

The Human RBPome : from genes and proteins to human disease

Yaseswini Neelamraju¹, Seyedsasan Hashemikhabir¹, Sarath Chandra Janga^{1, 2, 3,*}

¹Department of Biohealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Indianapolis

²Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, HITS, Indianapolis

³Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, Indiana



Abstract

RNA Binding Proteins (RBPs) play a central role in mediating post transcriptional regulation of genes. However, less is understood about them and their regulatory mechanisms. In this study, we construct a repertoire of 1344 genes encoding RBPs identified from several experimental studies and present a comprehensive analysis to understand their characteristics at a global scale. The domain architecture of RBPs enabled us to classify them into three groups - Classical (29%), Non-classical (19%) and Unclassified (52%). A higher percentage of proteins with unclassified domains reveals the presence of various uncharacterised motifs that can potentially bind RNA. In addition, enrichment of various unconventional superfamilies' suggest that RBPs could form an integral part of the cellular architecture. RBPs were found to be highly disordered compared to non-RBPs ($p < 2.2e-16$, Fisher's exact test), indicating a dynamic regulatory role of RBPs in cellular functioning. Evolutionary analysis in 62 different species showed that RBPs are highly conserved compared to non-RBPs ($p < 2.2e-16$, Wilcoxon-test), reflecting a conservation of various biological processes like mRNA splicing, ribosome biogenesis. Expression patterns of RBPs from human proteome map revealed that majority (~60%) of the RBPs are tissue-specific. Additionally, non-classical proteins were found to be highly expressed than the classical proteins ($p < 0.05$, Wilcoxon test) in ~50% of the tissues. RBPs were also seen to be highly associated with several neurological disorders, cancer and inflammatory diseases. Further, anatomical context like B cells, T-cells, Fetal Liver and Fetal Brain were found to be enriched, implying a prominent role of RBPs in mediating immune responses and different developmental stages. These analyses are made accessible to researchers in the form of a database called RNA Binding protein expression and disease dynamics database (READ DB).

Materials and Methods

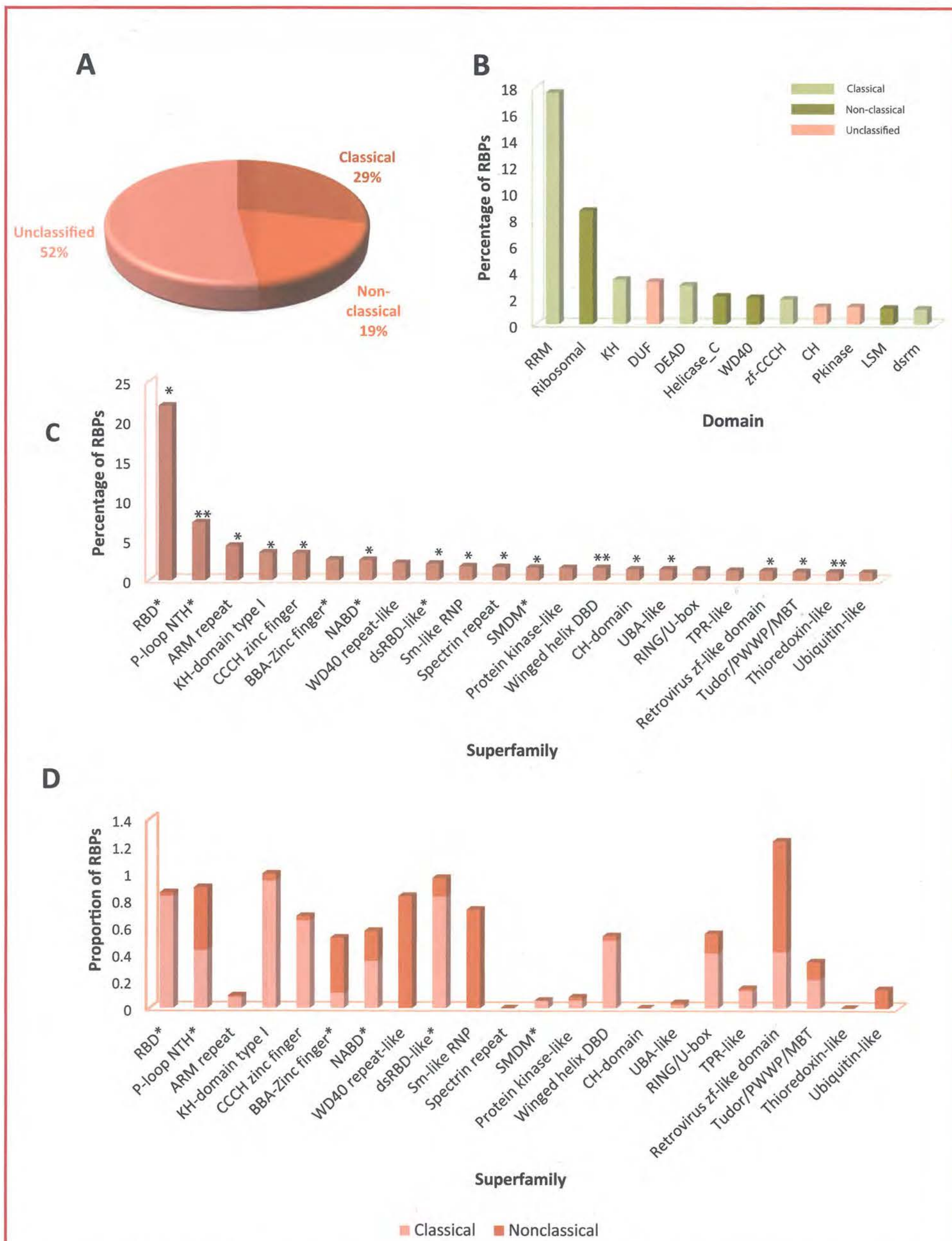
Data collection:

We constructed a catalogue of 1344 genes encoding RBPs that were identified from recent studies [1-5].

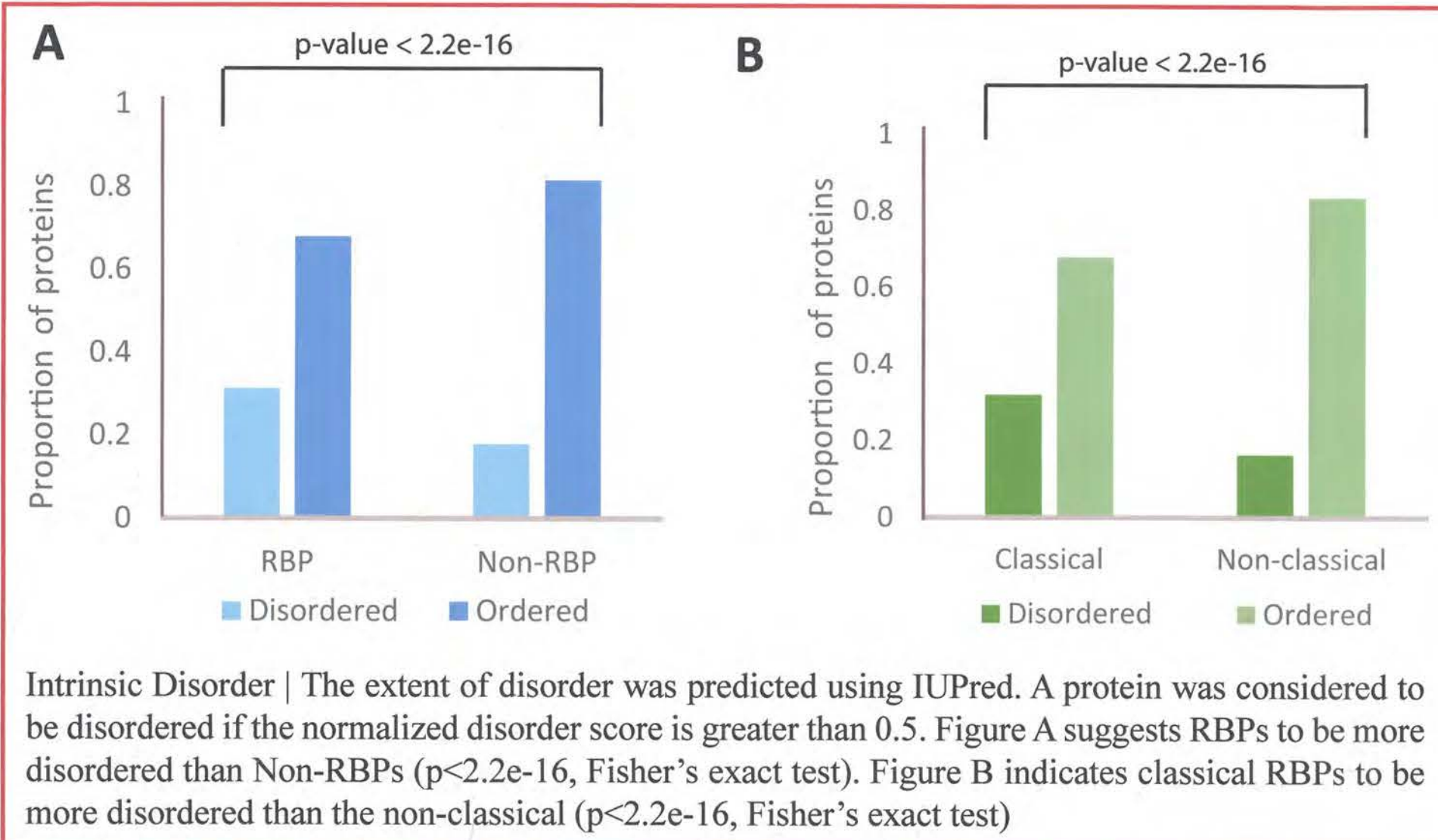
Methods:

The domain architecture of RBPs was analysed using the annotations reported in Pfam and Superfamily databases. Further, the intrinsic disorder nature was predicted using IUPred, which predicts disorder on a per residue basis. To understand the evolutionary conservation, we obtained the orthologs in 62 different species from the ENSEMBL compara (v73). The expression patterns in 25 different tissues were studied using the proteome data from the Human proteome map. Lastly, diseases associated with RBPs and the related anatomical context were extracted from the Malacards database.

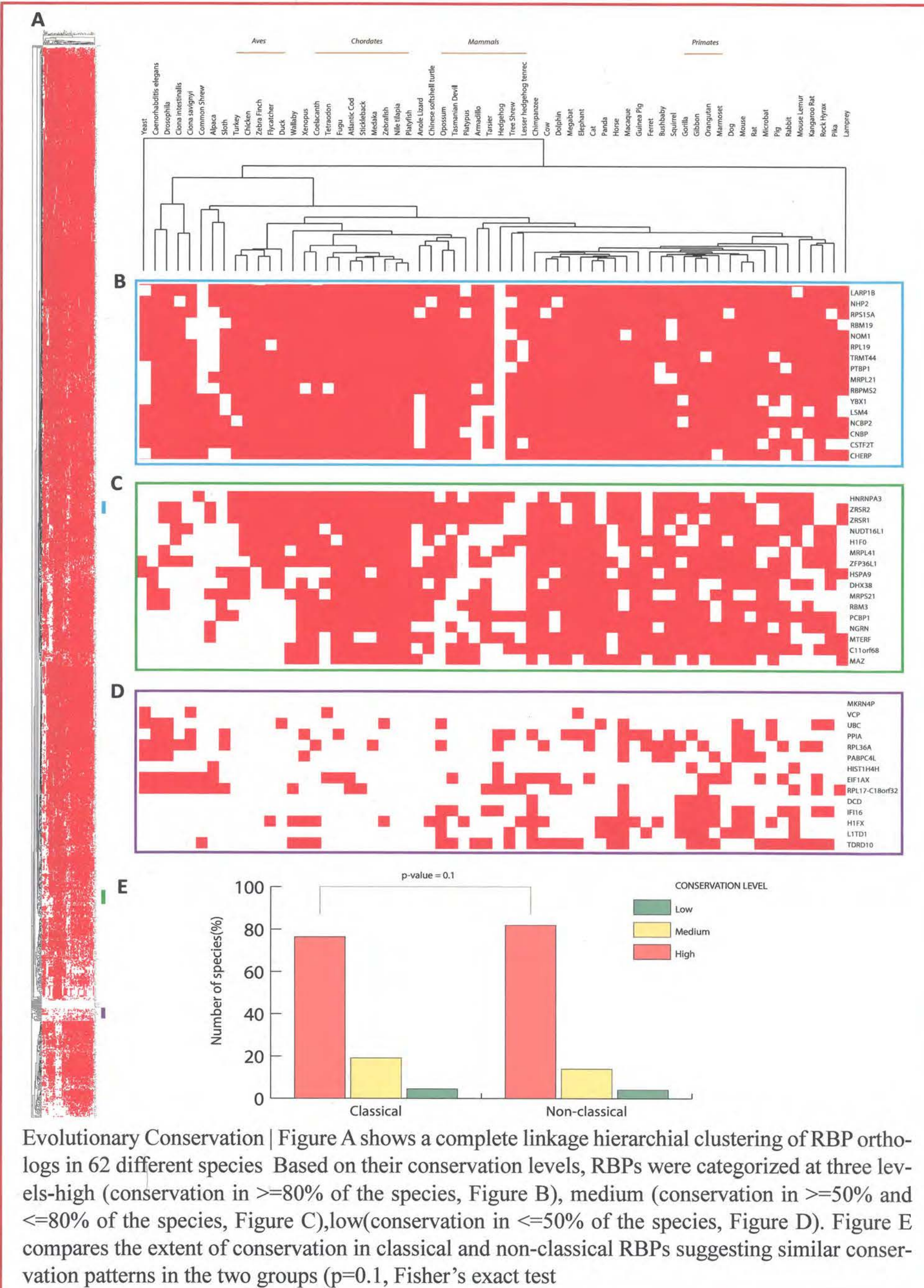
Results



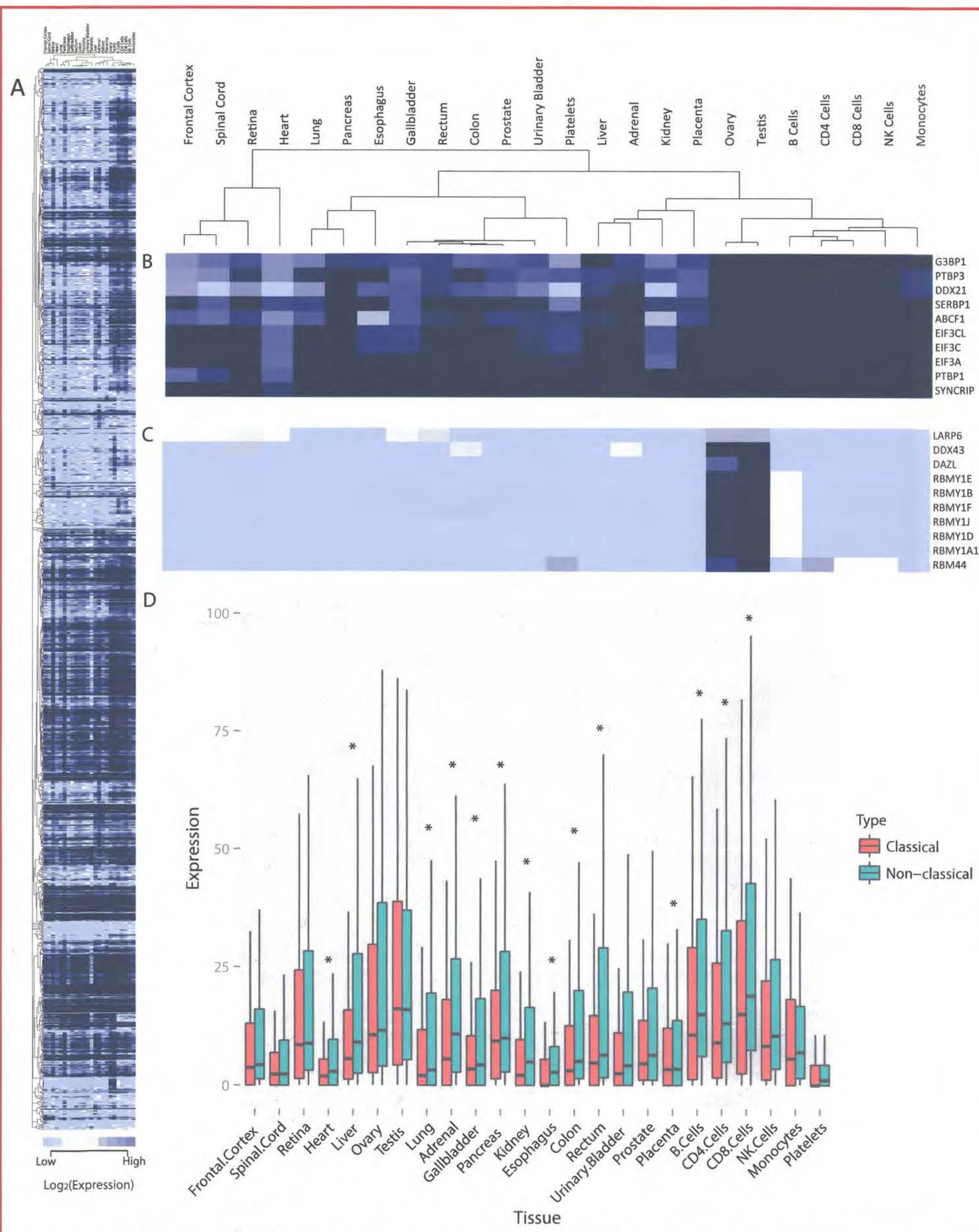
Domain Architecture | Figure A shows the overall domain distribution in the RBP repertoire. Definitions of classical and non-classical were obtained from a recent study [1]. RBPs that could not be classified in either classes were termed 'Unclassified'. Figure B shows the Pfam domain distribution. Figure C shows the superfamily distribution (* indicates $p < 1E-06$ and ** indicate $p < 0.05$, Fisher's exact test). Figure D shows the proportion of RBPs belonging to each of the superfamily.



