 or more selective genes were shown. Significant differences in protein disorder content between the different groups of tissue selective genes were evaluated through Kruskal Wallis *H* test shown with *P*-values.

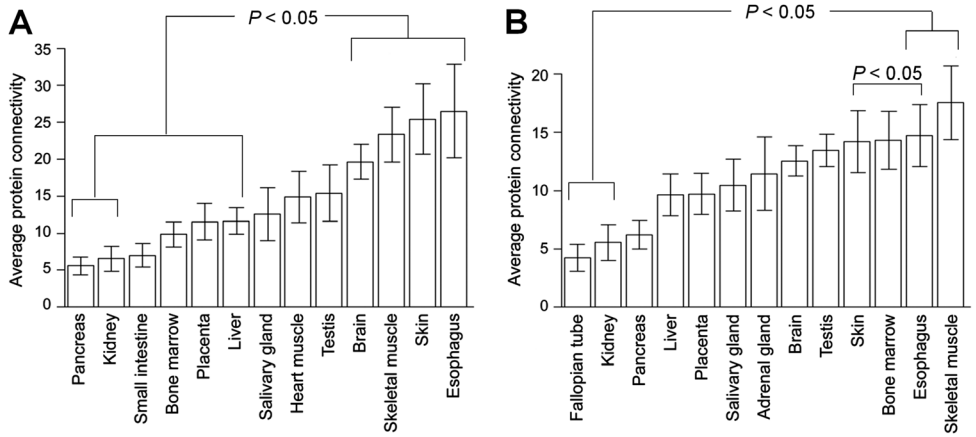

Fig. 4 Average protein connectivity of different groups of tissue-enriched genes. Only tissues with 30 or more tissue enriched genes were shown. (A) Tissue-selective genes retrieved from Uhlen *et al.* and (B) those identified following Chang *et al.*, and Greco *et al.* Significant differences between the different groups of tissue-selective genes were evaluated through Kruskal Wallis *H* test and shown with *P*-values.

former groups of tissue-selective genes is a prerequisite for forming protein–protein interaction sites. Next, we tested the influence of gene functionalities. Previous studies have grouped different functional keywords according to their ordered and disordered tendencies.[45–47] In particular, proteins involved in signal transduction, regulation, protein transport and development and differentiation-related processes were shown to be more disordered as compared to the proteins which mainly function in ion transport, metabolic and enzymatic activities.[4,45–47] When we analyzed the functional association of different groups of tissue-selective genes we noticed that genes that are selectively expressed in the testes, brain, and ovaries are enriched with disorder-related functions (cell cycle, reproductive processes, signaling, regulation, and cell differentiation, *etc.*) while the genes that are expressed mainly in the liver and kidneys are enriched with terms that rely on globular proteins (ion transport, transmembrane transport, metabolic processes, and regulation of metabolic processes, *etc.*) (Supplementary file 4, ESI†). These inherent biases towards disorder related functions may also account for the higher disorder content among the former groups of tissue-selective genes.

Discussion

Analysis of the gene expression pattern across tissues and organs was considered to be crucial for the understanding of human disease and biology. Expression levels can provide important clues about the phenotypes and functionalities of genes across different tissues and their regulatory mechanisms.[23,30,31] Although disordered proteins are considered as a predominant class, specifically among higher eukaryotes,[2,11,16,17,19] to date little attention has been paid to investigating their gene expression profile. In this study, we retrieved high-throughput gene expression data for more than 15 000 human proteins from published literature and analyzed their gene expression signature across 32 normal human tissues. Since disordered proteins are vulnerable towards protein aggregation, previously it was suggested that cells need

