INTEGRATIVE COMPUTATIONAL GENOMICS BASED APPROACHES TO UNCOVER THE TISSUE-SPECIFIC REGULATORY NETWORKS IN DEVELOPMENT AND DISEASE

Rajneesh Srivastava

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing,
Indiana University

March 2020

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

____

Sarath Chandra Janga, PhD, Chair

____

Xiaowen Liu, PhD

December 18, 2019

____

James A. Marrs, PhD

____

Mark H. Kaplan, PhD

DEDICATION

I dedicate my dissertation work to my parents who unconditionally supported me to follow my passion for science. I also dedicate this work to my brother who always encouraged me to follow my dreams. I especially thank and dedicate this work to my dear wife for her endless love and support throughout this journey.

ACKNOWLEDGEMENT

First and foremost, I thank Almighty GOD, for giving me strength to complete this thesis. Undertaking PhD has been a life changing experience for me, and it would not have been possible without the support and assistance of many people.

I would like to express my sincere gratitude to my PhD supervisor, Dr. Sarath Chandra Janga for his tremendous guidance, motivation, advice and support in completing this thesis. I am greatly indebted to Dr. Janga for providing me every bit of assistance I needed during the first few semesters of PhD when I was transitioning from a core experimental lab to a computational lab. I am thankful to him for his timely suggestions and constructive criticisms that kept me moving forward. He has graciously taught me how to define a research problem and find a solution to it with consistent efforts. I appreciate all his contributions of time, ideas and endless motivation that made my PhD experience productive and stimulating. I thank him for encouraging my research and enabling me to grow as an efficient researcher. His advice on my research as well as on my career have been invaluable.

I am thankful to Dr. Huanmei Wu, Department Chair for providing a conducive working environment for completing this study. I am grateful to my committee members, Dr. Xiaowen Liu, Dr. James A. Marrs and Dr. Mark H Kaplan for their brilliant comments and insightful suggestions that helped me to widen my research from various perspectives and gave the current shape to this thesis.

I humbly acknowledge the faculty members of IU School of Informatics and Computing, Dr. Xiaowen Liu, Dr. Meeta Pradhan, Dr. Yunlong Liu, Dr. Jake Chen and

Dr. Huanmei Wu who taught me the courses, that helped me to broaden my vision for my research study.

I thank the SOIC staff, Elizabeth Cassell, Robyn Hart, David Tauriainen and Kimberly Melluck for extending their timely help and technical support. I sincerely appreciate the UITS Research Technologies at Indiana University for providing the computing support that enabled the execution of this study.

I extend my sincere thanks to Dr. Tarek M. Ashkar, for his esteemed guidance on Umod project. I also thank Dr. Charlie X. Dong, Dr. Suthat and their research team at IU medicine for experimental validation and guidance for Sesn3 project. I am grateful to Dr. Salil Lachke and Dr. Soma Dash, for their significant contribution in eye project. I extend deep gratitude to Dr. James A. Marrs and Dr. Mark H Kaplan, for providing me learning opportunity while analyzing their research data.

I would like to thank my fellow labmates, Sasan, Aarti, Sasank, Dr. Raja, Gungor, Quoseena, Vidhur, Praneet, Kasish, Vishal, Arun, Sneha and Yasheswini who directly and indirectly worked with me and contributed to the progress of the projects by their stimulating discussions, and for all the fun we have had in the last five years.

I gratefully thank my undergrad faculties at BHU, India, Dr. Ravi Kumar Asthana and Late Dr. Surinder Singh for encouraging me to pursue my dreams. I would also like to say a heartfelt thanks to Dr. Sanjeeva Srivastava, for providing me opportunity to expand my expertise in proteomics research at IIT Bombay, India.   I am indebted to all my friends at IIT Bombay, Dr. Sandipan Ray, Dr. Jaipal Panga, Dr. Rekha Jain, Dr. Darpan Malhotra and Dr. Aishwariya Rao for their feedback on my research and for their support.

I owe gratitude to my friends during PhD journey, Dr. Vachik, Subbu, Vijay, Kartik, Masood, Aman, Keshav, Nikhil and many more, at Indianapolis who gave me stressbuster company and for the great times we have shared. I thank Subhamoy Das, Manjesh Dubey, Dr. Neha Agrawal and Dr. Komal Sanghwan for being supportive friends and for always being there for me during ups and downs.

Last but not the least, I am grateful to my parents Mr. Subhash Chandra Lal and Mrs Poonam Srivastava, for bringing me to this world and for showing their endless love and support in all my pursuits. This thesis would not have been possible without their warm love, continued patience, and endless support. I am thankful to my brothers; Mr. Manish Srivastava, Sachin Srivastava, Raveesh Srivastava, Rishabh Srivastava, sisters; Isha Srivastava, Sikha Srivastava and sister-in-law; Priya Srivastava, and Manisha Srivastava for their everlasting care and for the roles they played on my behalf while I was away from home. I deeply express my thanks to my in-laws, Mr. Brij Kishore Srivastava and Dr. Meena Srivastava for their unconditional love and care. I thank my brother-in-law Dr. Pranay Srivastava and sister-in-law, Dr. Aakriti Srivastava for their smiley support and for sharing their experiences with me to help me overcome challenges. In the end, I would like to extend my heartfelt special thanks to my loving, supportive, encouraging, and wonderful wife, Dr. Mansi Srivastava who has not only supported me morally but also made me smile during the most difficult time. Her faithful support during the final stages of PhD is commendable.

Finally, I thank and appreciate the contribution of each person who made this journey enjoyable and helped me to shape this thesis in its current form.

Rajneesh Srivastava

INTEGRATIVE COMPUTATIONAL GENOMICS BASED APPROACHES TO UNCOVER THE TISSUE-SPECIFIC REGULATORY NETWORKS IN DEVELOPMENT AND DISEASE

Regulatory protein families such as transcription factors (TFs) and RNA Binding Proteins (RBPs) are increasingly being appreciated for their role in regulating the respective targeted genomic/transcriptomic elements resulting in dynamic transcriptional (TRNs) and post-transcriptional regulatory networks (PTRNs) in higher eukaryotes. The mechanistic understanding of these two regulatory network types require a high resolution tissue-specific functional annotation of both the proteins as well as their target sites. This dissertation addresses the need to uncover the tissue-specific regulatory networks in development and disease. This work establishes multiple computational genomics based approaches to further enhance our understanding of regulatory circuits and decipher the associated mechanisms at several layers of biological processes. This study potentially contributes to the research community by providing valuable resources including novel methods, web interfaces and software which transforms our ability to build high-quality regulatory binding maps of RBPs and TFs in a tissue specific manner using multi-omics datasets. The study deciphered the broad spectrum of temporal and evolutionary dynamics of the transcriptome and their regulation at transcriptional and post transcriptional levels. It also advances our ability to functionally annotate hundreds of RBPs and their RNA binding sites across tissues in the human genome which help in decoding the role of RBPs in the context of disease phenotype, networks, and pathways.

The approaches developed in this dissertation is scalable and adaptable to further investigate the tissue specific regulators in any biological systems. Overall, this study contributes towards accelerating the progress in molecular diagnostics and drug target identification using regulatory network analysis method in disease and pathophysiology.

Sarath Chandra Janga, PhD, Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AML	Acute myeloid leukemia
ANOVA	Analysis of varience
API	Application Programming Interface
AS	Alternative splicing
BAM	Binary Alignment Map
BCM	Binary conservation matrix
BMo	Binding Motif
Cas9	CRISPR-associated endonuclease 9
cDNA	Complementary DNA
ChIP	Chromatin Immunoprecipitation
CLIP	Cross-linking and Immunoprecipitation
CRISPR	Clustered regulatory interspaced short palindromic repeats
DHS	DNase I hypersensitive site
DNase	Deoxyribonuclease
EEM	Exon expression matrix
ENA	European Nucleotide Archive
ENCODE	Encyclopedia of DNA Elements
GBM	Glioblastoma
GEO	Gene Expression Omnibus
GTEx	Genotype-Tissue Expression
GTF	HITS-A10:B11CLIP
HBM	Human BodyMap
HGNC	HUGO Gene Nomenclature Committee
HISAT	Hierarchical Indexing for Spliced Alignment of Transcripts
HITS-CLIP	High-throughput sequencing of RNA isolated by cross-linking immunoprecipitation
JSON	JavaScript Object Notation
KIRC	Kidney Renal Clear Cell Carcinoma
LIHC	Hepatocellular Carcinoma

lncRNA	Long non-coding RNA
MAF	Multiple alignment file
MEME	Multiple Em for Motif Elicitation
MGI	Mouse Genome Informatics (www.informatics.jax.org)
mRNA	Messenger RNA
MySQL	My Structured Query Language
ncRNA	Non-coding RNAs
NGS	Next-generation sequencing
ORF	Open reading frame
PAR-CLIP	Photoactivatable ribonucleotide-enhanced cross-linking and immunoprecipitation
PEEK	Prioritization of RBP-Binding sites using Expression and Evolutionary Constraints
PHP	Hypertext Preprocessor
PTRN	Post-Transcription Regulatory Network
PWM	Position Specific Matrix
qPCR	Quantitative polymerase chain reaction
RBP	RNA Binding Protein
RNA-seq	RNA-sequencing
RPKM	Reads Per Kilobase per Millions of reads
SAM	Sequence Alignment Map
SESN3	Sestrin 3
sgRNA	Single-guide RNA
SRA	Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra)
TCGA	The Cancer Genome Atlas
TF	Transcription factor
TFBS	Transcription Factor Binding Site
TPM	Transcripts Per Million reads sequenced
TRN	Transcription Regulatory Network
TSS	Transcription start site
UMOD	Uromodulin