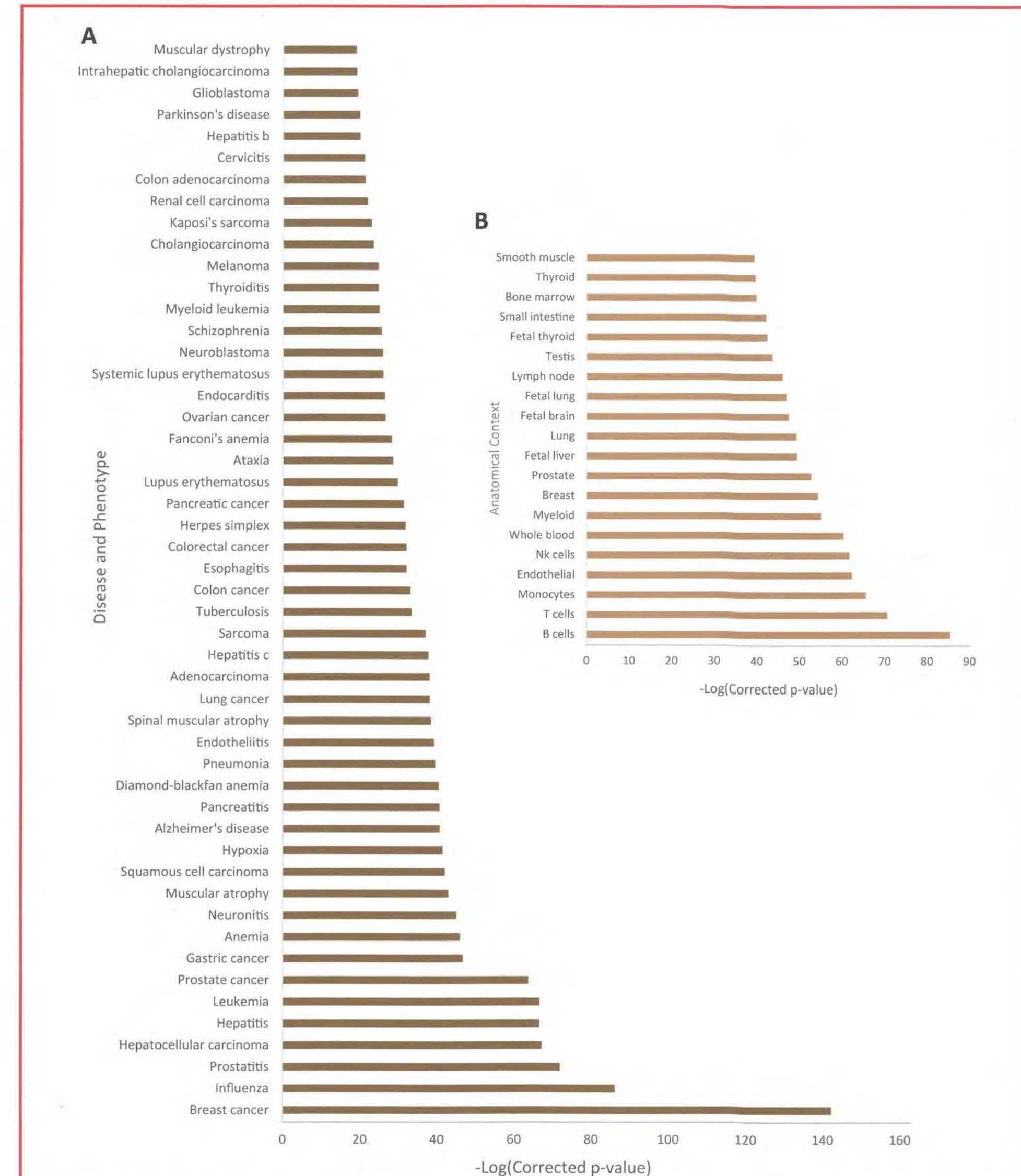
Intrinsic Disorder | The extent of disorder was predicted using IUPred. A protein was considered to be disordered if the normalized disorder score is greater than 0.5. Figure A suggests RBPs to be more disordered than Non-RBPs ($p < 2.2e-16$, Fisher's exact test). Figure B indicates classical RBPs to be more disordered than the non-classical ($p < 2.2e-16$, Fisher's exact test).