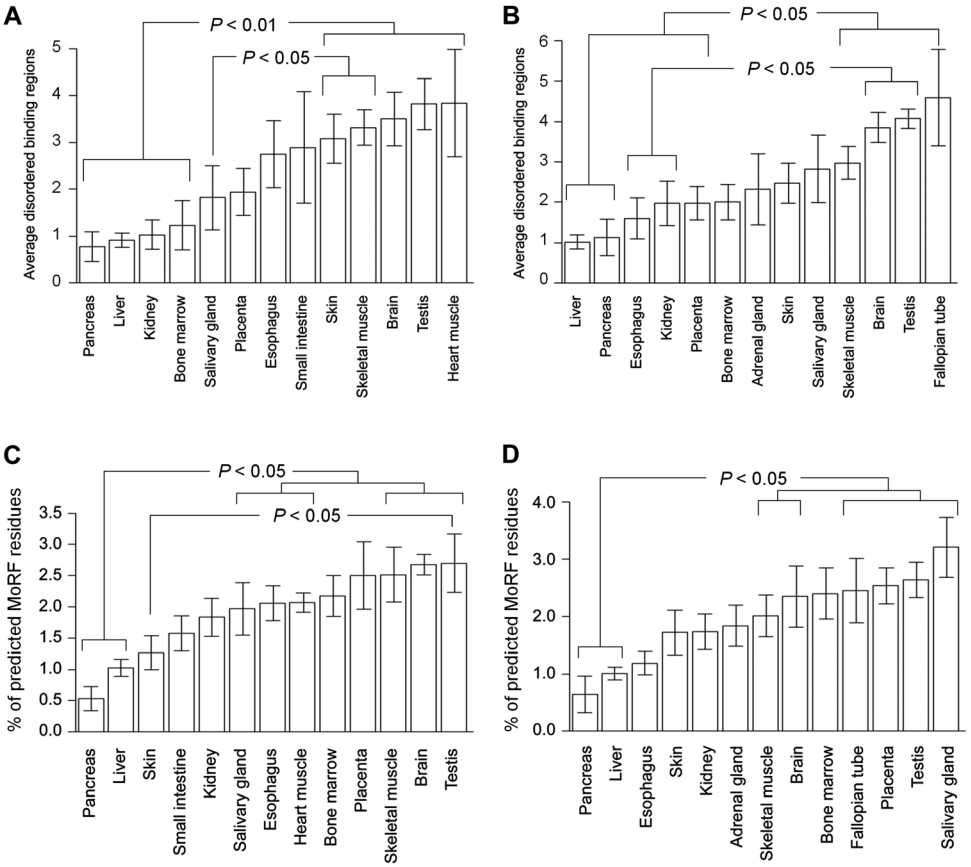

Fig. 5 Average protein disordered binding regions of different groups of tissue enriched genes. Only tissues with 30 or more tissue enriched genes were shown. (A and B) Using ANCHOR for (A) Uhlen *et al.* and (B) Chang *et al.*, and Greco *et al.* datasets. (C and D) Using fMoRFpred for (C) Uhlen *et al.* and (D) Chang *et al.*, and Greco *et al.* datasets. Significant differences between the different groups of tissue selective genes were evaluated through the Kruskal Wallis *H* test shown with *P*-values.

intricate regulatory mechanisms to maintain their concentration below a certain limit.[25,26,48] However, here we did not find any general trend of low expression of disordered proteins except in a few specific tissues (Fig. S2–S6 and Table S1, Supplementary file 1, ESI†). Moreover, our results suggested that in a number of human tissues disordered proteins tend to be expressed at a higher level than ordered globular proteins. Based on the tissue averaged gene expression intensity, previously Gsponer *et al.*,[26] have shown that human disordered proteins tend to be expressed at a lower level than globular proteins. Consequently, disordered proteins were shown to contain a higher proportion of ubiquitination and micro-RNA target sites and high mRNA decay rates suggesting a complex association between gene expression level and protein intrinsic disorder content.[25] Considering the mean gene expression level of 32 human tissues, here we observed a similar trend. However, we did not find any significant difference when we calculated mean values without considering three tissues (liver, pancreas, salivary gland) which may imply that the lower tissue-averaged values are caused by the low gene expression level of disordered proteins in some specific tissues. Previous studies suggested that longer genes tend to be expressed at a lower level than shorter genes.[41,42] Accordingly, here we noticed a similar trend in each and every individual tissue considered in this study. Considering this, together with the fact that moderately and highly disordered proteins are longer than ordered globular proteins (Fig. 2) here we assumed that protein length may have some influence on the correlation between protein disorder and gene expression level. This became clear from partial correlation analysis where the correlation between gene expression and disorder content disappeared after controlling protein length. In addition, comparing the expression levels of genes having a similar protein length, no significant difference was observed between the disorder groups, suggesting that the protein length, rather than protein disorder content is the major determinant of gene expression level here. Therefore, overall this study suggests that the previously accepted impression that disordered proteins are expressed at a lower level than ordered protein holds true for only a few tissues, and is mostly influenced by their higher protein length.