CHAPTER 1

BACKGROUND

1.1 Basic gene regulation model

Gene regulation is a multistep dynamic process that occurs through a highly controlled and concerted regulatory network (1, 2). This mechanism occurs at all levels of biological processes including chromatin remodeling (3), transcriptional regulation (4), mRNA processing (5), modification (6), post transcriptional regulation (7), stability and degradation (8, 9), localization, translation and post-translational modification (10). Interestingly, several transcriptional and post transcriptional regulatory mechanisms are governed by special class of proteins (Figure 1). Genes are transcribed into RNAs under the transcriptional control by one or many transcription factors (TFs). TFs are the proteins that bind specifically to gene promoters at regulatory positions (binding motifs) and thus contribute to cell identity, physiology and development. The RNAs thus

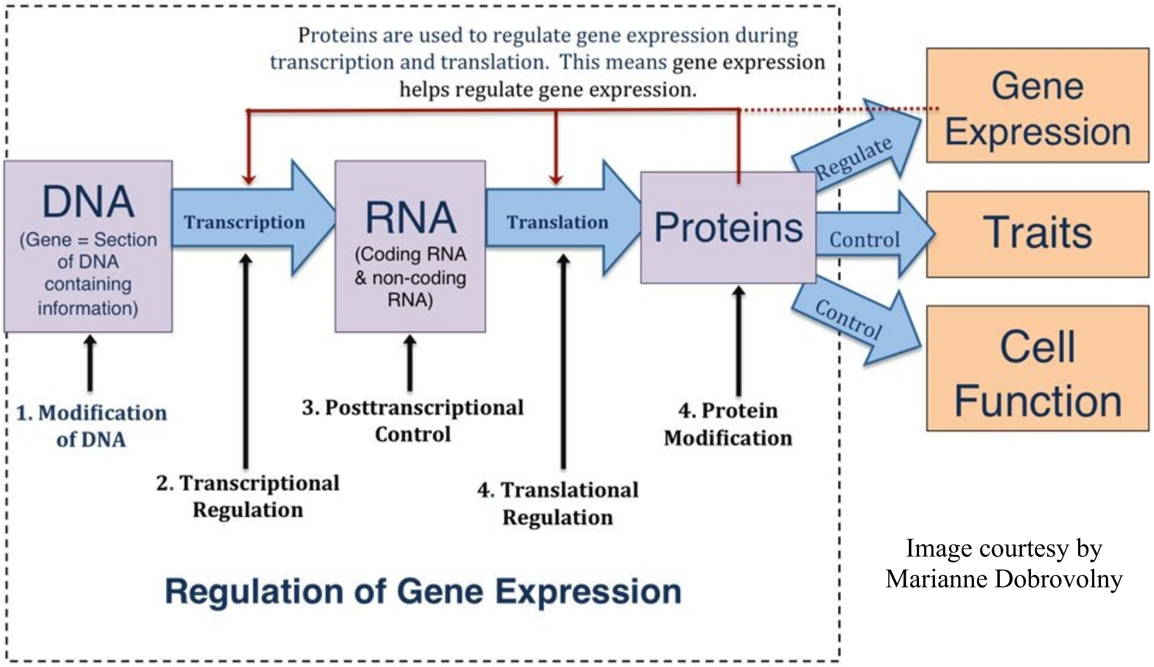

Proteins are used to regulate gene expression during transcription and translation. This means gene expression helps regulate gene expression.
DNA
(Gene = Section of DNA containing information)
Transcription
RNA
(Coding RNA & non-coding RNA)
Translation
Proteins
Regulate
Control
Control
Gene Expression
Traits
Cell Function
1. Modification of DNA
2. Transcriptional Regulation
3. Posttranscriptional Control
4. Translational Regulation
4. Protein Modification
Regulation of Gene Expression

Image courtesy by
Marianne Dobrovolny

Figure 1: Overview of gene expression and regulation. Fundamental role of regulatory proteins at several steps concatenating to the functional phenotype.

produced, undergo several maturation processes, especially in eukaryotes that involves post-transcriptional regulation such as 5' capping, 3' polyadenylation, splicing and possibly RNA editing. These processes are mediated by RNA-binding proteins (RBPs). RBPs are another class of regulatory proteins that have a wide range of functions (such as post transcriptional regulation and metabolism of RNA) in eukaryotes, including splicing, poly-adenylation and capping as well as their localization, translation, stability and degradation (7). These proteins are implicated in several human diseases, including HIV/AIDS, cancer, and neurodegenerative disorders.

1.2 High-throughput sequencing technology and data explode

High-throughput sequencing technologies have rapidly evolved since the last few decades. With the advent of next generation sequencing technology (NGS, considered as 2nd generation technology), the horizon of molecular biology research has been expanded from "epi" gene to "proteo-genomics" with reduced cost (11, 12). This technology has been widely used to deal with a variety of biological questions. It quantitatively helps in deciphering the fundamental challenges priming in expression, regulation and conformation studies. With this technology, researchers are able to pierce the dark matter of the genome, and set up a rationale to believe the 'junk' is no more junk (13-15). Such techniques have been utilized to generate tons of sequence data in the field of genetics, genomics, transcriptomics, regulomics and structuromics (16-20). It enables the researchers to establish a landmark in mentioned field including the temporal, condition specific, tissue specific or cell type specific incomprehensible dilemma. This technology has further evolved to third generation, i.e. long-read sequencing as introduced by Pacbio (21), and fourth generation with the invention of nanopore (22) based long read sequencing technology. These upgrades have delimited the molecular approach to investigate the biological problems at the single-molecule scale (15, 23).

1.3 Overview of the study

Various in vitro, in vivo and in silico approaches have been developed so far to identify the functional motif of regulatory proteins genome wide, the tissue specific binding pattern of most of TFs and RBPs and their dynamic transcriptional (TRNs) and post-transcriptional regulatory networks (PTRNs) in higher eukaryotes is still illusive. This dissertation is committed to uncover the tissue-specific regulatory networks in development and disease. It aims to establish an integrative computational genomics based approaches to further enhance the current understanding on regulatory circuits and decipher the associated mechanism at several layers of biological processes.

This dissertation work consists of five chapter, with Chapter 1 being the background of the study. Chapter 2 is aimed to study the compendium of transcription factors regulating the gene associated to disease phenotype. In this section, I established a gene centric intricate network of conserved DNA upstream motifs and associated transcription factors and investigated how these TRNs modulates the expression of targeting gene using an integrated computational and experimental approach. This chapter includes two gene centric case study, where an in silico phylogenetic foot printing approach was implemented for genes such as Uromodulin (highly expressed in kidney) and Sestrin3 (functionally important gene in maintaining homeostasis in liver). The highly connected TRNs established in this chapter were further investigated to suggest their central role in controlling the gene expression. Several transcription factors regulating target gene were further supported by known literature or verified by CRISPR-Cas9 knock out experiments.

Chapter 3 is aimed to investigate the complete transcriptome architecture of developing mouse eye (at current lens and retina) and to develop a resource for easy navigation of transcriptome profiles encompassing known and novel transcripts across multiple development stages in eye tissues. In this chapter, a total of 35 RNA sequence data encompassing 7 developmental stages of lens and 11 developmental stages of retina from publicly available wild-type mouse datasets were included. These datasets were processed, aligned, quantified and analyzed with in-house RNA-Seq analysis pipeline and hosted a total of >81,000 transcripts in the lens and >178,000 transcripts in the retina across all the included developmental stages. This study revealed an abundance of novel transcripts and extensive splicing alterations (especially in lens) with significantly decreased extent of novelty (of expressed transcripts) in post-natal lens compared to embryonic stages. Several of the novel transcripts and splicing events are verified using RT-PCR and Sanger sequencing.

