Discovery and evolutionary dynamics of RBPs and circular RNAs in mammalian transcriptomes

ABHIJIT BADVE

Submitted to the faculty of the Bioinformatics Graduate Program in partial fulfillment of the requirements for the degree Master of Science in Bioinformatics in the School of Informatics and Computing Indiana University March 2015

Accepted by the Graduate faculty, Indiana University, in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Master's Thesis Committee

Sarath Chandra Janga, Ph.D.

Matthew Hahn, Ph.D.

Yunlong Liu, Ph.D.

Copyright page

To my parents, Mr. Sadanand Badve and Mrs. Rajani Badve and my family. They raised me, supported me, taught me and loved me. This thesis is dedicated to them.

Acknowledgement

This thesis was made possible due to the guidance and encouragement from many people. So it gives me a great pleasure to thank these people and acknowledge their contribution. I owe a sincere appreciation to my thesis advisor Dr Sarath Chandra Janga for his motivation and immense knowledge. Without his thoughtful guidance, energy and constructive criticism this thesis would have never been possible. Besides my lab members, I would also like to thank the other members of my committee Dr. Matthew Hahn and Dr Yunlong Liu for supporting my work, reading my thesis and providing helpful suggestions.

I wish to express sincere appreciation to School of Informatics at Indiana University Purdue University Indianapolis for providing me an opportunity to pursue a bright carrier in bioinformatics.

I thank my family for the constant support, love and blessing. Their teachings have made me the person I am today.

Table of Contents

List of Figures:

STUDYING POST-TRANSCRIPTIONAL NETWORKS CONTROLLED BY RNA-BINDING PROTEINS IN MAMMALIAN TRANSCRIPTOMES AND DISCOVERING AND CHARACTERIZING CIRCULAR RNA USING LONG POLY (A) SEQUENCING

Chapter 1: Introduction: Abstract and Background

RNA-binding proteins (RBPs) are vital post-transcriptional regulatory molecules in transcriptome of mammalian species. It necessitates studying their expression dynamics to extract how post-transcriptional networks work in various mammalian tissues. RNA binding proteins (RBPs) play important roles in controlling the post-transcriptional fate of RNA molecules, yet their evolutionary dynamics remains largely unknown. As expression profiles of genes encoding for RBPs can yield insights about their evolutionary trajectories on the post-transcriptional regulatory networks across species, we performed a comparative analyses of RBP expression profiles across 8 tissues (brain, cerebellum, heart, lung, liver, lung, skeletal muscle, testis) in 11 mammals (human, chimpanzee, gorilla, orangutan, macaque, rat, mouse, platypus, opossum, cow) and chicken & frog (evolutionary outgroups). Noticeably, orthologous gene expression profiles suggest a significantly higher expression level for RBPs than their non-RBP gene counterparts - which include other protein-coding and non-coding genes, across all the mammalian tissues studied here. This trend is significant irrespective of the tissue and species being compared, though RBP gene expression distribution patterns were found to be generally diverse in nature. Our analysis also shows that RBPs are expressed at a significantly lower level in human and mouse tissues compared to their expression levels in equivalent tissues in other mammals chimpanzee, orangutan, rat, etc. which are all likely exposed to diverse natural habitats and ecological settings compared to more stable ecological environment humans and mice might have been exposed, thus reducing the need for complex and extensive post-transcriptional control. Further analysis of the similarity of orthologous RBP expression profiles between all pairs of tissue-mammal combinations clearly showed the grouping of RBP expression profiles across tissues in a given mammal, in contrast to the clustering of expression profiles for non-RBPs, which frequently grouped equivalent tissues across diverse mammalian species together, suggesting a significant

evolution of RBPs expression after speciation events. Calculation of species specificity indices (SSIs) for RBPs across various tissues, to identify those that exhibited restricted expression to few mammals, revealed that about 30%of the RBPs are species-specific in at least one tissue studied here, with lung, liver, kidney & testis exhibiting a significantly higher proportion of specie-specifically expressed RBPs. We conducted a differential expression analysis of RBPs in human, mouse and chicken tissues to study the evolution of expression levels in recently evolved species i.e. humans and mice than evolutionarily distant specie i.e. chicken. We identified more than 50% of the orthologous RBPs to be differentially expressed in at-least one tissue compared between human and mouse but not so between human and an outgroup chicken in which RBP expression levels are relatively conserved. Among the studied tissues brain, liver and kidney showed a higher fraction of differentially expressed RBPs, which may suggest hyper regulatory activities by RBPs in these tissues with species evolution. Overall, this study forms a foundation for understanding the evolution of expression levels of RBPs in mammals, facilitating a snapshot of the wiring patterns of post-transcriptional regulatory networks in mammalian genomes.

In our second study we focused on elucidating novel features of post-transcriptional regulatory molecules called as circRNA from LongPolyA RNA-seq data. The debate over presence of non-linear exon splicing such as exon-shuffling or formation of circularized forms has finally come to an end as numerous repertoires have shown of their occurrence and presence through transcriptomic analyses. It is evident from previous studies that along with consensus-site splicing non-consensus site splicing is robustly occurring in the cell. Also, in spite of applying different high-throughput approaches (both computational and experimental) to determine their abundance, the signal is consistent and strongly conforming the plausible circularization mechanisms. Earlier studies hypothesized and hence focused on the ribo-minus non-polyA RNA-seq data to identify circular RNA structures in cell and compared their abundance levels with their linear counterparts. Thus far, the studies show their conserved nature across tissues and species also that they are not translated and preferentially are without poly (A) tail with one

to five exons long. Much of this initial work has been performed using non-polyA sequencing thus probably underestimates the abundance of circular RNAs originating from long poly (A) RNA isoforms. Our hypothesis is if the circular RNA events are not the artifact of random events but has a structured and defined mechanism for their formation then there would not be biases on preferential selection / leaving of polyA tails while forming the circularized isoforms. We have applied an existing computational pipeline from earlier studies by Memczack et.al on ENCODE cell-lines long poly (A) RNA-seq data. With same pipeline we achieve a significant number of circular RNA isoforms in the data some of which are overlapping with known circular RNA isoforms from the literature. We identified an approach and worked upon to identify the precise structure of circular RNA which is not plausible from the existing computational approaches. We aim to study their expression profiles in normal and cancer cell-lines and see if there exists any pattern and functional significance based on their abundance levels in the cell.

Chapter 2 Studying and Elucidating Post-Transcriptional Networks Controlled by Rna-Binding Proteins in Mammalian Transcriptomes

2.1 Introduction

With the advent of high-throughput techniques in RNA-sequencing, studying mammalian genomes for uncovering evolution by examining gene expression profiles has become feasible. Earlier studies focused on identifying selectively driven expression switches which explicate variations in organs, lineages and chromosomes among mammalian species. They compared six organs that represent all major mammalian species to unravel evolutionary intricacies of mammalian transcriptomes. Though overall mammalian genes are conserved and homologous; their differential rates of expression changes owing to differential selective pressures contribute to phenotypic changes in organs of mammals.[1] In another study, a large-scale comparative analysis with perspective of studying long non-coding RNA (lncRNA) repertoire of mammalian genomes can be characterized. This study showed several classes of lncRNA based on their analysis of expressions patterns within lncRNA such as primate-specific lncRNA, ancient lncRNA and conserved lncRNA. Also through co-expression network analyses of lncRNA, varied potential novel functions for studied lncRNA were established.[2]

Other studies also concentrated on studying mammalian tissue-specific conservation of splicing patterns. These studies provided novel insights into how mammalian genomes splicing patterns vary across primate and non-primate lineages. These studies also showed unlike tissue-specific gene expression programs which are conserved across mammalian transcriptomes, alternative splicing is a lineage-specific event and is conserved only in specific set of tissues.[3] From these studies several novel, conserved and lineage-specific alternatively spliced exon signatures were identified. They exhibited how species-specific cis-directed splicing patterns are prevalent in vertebrate species and also how various other splicing events lead to diversification of splicing and underlie a phenotypic differences within mammalian species.[3, 4]

In eukaryotes, post-transcriptional regulation of gene expression is intricate and it is essential to gain full understanding of vital steps of complex and yet well-coordinated gene regulation. RNA

binding proteins (RBPs) and Ribonucleoprotein complexes (RNPs) control extensive post-transcriptional processing of pre-mRNA that produces a diverse collection of mRNAs in a genome and thus facilitate an addendum of gene regulation. RBPs have specific RNA binding affinities and specificities and in turn RBPs preferentially bind to only specific RNA molecules. Cells are able to generate numerous RNPs whose composition and arrangement of components is unique to each mRNA and the RNPs are further remodeled during the course of the maturation of the mRNA into its functional form[5]. Hence it is preeminent to note that during course of evolution, RBP structural domains and motifs undergo diverse changes in different species which enables them for their mRNA sequence binding specificity in a species. Various studies focusing on decoding one or the other steps of post-transcriptional regulation and gene dysfunctions in various disorders especially in cancers have been conducted and it has been shown in multiple studies how the interplay between different mechanisms and extensive involvement of RNA binding molecules occur which in turn control gene expressions. However there is no extensive study involving how RBPs expressions evolve in mammalian transcriptomes. We present a comprehensive comparative analyses of RBP expression profiles across 8 tissues (brain, cerebellum, heart, lung, liver, lung, skeletal muscle, testis) in 11 mammals (human, chimpanzee, gorilla, orangutan, macaque, rat, mouse, platypus, opossum, cow) and chicken & frog (evolutionary outgroups). We specifically addressed three major points while conducting these analyses. By studying global expression patterns of orthologous RBPs across mammals with respect to humans in various tissues, if the variations can across species be explained based on evolutionary distances. When tissue-wide expression profiles across species are compared, if we can uncover whether RBPs are species-specific or they are conserved in their expression levels across species. We also wished to study functions, domains and expression levels of RBPs which are species-specific versus widely expressed across the mammalian tissues. Also by conducting differential expression analysis of RBPs between recent mammalian lineages such as primates and rodents to ancient non-mammalian species such as birds, we uncovered signature RBP

clusters which are categorically only expressed in ancient species, while some are expressed only in recent species while most of the RBPs have conserved expression profile across the mammalian species. In all this study furnishes a snapshot of how the expression patterns of post-transcriptional regulatory molecules are evolving in mammalian genomes.

2.2 Materials and Methods

Data for expression profiling of RNA-binding proteins in mammalian tissues

We have illustrated an overall workflow design in Figure 1. In our study, we collected 311 RNA-seq data samples published from previous works by Fietz *et al*, Brawand *et al*, Merkin *et al* and Necsulea *et al* for 11 mammalian species (human, chimpanzee, gorilla, orangutan, macaque, mouse, rat, platypus, opossum, and cow) and 2 evolutionary out-groups (chicken and frog) available from NCBI SRA resource[6 1 3 2]. This data represents 8 tissues brain, cerebellum, heart, kidney, liver, lung, skeletal muscle, and testis. Raw RNA-seq reads were subjected to quantification using Sailfish- a tool for alignment free quantification. Sailfish generates k-mer based indexes of the reference genomes and then employs expectation maximization (EM) algorithm for quantification of relative transcript abundance for both paired-end or single-end reads.[7] We ran Sailfish with latest ENSEMBL releases of reference annotations for the species we selected for the study.[8] The details of which are mentioned in supplementary materials.

We used transcripts per million (TPM) metric for comparison of relative abundances across and within tissues of mammalian species. As it was reported in previous studies, reads per kilobase per million reads (RPKM) cannot be the true measure of relative molar RNA concentration (RMC), we used TPM metric which respects invariance property and also eliminates statistical biases inherent while comparing data across tissues of various species[9, 10]. As the orthology can be extracted at only gene level and not transcript level, we calculated mean TPM values of each transcript of orthologous genes and considered this value for comparison of expressions in all 8 tissues across 13 species [Selection of orthologous genes is explained in detail in Methods Section#2]. While constructing expression profiles of RBPs and non-RBPs we measured the

evolutionary distance of each species from humans. We utilized data and phylogenetic trees inferred from annotated ribosomal RNA sequence alignments from Ribosomal Database Project (RDP II).[11] According to the phylogenetic tree, species evolved earlier than humans e.g. chicken originated around ~300 million years ago; are placed distant to humans. While species evolved later and are closer to humans e.g. chimpanzee which are originated ~8o million years ago; are closer to humans in an evolutionary tree. [Figure 2].

Non-parametric Kolmogorov-Smirnov (KS) tests were performed to compare RBP vs non-RBP gene expression distributions in all 8 tissues across all species. We further calculated spearman correlation coefficients (ρ) for all vs all tissue-species combinations RBP genes expression profiles. The final matrix consisted spearman coefficients of all combinations of tissues of each species RBP genes expression data compared against all tissues of other species in our study. To construct a correlation matrix we considered only primates and rodents tissues data. We further performed hierarchical clustering using hclust package in R and plotted results as a heatmap. Similar plot was constructed for non-RBP genes comparisons across primates and rodents to observe differential clustering results in case of RBPs and non-RBPs.

Tree constructed from correlation coefficients comparisons of non-RBPs with tissue-species combinations yield similar tissues of closer species are clustered together moderately with only few exceptions of mouse and human tissues. The hierarchical clustering results for RBPs across tissues-species combinations yield a significantly different phylogenetic tree where tissues of same species are clustered together which means expression profile of RBPs is conserved within different tissues of same species. To elucidate this behavior of RBPs we tested the correlation coefficients of RBPs and non-RBPs by classifying the combinations of tissues and species into three different categories. The categories are as follows: i) correlation coefficients between different tissues and different species. ii) Correlation coefficients between same tissues of different species. iii) Correlation coefficients between different tissues of same species. [Figure 3].