Evolutionary Conservation | Figure A shows a complete linkage hierarchical clustering of RBP orthologs in 62 different species. Based on their conservation levels, RBPs were categorized at three levels-high (conservation in $\geq 80\%$ of the species, Figure B), medium (conservation in $\geq 50\%$ and $\leq 80\%$ of the species, Figure C), low (conservation in $\leq 50\%$ of the species, Figure D). Figure E compares the extent of conservation in classical and non-classical RBPs suggesting similar conservation patterns in the two groups ($p = 0.1$, Fisher's exact test).



Expression patterns of RBPs | Protein expression levels of RBPs across 25 different tissues were obtained from the human protein catalogue (Kim et al, 2014). Complete linkage hierarchical clustering of the expression data is shown in Figure A. Based on the expression patterns, RBPs were classified as ubiquitous and tissue-specific (Yanani et al, 2005). Figure B and C show a subset of RBPs that are ubiquitous and tissue-specific respectively. Figure D compares the expression levels of classical and non-classical RBPs in each tissue (* indicates $p < 0.05$, Wilcoxon test).

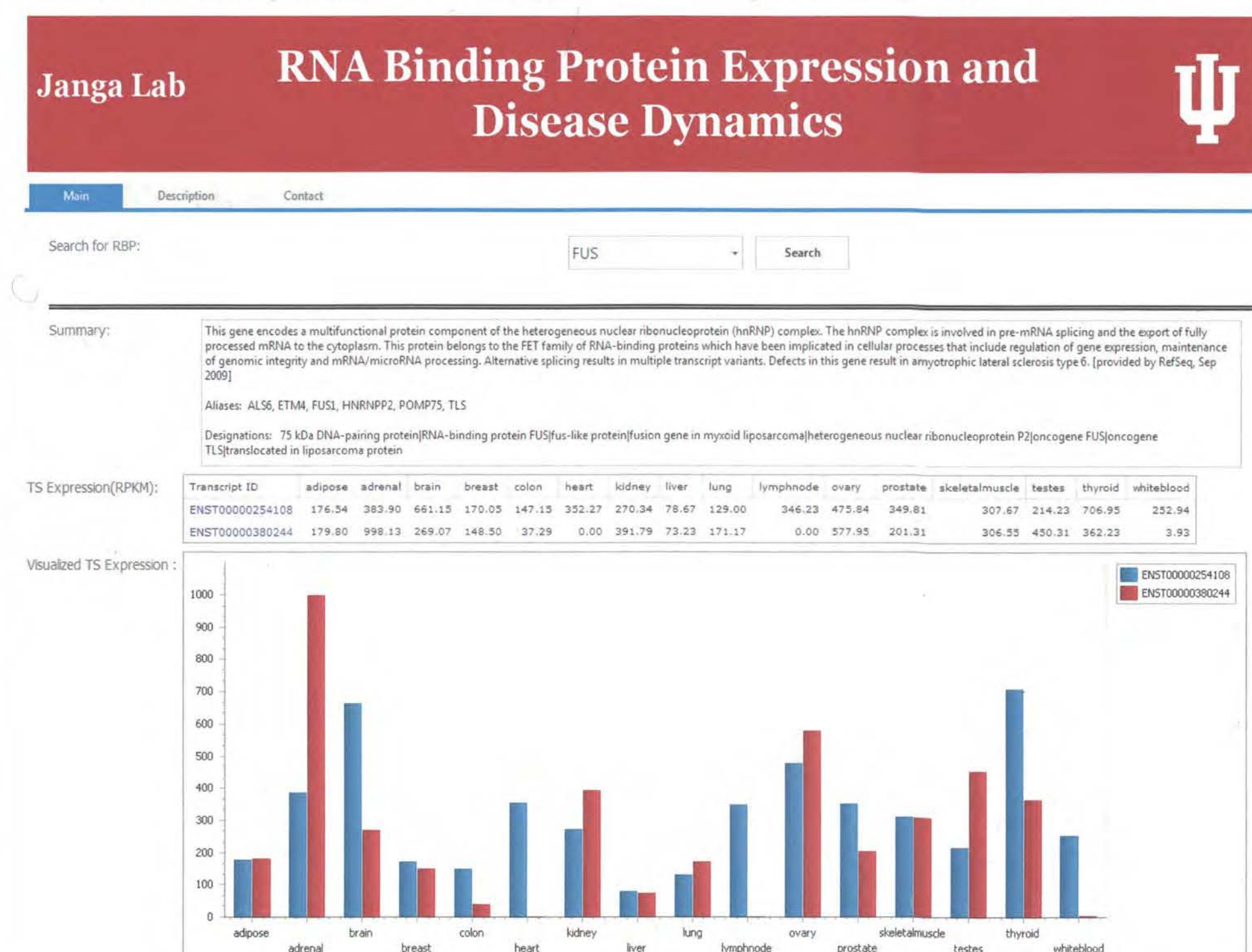


Disease associations | Diseases associated with RBPs were obtained from malacards. 165 diseases (Number of associations = 10, $p < 1E-05$, $FDR < 1$) were found to be enriched for RBPs. The anatomical context associated with these 165 diseases were identified and 70 anatomical contexts were found to be enriched.

Conclusion