In the next part, we tried to explore the functional significance of disordered proteins in human tissues by considering the disorder content of tissue-enriched genes. Previously, great interest has been paid to characterizing different human tissues in terms of their transcriptome profile.[22,24,49,50] These studies suggested that most, if not all, of the human tissues express a few genes predominantly which are crucial for maintaining

their functional differences with other tissues as well as for their development and differentiation.[22,24,30,31,43] Several methods have been proposed earlier to evaluate whether a gene has an affinity to be expressed in a particular tissue selectively.[51] To underscore the proteins that are important for the proper functioning of different human tissues here we considered two such approaches and identified the genes which show predominant expression in each tissue individually. Functional analysis of selectively expressed genes for the tissues where we found an adequate number of such genes suggested an overall concurrence with the function of the respective tissues. Comparing their protein disorder content, here we noticed a higher fraction of disordered residues among the genes expressed mainly in the testes, brain *etc.* tissues as compared to those expressed predominantly in the liver, pancreas, kidney *etc.* tissues suggesting that disordered proteins may have important functional consequences for the former group of tissues. Consequently, our analysis suggested that the proteins encoded by the former group of tissue-selective genes interact with a higher number of partners in their protein interaction network than the latter group of tissue-selective genes (liver, pancreas, kidney, esophagus, *etc.*). Disordered proteins provide internal flexibility during protein–protein interaction and facilitate promiscuous binding.[1,2,11] Therefore, highly connected proteins (hub-proteins) were shown to be enriched with intrinsically disordered regions.[11] Higher protein disorder among the former group of tissue-selective genes may suggest that disordered regions are crucial to maintain their higher protein connectivity. In order to further explore the role of disordered proteins in tissue functionalities, here we carefully examined the presence of disordered binding sites among the different groups of tissue-selective proteins. Disordered proteins interact through fly-casting mechanisms where they undergo folding upon binding. Disordered binding regions act as elementary units in molecular recognition that facilitate high-specificity and low-affinity interaction, a specific signature of disordered proteins.[32] Thus, the higher proportion of disordered binding regions among the former group of tissue-selective genes (Fig. 5) may be considered as an indication that disordered residues help these tissues to sustain their functional specificity by providing structural flexibility for binding promiscuity. We also observed that in tissues where tissue-selective genes are enriched in protein disorder, the disorder associated functions like cell cycle, reproductive processes, signaling, regulation, and cell differentiation, *etc.* are overrepresented. In contrast, in tissues having low disorder content in tissue-selective genes, the globular protein-associated terms like ion transport, transmembrane transport, metabolic processes, and regulation of metabolic processes, *etc.* are overrepresented[1–5] (Supplementary file 4, ESI†). Our results suggested a strong deterioration in mean gene expression level of disordered proteins only in the liver and kidneys. The liver is the most metabolically active tissue in the human body[53] and the kidneys are also associated with the elimination of metabolic wastes. The pancreas is composed of both endocrine and exocrine glands whose main function is to produce enzymes and hormones.[54] Functional analysis of the genes specific to these two tissues suggested that these genes are mostly involved in functions which need a relatively lower fraction of disordered residues. Indeed, when we look closely into the genes selectively expressed in these tissues $>80\%$ of the genes were found to have predicted disorder content $<20\%$ (by consensus of five algorithms). Among the liver-expressed genes, there were 11 genes (AADAC, ADH6, CFHR3, CFI, CPB2, CYP8B1, F11, FGL1, FMO3, PON1, SERPINA6) with $<1\%$ of predicted disordered residues most of which are enzymes involved in different catalytic mechanisms. Among the pancreas-enriched genes, we found six (AMY2A, AMY2B, CPA1, CPB1, CTRB2, FBXW12, GRPR, PNLIP) genes with predicted disordered residues $<1\%$ four of which encode proteins with enzymatic functions. On the other hand, $>50\%$ of genes selectively expressed in the testes and brain fall into different disorder categories with a predicted disorder content of more than 20%. The testes are male reproductive organ whose main function is to develop male-specific characteristics.[50] Most of the proteins expressed selectively in the testes are involved in spermatogenesis, a process that needs intricate regulation.[55] Genes showing elevated expression in the testes are tightly regulated starting from synthesis to degradation and are mostly involved in different types of molecular binding.[52] Proteins involved in a binding mechanism will certainly need a high amount of disordered residues to interact with a large number partners as we observed in our study. Among the testis-expressed genes, we noticed 15 completely disordered proteins (PAGE1, TNP1, PRM2, VCY1B, VCY, PAGE5, VCX3A, PCP2, PRM1, VCX2, TNP2, VCX, GAGE2A, VCX3B, SRRM5) which play key roles in different phases of spermatogenesis and are involved in nuclear signaling and regulatory processes. The brain is the most complex organ of the human body which expresses genes mostly associated with developmental processes and synaptic signaling.[56] Here we found 11 brain-specific genes (AMER2, FAM107A, MAPT, BAALC, ERMN, VGF, CPLX1, SRRM4, CPLX2, NRGN, MBP) with predicted disorder content $>90\%$ which are involved in various neurological processes. Altogether, our study relates to the specialized functionalities of the tissue enriched genes of both groups, from the reproductive process or the cellular differentiation in the testes[52] to the cellular signaling indispensable for the functionality of brain[56] in the disorder-rich class and from the metabolic processes and their regulation in tissues like the liver[53] in the disorder poor class.