Prediction of orthologous RBP and non-RBP genes using ENSEMBL Compara

Further we classified the genes based on their human annotations as RBPs and non-RBPs. RBPs set comprised of 1344 genes constituting 12788 transcripts characterized experimentally from various repertoires.[12] [13] [14] [15] All other genes including non-protein-coding genes were classified as non-RBPs in this study. This human dataset was used as a reference for deciphering a set of orthologous genes across other mammalian species and out groups. We used ENSEMBL Compara datasets to map and predict human orthologous RBPs and non-RBPs for each mammal. Compara is a rich data source from ENSEMBL which utilizes gene tree-based phylogenetic mapping of protein-coding genes across multiple vertebrate species.[16] The parameters used for the selection of orthologous genes were %identity, biotype (strictly protein-coding in case of RBPs) and orthology confidence score. We were able to map on an average 80% high confidence and low confidence orthologous genes across all the species considered in this study. In certain cases genes could not be mapped and were discarded from study subject to lacking strong evidence [Supp. Fig. 1]. We considered only mapped orthologous genes (RBPs and non-RBPs) for expression analyses and species-specificity analyses.

Species Specificity Index (SSI) calculations of RBPs and non-RBPs in mammalian species

Earlier studies by Yanai *et al* defined a tissues specificity index (τ) which is calculated to get insights of gene expression patterns across tissues: one-tissue specific, housekeeping genes or midrange expressions of genes meaning expressed in subset of tissues.[17] This index values vary between 0 signifying housekeeping genes to 1 meaning strictly tissue specific, thereby giving unique impression of gene expression profiles to infer evolutionary diversion of genes based on their expression values. From our expression profile analyses, it was seen that RBPs exhibit diverse expression patterns across mammals in all 8 tissues being compared. We extended the usage of this index analogously to calculate species specificity index which we contemplate will provide insights into how RBPs and other protein coding genes are evolving in mammalian

species considering each tissue at a time. The species specificity index for any tissue is calculated as:

$$\frac{\sum_{i=1}^{N}(1-x_i)}{N-1} \; where\ N\ are\ number\ of\ species\ compared\ and\ \ x_i\ is\ gene\ expression\ in\ species\ i$$

Similar to tissue-specificity index, SSI also interpolates values between 0 being expressed generically in multiple species while 1 suggesting species-restricted expressions. To be able to classify RBPs robustly we categorized RBPs based on their species specificity indices in multiple tissues. We term RBPs to be single-tissue species specific if they exhibit species specificity patterns of expressions in only one or two tissues being compared. While RBPs which are expressed in >3 tissues simultaneously, we term them multi-tissue species specific RBPS.

We further employed kernel density function on SSI values of RBPs to construct kernel-density plots across different tissues using SM package in R. We wish to infer the global patterns of RBPs' SSI in multiple tissues under study. We compared kernel density values of SSI in RBPs with other protein coding genes. This analysis assisted in understanding how expression patterns of RBPs are preferentially selected or conserved in particular tissue or set of tissues under study across mammals.

Identifying differentially expressed RBPs between Human, Mouse and Chicken to uncover evolutionary trends

The RBPs expression data for six tissues (brain, heart, kidney, liver, lung and testis) across three species (human, mouse and chicken) was subjected to differential expression analysis using DESeq2 package in R. DESeq2 implements a statistical inference model which takes into account raw read-counts for calculating log-fold changes of expression within condition specific data and assigns a FDR corrected p-value to each calculation. We calculated mean read-counts of gene from transcript levels and provided as input. We infer a gene to be differentially expressed (either up-regulated or down-regulated) between two species for each tissue being studied; if the log-fold change is >1.5 and an adjusted p-value <0.05. We compare expression profiles for mouse

which is intermediately placed with respect to humans in mammalian species evolutionary tree and an evolutionary outgroup chicken. Based on p-value and fold-change filters we assign a binary value 1 or 0 if the gene is dysregulated or non-dysregulated respectively on comparison across tissues. We construct a heatmap to visualize the patterns across tissues and species; it clearly elucidates four distinct classes of RBPs based on their dysregulation in at-least 4 tissues under comparison. The classes can be termed as I. Continuously evolving RBPs which are dysregulated across human, mouse and chicken II. Recently evolved RBPs which are changing in majority of tissues in mouse and human but not in chicken III. Ancient RBPs which are only dysregulated in chicken on comparisons with mouse and human IV. Non-changing RBPs which do not show specific trends of dysregulation in any species being compared.

2.3 Results and Discussion

RBPs are expressed significantly higher than non-RBPs across species and tissues

Advances in expression profiling using high-throughput techniques such as RNA-seq have enabled us to get insights into transcriptomic expression dynamics. In various studies conducted earlier it was shown that RBPs play very important role in post-transcriptional and translational regulation of human transcriptome[18, 19, 20, 21]. Also it was shown that they are expressed at significantly higher levels than non-RBPs in context of human TCGA cancer versus healthy genomes. [22 23 24 25] However it is still uncertain how the post-transcriptional networks involving RBPs must be evolving in mammalian species. We present here a first comprehensive analysis showing RBP expression dynamics in mammalian species across various tissues. We selected six tissues (brain, cerebellum, heart, liver, kidney and testis) RNA-seq data of four mammalian orders namely primates such as human, chimpanzee, orangutan, gorilla, macaque; rodents such as mouse and rat; marsupial such as opossum; primitive egg laying mammal such as platypus and two outgroups chicken and frog for our expression analyses. We classified genes encoding proteins which have reports of RNA-binding from various literature studies into RBPs and other genes as non-RBPs. Then we compared expression values (TPM) of RBPs and non-RBPs in six

tissues across 11 mammalian species and 2 outgroups (chicken and frog) studied here. As explained earlier transcripts per million mapped reads (TPM) metric provides an invariant and unbiased measure of relative abundances of transcripts in samples. We observe that RBPs expression values when plotted against their non-RBP complements show a significantly higher expression patterns in all the mammalian and non-mammalian species studied. This trend is generically significant for all the tissues being compared, though the expression levels of RBPs are diverse in nature. We compared the distributions of RBPs expression versus non-RBPs across species (Kolmogorov and Smirnov p-value at $2.2e^{1.16}$). As compared to other species, human and mouse expression profiles are at lower levels for both RBPs and non-RBPs across all tissues being compared. We speculate that human and mouse show unique expression patterns compared to other species as they have evolved across more diverse natural habitats, environments and ecological settings. On the other hand, other mammalian species have a restricted and stable environments and ecologies, thus reducing the need of extensive post-transcriptional regulation by RBPs in these species. Also it has been confirmed from studies performed earlier that human and mouse transcriptomes have high correlation with respect to their gene expression levels in multiple tissues for numerous genes.[26]

It is shown in previous studies how the correlation between expressions levels of protein coding genes be accurately used to construct an evolutionary tree of mammalian lineages. It is also established how tissue-type and species-type are primary components of variability in gene expression profiles in vertebrates.[27] [28] This analysis helps in gaining insights into how proteins evolve after speciation events in various tissues of mammals. It has been shown that primates and rodents have a complex transcriptome and hence to decipher RBPs' species-specific post-transcriptional regulation has advanced in those species, we limited our analysis to include only higher mammals i.e. primates (human, chimpanzee, orangutan, gorilla, macaque) and rodents (mouse and rat) which are spread across ~90 million years in evolution. From the expression patterns of RBPs and non-RBPs we wanted to infer correlations between expression levels within

various tissues and species combinations hence we subject expression profiles of RBPs and non-RBPs to hierarchical clustering.[Fig 2a, 2b] We calculated spearman correlation coefficients (σ) between every tissue of each mammalian species versus every other mammalian tissue-species combinations. So from all vs all comparisons of correlation coefficients of expression values, we infer that RBPs and non-RBPs cluster differentially. As expected, non-RBPs cluster the relative tissues of evolutionarily close species together confirming the observations found in earlier studies [Figure 2b]. On contrary, RBPs cluster within same species different tissues together [Fig. 2a]. The variability in RBP gene expression profiles owe primarily to after-speciation events and factors like habitat, ecological and environments play a huge role contributing to evolution in their expression patterns, while non-RBPs gene expression variability owes primarily to species-type first and then tissue-type variation based on spacing of species on evolutionary tree. We can infer that in case-of non-RBPs species evolution and tissue development is complementary. In case of RBPs relative or similar tissues of one species will be always clustered together. We note here that above patterns are distinctively evident in many tissues of mammalian species with few exceptions. Clustering pattern of human and mouse tissues for non-RBPs suggests that clusters relative tissues are formed and also among those species there is a higher correlation than between any other species placing them close to each other. In summary from clustering analysis we infer that RBPs express in species-specific manner rather than tissue-specific manner unlike non-RBPs.

Further to elucidate how correlation patterns are distributed we classify correlation coefficients between all vs all tissue-species combinations into three mutually exclusive sets for both RBPs and non-RBPs. We consider correlation coefficients between different tissues of each species and it forms a set I. Set II constitutes correlation values between relative tissues of different species e.g. correlation coefficients between kidneys of each species or livers of each species. Lastly rest of all correlations of tissue-species combinations constitute a set III. [Fig. 3]. Set I constituting correlation values between relative tissues of same species of RBPs shows highest

correlation among them which suggests that evolution in expression values of RBPs are species driven rather than tissue driven. Set II and set III show relatively lower correlation as compared to set I which suggests that RBPs are evolutionarily classified per species. Inversely, as it is established that non-RBPs cluster relative tissues of mammalian species together, trends were confirmed from their correlation distributions. Set III comprising different tissues of different species shows relatively lower correlation among them. Also Non-RBPs show highest correlation between relative tissues of different species i.e. for Set II. Also for non-RBPs, set I and set III correlation is relatively lower. This observation clearly demarcates between expression trends of RBPs and non-RBPs and thereby their evolutionary selection (Wilcoxon test p-value significance at 0.05). This analysis also fortifies the hypothesis that RBPs expression evolution is species-specific while non-RBPs expression evolution is majorly tissue-specific.

Evolution of genes encoding for RBPs expressions in mammals is species-specific

From the expression analyses it is clear that RBPs evolution is driven by species-specific events. It compelled us to calculate species-specificity index (SSI) of RBPs across tissues studied. We customized a tissue-specificity index (TSI) and developed a specialized method to calculate species-specificity index. We calculated SSI values for 8 tissues (brain, cerebellum, heart, kidney, liver, lung, skeletal muscle and testis) for higher mammals (primates and rodents) being studied. We found that about 30% of the RBP repertoire is species specific in at-least one tissues studied here, with several tissues exhibiting a significantly higher proportion of specie-specifically expressed RBPs. We further established similar calculations for non-RBPs in order to compare species specificity trends in RBPs vs non-RBPs. We plotted SSI density distributions for 8 tissues across selected mammalian species for RBPs and non-RBPs and observed that out of 8 tissues lungs, kidney, testis and brain show significantly higher species specificity levels than non-RBPs. While in other tissues (cerebellum, liver, heart) the trends are still significant for RBPs compared to non-RBPs except for skeletal muscle (p-value 0.98). It is believed that RBPs bind with multiple

RNA targets in a coordinated post-transcriptional regulatory manner in complex metabolic pathways especially during development. Hence there is no ambivalence in believing that the mechanisms with which they control their targets can be divergent in different organisms thereby leading to varied trends of expression patterns across organisms.

The SSI is defined on the scale 0 to 1 which is calculated considering expression value of each RBP in each species in a tissue for which SSI is calculated. The higher the value of SSI, higher is the preferential expression of RBPs in specific species while lower values devise a class of RBPS which are significantly expressed across species in a tissue being considered but no preferential expression in any species. Though SSI gives a broad impression of how RBPs must be evolving in particular tissue, it does not specifically distinguish the species in which it is expressed preferentially. Based on the earlier studies where TSI values were used to study the tissue specific expression of genes, we used the same threshold (>0.85) to classify RBPs based on their SSI values in two distinct classes: single-tissue species-specific RBPs which are only expressed in single tissue out of all tissues being studied and multiple-tissue species specific RBPs which are with higher SSI in multiple tissues. We speculate that RBPs which show distinct species specificity patterns undergo differential evolution in those species and would enrich more diverse functions in case of multi-tissue species specific RBPs vs more specific and restricted functions in case of single tissue species specific RBPs. In overall analyses we find around 3 fold more single-tissue species-specific RBPs (16%) than multi-tissue species-specific RBPs (6%) of total RBPs considered in the study. Also lungs, liver, kidney and testis exhibit highest proportion of species-specific RBPs. On closer look at the function enrichment of two classes show that single-tissue species specific RBPs enrich for varied specialized functions related to regulation of RNA/DNA conformation change, RNA stability, regulating histone H3-H4 methylation and ATP catabolic process. Other class of RBPs which is multi-tissue species specific RBPs show more generic roles such as mRNA nuclear transport, RNA processing and regulation of RNA processing, regulation of RNA splicing etc. Apart from these functions both categories enrich core

post-transcriptional regulatory functions such as Ribonucleoprotein complex formation, PolyA binding and mRNA 5' UTR binding. [Supplementary Figures 1a, 1b, 2a, 2b].