- This study revealed several unconventional domains to be associated with RBPs.
- It was seen that RBPs are more intrinsically disordered than the rest of the proteome and classical proteins are more disordered when compared to the non-classical proteins.
- RBPs are highly conserved than the remaining genome although, the classical and non-classical proteins did not show a significant difference in their conservation levels.
- Expression analysis showed that ~40% of the repertoire is ubiquitously expressed and the remaining are tissue specific whereas non-classical were seen to be highly expressed in ~50% of the tissues.
- Several novel diseases and anatomical contexts associated with RBPs were found to be enriched.

These analyses are made available through our database READ DB



References

- Castello, A., et al., Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell, 2012. 149(6): p. 1393-406.
- Kwon, S.C., et al., The RNA-binding protein repertoire of embryonic stem cells. Nat Struct Mol Biol, 2013. 20(9): p. 1122-30.
- Cook, K.B., et al., RBPDB: a database of RNA-binding specificities. Nucleic Acids Res, 2011. 39(Database issue): p. D301-8.
- Baltz, A.G., et al., The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol Cell, 2012. 46(5): p. 674-90.
- Ray, D., et al., A compendium of RNA-binding motifs for decoding gene regulation. Nature, 2013. 499(7457): p. 172-7.

Contact

Sarath Chandra Janga, PhD, Assistant Professor
e-mail : scjanga@iupui.edu Lab website : <http://www.iupui.edu/~janggalab>