Chapter 4 is aimed to develop a computational framework for systematic tissue-specific annotation of protein-RNA interaction in the human genome to uncover disease associated binding events. In this section, I develop a tool for systematic identification and comparison of processes, phenotypes, and diseases associated with RNA-binding proteins from condition-specific CLIP-seq profiles. The study further aimed to develop a computational framework for systematic tissue-specific annotation of functional binding sites of RBPs in the human genome and to uncover disease associated binding events. The proposed computational framework employed a tissue-specific cross-species RNA-seq information from more than 100 samples encompassing 4 vital tissues (Kidney, Liver, Brain, Heart) and 10 species to prioritize and evolutionarily annotate the binding sites of RBPs across tissues. Several of these high confidence functional binding sites predicted to control the proximal exons in human cell lines were validated using Crispr/Cas9 screening.

Chapter 5 discusses the significance of the study along with innovation, achievements and limitations of the project. It also discusses the projected future work of this study.

1.4 Impact of the study

This dissertation aims to investigate several genomic features such as conservation and chromatin accessibility that could help to identify novel regulatory motifs and associated TFs. It also investigates the temporal and evolutionary dynamics of transcriptome in developmental stages that could further expand the current understanding of the complete transcriptomic architecture and their regulation in developing mouse eye.

The study deciphers a high resolution regulatory network of RBPs using multi-OMIC datasets. It also investigates the positional binding of RBPs and their impact on proximal functional transcriptomic elements. This study will enhance the knowledge of RBPs and their target RNA in the context of disease phenotype, networks and pathways. Overall, this study could help accelerating the progress in molecular diagnostics and drug target identification.

CHAPTER 2

TO STUDY THE COMPENDIUM OF TRANSCRIPTION FACTORS REGULATING THE GENE ASSOCIATED TO DISEASE PHENOTYPE

2.1 An intricate network of conserved DNA upstream motifs and associated transcription factors regulate the expression of uromodulin gene

2.1.1 Introduction

Uromodulin, also called THP (Tamm–Horsfall protein), is the most abundant protein excreted in the urine under physiological conditions. It is highly produced in the kidney and secreted into the urine via proteolysis of its GPI (Glycosylphosphatidylinisotol) anchored domain (24). It has also been reported in blood as a secretory product in circulation (25). Although the biological function of Uromodulin had been elusive for many years, the last decade witnessed significant advancements in understanding the function of this protein in health and disease (25, 26). In fact, uromodulin is now thought to facilitate electrolyte transport across thick ascending limb (27), modulate the inflammatory response during kidney injury (28), inhibit stone formation (29) and protect the bladder from invasive ascending infections (30). Interestingly, the correlation of rate of uromodulin excretion with disease models is not well established (26), although recent studies suggest that uromodulin expression and excretion increases in injury states and may serve as biomarker for developing chronic kidney disease (31). Because of its protective function in vivo during injury, several studies have previously suggested that this increase in excretion is reactive (26), but the factors that control uromodulin expression are not well studied. Indeed, understanding the complex regulatory mechanisms that regulate the Uromodulin gene is essential to advancing this field.

Several studies have reported that any variation or mutation occurring in this gene directly or indirectly is linked to kidney disorders like glomerulocystic kidney disease, medullary cystic kidney disease type 2 (ADMCKD2), familial juvenile hyperuricemic nephropathy disease (FJHN) etc. with the autosomal dominant tubulointerstitial kidney diseases collectively known as uromodulin-associated kidney diseases (UAKD) (32). Many SNPs on UMOD are linked to chronic kidney disease (CKD) by Genome-wide association studies in the general population (25).

Transcription factors (TFs) are known to bind specifically to gene's promoters at regulatory positions (binding motifs) and thus contribute to cell identity, physiology and cell development. Various *in vitro* (33), *in vivo* (34) and *in silico* (35) approaches have been developed so far for the regulatory motif discovery of genes. Typically, potential TF binds to its high affinity binding site however little is known about the tissue specific binding pattern (represented as a weight matrix) of most TFs in higher eukaryotes (36). In this study, I used the upstream regulatory regions of human UMOD orthologs from a diverse set of 8 primates and 7 rodents to perform phylogenetic foot-printing (37) by employing the MEME-SUITE of tools (http://meme.nbcr.net/meme/intro.html), which allowed the identification of high confident conserved binding motifs and corresponding position specific weight matrices. I also tested the feasibility (i.e. TF binding tendency) of BMo in open chromatin region of mouse using DNAse Hypersensitive sites in UMOD upstream region. Further, the predicted binding motifs were analyzed by a motif comparison tool from MEME-SUITE, TOMTOM- which compares discovered motifs with currently annotated motifs, to identify transcription factors, which have a high tendency and specificity to bind to these discovered motifs. Predicted TFs were integrated with existing protein-protein interaction databases like BioGRID (http://thebiogrid.org/) and tissue-specific protein expression information available for specific TFs (http://www.genecards.org) to delineate the important regulators and the network of interactors controlling the expression of UMOD in kidney.

2.1.2 Materials and Methods

I analyzed the RNA seq data for all available transcripts of uromodulin released in HBM to profile their expression across tissues. Human-UMOD orthologs and their upstream regulatory regions were extracted (fasta sequences) from ENSEMBL. These UMOD sequences from human and its orthologs (8 Primates, 7 Rodents) were taken and executed using MEME-SUITE (an open source hub of bioinformatics tools. Prediction of novel regulatory motifs was performed by using phylogenetic footprinting, an in silico method coupled with downstream computational analysis. Based on this, consensus sequences in upstream region were discovered by MEME analysis. These consensus sequences were further analyzed using TOMTOM tool which enables the comparison of predicted motifs with PWM's of TFs for overlap. Further, protein interactions network was constructed between the potential TFs by utilizing the available physical interactions in BioGRID and tissue-specific protein expression information available for specific TFs (http://www.genecards.org) to delineate the important regulators and the network of interactors controlling the expression of UMOD in kidney. Step wise methodology implemented in the study is described below.

2.1.2.1 UMOD transcripts and their expression profiles across tissues

UMOD gene is located on chromosome 16. I obtained human UMOD gene (Ensembl ID ENSG00000169344) and its sequence from the ENSEMBL database. There are 15 transcripts reported in Ensembl database for human UMOD gene (in Ensembl genome build GRCh37.p7). Full length transcripts of this gene which had expression data available were used for expression profiling. RNA-seq data available for 16 different human tissues (viz. adipose, adrenal, brain, breast, colon, heart, kidney, liver, lung, lymph, muscle, ovary, prostate, testes, thyroid and white blood cells) from ArrayExpress (38) (Accession no. E-MTAB-513) as part of the Human Body Map (HBM) 2.0 project (http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/) (39), was obtained for expression profiling the transcripts of interest. Expression data from the HBM project is quantified per transcript using the current annotations of the human genome from the Ensembl and is available as Reads Per Kilobase per Millions of reads (RPKM) for each sample and hence can be compared across tissues. Expression profiles of UMOD transcripts were visualized using matrix2png (40).

2.1.2.2 Identification of human-UMOD orthologs and their upstream regulatory regions for phylogenetic footprinting

Phylogenetic foot printing is one of the classical methods applied for DNA binding motif discovery (37, 41). It involves using the upstream regulatory sequence of a gene of interest across possible orthologs to search for highly conserved consensus DNA binding sites. I selected 15 orthologs of human UMOD gene including eight primates and seven rodents using Ensembl Compara gene trees(42) which allowed the identification of orthologous sequences across species with high sequence resemblance shown in Table 1.