Conclusions

Disordered proteins provide flexibility in protein functionalities. Due to their binding promiscuity, IDPs are considered as hubs in protein interaction networks where they interact with several other proteins. Considering the risk associated with increased use of disordered proteins, previously it was suggested that the gene expression level of disordered proteins is tightly regulated at multiple layers of transcriptional control machinery.[26] However, the probability of interaction largely depends upon the availability of interacting proteins.[57] Therefore, reduction of the gene expression level of disordered proteins may prove detrimental

to the interaction network. So, the negative correlation between protein intrinsic disorder and gene expression level in humans as was obtained by previous studies seems debatable. In this study, we explored the gene expression profile of human disordered proteins across 32 normal human tissues. Our results indicated that disordered proteins do not have any definite association with gene expression levels, instead lower gene expression of these proteins resulted from their biased gene expression in some specific tissues and their higher protein length. Moreover, here we found evidence that tissues like the testes, ovaries, brain, *etc.* predominantly express genes encoding disordered residues to sustain their high protein connectivity through a higher number of disordered protein binding sites and are associated with functions that are signatures of disordered proteins.

Abbreviation

IDP	Intrinsically disordered proteins
FPKM	Fragments per kilo base of exon model per million mapped reads.
D2P2	Database of Disordered Protein Predictions
MoRF	Molecular Recognition Features
GO	Gene Ontology

Conflicts of interest

The authors declare that they have no conflict of interest.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive suggestions and useful comments. The authors thank Bose Institute and the Department of Science and Technology, Government of India for their financial support. The authors are also thankful to Mr Sanjib Gupta for technical help and Dr Tina Begum and Dr Sandip Chakraborty for useful discussions.