Profiling of differential RBP expressions in Human, Mouse and evolutionary outgroup Chicken and its evolutionary dynamics

In previous analyses of gene expressions evolution in mammals, various authors show that the rate of gene expression evolution varies among organs and lineages. Authors also show that purifying selection is primary factor for evolution in gene expressions and also identify numerous potentially selectively driven expression switches, which occurred at different rates across lineages and tissues which contribute to evolution of organs in mammalian species[1, 29]. We conduct a differential expression analysis of RBPs between mouse (~90 million years) and chicken (evolutionary outgroup ~300 million years) with respect to human (~15 million years). We strive to find different patterns of RBPs evolution based on their expression analyses. We employed DESeq2[30] package of R to conduct this analysis taking into account the read-count metric of RBPs across 6 tissues(kidney, brain, liver, heart, cerebellum and testis) between 3 species named above. We assign the binary value to the orthologous genes based on their differential expression (1) (either upregulated or down regulated) or not changing (0) in any of the compared species at filter of (FDR corrected p-value < 0.05 and logfoldchange > 1.5). We identified more than 50% of the orthologous RBPs to be differentially expressed in at-least one tissue compared between human and mouse but not so between human and chicken in which RBP expression levels are relatively conserved. Among the studied tissues brain, liver and kidney showed a higher fraction of differentially expressed RBPs, which may suggest hyper regulatory activities by RBPs in these tissues with species evolution. Figure 6a shows the differentially expressed genes between human and mouse while figure 6b shows the differentially expressed genes between human and chicken. From the visual interpretation of heatmaps, it is evident that RBPs based on their DE status can be classified into 4 major evolutionary classes as:

1.Continuously changing RBPs (8%) 2. Recently evolving RBPs (12%) 3. Ancient RBPs (5%) and 4. Non-changing or conserved RBPs (75%). Continuously evolving RBPs are those which are differentially expressed in outgroup chickens as well as mouse, recently evolving RBPs are the ones which are only changing between closer species i.e. human and mouse but not in chicken, and inversely RBPs which are only expressed differentially in chicken w.r.t humans are termed ancient RBPs. The non-changing RBPs as the name suggest do not differentially express in any of the species. Figure 7a and 7b show the top 5 tissues in which the RBPs are differentially expressed. The analysis shows that major cohort of RBPs change between human and mouse exclusively in brain (20%) suggesting brain is undergoing major changes evolutionary in both of those species. Incidentally majority of RBPs changing expression levels in brain fall into class of recently evolving RBPs. Kidney and testis in chicken and human make up for majority of RBPs to be differentially expressed belonging to 2 classes majorly i.e. of ancient RBPs and continuously evolving RBPs. Also numerous RBPs (5%) which are dysregulated between human and mouse are changing in all tissues suggesting the collective roles of post-transcriptional regulatory control functions for those RBPs. When studying dysregulated RBPs exclusively in brain between human and mouse functional annotation we found abnormality of nervous system morphology, mental function, erythroid lineage cell's abnormality to be enriched. In summary, differential expression analysis of orthologous RBPs in human, mouse and chicken classify RBPs evolutionarily owing to differential rates of their expressions leading to specialized roles in tissues and lineages of mammals.

2.4 Conclusion

In our analysis we focused on gene expression profiles of orthologous RNA binding proteins in mammals and outgroups to get implications of their evolutionary dynamics in them based on their RNA expression levels. We explored RBPs expressions in 8 tissues across 11 mammalian lineages (Primates, Rodents, Marsupials, and Monotremes) and two evolutionary outgroups Chicken and Frog. From the literature studies, we established a set of orthologous RBPs across

the mammals selected for this study using orthology confidence threshold, strictly selecting a genes of protein-coding biotype and percent identity with query proteins. Human RBPs against which the orthology searches were conducted, were selected based on the known evidence of their binding to RNA. Expression analyses of orthologous RBPs noticeably suggest a significantly higher expression levels for RBPs than their non-RBP gene counterparts - which include other protein-coding and non-coding genes, across all the mammalian tissues studied here. This trend is significant irrespective of the tissue and species and also the RBP gene expression distribution patterns were found to be generally diverse in nature. Also human and mouse tissues were significantly less expressed compared to other higher mammalian species which suggest that in evolutionary progression, regulatory roles of RBPs seem to have limited in those species meaning extensive regulatory control by RBPs might be plummeting in human and mouse. Also this observation is in agreement with previous studies conducted which show that human and mouse gene expressions has highest correlation among them.[6, 26, 31, 32] We also speculate that RBPs are expressed lowly in human and mouse as they are limited in exposure to diverse ecological and environmental settings and are living in more controlled environment than other mammals. Correlation studies of RBPs and non-RBPs unleash interesting trends for RBPs. Non-RBPs evolution is driven by their tissue-specific nature and cluster relative tissues of close species together suggesting non-RBP genes have differential evolutionary trends than RBPs which show that their expression profiles are more central to species rather than tissues. From the correlation studies of RBPs it's clear that majority of RBPs are expressed species-specific with their spearman correlations when clustered into different classes show that the correlation is highest between relative tissues of same species while lowest when different tissues of different species are compared.

When species specificity indices are compared for different tissues in higher mammals it is seen that after speciation events contribute to the behavior of RBPs rather than organ development contributing to RBPs' evolution. About 30% RBPs are showing species-specificity indices above

threshold of 0.85 in at least one tissue being compared with highest proportion of RBPs in liver, lung, kidney and testis as species-specific. SSIs are higher in multiple tissues or single tissue depending on which they can be classified as multi-tissue species specific or single-tissue species-specific RBPs. When the two classes of RBPs are compared for their functional enrichment tests, both the classes enrich certain core functions which are shared in both categories but numerous specialized functions are enriched in case of single-tissue species-specific RBPs. Also multi-tissue species-specific RBPs exhibit more generalized regulatory roles. Single-tissue species-specific RBPs are 3 fold more than multi-tissue species-specific RBPs. (See Results).

Differential expression (DE) analyses between human, mouse and chicken gives insights into how RBPs must be evolving in mammalian species. DE analysis gives exactly which RBPs are contributing to forming different classes of RBPs based on their expressions. We found around 50% RBPs to be differentially expressed in at-least one tissue between the three species. The important classes as continuously evolving, ancient, recently evolving and non-changing RBPs formed based on DE analysis gives differential functional categories to be enriched in those RBPs. Majorly between human and mouse RBPs in brain are changing faster giving implications of major post-transcriptional control of RBPs in brain tissues. This set of RBPs are exclusive to only human and mouse species and are not expressed differentially in chicken suggesting the recent evolution of such RBPs in human and mouse brain tissues. Human and chicken majorly show the differentially expressed genes exclusively in kidney, liver and lung tissues which form a class of ancient RBPs. Thus, in all RBPs undergo significant divergence in their expression as they evolve in evolutionary timeline from ancient species like chicken to recent ones like humans and mouse. Differential expression analysis thereby gives numerous insights into how the post-transcriptional regulation might be occurring in mammalian species and also help us understand their evolutionary dynamics in mammals.

Overall, this study forms a foundation for understanding the evolution of expression levels of

RBPs in mammals, facilitating a snapshot of the wiring patterns of post-transcriptional regulatory networks in mammalian genomes.

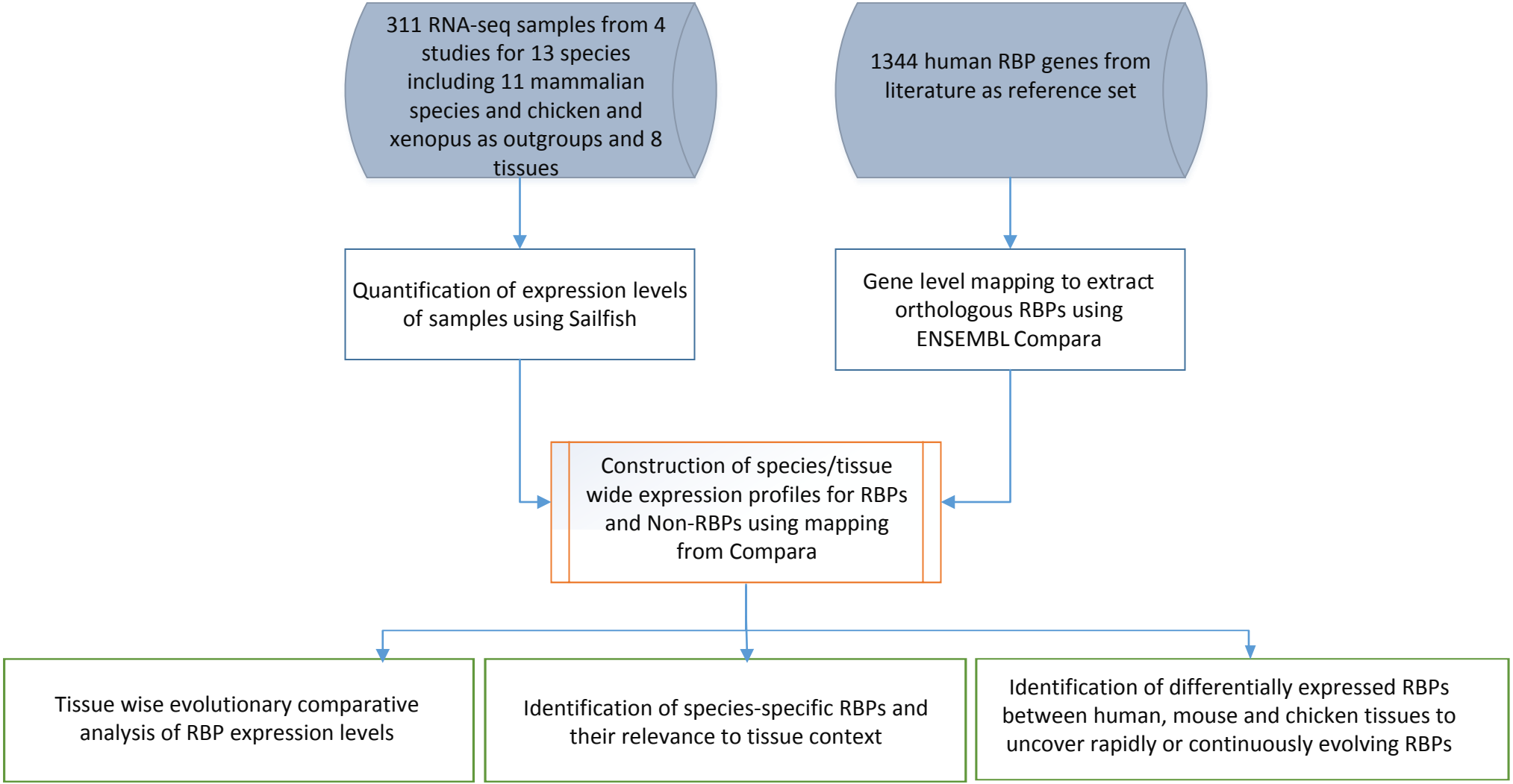
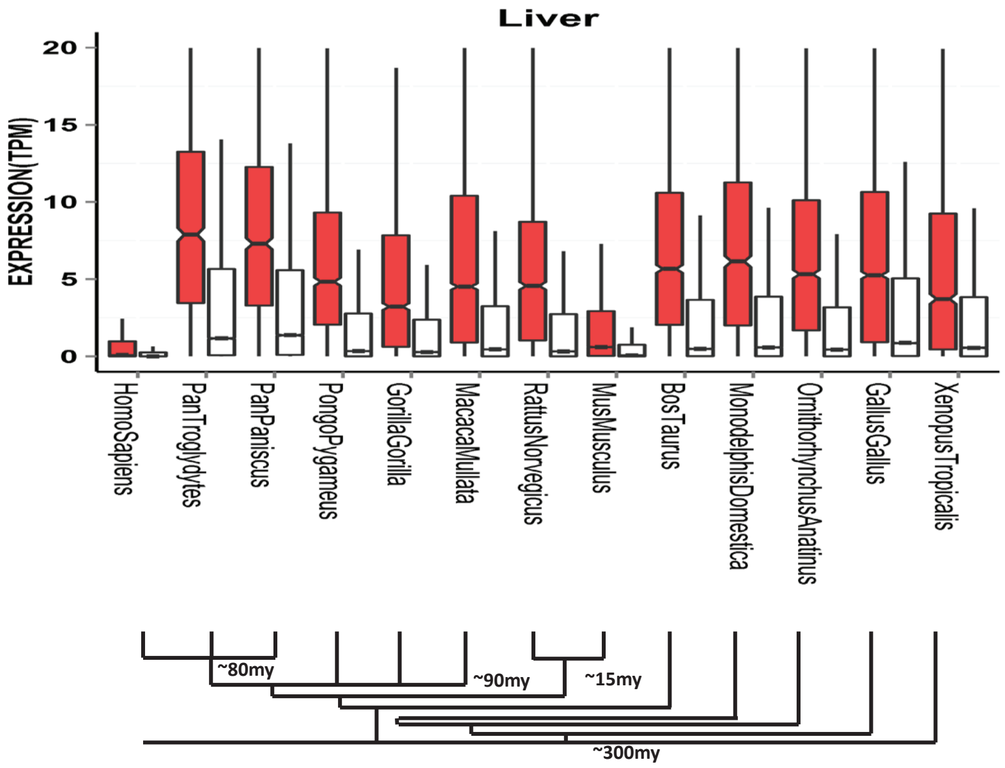
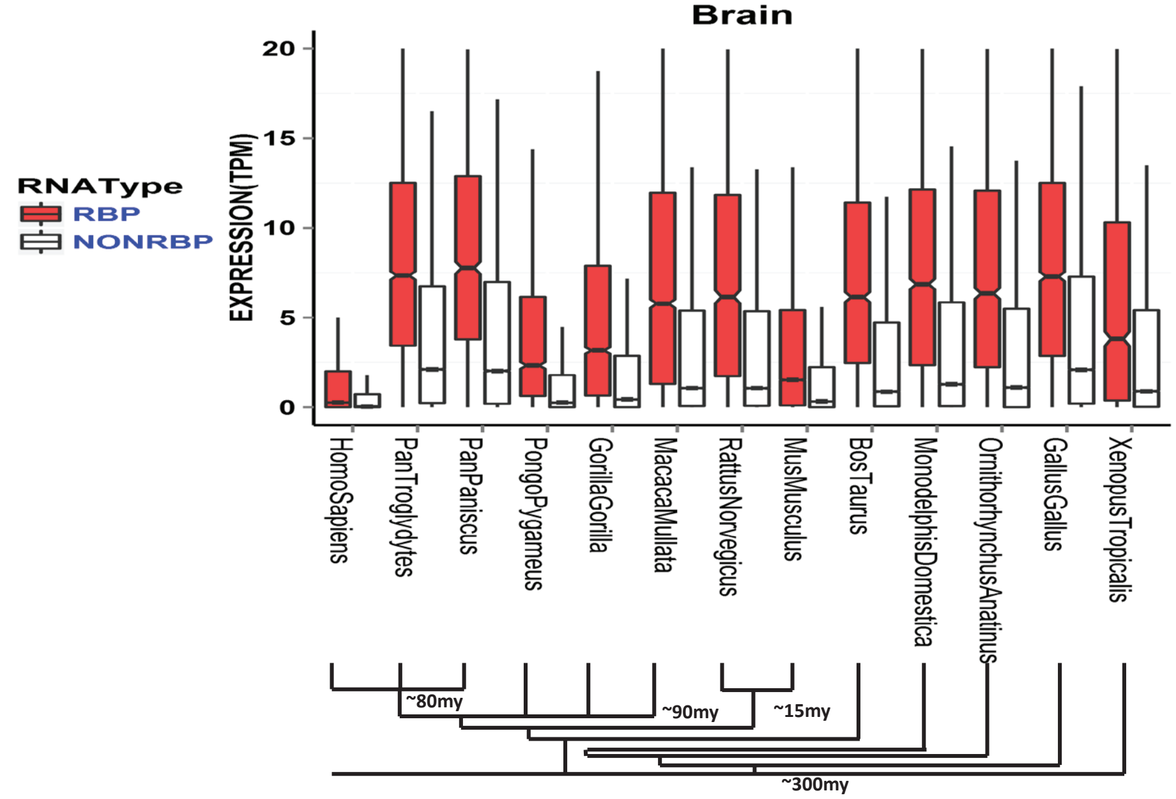
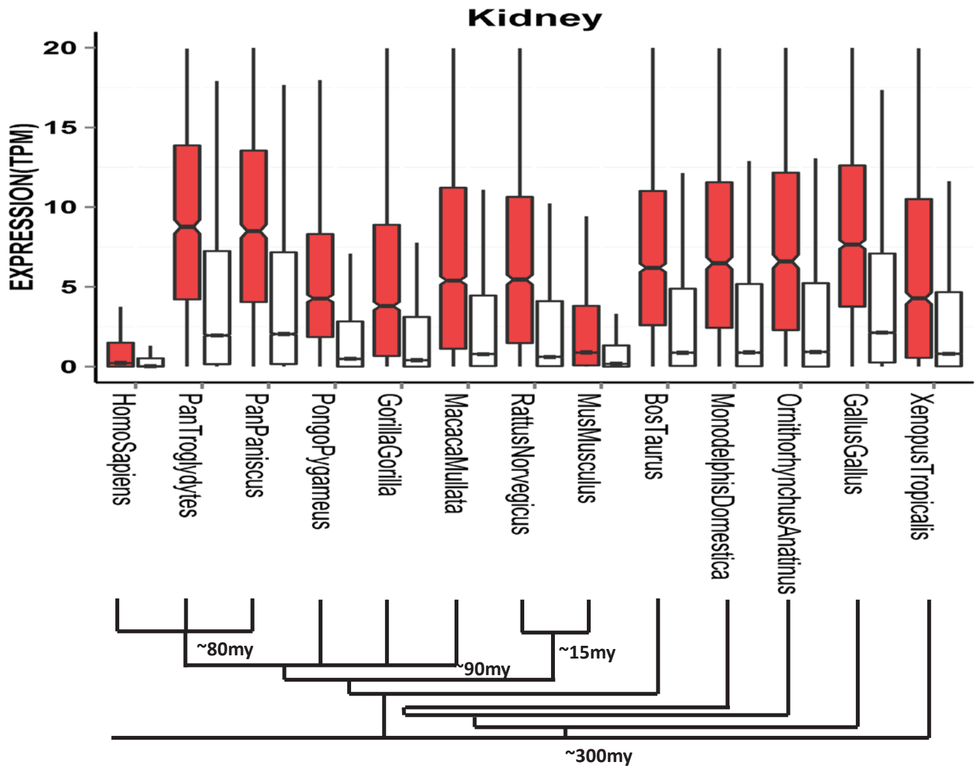
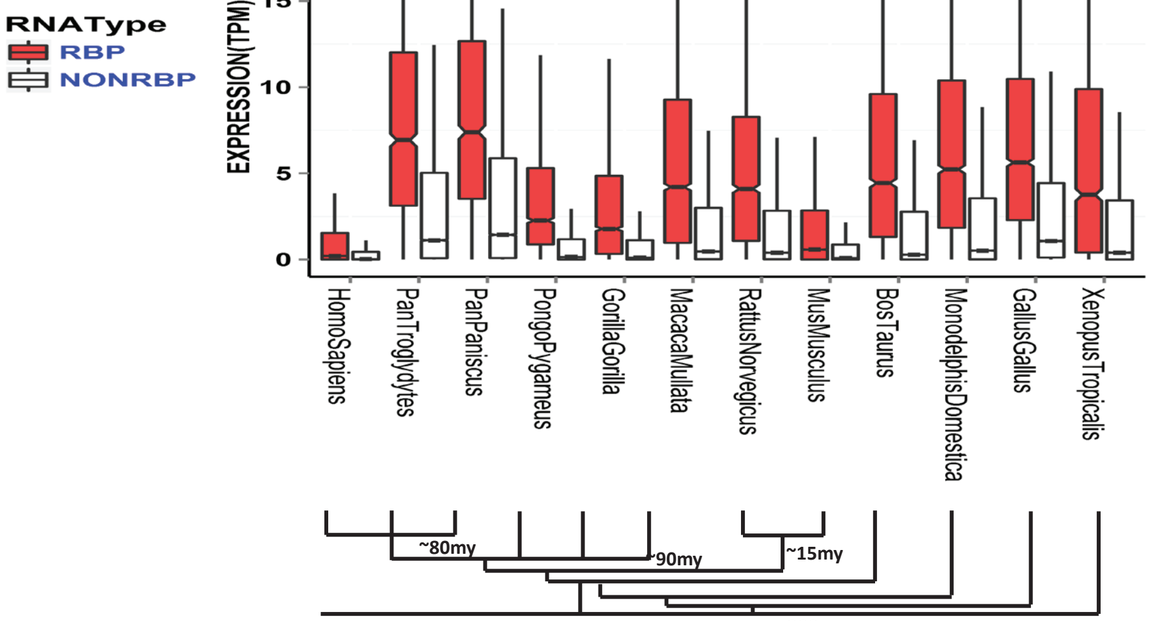