Cat.	Species	Type	Location	Target %id	Query %id
P	Chimpanzee (*Pan troglodytes*)	1-to-1	16:20210632-20223731:-1	99	49
P	Orangutan (*Pongo abelii*)	1-to-1	16:19749545-19771699:-1	98	98
P	Gorilla (*Gorilla gorilla*)	1-to-1	16:20941424-20961082:-1	94	99
P	Macaque (*Macaca mulatta*)	1-to-1	20:19328933-19349164:-1	94	94
P	Gibbon (*Nomascus leucogenys*)	1-to-1	GL397283.1:7871691-7891404:1	93	98
P	Marmoset (*Callithrix jacchus*)	1-to-1	12:19917641-19961869:-1	89	94
P	Bushbaby (*Otolemur Garnettii*)	1-to-1	GL873563.1:1685537-1696599:1	85	81
R	Squirrel (*Ictidomys tridecemlineatus*)	1-to-1	JH393311.1:5833391-5846021:-1	84	84
R	Tree Shrew (*Tupaia belangeri*)	1-to-1	GeneScaffold_4852:239474-251289:-1	82	83
R	Guinea Pig (*Cavia porcellus*)	1-to-1	scaffold_4:33749021-33758533:1	80	74
R	Rabbit (*Oryctolagus cuniculus*)	1-to-1	6:8027968-8041363:1	78	78
R	Rat (*Rattus norvegicus*)	1-to-1	1:177729221-177742566:-1	77	78
R	Mouse (*Mus musculus*)	1-to-1	7:119462866-119479255:-1	76	76
P	Mouse Lemur (*Microcebus murinus*)	1-to-1	GeneScaffold_865:812142-824452:-1	75	75
R	Kangaroo rat (*Dipodomys ordii*)	1-to-1	GeneScaffold_5176:37235-51592:-1	64	68

Table 1: Human-UMOD orthologs and their upstream regulatory regions for phylogenetic footprinting. Category P=Primates, R=Rodents

Gene expression is controlled by various cis-acting transcriptional regulatory factors by binding mostly in close proximity to the transcription start sites in the promoter regions of a gene (43). Based on previous computational studies from other groups (1, 44) and my analysis (data not shown) I found that most functional TF binding sites occur with-in the 5kb upstream region of the gene starts. Initially, I focused on investigating the 2kb upstream regions of UMOD for motif discovery and later extended to 5kb region. Upstream regulatory regions for human and its 15 listed (Table 1) UMOD orthologs were obtained from Ensembl database.

2.1.2.3 MEME analysis for discovering DNA binding motifs

DNA binding motif discovery using phylogenetic footprinting approach uses regulatory regions in the promoters of orthologous genes from multiple species under the notion that regulatory elements would be conserved in the background of non-functional sequences and hence can be discriminated as footprints contributing to regulatory control. To facilitate the motif finding in these regions, MEME-suite of tools (45) was implemented. MEME is a tool for discovering motifs in a group of related DNA or protein sequences, which detects the frequently occurring conserved sequence across a group of related DNA sequences, using expectation maximization (46). These motifs are typically represented as position-dependent letter-probability matrices in logos which describe the probability of each possible letter at each position in the pattern to incorporate the variation in the detected motif instances across sequences. In this study, both 2kb and 5kb upstream sequences of human UMOD and its 15 orthologs (12 orthologs for 5kb regions due to limitations on the total length of the sequences) were compiled as a fasta file and used as an input for MEME to identify significantly over-represented motifs (p <1E-28).

2.1.2.4 Prediction of TFs associated with discovered motifs

Transcription Factors (TFs) are proteins which bind specifically to their corresponding binding motif and regulate the expression of a gene. DNA binding motifs were represented as PWM (Position-Specific Weight Matrix) based logos. Nucleotide constituent of each consensus motif has its own probability of occurrence within the site. Since PWMs for various TFs have already been reported in JASPAR (47), Uni-PROBE (48) and Jolma et al (4) public databases, based on a comparison of the similarity between the reported PWM of a TF to the footprinted PWM in the orthologous upstream regions, it is possible to predict the TFs which are most likely to bind to these predicted binding sites. Tomtom (49) is a tool in the MEME-suite which compares discovered

DNA motifs to known motifs of such databases. PWMs of various discovered motifs were used as input file for Tomtom and compared with already reported PWMs of TFs from Jolma2013 (4), JASPAR_CORE_2009 (47) and Uniprobe_Mouse(48) databases to identify the potential TFs binding to the UMOD upstream regions. Only the TF associations which are identified at $p \leq 0.02$ are considered significant for both the 2kb and 5kb regions.

2.1.2.5 Analysis of DNase I hypersensitive site in UMOD upstream region

DNase I hypersensitive sites are open chromatin region of DNA, sensitive to DNase I cleavage. After enzymatic cleavage, this site is accessible to binding of protein such as transcription factor. It is believed that, occurrence of DHS in a region, especially in promoter region (50) is an indicator of potential binding of transcription factor. The DHS data available for adult mouse kidney were extracted from ENCODE project (51) and visualized for upstream region of UMOD gene in UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDnase). The image generated from the browser was positioned according to the coordinate of UMOD upstream region of block diagram and studied for active BMo.

2.1.2.6 Calculating motif abundance similarity across genomes

To quantitatively compare the number of instances of a given motif across various genomes, a matrix comprising the number of instances of a motif across the genomes was constructed and then each value was divided by the maximum number of times it was identified in a genome. Such a motif centric normalized matrix was used as input to the cluster algorithm (52) to hierarchically cluster the motifs using uncentered correlation as the distance metric and complete linkage as the clustering method. Resulting data matrix was used to generate a heatmap using the javatree view package (53). To have an identity for each motif, potential TF likely to bind the motif based on Tomtom analysis was used as a reference name, along with the motif ID. Similar approach was adopted to hierarchically cluster the protein-protein interaction network between TFs by constructing a matrix of physical interactions between all pairs of TFs.

2.1.2.7 Mapping protein interactions between the potential TFs

Eukaryotic TFs often regulate the expression of genes by forming protein complexes and several examples have been documented in the literature including that of SP1 interacting with SMAD3 (54), KLF4 (55) and GATA3 (56) in kidney/kidney cell line to modulate the transcription of target genes. To map the physical associations between the predicted TFs from the Tomtom analysis for the 5kb region, manually curated set of protein-protein interactions for the human genome were employed from the BioGRID database (57). This not only allowed the construction of a protein interaction network between the predicted TFs but allowed the dissection of the major TFs based on their number of protein interactions in the network. TFs which had high degree were analyzed for their protein expression across cell types available from gene cards (58).

2.1.3 Results and Discussion

Uromodulin is one of the most abundant proteins in urine. Although a total of 15 transcripts are currently annotated for human UMOD gene in the Ensembl Database (GRCh37.p13), major transcript forms of UMOD are expressed exclusively in the kidney (Figure 2) and hence UMOD is likely to exhibit a specific cis-regulatory signature not prevalent in other non-kidney specific genes.

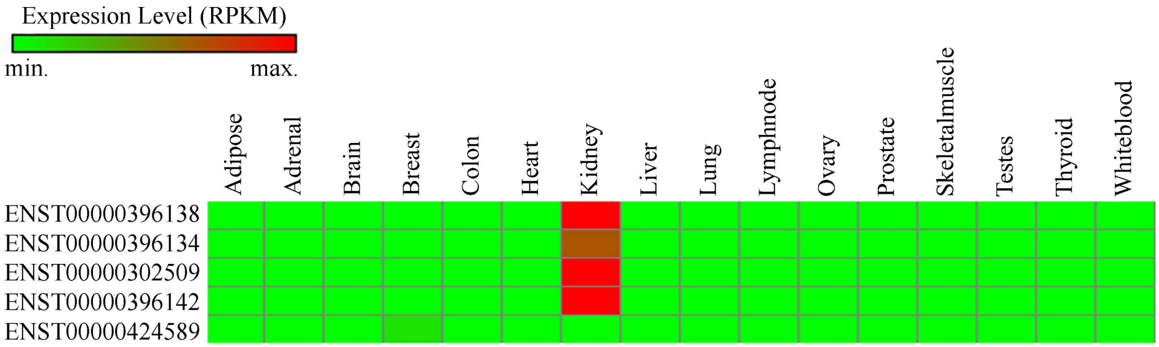


Figure 2: Transcript expression profile of UMOD. Heatmap showing the expression profile of all full length coding transcripts of UMOD gene across the 16 human tissues using the Illumina generated RNA-seq data from the Human Body Map 2.0 Project(38).

Uromodulin is known to be involved in various biological processes like regulation of ion homeostasis, cellular defense response and kidney injury (25, 26). However, little is known about the factors and mechanisms controlling its expression at the transcriptional level. It is also unclear as to their contribution and involvement (i.e. direct or indirect) in kidney disorders (59, 60). This study attempts to identify the cis-regulatory binding sites controlling UMOD and all possible regulatory proteins which may be involved in regulating the expression of UMOD gene at transcriptional level.

2.1.3.1 Identification of potential binding motifs by phylogenetic footprinting the regulatory regions of UMOD across primates and rodents

Since UMOD was found to be tissue-specifically expressed, I postulated that it's cis-regulatory signature might be governed by an intricate interplay between TFs whose network might control its expression in a tissue-specific manner. Hence, to uncover the set of binding sites and the TFs controlling the UMOD gene, I implemented motif discovery based on phylogenetic alignments of orthologous sequences from a diverse set of eight primates and seven rodents using the human UMOD gene as a reference (see Materials and Methods, Table 1). Phylogenetic footprinting is a method for the discovery of regulatory elements in a set of orthologous regulatory regions from multiple species. It does so by identifying the best conserved motifs in those orthologous regions (37). It can be argued, this approach, may miss some of the binding motifs not conserved in UMOD upstream, however I believe this approach as the best fit because of little knowledge about transcription regulator for this gene and also to limit false discovery of motifs and associated TFs.