References

1 J. Gsponer and M. M. Babu, *Prog. Biophys. Mol. Biol.*, 2009, **99**, 94–103.
2 J. Habchi, P. Tompa, S. Longhi and V. N. Uversky, *Chem. Rev.*, 2014, **114**, 6561–6588.
3 V. N. Uversky, *Int. J. Biochem. Cell Biol.*, 2011, **43**, 1090–1103.
4 A. K. Dunker, I. Silman, V. N. Uversky and J. L. Sussman, *Curr. Opin. Struct. Biol.*, 2008, **18**, 756–764.
5 A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, *Biochemistry*, 2002, **41**, 6573–6582.
6 P. E. Wright and H. J. Dyson, *Nat. Rev. Mol. Cell Biol.*, 2015, **16**, 18–29.
7 G. J. P. Rautureau, C. L. Day and M. G. Hinds, *Int. J. Mol. Sci.*, 2010, **11**, 1808–1824.
8 S. C.-C. Chen, T.-J. Chuang and W.-H. Li, *Mol. Biol. Evol.*, 2011, **28**, 2513–2520.
9 A. Panda, T. Begum and T. C. Ghosh, *PLoS One*, 2012, **7**, e48336.
10 M. Arai, K. Sugase, H. J. Dyson and P. E. Wright, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 9614–9619.
11 C. Haynes, C. J. Oldfield, F. Ji, N. Klitgord, M. E. Cusick, P. Radivojac, V. N. Uversky, M. Vidal and L. M. Iakoucheva, *PLoS Comput. Biol.*, 2006, **2**, e100.
12 E. Schad, P. Tompa and H. Hegyi, *Genome Biol.*, 2011, **12**, R120.
13 B. Xue, A. K. Dunker and V. N. Uversky, *J. Biomol. Struct. Dyn.*, 2012, **30**, 137–149.
14 J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J. Mol. Biol.*, 2004, **337**, 635–645.
15 A. K. Dunker, P. Romero, Z. Obradovic, E. C. Garner and C. J. Brown, *Genome Inf.*, 2000, **11**, 161–171.
16 M. Y. Lobanov and O. V. Galzitskaya, *Int. J. Mol. Sci.*, 2015, **16**, 19490–19507.
17 Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky and L. Kurgan, *Cell. Mol. Life Sci.*, 2015, **72**, 137–151.
18 A. Panda and T. C. Ghosh, *Gene*, 2014, **548**, 134–141.
19 N. Pietrosemoli, J. A. García-Martín, R. Solano and F. Pazos, *PLoS One*, 2013, **8**, e55524.
20 A. Panda, S. Podder, S. Chakraborty and T. C. Ghosh, *Genomics*, 2014, **104**, 530–537.
21 S. Chakrabortee, J. S. Byers, S. Jones, D. M. Garcia, B. Bhullar, A. Chang, R. She, L. Lee, B. Fremin and S. Lindquist, *Cell*, 2016, **167**, 369–381.
22 L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson and K. Edlund, *Mol. Cell. Proteomics*, 2014, **13**, 397–406.
23 D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden and S. A. Teichmann, *Mol. Syst. Biol.*, 2011, **7**, 497.
24 M. Uhlen, P. Oksvold, L. Fagerberg, E. Lundberg, K. Jonasson, M. Forsberg, M. Zwahlen, C. Kampf, K. Wester and S. Hober, *Nat. Biotechnol.*, 2010, **28**, 1248–1250.
25 Y. J. K. Edwards, A. E. Lobley, M. M. Pentony and D. T. Jones, *Genome Biol.*, 2009, **10**, R50.
26 J. Gsponer, M. E. Futschik, S. A. Teichmann and M. M. Babu, *Science*, 2008, **322**, 1365–1368.
27 M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen and F. Pontén, *Science*, 2015, **347**, 1260419.
28 P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. García Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo,

S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. P. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino and S. M. J. Searle, *Nucleic Acids Res.*, 2014, **42**, D749–D755.

29 M. E. Oates, P. Romero, T. Ishida, M. Ghalwash, M. J. Mizianty, B. Xue, Z. Dosztányi, V. N. Uversky, Z. Obradovic, L. Kurgan, A. K. Dunker and J. Gough, *Nucleic Acids Res.*, 2013, **41**, D508–D516.