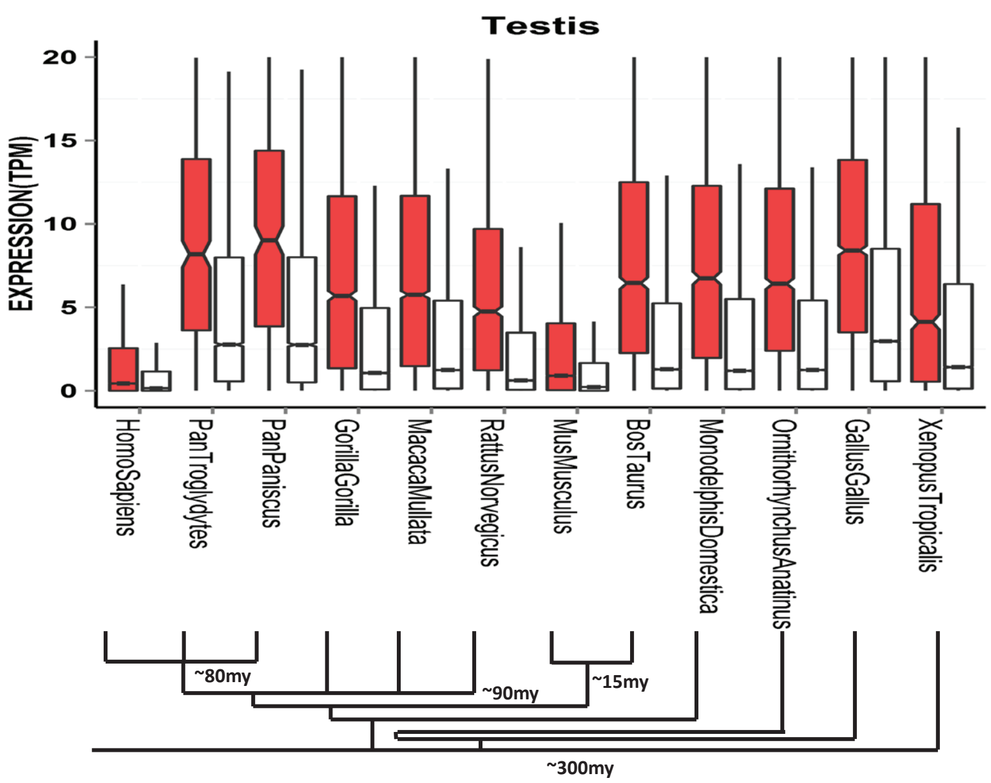
Figure 1: A chart representing analysis pipeline for studying evolutionary dynamics RBPs expression levels in mammalian tissues