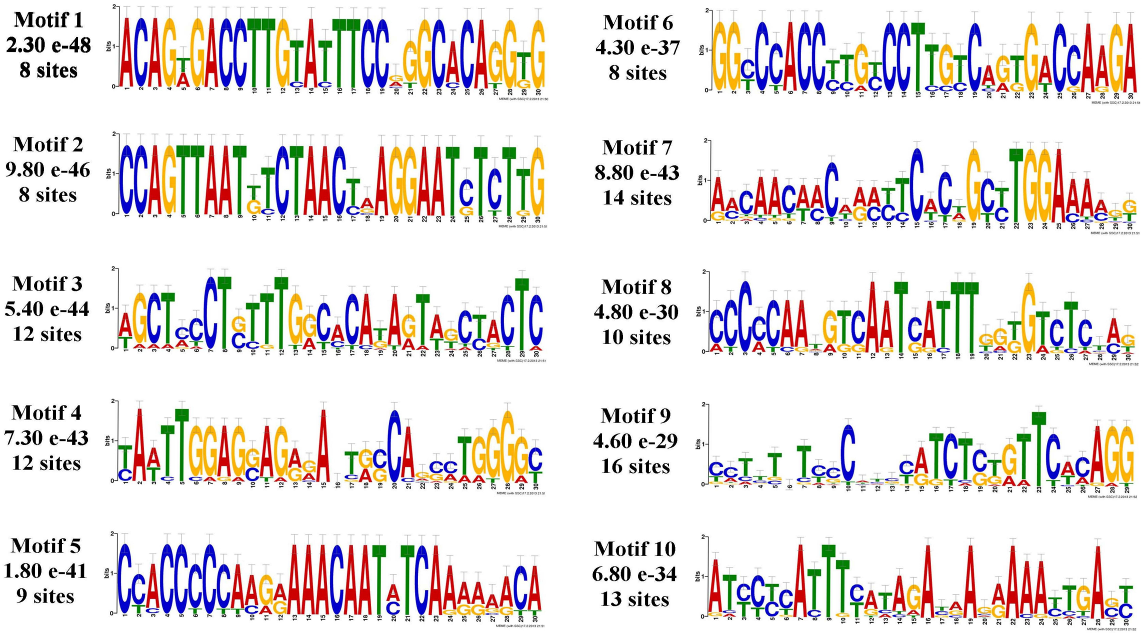


Figure 3: Identification of potential binding motifs by phylogenetic footprinting of 2kbupstream regulatory regions of UMOD gene. Ten phylogenetically conserved and statistically significant (indicated by e-value) novel motifs with the number of sites contributing to their identification were shown for UMOD 2kb upstream. These motifs were displayed as sequence LOGOs representing position weight matrices of each possible letter code occuring at particular position of motif and its height representing the probability of the letter at that position multiplied by the total information content of the stack in bits.

Briefly, 2kb upstream sequence of UMOD gene for human and its orthologs (Table 1) were analyzed by MEME (http://meme.nbcr.net/meme/intro.html) that uses an expectation maximization-based motif-finding algorithm, to identify the potential binding sites conserved across the species. Based on the alignments, position-specific weight matrices (PWMs) (61) representing each of the 10 most significant motifs enriched across the analyzed sequences were identified. Motif logos (62) corresponding to each of these 10 significantly conserved ones along with the number of occurrences of the motifs across the 16 sequences are shown in Figure 3. As evident from Figure 3, I found that all of these motifs exhibited a frequency of at least eight occurrences among the 16 sequences analyzed, with Motifs 7, 9 and 10 exhibiting the highest number of instances.

2.1.3.2. Distribution of binding motifs for UMOD across species

In all cellular systems, DNA-binding transcription factors mediate the activation or repression of gene expression by binding specific regulatory sequences associated with a given target gene. Genes of many eukaryotes display a more complex architecture of associated regulatory elements, which include proximal promoter elements with binding sites for basal transcription factors, and several distal or upstream elements with binding sites for a host of specific transcription factors (63). Several elegant studies on developmentally regulated (64) and immune-response genes (65, 66) have revealed an important role for combinatorial interactions between different transcription factors (TFs) in establishing the complex temporal and spatial patterns of gene expression. Hence, increasing evidence now suggests the importance of not only knowing the binding location of a eukaryotic TF (67) but also the complex combinatorial interplay between them, dominant in eukaryotic transcriptional networks (68). Therefore, I first mapped the locations of the discovered conserved and novel motif sequences across multiple species. These binding motifs were found quite different from each (Pearson correlation coefficient). Relative positions of the discovered binding sites in the 2kb upstream regulatory sequences across the species organized by phylogenetic distance along with the combined significance of motif co-occurrence is shown as a block diagram (Figure 4A). I discovered a total of five binding motifs in Human-UMOD 2kb upstream sequence - most of them dispersed on the chromosome compared to other species (Figure 4A), possibly suggesting evolutionary divergence of the cis-regulatory signature in humans and other close relatives. In particular, I found that as the evolutionary distance of the species with respect to human increased, the extent of conservation and clustering of the binding sites increased, indicating either the gain of the binding site clusters or the lack of these signals in the 2kb region of some of these species. The results suggest the possibility of altered wiring of the transcriptional regulatory network controlling UMOD gene across primates and rodents either due to its altered functionality in kidney or due to increased complexity of the genome in some cases. To further address the functional importance of each of the motifs conserved in humans compared to those which were identified in other species but not in humans, the functional enrichment analysis of the genes (in the whole genome) containing each of these motifs were performed using GOMO (69) (Table 2).

This analysis indicated that Motifs 2 and 6 associated with the functional theme G-protein coupled receptors and associated signaling were not found in humans while the Motifs 3, 9 and 10 discovered in humans and are well conserved were found to be enriched in genes annotated with signaling, translational control and immune response associated processes, suggesting that post-transcriptional regulatory control and immune response related binding motifs might be highly preserved across the species in UMOD upstream regions.