30 C.-W. Chang, W.-C. Cheng, C.-R. Chen, W.-Y. Shu, M.-L. Tsai, C.-L. Huang and I. C. Hsu, *PLoS One*, 2011, 6.

31 D. Greco, P. Somervuo, A. Di Lieto, T. Raitila, L. Nitsch, E. Castrén and P. Auvinen, *PLoS One*, 2008, **3**, e1880.

32 B. Mészáros, I. Simon and Z. Dosztányi, *PLoS Comput. Biol.*, 2009, **5**, e1000376.

33 Z. Dosztányi, B. Mészáros and I. Simon, *Bioinformatics*, 2009, **25**, 2745–2746.

34 J. Yan, A. K. Dunker, V. N. Uversky and L. Kurgan, *Mol. BioSyst.*, 2016, **12**, 697–710.

35 C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, *Nucleic Acids Res.*, 2006, **34**, D535–D539.

36 S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. B. Gandhi, K. N. Chandrika, N. Deshpande and S. Suresh, *Nucleic Acids Res.*, 2004, **32**, D497–D501.

37 H.-W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd and B. Weil, *Nucleic Acids Res.*, 2002, **30**, 31–34.

38 P. McQuilton, S. E. S. Pierre, J. Thurmond and C. FlyBase, *Nucleic Acids Res.*, 2011, gkr1030.

39 E. Eden, R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini, *BMC Bioinf.*, 2009, **10**, 48.

40 E. Eden, D. Lipson, S. Yogev and Z. Yakhini, *PLoS Comput. Biol.*, 2007, **3**, e39.

41 A. O. Urrutia and L. D. Hurst, *Genome Res.*, 2003, **13**, 2260–2264.

42 C. I. Castillo-Davis, S. L. Mekhedov, D. L. Hartl, E. V. Koonin and F. A. Kondrashov, *Nat. Genet.*, 2002, **31**, 415–418.

43 S. Liang, Y. Li, X. Be, S. Howes and W. Liu, *Physiol. Genomics*, 2006, **26**, 158–162.

44 A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky, *FEBS J.*, 2005, **272**, 5129–5148.

45 S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1899.

46 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J. Proteome Res.*, 2007, **6**, 1882.

47 H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1917.

48 M. M. Babu, R. van der Lee, N. S. de Groot and J. Gsponer, *Curr. Opin. Struct. Biol.*, 2011, **21**, 432–440.

49 M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle and A. Teufel, *Bioinformatics*, 2012, **28**, 1184–1185.

50 D. Djureinovic, L. Fagerberg, B. Hallström, A. Danielsson, C. Lindskog, M. Uhlén and F. Pontén, *Mol. Hum. Reprod.*, 2014, gau018.

51 N. Kryuchkova-Mostacci and M. Robinson-Rechavi, *Briefings Bioinf.*, 2016, bbw008.

52 M. T. Anand and B. V. L. S. Prasad, *J. Hum. Reprod. Sci.*, 2012, **5**, 266.

53 C. Kampf, A. Mardinoglu, L. Fagerberg, B. M. Hallström, K. Edlund, E. Lundberg, F. Pontén, J. Nielsen and M. Uhlen, *FASEB J.*, 2014, **28**, 2901–2914.

54 M. J. Muraro, G. Dharmadhikari, D. Grün, N. Groen, T. Dielen, E. Jansen, L. van Gurp, M. A. Engelse, F. Carlotti and E. J. P. de Koning, *Cell Syst.*, 2016, **3**, 385–394.

55 S. R. Grimes, *Gene*, 2004, **343**, 11–22.

56 E. Sjöstedt, L. Fagerberg, B. M. Hallström, A. Häggmark, N. Mitsios, P. Nilsson, F. Pontén, T. Hökfelt, M. Uhlén and J. Mulder, *PLoS One*, 2015, **10**, e0130028.

57 A. Bossi and B. Lehner, *Mol. Syst. Biol.*, 2009, **5**, 260.