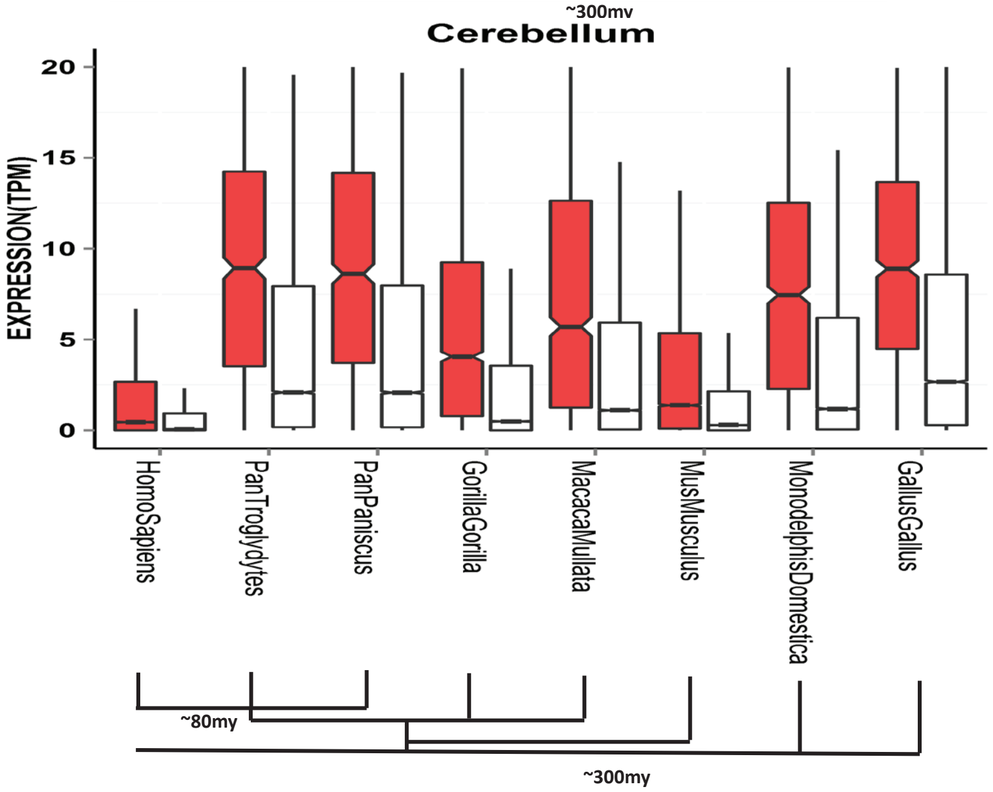




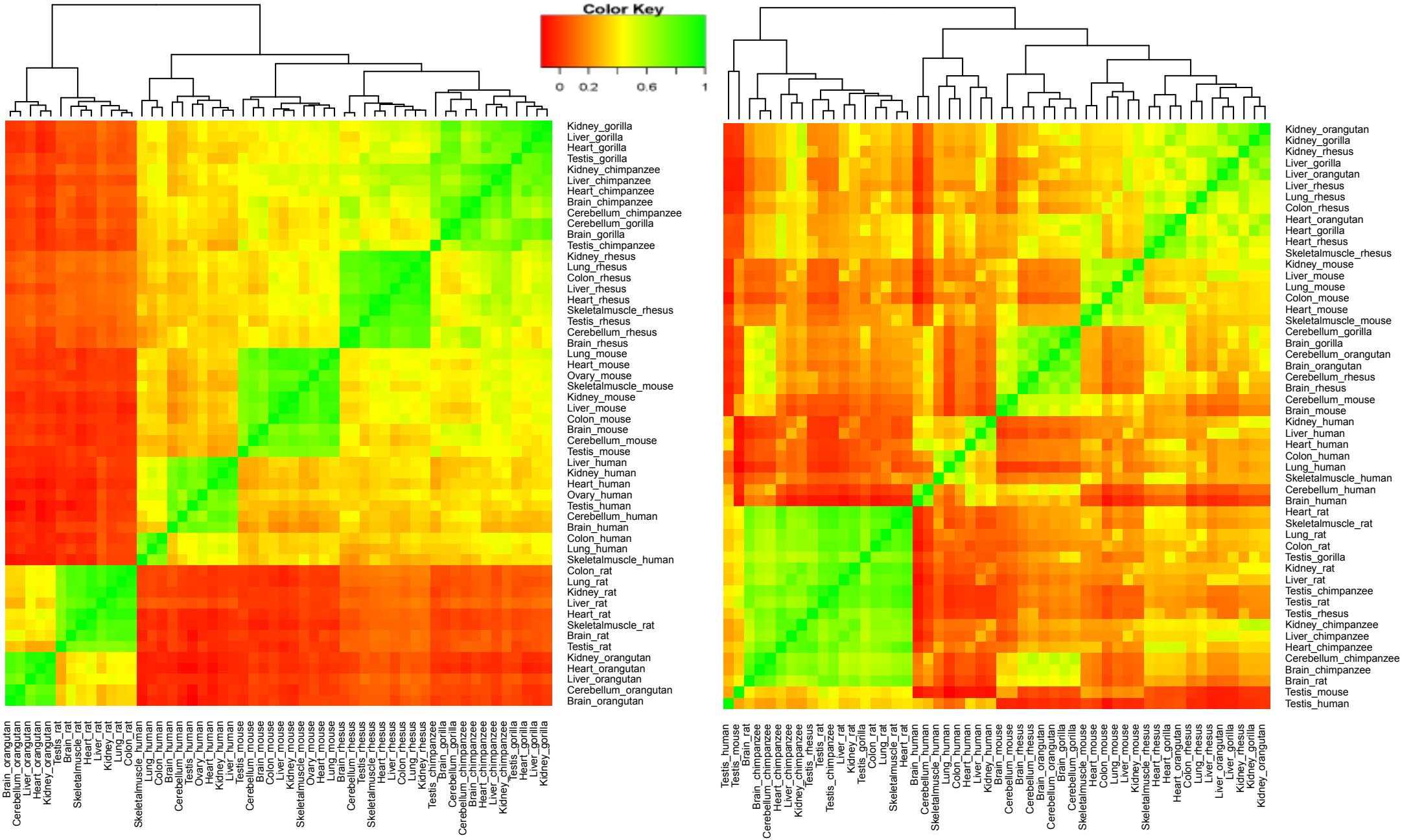




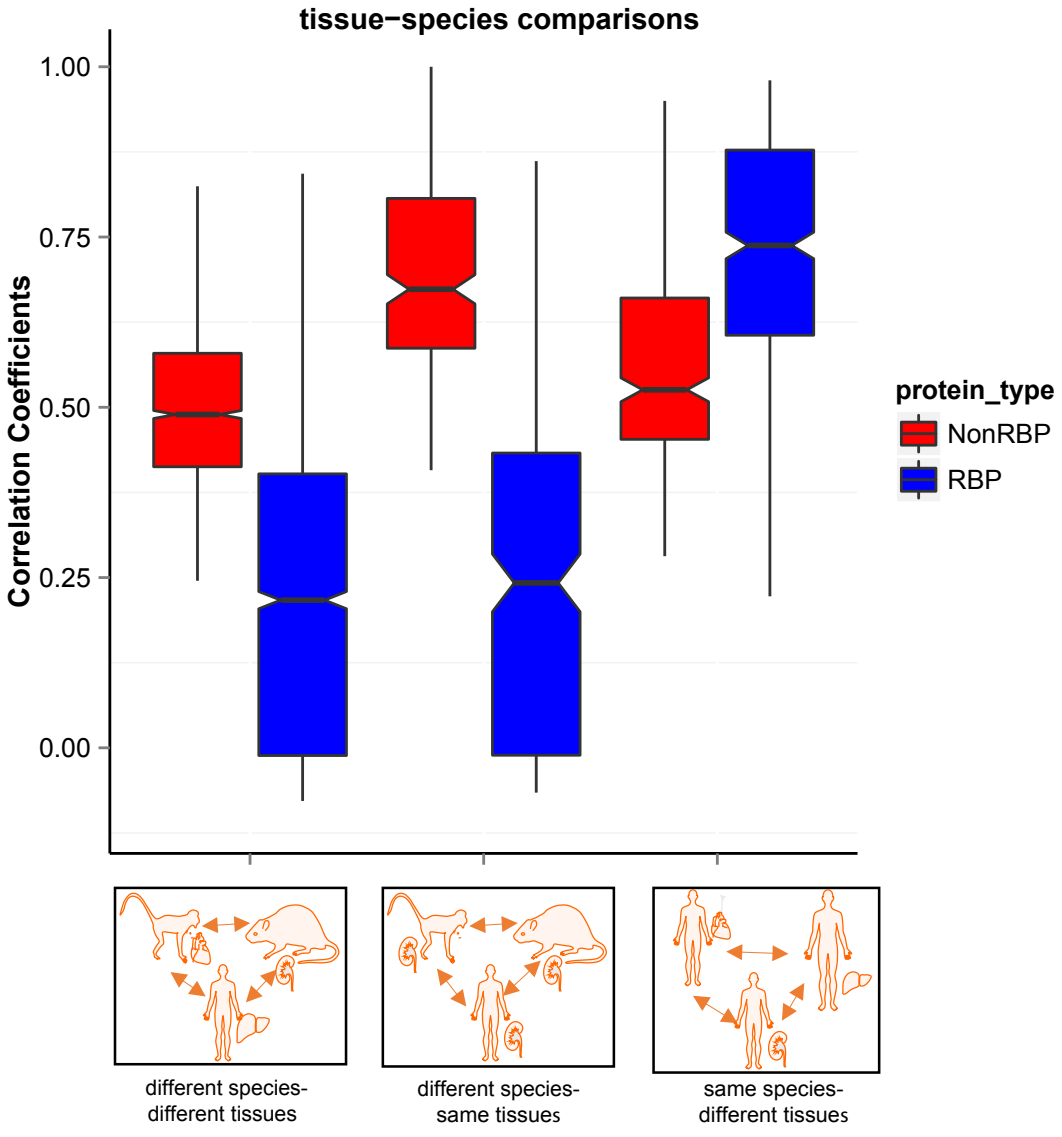




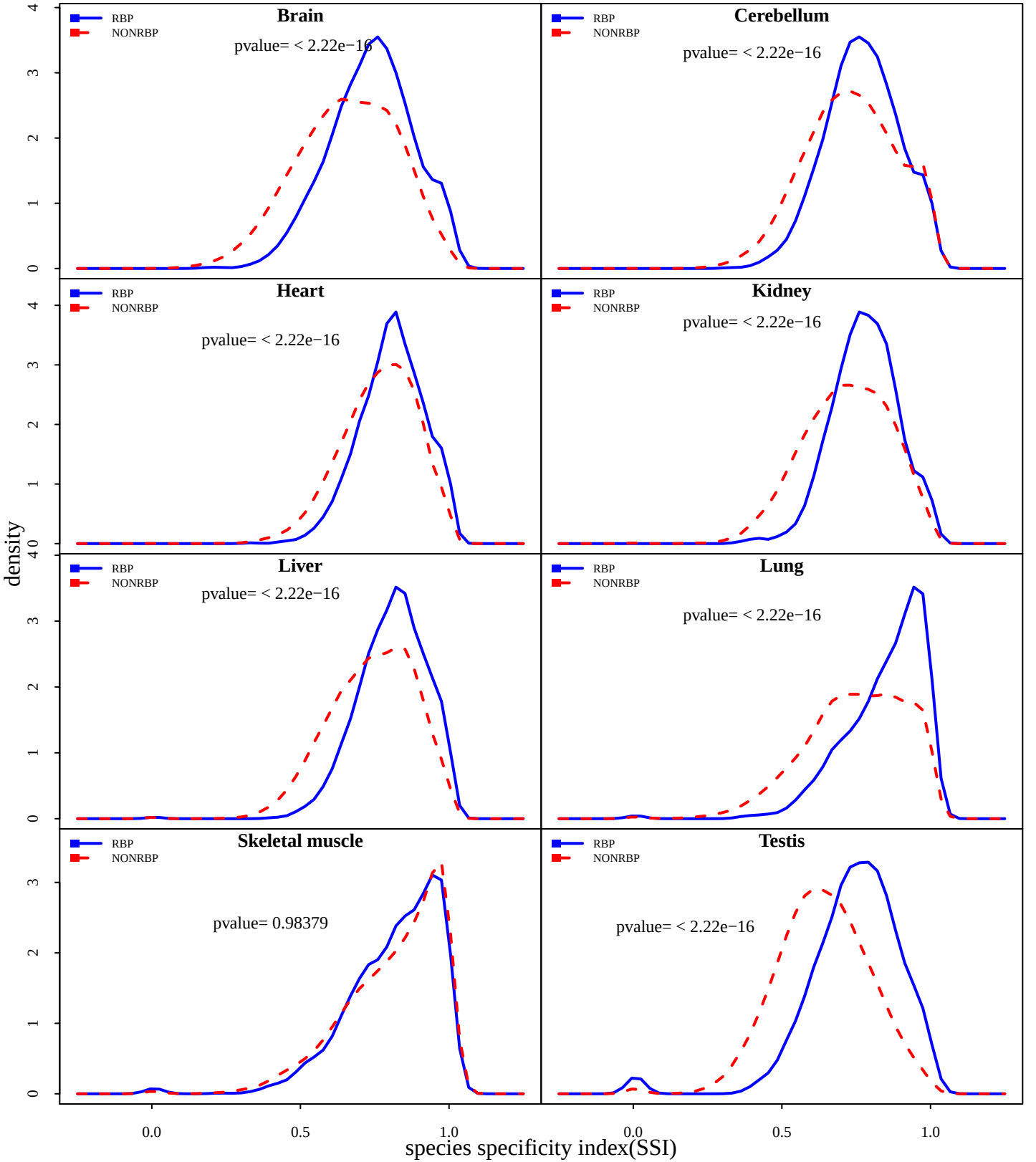
Expression Profiles| Figure 2: Multi-panel boxplots showing the expression level (TPM) comparisons between orthologous RBPs vs Non-RBPs across 6 tissues studied here (KS test p-values 2.2e^-16)



Clustering of expression correlations for RBPs and non-RBPs| Figure 3a: Heatmap shows clustering based on spearman correlation coefficients calculated from expression profiles of each tissue across species for RBPs (A) and non-RBPs (B)

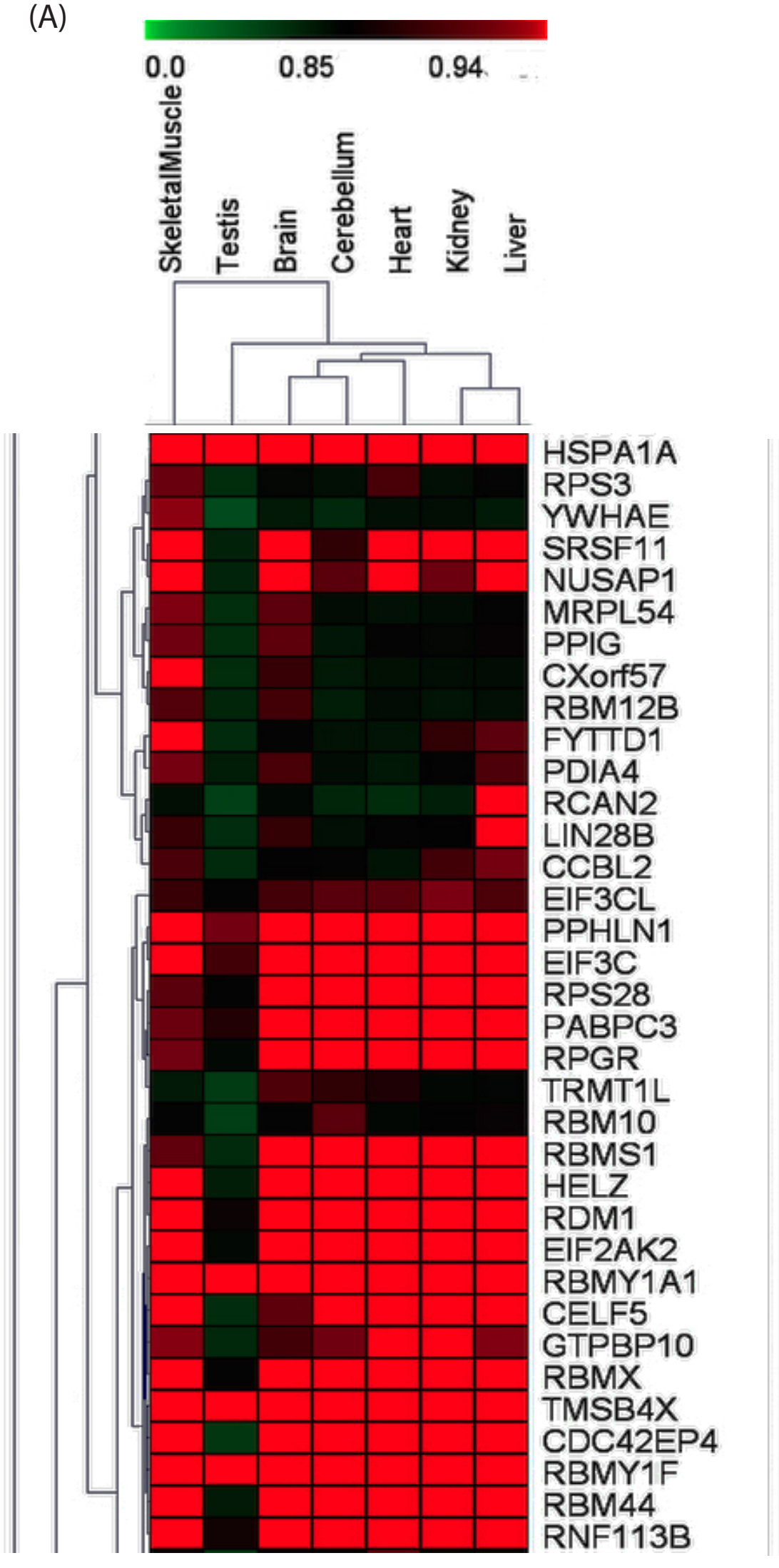


Comparison of Expression Profiles | Figure 3b shows comparisons of expression correlation coefficients for specie-tissue combinations classified into 3 mutually exclusive sets A) of different species and different tissues B) of different species and same tissues and C) of same species and different tissues for RBPs and non-RBPs.