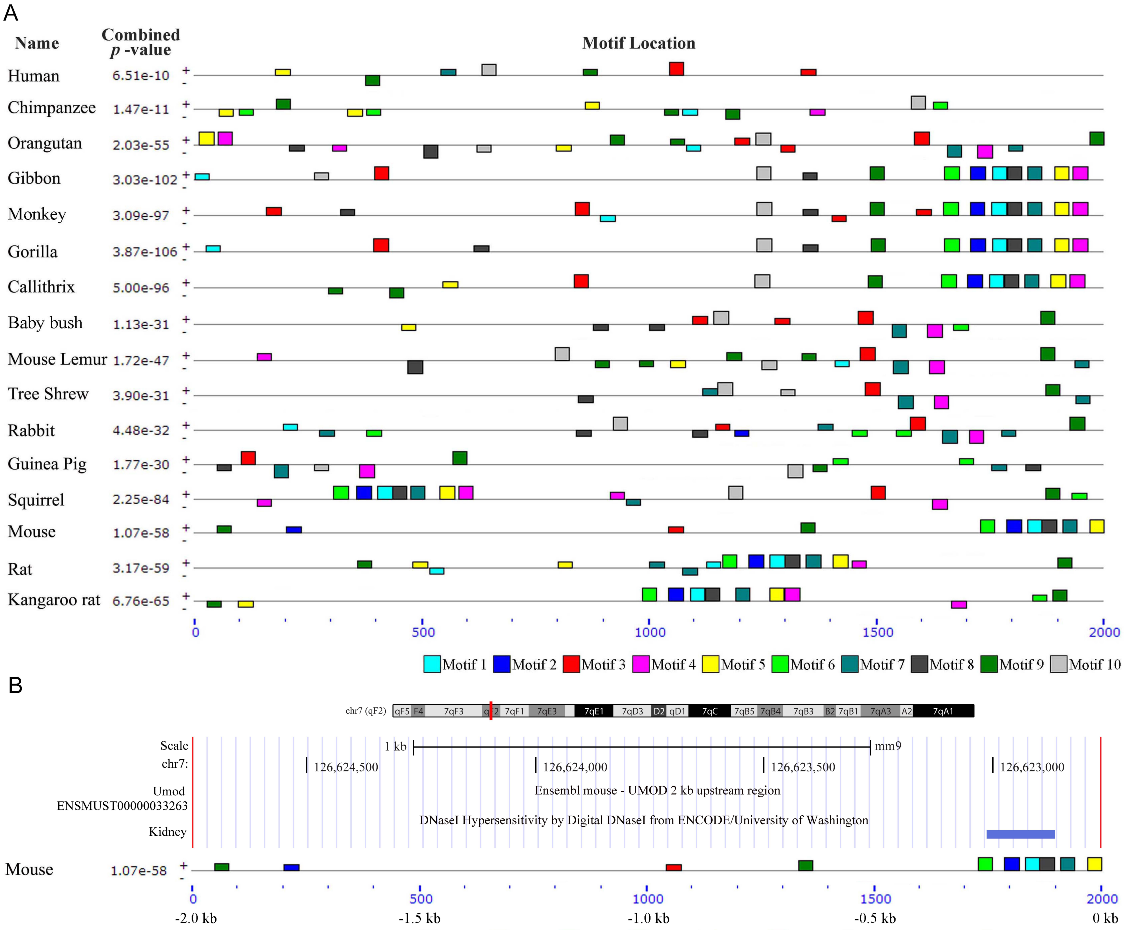


Figure 4: Block diagram showing occurrence of conserved motifs. (A) Location of ten motifs identified and their distribution in 2 kb upstream sequences across human-UMOD & its 15 other primate/rodent orthologous species are shown in the block diagram. The combined best matches of a sequence to a group of motifs were shown by combined p value. Sequence strand specified as "+" (input sequence was read from left to right) and "-" (input sequence was read on its complementary strand from right to left) with respect to the occurrence of motifs. Coordinates of each motif across species is shown as a sequence scale below the diagram. (B) DNase I hypersensitive region was shown in 2kb upstream region of mouse UMOD using ENCODE project coupled with UCSC browser visualization tool. An overlap of DHS signal was found and shown as blue band over motif 1, 2, 6 and 8 near ~0.25 kb UMOD transcription start site (TSS) in the mouse block diagram.

Motif ID	Top 5 specific predictions				
1	NA	NA	NA	NA	NA
2	MF olfactory receptor activity	BP sensory perception of smell	BP G-protein coupled receptor protein signaling pathway	CC extracellular region	BP defense response to bacterium
3	MF olfactory receptor activity	BP sensory perception of smell	BP G-protein coupled receptor protein signaling pathway	BP immune response	BP inflammatory response
4	NA	NA	NA	NA	NA
5	NA	NA	NA	NA	NA
6	CC cytoplasm	CC intracellular organelle	BP heart contraction		
7	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA
9	MF structural constituent of ribosome	MF RNA binding	CC cytosolic small ribosomal subunit	BP translational elongation	
10	MF olfactory receptor activity	BP sensory perception of smell	BP G-protein coupled receptor protein signaling pathway	BP immune response	

Table 2: GOMO analysis of discovered enriched motifs in 2 kb upstream region of UMOD gene showing the over-represented Gene Ontology annotations. MF- Molecular Function, BP- Biological Process and CC- Cellular Component.

2.1.3.3 DHS profile confirming the feasibility of predicted binding motifs

DNase I hypersensitive sites (DHSs) are DNAse I enzyme sensitive region of chromatin, where chromatin has lost its condensed structure due to cleavage and get accessible to binding proteins such as TFs (3). I used the DHS data available for adult mouse kidney, generated from University of Washington and consigned in ENCODE project (70). This analysis strongly suggested motif 1, 2, 6 and 8 in 2 kb upstream region of mouse-UMOD to be active and open for transcription factor binding as shown in Figure 4B. It might be noteworthy to promote the prediction of TF-TF combination likely to be feasible as the cluster of 4 motifs coupled with DHS signal.

2.1.3.4 Prediction of transcriptional regulatory apparatus targeting discovered motifs of UMOD upstream region

In order to further dissect the regulatory factors that bind the discovered novel regulatory protein binding sites by phylogenetic foot-printing analysis, Tomtom (http://meme.nbcr.net/meme/cgi-bin/tomtom.cgi), a motif comparison tool was used, which aligns and compares the already reported PWMs for well-studied TFs available from motif databases with the discovered motifs (see Materials and Methods). All possible TFs predicted to significantly bind to the discovered motifs were shown in Figure 5A. Best predicted and highly aligned PWMs of TFs for all 10 motifs include GATA 3, HNF 1, SP1, SMAD3 and STAT3. In addition to these key TFs, several significant and reliable list of transcription factors were identified that could potentially bind to these 10 discovered motifs of UMOD (Figure 5A). Figure 5B shows a subset of those TFs which exhibited statistically significant alignment with the discovered motifs or those with literature evidence in support of their functional role in controlling biological processes that might be directly or indirectly associated to UMOD expression in normal/diseased state. Some of the listed TFs have no functional evidence in human but have been documented in other species.

Motif	Consensus sequence	Transcription factors
1	ACAGAGACCTTGTATTTCCGGGCACAGGTG	ELF3, sna, EHF, STAT1, ESRRA, ETV2
2	CCAGTTAATGTCTAACTAAGGAATCTCTTG	HNF1B, Hoxa4, Vsx1, HNF1A, Hoxc5
3	AGCTCCCTCTTTGGCACATAGTAGCTACTC	TBX19, HAP3, NFIC, HAP2, SPIC, Gata3_secondary
4	TAATTGGAGGAGAGAGTGCCAGCCTGGGGC	TFAP2A, MSX2, Tcfap2a, ALX4, opa
5	CCACCCCCAAGAAAACAATATCAAAAAACA	Sox5, SRY, Klf4, Zfp740, foxj3, SP3, SP1
6	GGCCCACCTTGCCCTTGTCAGTGACCAAGA	ESR2, Gata5_primary, AR, Klf7_primary, TP53
7	AACAACAACAAACTCACAGCTTGGAAAAGG	NFATC2, STAT1, STAT3, Foxk1_secondary, Smad3
8	CCCCCAATGTCAATCATTTGGTGTCTCTAG	HAT5, Lef1_secondary, Esrrb, ADR1, NR2E1
9	CCTTTCTCCCATCCATCTCTGTTCACAGG	Su(H), INO4, MEIS2, Sox-4, Gabpa_secondary
10	ATCCCCATTTCATAGACAAGAAAATTGACC	NHP6B, Gata3_primary, Hbp1_secondary, Hoxa10_2318.1, Gata1

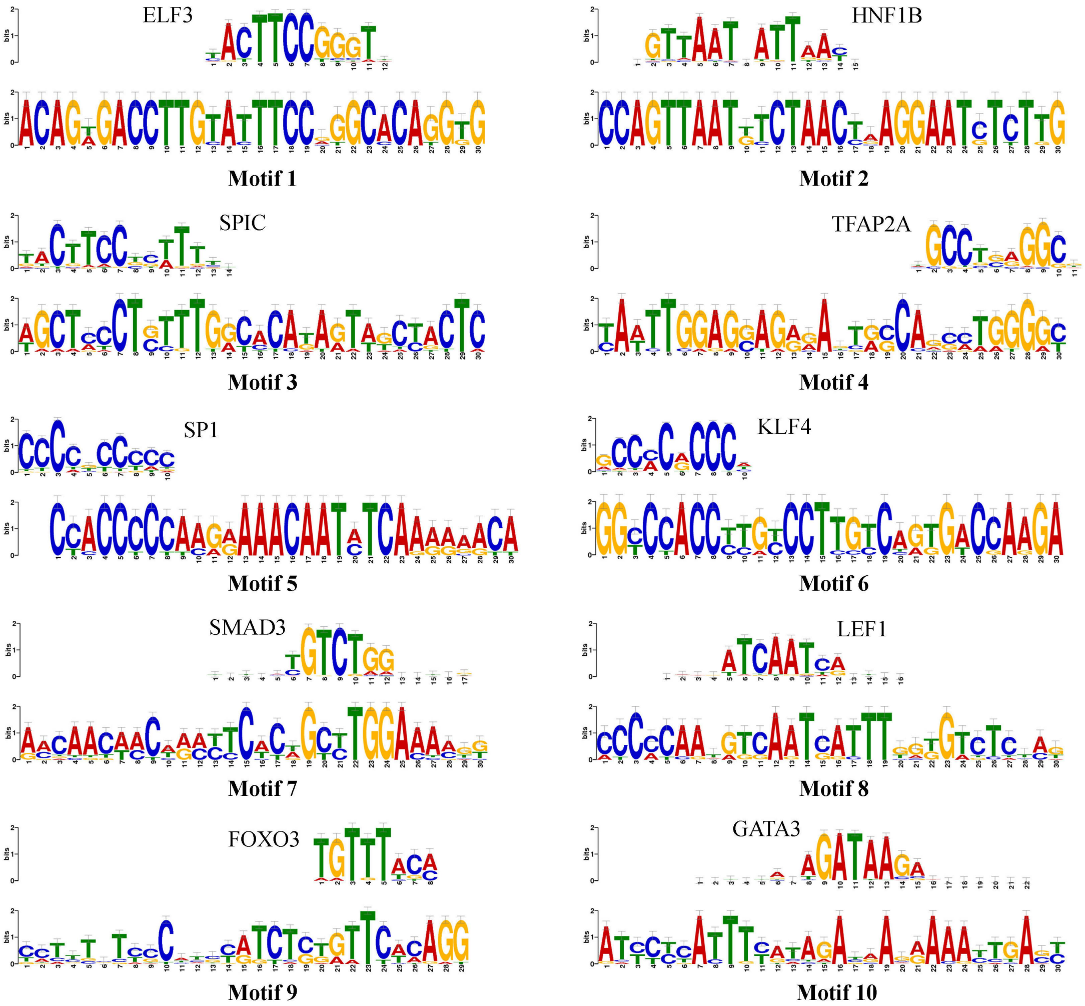


Figure 5: TOMTOM analysis results for conserved motifs. (A) Transcription factors predicted for 10 consensus sequences (as query motif) by TOMTOM analysis (B) Selected set of motif alignments for each of the 10 significant motifs with the matched TF's PWM (top) and query motif (bottom). Binding specificity of TF (1st mammalian TF hit found) was shown for all 10 regulatory motifs.

2.1.3.5 An investigation of all possible regulatory motifs and associated TFs beyond 2 kb upstream sequence of UMOD

Transcription factors' binding specificity depends on the similarity of a target motif sequence to its consensus (61, 67). However, the transcriptional regulatory region of a particular gene may not be restricted to the immediate 2kb upstream sequence but rather its cis-regulatory code might be embedded in regions much further upstream (3). Indeed, several genomic analysis suggests that majority of the TF binding sites occur beyond the 2kb region of the gene start (3, 71) with most of them acting as proximal binding sites with respect to TSS (Transcription Start Site) to define the assembly of transcription pre-initiation complex or some at distal site of TSS to define the rate of transcription (43). Despite, 2kb upstream region provides significant information and coverage about the regulatory motifs contributing to the transcriptional control, are not sufficient to encompass all possible motifs involved in regulation of UMOD gene expression. Thus, the foot printing analysis was extended to include the 5 kb upstream sequence of UMOD gene from Human and its orthologs to discover an extended set of 20 most significant novel regulatory binding motifs across the species and their corresponding TFs using the MEME suite of tools (See Materials and Methods). Further DHS signal analysis in this region was performed and identified active BMo (72) – 1, 2, 3 and 9 (Figure 6A). It is important to note that the motif IDs for the 2kb region (Figures 3-5) do not necessarily correlate with the motif IDs annotated for the 5kb region as the motif numbering is arbitrary. It is also worth mentioning that some of the motifs not detected in the 2kb region either due to the stringent thresholds or due to their occurrence in fewer species (with in 2kb) might be identified in this extended analysis. Block diagram in Figure 6A shows the distribution of the 20 motifs across 12 different species analyzed. It is evident from the block diagram that the cluster of binding sites formed by the motifs 3, 1, 2, 9, 7, 12 and 4 which is prevalent in most species is nearly absent in the human genome. Several motif clusters such as those surrounded by motif 5 were found conserved and drifted across the region between species.

TF and Transcription Factor Binding Site (TFBS) interaction exist in a co-evolutionary relationship within the eukaryotes (73). So, I wanted to learn if the discovered motifs can be grouped based on their frequency of occurrence across species to identify potential co-regulatory relationships between motifs. Figure 6B shows hierarchical clustering of the normalized motif occurrence profiles by selecting a representative TF for each motif, identified based on the Tomtom analysis (see Materials and Methods). This analysis suggested that developmental factors like FOX family exhibited a similar abundance profile as the HNF family, while the binding sites of TFs, SMAD3 and Gata3 co-occurred across the studied genomes. Overall, this analysis provides higher order evolutionary relationships between motifs across the species based on their abundance, suggesting different motifs might be selectively enriched in various species.

Based on Tomtom analysis, I identified a set of predicted TFs which can bind to these 20 discovered motifs. In order to prioritize these associated TFs and to know potential protein complexes that might be responsible for regulation, I integrated the currently available human protein interaction network available from the BioGRID database (http://theBioGRID.org/) to construct a network of physical associations between TFs predicted to be binding to the UMOD regulatory regions (see Materials and

Methods). This resulted in a network of 64 TFs with 112 associations, with TFs like SP1, SMAD3, TP53, SP3, RXRA, RARA and SPI1 exhibiting high degree of associations (Figure 6C). This TF-TF interaction network was analyzed and hierarchically clustered using Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/) using uncentered correlation as distance metric and complete linkage as clustering method to identify potential protein complexes (Materials and Methods). This resulted in 5 major group of physically associated TFs i.e. KLF4 (ESRRG, EGR1, ESRRA, ESRRB, GATA1, GATA3, NFYA, NR2E1, HNF4A, SMAD3), POU2F1 (CREB1, HOXB13, RARA), RUNX3 (RUNX2, STAT1), SP3 (SP1, ZBTB7B) and HNF1B (HNF1A, STAT3), strongly recommend to consider as functionally important TF complexes which help in deciphering the mechanism of UMOD gene regulation. TFs with such higher degree of association with other TFs in the TF-TF protein interaction network were further investigated in Gene Cards (http://www.genecards.org) database to dissect their protein expression profile across reported tissue/fluid and cell lines (Figure 6D). I found that majority of the TFs including SP1, SMAD3, SP3 and RXRA which had high degree of associations were expressed significantly higher in the HEK293 cells compared to other cell types or body fluids except TFs like STAT3 which were found to be higher expressed across a range of cell types including kidney suggesting that most of these identified TFs are not only active in kidney cells but are likely to form a dense network of physical associations with other TFs to regulate the expression of UMOD gene. For instance, some TFs like RUNX2, Pou2f1 (Oct1) which were not identified in the 2kb analysis, were discovered exclusively in this analysis.