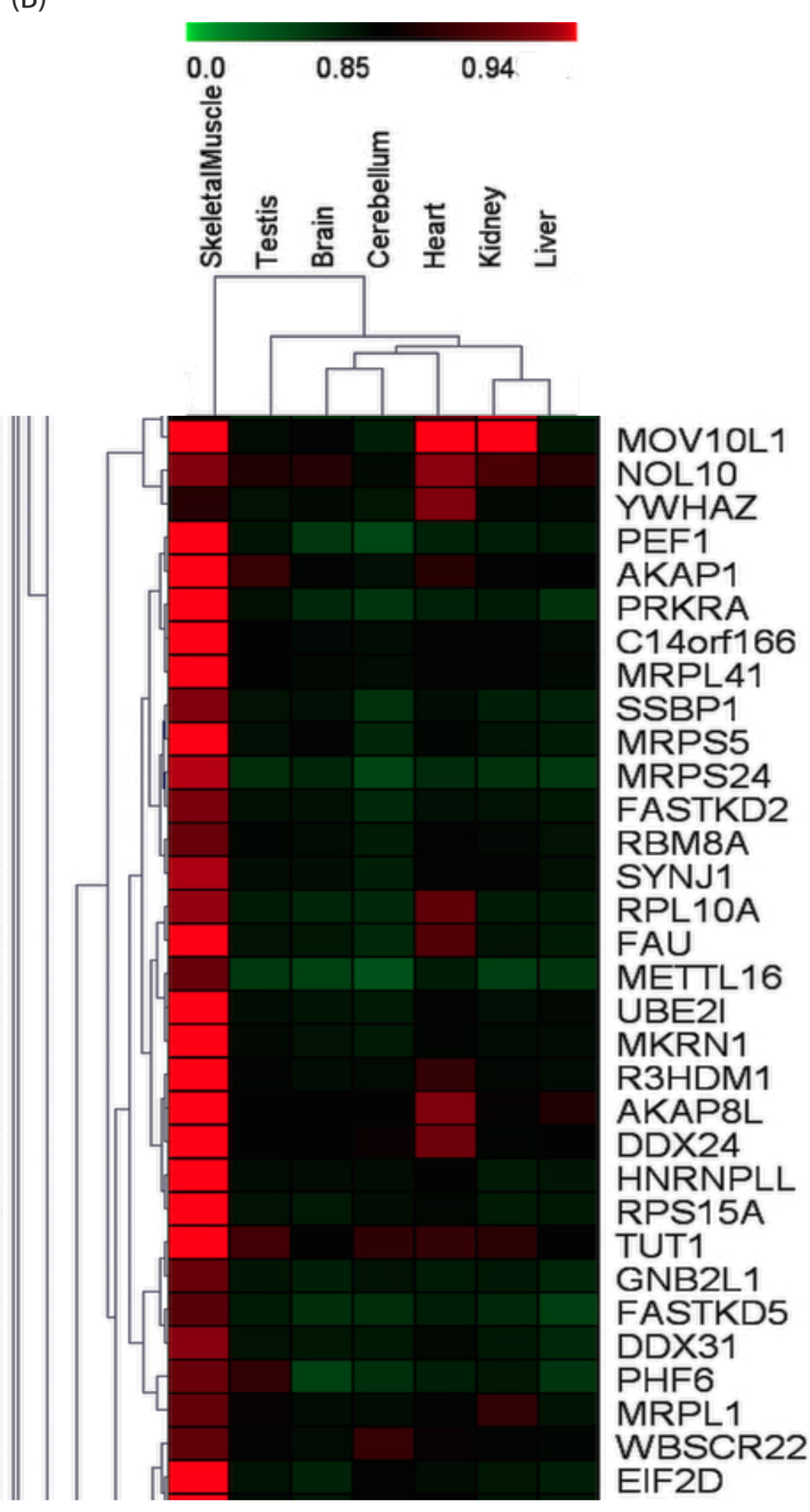


Species Specificity | Figure 4a shows pairwise density distributions of species specificity indices (SSIs) for RBPs vs non-RBPs across various tissues under study

(A)

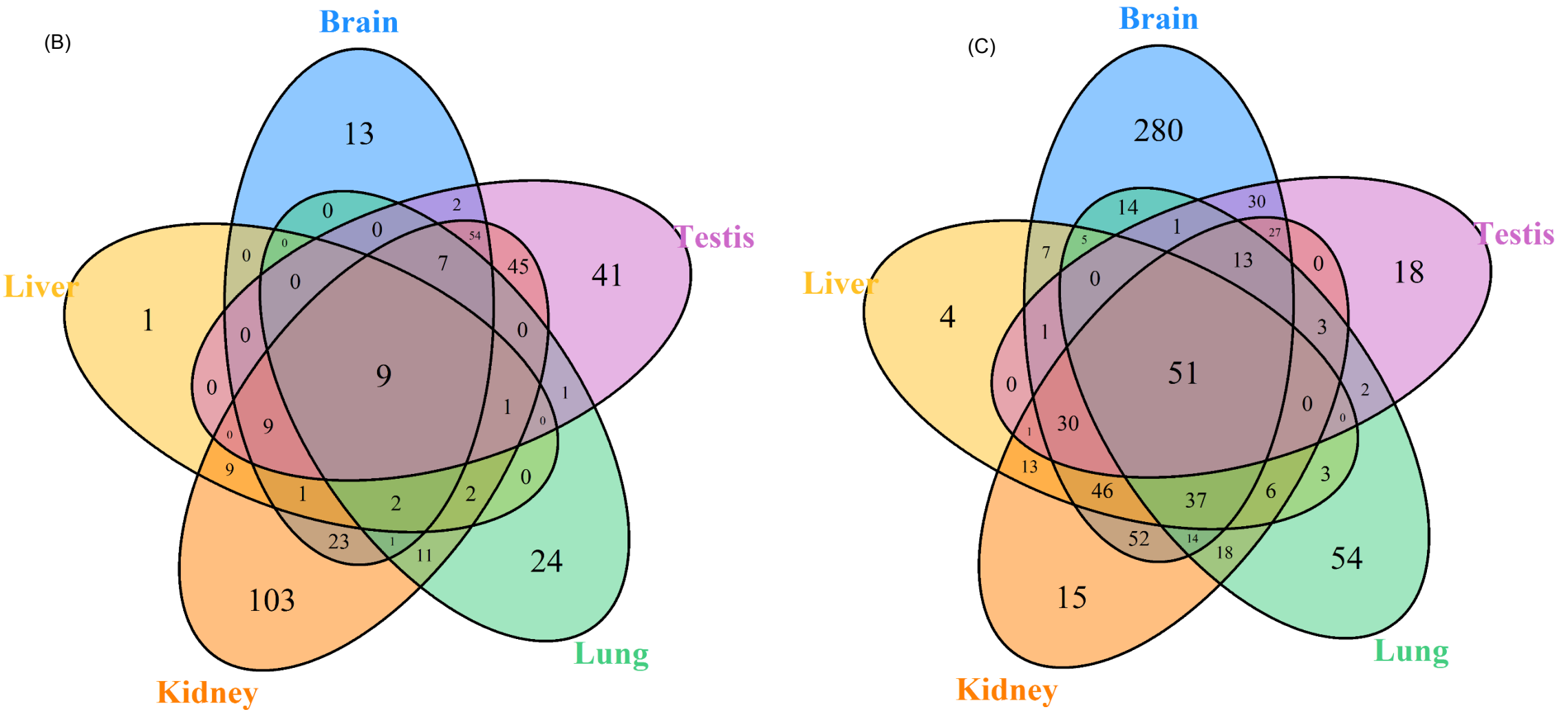


(B)



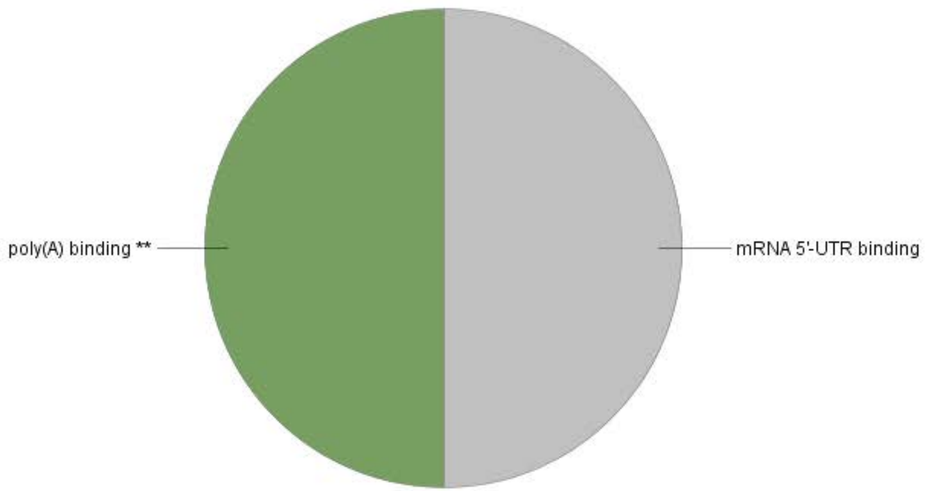
Species Specificity | Figure 4b shows species specific RBPs distinguished into two classes A) Multi-tissue species specific RBPs & B) Single-tissue species specific RBPs

Differential Expression | Figure 5a: Heatmap shows differentially expressed RBPs for six tissues between mouse-human (A) and chicken-human (B). RBPs form four classes: RBPs 1. Continuously evolving RBPs 2.recently evolved RBPs 3. Ancient RBPs 4. Non-changing RBPs.

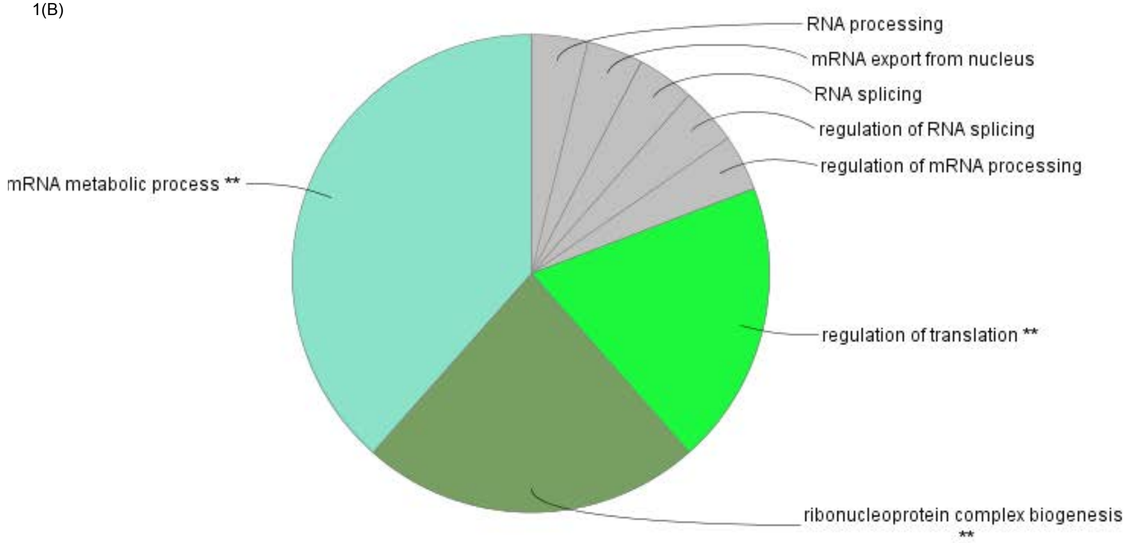


Differential Expression | Figure 5b & 5c: Venn diagrams showing differentially expressed genes in human and chicken (b) and human and mouse(c) in top 5 tissues

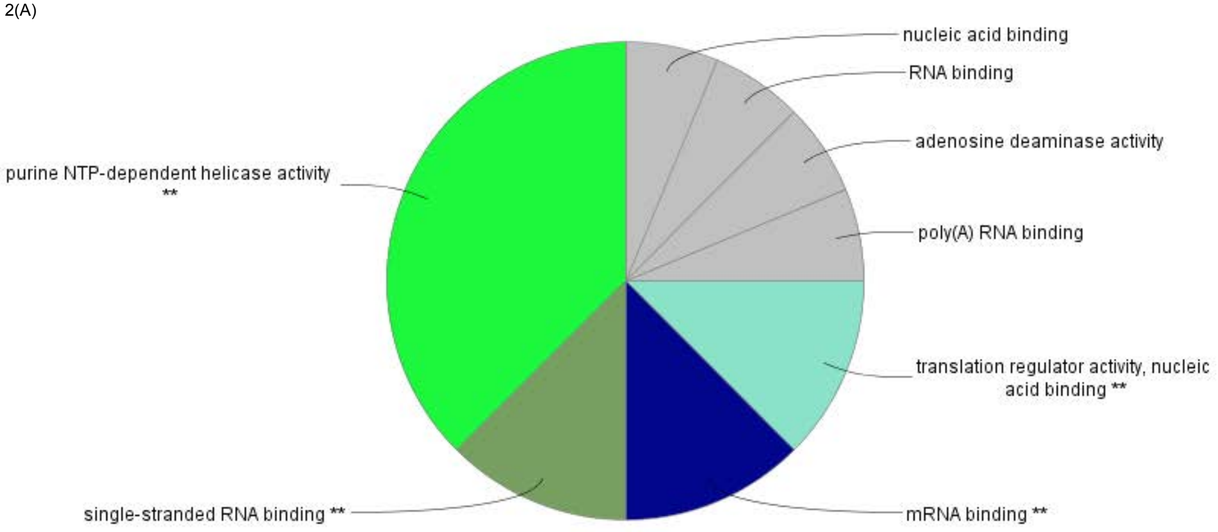
1(A)



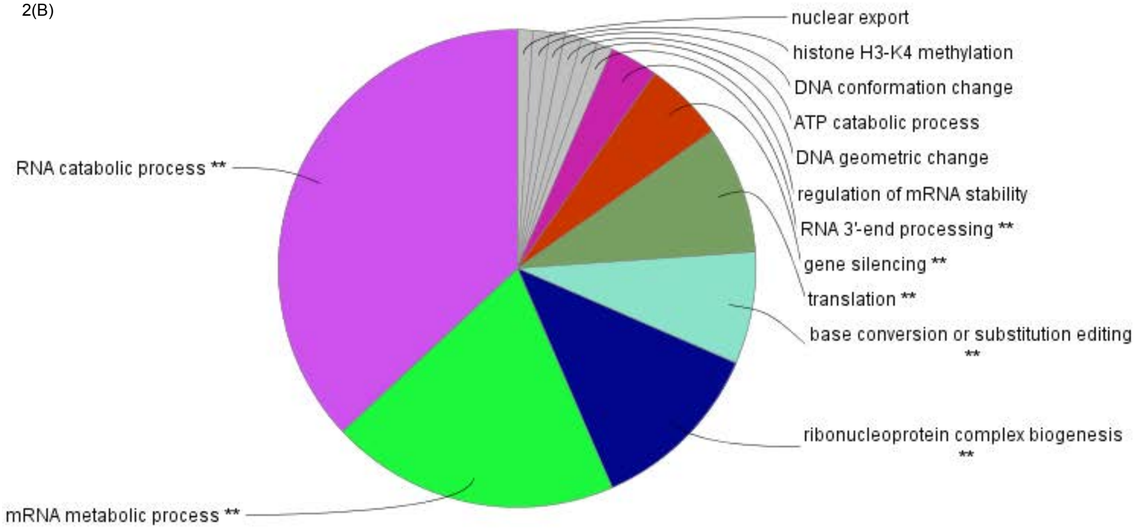
1(B)



2(A)



2(B)



Supplementary Figures 1a, 1b, 2a, 2b: Showing functional enrichment of multi species specific RBPs and single species specific RBPs in mammalian tissues (Human orthologous RBPs were used for conducting functional enrichment)

Chapter 3 Identification and Characterization of Circular RNA in Human Transcriptomes Using longPoly (A) Sequencing

3.1 Introduction

The debate over presence of non-linear exon splicing such as exon-shuffling or formation of circularized forms has finally come to an end as numerous repertoires have shown of their occurrence and presence through transcriptomic analyses.[33, 34, 35] It is evident from these studies that along with consensus-site splicing non-consensus site splicing is robustly occurring in the cell. Also, in spite of applying different high-throughput approaches (both computational and experimental) to determine their abundance, the signal is consistent and strongly conforming the plausible circularization mechanisms. Earlier studies hypothesized and hence focused on the ribo-minus / non poly (A) RNA-seq data to identify circular RNA structures in cell and compared their abundance levels with their linear counterparts. Thus far, the studies show their conserved nature across tissues and species also that they are not translated and preferentially are without poly (A) tail with one to five exons long.

Much of this initial work has been performed using non-polyA sequencing thus probably underestimates the abundance of circular RNAs originating from long poly (A) RNA isoforms. Our hypothesis is if the circular RNA events are not the artifact of random events but has a structured and defined mechanism for their formation then there would not be biases on preferential selection / leaving of polyA tails while forming the circularized isoforms. We have applied an existing computational pipeline from earlier studies by Memczack et.al[35] on ENCODE cell-lines long poly (A) RNA-seq data. With same pipeline we achieve a significant number of circular RNA isoforms in the data some of which are overlapping with known circular RNA isoforms from the literature. We identified an approach and worked upon to identify the precise structure of circular RNA which is not plausible from the existing computational approaches. We aim to study their expression profiles in normal and cancer cell-lines and see if there exists any pattern and functional significance based on their abundance levels in the cell.

3.2 Material and Methods

We accessed ENCODE5 (The Encyclopedia of DNA Elements) project cell-line longPolyA and non-PolyA RNA sequencing data for 6 cancer and 4 normal cell-lines (Table 1). ENCODE is a collaborative consortium among research groups across the globe which maintains integrated repository of cell lines and primary cell types. The data is grouped and prioritized as Tier1 and Tier 2, Tier 1 having high priority and more common cell type. We collected both longPolyA and corresponding non-polyA RNA-sequencing raw sequencing fastq files from ENCODE. We accessed the level one raw reads from The Cancer Genome Atlas (TCGA) for three major cancers (i.e. Liver, Lung, and Breast Cancers). We downloaded RNA and Total RNA samples of solid tissue normal and tumor for each patient. We wished to study whether between cancer and normal samples there is a differential expression of circular RNA transcripts. UCSC hg19 human reference indices were used for detecting the circular RNA origination locations. For the purpose of this thesis the findings from differential expressions of circular RNA in cancers and normal samples are not reported.

Cell	Tier	Description	Lineage	Karyotype
HeLa-S3	2	cervical carcinoma	Cervix	cancer
HepG2	2	liver carcinoma	Liver	cancer
A549	3	Epithelial cell line derived from a lung carcinoma tissue.	Epithelium	cancer
MCF-7	3	mammary gland adenocarcinoma	Breast	cancer
K562	1	Leukemia continuous cell line K-562	Blood	cancer
SK-N-SH_RA	3	neuroblastoma cell line,	Brain	cancer
H1-hESC	1	embryonic stem cells	Embryonic Stem Cell	normal
AG04450	3	fetal lung fibroblast	Lung	normal
HUVEC	2	umbilical vein endothelial cells	Endothelium	normal
NHLF	3	Normal Human Lung Fibroblasts	Lung	normal
HMEC	3	Human Mammary Epithelial Cells	Breast	normal
HSMM	3	Normal Human Skeletal Muscle Myoblasts	Muscle	normal

Table 1: The input ENCODE cell-lines data distribution for which long-polyA and non-polyA RNA-seq data analyzed

3.3 Computational Pipeline

The computational pipeline that we used to analyzed and identify the circular RNA isoforms in longPolyA and non-polyA data is shown as Figure 7. The computational pipeline employs use of existing algorithms Bowtie and Samtools for accomplishing the alignment and mapping steps to the hg19 human reference genome. Then the unmapped reads were extracted to be able to

extract the potential candidates of circular splicing sites. The custom script from Memczack et al[35] was applied to align and extend the anchor positions in the unmapped reads in the head-to-tail orientation to detect the circRNA reads. The reads obtained from this step are again mapped and aligned to identify how many reads are falling into the region undergoing circularization. The final output is a standard bed file with chromosomal locations, strand information and reads count statistics and length of each circular RNA transcript. But the script is not designed to predict exact internal structure of circRNA transcript giving exons and introns organization. This is a major challenge to come to a concluding step of predicting their expression patterns. As the structure is not known of circular RNA transcript we are unable to estimate their abundance levels in cell. The computational pipeline was tested on two replicates of one sample from Hela-S3 cell line to be able to validate the candidates are fractionally overlapping and merely not detected randomly. We found that most of the circRNA are overlapping in two replicates of HeLa-S3 cell-line. Hence we decided to select a single replicate for the detection of circRNA.

In the study we focused on detection of circular RNA formed from head-to-tail orientation using Memczack et al[35] pipeline and not any other form of circRNA such as one forming due to inverted repeat homology of ALU repeats in the transcripts having longer intronic sequences comprising ALU repeats. There is a need of other robust approaches to be developed to detect such circRNA isoforms.

Further to identify the internal exonic structure of circRNA, we used bedtools intersect option and identified known exons from reference human index. We considered the exons completely lying within the circRNA transcript region. To identify exons identifying circRNA transcript circRNA candidate co-ordinates were unchanged to maintain uniformity. We imported the output bed files at Galaxy Workbench. We utilized UCSC human reference genome (hg38, GRCh38) bed file which was imported into the workbench. We used the “organize on genomic intervals” option available at Galaxy Workbench. We intersected the intervals of two datasets option with minimum overlap of ~500nt as mean lengths of circRNA is ~1kb.

Then the bed2bam utility was applied to calculate the expression levels of exons contributing to circRNA transcripts using cufflinks framework. We compared the transcript abundances between circRNA in longPolyA against those predicted in non-polyA cell-lines yet we couldn't find a profound signal for their differential expression patterns. We speculate that lack of enough replicates to calculate variances among samples may be responsible for lack of identifying significant log fold changes.

3.4 Results and Discussion

Our study identifies circular RNA even in longPolyA RNA sequencing data which by earlier hypotheses was clearly underestimated. The ribominus non-polyA data achieves about 10 fold number of circRNA candidates' (on average) detection using the same pipeline than longPolyA data (Figure 8).Though the percentage occurrence of circular reads out of total reads in both data is limited below 0.1%. (Figure 9). Also there is no clear pattern as to circular reads ratio being higher in non-polyA data than in longPolyA, which suggest that we can't override the hypothesis that circRNA detection is plausible in longPolyA data. Remarkably it is seen that indeed the circular reads ratio is higher in case of certain cell-lines such as Nhlf, Hmec, K562 and Huvec. The initial statistics of spliced reads ratio with circular reads ratio in longPolyA data suggests that occurrence of higher splicing reads ratio leads to detection of higher ratio of circRNA which demarcates circRNA identified in longPolyA from ones in non-polyA data, which shows opposite trend of lesser spliced reads in data to lead to detection of higher ratio of circular reads (t-test p-value 4.671e-06). Earlier studies[35, 36, 37] pointed out that circRNA involves lower splicing events to be able to form circRNA, while the trend in our data shows the contrary observations. Average lengths of circRNA spans about 3-5 exonic distances in our data which we extracted by intersect option of bedtools with reference human genome. Also the frequency of identified circular RNA per million base length of human chromosome is varying in range 0.1 to 0.7 across the 10 cell-lines in longPolyA data. Though there is no bias is seen towards any particular genomic region, average frequency of circRNA is typically high in case of chromosome 19 seen for all cell-lines

for longPolyA data. Also after intersection with human reference file, in consensus we observe that average biotype of circular reads is owing to majorly protein-coding regions (50%). Though significant percentage of circular reads originate from various non-coding regions (i.e. lincRNA, pseudogenes, processed transcripts, etc). (Figure 11)

We also performed quantification of the identified circRNA candidates using Cufflinks (Methods). We were unable to establish a particular expression pattern among various cell lines selected for stud and also among non-polyA and long polyA data.

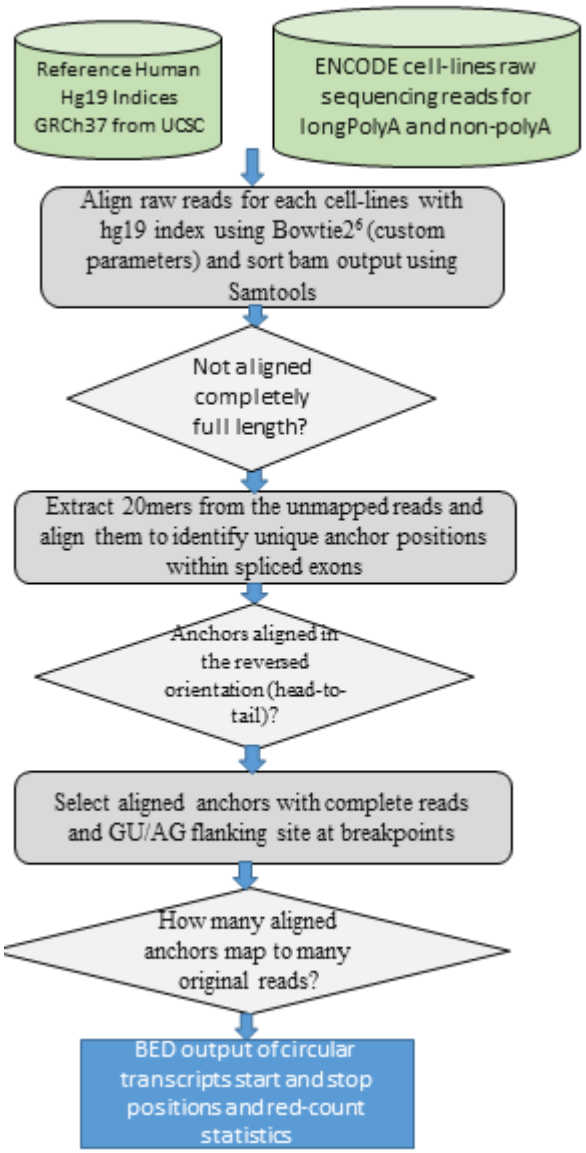


Figure 7: Computational pipeline for detection of circular RNA candidates

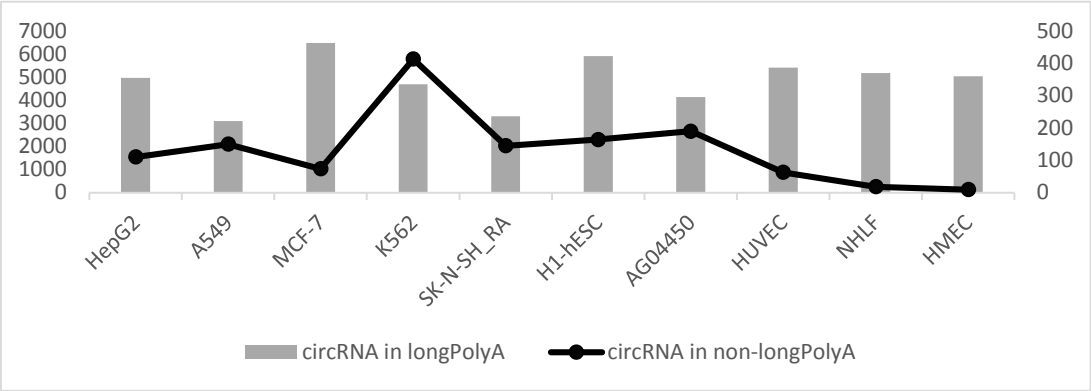


Figure 8: detected circRNA in longPolyA vs non-polyA data (longPolyA (right axis), non-PolyA (left axis)) across cell-lines

This study opens new avenues for working on yet another vital molecule in the cell machinery which may altogether alter our perspective of working transcription and post-transcription regulation machinery.

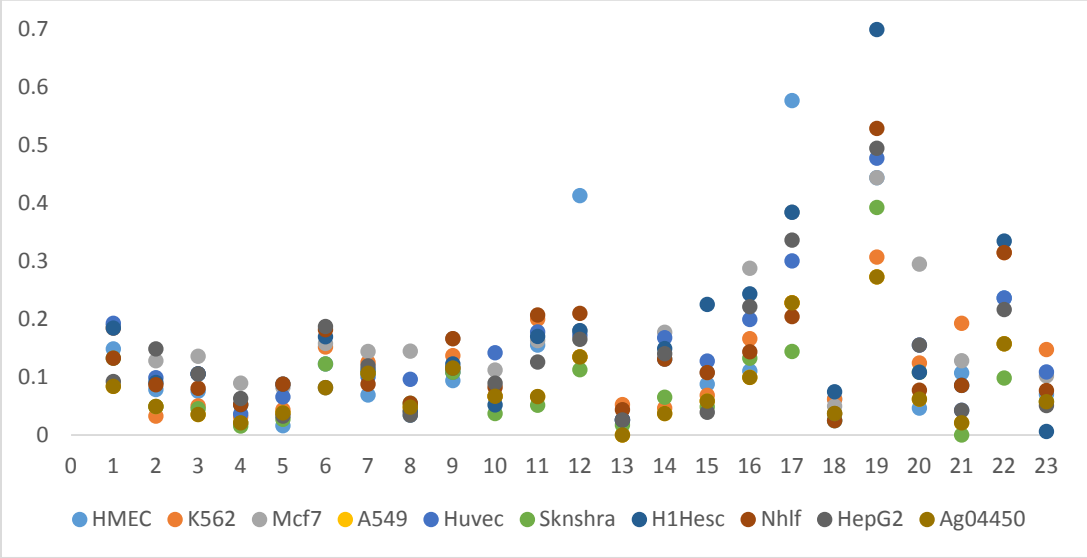


Figure 9: Shows distribution of circRNA candidates detected per million length per chromosome

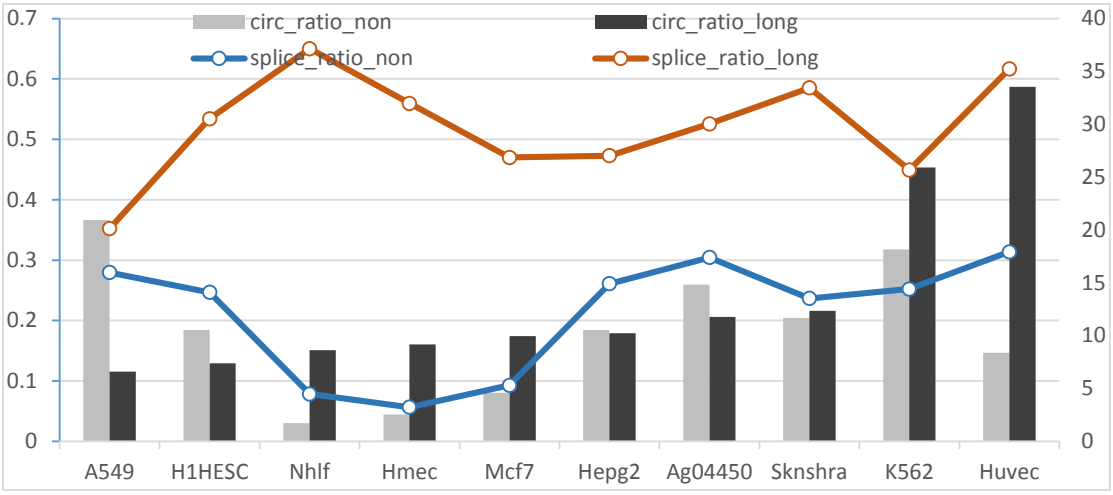


Figure 10: SpliceReads / TotalReads Ratio (Right Vertical Axis) and CircularReads / TotalReads Ratio (Left vertical Axis) for longPolyA and non-Poly A data for the 10 ENCODE cell-lines

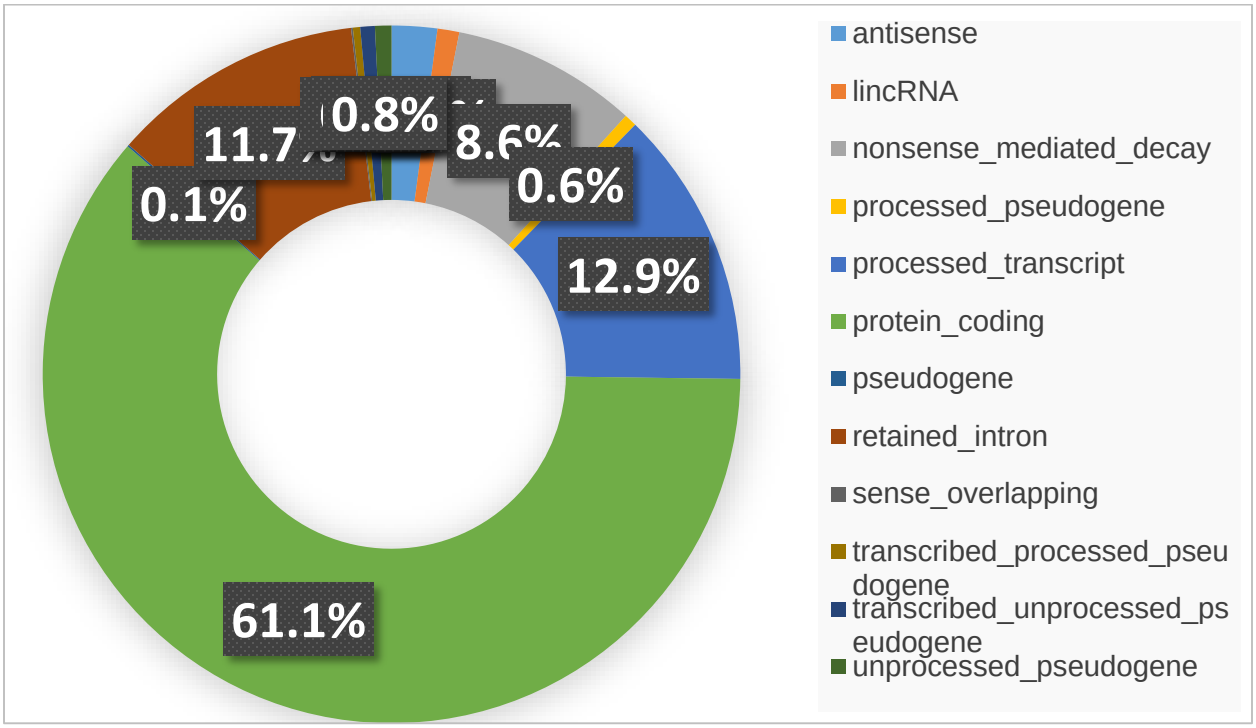


Figure 11: Average Biotype Constitution in predicted transcripts of from circular reads.

Chapter 4: References

1. Brawand D*, et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-348 (2011).

2. Necsulea A*, et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**, 635-640 (2014).

3. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338**, 1593-1599 (2012).

4. Barbosa-Morais NL*, et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338**, 1587-1593 (2012).

5. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters* **582**, 1977-1986 (2008).

6. Fietz SA*, et al.* Transcriptomes of germinal zones of human and mouse fetal neocortex suggest a role of extracellular matrix in progenitor self-renewal. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 11836-11841 (2012).

7. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology* **32**, 462-464 (2014).

8. Hubbard TJ*, et al.* Ensembl 2009. *Nucleic acids research* **37**, D690-697 (2009).

9. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences = Theorie in den Biowissenschaften* **131**, 281-285 (2012).

10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).

11. Cole JR*, et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic acids research* **42**, D633-642 (2014).

12. Kwon SC*, et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nature structural & molecular biology* **20**, 1122-1130 (2013).

13. Ray D*, et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172-177 (2013).

14. Castello A*, et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393-1406 (2012).

15. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: a database of RNA-binding specificities. *Nucleic acids research* **39**, D301-308 (2011).

16. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome research* **19**, 327-335 (2009).

17. Yanai I*, et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650-659 (2005).

18. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nature reviews Genetics* **15**, 829-845 (2014).

19. Kishore S, Luber S, Zavolan M. Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Briefings in functional genomics* **9**, 391-404 (2010).

20. Mittal N, Roy N, Babu MM, Janga SC. Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 20300-20305 (2009).

21. Mittal N, Scherrer T, Gerber AP, Janga SC. Interplay between posttranscriptional and posttranslational interactions of RNA-binding proteins. *Journal of molecular biology* **409**, 466-479 (2011).

22. Musunuru K. Cell-specific RNA-binding proteins in human disease. *Trends in cardiovascular medicine* **13**, 188-195 (2003).

23. Kechavarzi B, Janga SC. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome biology* **15**, R14 (2014).

24. Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. *Trends in genetics : TIG* **29**, 318-327 (2013).

25. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673-684 (2009).

26. Su AI*, et al.* Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 4465-4470 (2002).

27. Bossi A, Lehner B. Tissue specificity and the human protein interaction network. *Molecular systems biology* **5**, 260 (2009).

28. Milinkovitch MC, Helaers R, Tzika AC. Historical constraints on vertebrate genome evolution. *Genome biology and evolution* **2**, 13-18 (2010).

29. Kosiol C*, et al.* Patterns of positive selection in six Mammalian genomes. *PLoS genetics* **4**, e1000144 (2008).

30. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106 (2010).

31. Odom DT*, et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature genetics* **39**, 730-732 (2007).

32. Yue F*, et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364 (2014).

33. Staff, P.O., Correction: Circular RNA Is Expressed across the Eukaryotic Tree of Life. PLoS One, 2014. 9(4): p. e95116.

34. Wang, P.L., et al., Circular RNA is expressed across the eukaryotic tree of life. PLoS One, 2014. 9(3): p. e90859.

35. Memczak, S., et al., Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 2013. 495(7441): p. 333-338.

36. Salzman, J., et al., Cell-type specific features of circular RNA expression. PLoS Genet, 2013. 9(9): p. e1003777.

37. Jeck, W.R., et al., Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA, 2013. 19(2): p. 141-57.

38. Hansen, T.B., et al., Natural RNA circles function as efficient microRNA sponges. Nature, 2013. 495(7441): p. 384-8.

39. Lukiw, W.J., Circular RNA (circRNA) in Alzheimer's disease (AD). Front Genet, 2013. 4: p. 307.

40. Hansen, T.B., J. Kjems, and C.K. Damgaard, Circular RNA and miR-7 in cancer. Cancer Res, 2013. 73(18): p. 5609-12.

41. Danan, M., et al., Transcriptome-wide discovery of circular RNAs in Archaea. Nucleic Acids Res, 2012. 40(7): p. 3131-42.

42. Salzman, J., et al., Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS One, 2012. 7(2): p. e30733.

43. Hansen, T.B., et al., miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. EMBO J, 2011. 30(21): p. 4414-22.

CURRICULUM VITAE

ABHIJIT BADVE
m: 317-378-9656 | e: asbadve@iupui.edu | address: 322,Canal Walk, Apt#375, Indianapolis, IN-46202

CAREER OBJECTIVE:

To seek a full-time bioinformatics analyst position which provides me avenues to utilize cross-disciplinary knowledge and analytical skills in Bioinformatics and contribute towards addressing complex biomedical research problems.

PROFESSIONAL SUMMARY:

- 3+ years of professional and research experience of bioinformatics analysis and software development using C++, Perl, Python and Matlab
- Building "**Next Generation Sequencing pipelines**" and using "**Statistical and analytical packages for R**" for analyzing large genomic sequencing data sets and cancer genotyping datasets

WORK EXPERIENCE:

***Research Assistant, Lab of Genomics and Systems Biology, Indiana University,* Since 08/2013**

- Studying evolutionary aspects of RNA binding protein expressions across mammalian tissues from RNA-sequencing platforms (Submitted)
- Deciphering post-transcriptional regulatory networks of long non-coding RNA(lncRNA) and microRNA (miRNA) interactions using RNA-sequencing and expression profiling
- Analyzing TCGA cancer genomic data for elucidating cancer-specific miRNA and TF interaction networks
- Computational method design for detection of circular RNA (circRNA) from PolyA+ RNA-sequencing data in ENCODE cell lines and TCGA level-1 RNA sequencing datasets
- Development of relational database with web interface in MySQL and PHP for demarking epigenetic events in cancer genes

Awarded departmental graduate research assistantship for entire tenure of master's degree

***Systems Engineer, Tata Consultancy Services*, 07/2010 - 07/2013,**

- Human machine Interface design and testing modules in C++, Python and Matlab. Creation of automated testing scripts using VB.net and Python.
- Web-based development for management and maintenance of client specifications and record delivery notes using Microsoft Sharepoint and MS-Access.

Awarded outstanding team member of the year for exceptional contribution towards leading a human-machine interface development project

EDUCATION:

MS in Bioinformatics, *Indiana University-Purdue University at Indianapolis* ***(Cumulative GPA: 3.67)*** *(*Expected 2015)

M.Sc in Bioinformatics, *Birla Institute of Technology, Mesra, Ranchi, India* ***(GPA: 3.6) (05/2010)***

ACHIEVEMENTS:

10/2014: Invited for brief and poster presentation at Rustbelt RNA Meeting, Pittsburgh, Pennsylvania (2014)

4/2010: Poster presentation at Microbial Society of India, Department of Biotechnology, Birla Institute of Technology, Mesra, India

07/2010: Co-authored a research article “Molecular evolution of virulence genes of Swine Influenza Virus Subtype-A H1N1 : An analysis of host radiation” at International Journal of Pharma and Biosciences

09/2009: Co-authored in a research communication “The 2009 Nobel prize in Chemistry: for studies of the structure and function of the Ribosome” at International Journal of Applied Biology and Pharmaceutical Technology

SKILLS:

Programming: Well versed with scripting languages such as Perl, Python, Unix-Shell, VB.net and High Performance computing using SGE or PBS systems

Databases and Web Technologies: design using MySQL, SQLite and PostgreSQL; PHP, ASP.NET, HTML, CSS, CGI-Perl

Bioinformatics: SAM tools, BED tools, PRINSEQ, FASTx Toolkit, BOWTIE2, BWA, TOHAT, SAILFISH, CUFFLINKS, RSEM; Galaxy, GenePattern, GATK, IGV, IGB; VCFTools; R/Bioconductor packages: Limma, EdgeR and DESeq2, Cummerbund; CUFFDIFF, MISO, SpliceR, AltAnalyze; Network analysis using R/iGraph, Cytoscape, MFINDER, MeV, GeneE, Java tree-view; MEME Suite, Homer, PHI-Blast; DAVID, gProfiler, Panther, Ingenuity IPA,GSEA; Bioinformatics resources at TCGA, dbGaP, dbSNP, Human Microbiome Project, SRA, NCBI, UCSC, ENSEMBL and Broad Institute.

NGS data-sets: Experience in analyzing DNA-Seq, RNA-Seq, Microbiome, Chip-Seq, whole-genome sequencing datasets, exome-sequencing datasets.

Ability to work well on a team: Developed team-working skills by working on collaborative research projects at Genomics and Systems Biology Lab under Dr. Janga (2013-14). Also received an award for team-working skills and exceptional contribution towards project completion at Tata Consultancy Services (2011).

Leadership capabilities: Actively taken responsibility to supervise a module development team at Tata Consultancy Services.

Dedication and commitment: I really enjoy stretching myself outside my core domain and am very interested in finding new ways to expand my knowledge and apply it to research questions