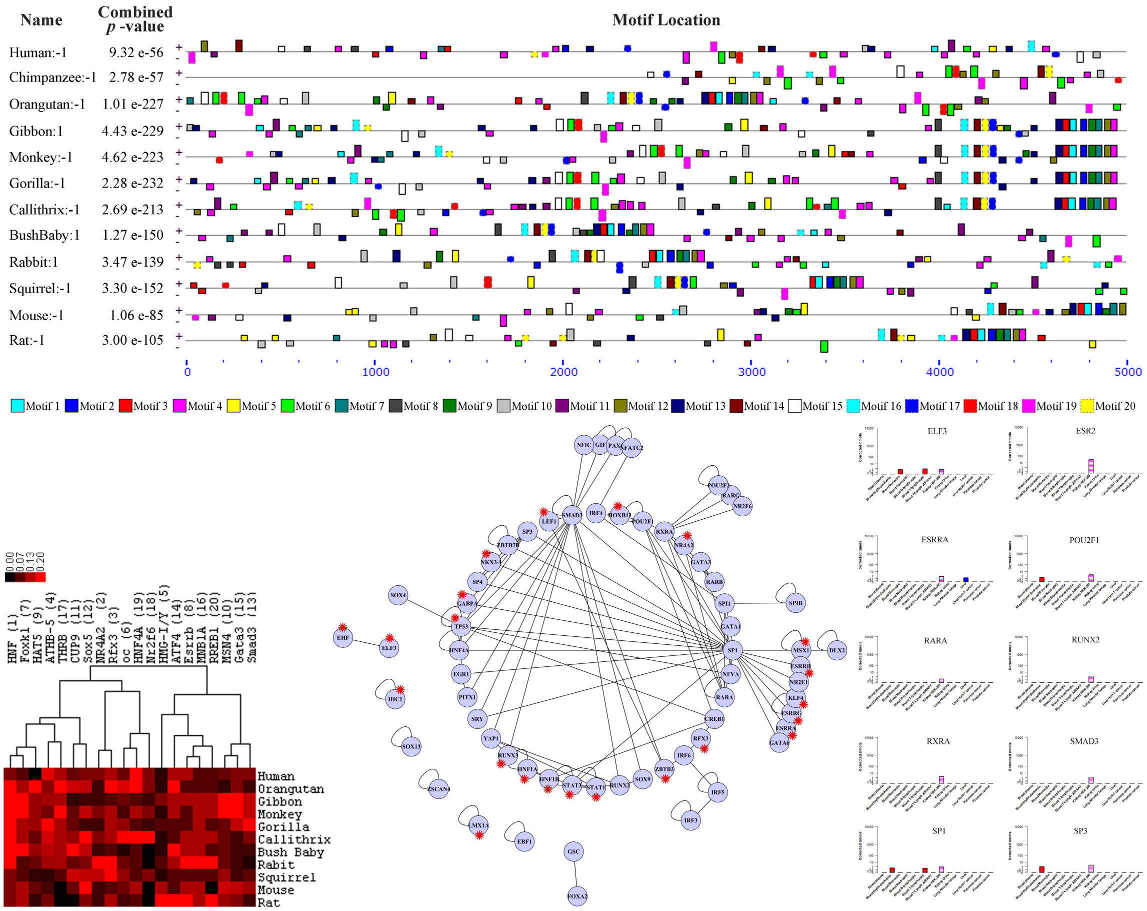


Figure 6: An investigation of all possible regulatory motifs and associated TFs beyond 2 kb upstream sequence of UMOD. (A) Distribution of 20 binding motifs across human-UMOD & its 15 primate, rodent orthologous species were shown as a block diagram for 5 kb upstream region. (B) Occurrence of each motif across the species were grouped and represented as clustered heatmap. A representative TF name is shown for each of the 20 motifs. (C) Protein interaction network between TFs constructed for all possible predicted transcription factors using BioGRID database with TFs belongs to DHS signaled BMo were shown in red asterisk “*”. (D) Protein expression profile of highly associated TFs in the protein interaction network.

2.1.4 Conclusion

In this study, I used a cross-genomic approach to mine the conserved set of binding sites and predicted the associated TFs which are likely responsible for binding these locations to control the expression of UMOD. The current approach not only revealed several novel binding sites to provide insights into their evolutionary dynamics across primates and rodents but also provided a compendium of TFs expressed in the human kidney which are responsible for controlling UMOD's expression thus providing a roadmap for characterizing the regulatory architecture of its promoter regions.

I integrated the predicted list of TFs from the 5kb region of UMOD with publicly available curated set of protein-protein interactions to build a TF-TF protein interaction network responsible for controlling UMOD expression. This study uncovers several highly connected TFs such as SP1, SP3, SMAD3, STAT3 and RARA as well as the likely protein complexes formed between them. The significant expression of these TFs in kidney cells compared to other tissues further suggested their central role in controlling UMOD expression.

2.2 Prediction and validation of transcription factors modulating the expression of sestrin3 gene using an integrated computational and experimental approach

2.2.1 Introduction

Sestrins belong to a small family of evolutionally conserved proteins. They are distinct from any other characterized eukaryotic protein families because they do not have any previously identified domain structures (74). Mammals express three sestrin genes (*SESN*1/2/3), while most invertebrates contain only a single sestrin gene (75). Sestrins do not contain any known structural domains/catalytic motifs; only a partial homologous sequence to bacterial oxidoreductases is identified, suggesting an antioxidant function of this protein (74). Sestrins regulate multiple signaling pathways for metabolic and cellular homeostasis (76). First, sestrins reduce oxidative stress through either their intrinsic oxidoreductase activity or *NRF2* (nuclear factor erythroid derived 2 like 2)-regulated pathway (77, 78). Second, sestrins modulate glucose and lipid metabolism through *AMPK* (AMP-activated protein kinase) and *mTORC1* (mechanistic target of rapamycin complex 1)(74). Third, Sestrins regulate autophagy through activation of AMPK and inhibition of *mTORC1* (75). Deletion of a single *SESN* gene in fruit fly leads to triglyceride accumulation in its body (75), equivalent to the liver in mammals. Several studies shows that ethanol suppresses *SESN3* gene expression and function in hepatocytes and mouse livers. For instance- over expression of *SESN3* dramatically reduces the ethanol-induced hepatic steatosis (79). In addition, *SESN2* and *SESN3* have also been shown to regulate insulin sensitivity and glucose homeostasis (80, 81). However, to date, the factors that control *SESN3* expression are not well studied.

To understand the complex regulatory mechanisms that regulate the *SESN3* is of importance, as new therapeutic targets for metabolic diseases might be discovered.
In this study, I used the upstream regulatory regions of human *SESN3* orthologs from a diverse set of primates and rodents (with at least 85% sequence homology with human) to perform phylogenetic footprinting (37). I employed the MEME-SUITE of tools (45),(82) which allowed the identification of high confidence conserved binding motifs and corresponding position specific weight matrices. The feasibility (i.e. TF binding tendency) of these binding motifs (BMo) were also tested in open chromatin region of human cell lines and mouse liver using DNase Hypersensitive Sites (DHS) in *SESN3* upstream region. Predicted binding motifs were further analyzed by Tomtom (a motif comparison tool from MEME-SUITE) to identify motif specific potential transcription factors. Predicted TFs were integrated with documented protein-protein interaction in BioGRID (83) to decipher the important regulators and the network of interactors controlling the expression of the *SESN3* gene.

2.2.2 Materials and Methods

Human-*SESN3* orthologs and their upstream regulatory regions were extracted (FASTA sequences) from ENSEMBL. These *SESN3* sequences from human and its 10 orthologs (Primates and Rodents) were taken and executed using MEME-SUITE, an open source hub of bioinformatics tools. Prediction of novel regulatory motifs was performed by using phylogenetic footprinting, an in silico method coupled with downstream computational analysis. Based on this, consensus sequences in upstream region were discovered by MEME analysis. These consensus sequences were further analyzed using the Tomtom tool which enables the comparison of predicted motifs with Position Weight

Matrices (PWM) of TFs for overlap. Further, protein-protein interaction network was constructed between the potential TFs by utilizing the available physical interactions in BioGRID to delineate the important regulators and the network of interactors controlling the expression of *SESN3* gene.

2.2.2.1 Sestrin3 transcripts and their expression profile

Human *SESN3* gene is located on chromosome 11. I obtained the DNA sequences for the human *SESN3* gene (Ensembl ID ENSG00000149212) from ENSEMBL database. There are 5 transcripts reported for the human *SESN3* gene, of which 4 have been reported to be protein coding.

2.2.2.2 Identification of human SESN3 orthologs and their upstream regulatory regions for phylogenetic footprinting

Phylogenetic footprinting is one of the classical methods applied for DNA binding motif discovery (37, 41, 72). It involves the upstream regulatory sequence of a gene of interest across possible orthologs to search for highly conserved consensus DNA binding sites. I selected orthologs of the human *SESN3* gene from primates and rodents using Ensembl Compara gene trees (42). These datasets allow the identification of orthologous sequences across species with high sequence resemblance as shown in Table 1. Gene expression is controlled by various cis-acting transcriptional regulatory factors by binding mostly in close proximity to the transcription start sites (TSS) in the promoter regions of a gene (43). Based on previous studies (1, 44, 72), I found that most functional TF binding sites occur within the 5kb upstream region of the gene TSS (data not shown). So, I focused mainly on *5kb upstream regions* of the *SESN3* gene for motif discovery.

2.2.2.3 MEME analysis for discovering DNA binding motifs

DNA binding motif discovery using the *in silico* phylogenetic footprinting approach covered regulatory regions in the promoters of orthologous genes from multiple species (72). This is under the notion that regulatory elements would be conserved in the background of non-functional sequences and hence could be discriminated as footprints contributing to regulatory control. In this study, I used 5kb upstream sequences of human *SESN3* and its 11 orthologs compiled as a FASTA file and used as an input data for MEME(45) to identify significantly over-represented motifs (E-value < e^{-34}). Here I limited the width of discovered binding motifs in MEME analysis to reflect the widths of most established PWMs - which typically vary in length between 4bp to 30bp (47).

2.2.2.4 Prediction of TFs associated with discovered motifs

I used a set of 2201 DNA motifs ranging between 4bp and 30bp in length (average length 12.7) from TRANSFAC, 843 DNA motifs ranging between 7bp and 23bp in length (average length 12.7) in Jolma et al and 979 DNA motifs ranging between 5bp and 30bp in length (average length 13.0) in JASPAR CORE and UniPROBE Mouse. I rationalized that a motif length between 4bp to 30bp for the discovered motifs, would be able to capture most of these recognition sequences in the *SESN3* upstream regions. PWMs of various discovered motifs were used as input file for Tomtom(49) and compared with already known PWMs of TFs in the above described databases to identify the potential TFs binding to the *SESN3* upstream regions. Only the TF associations which are identified at $p \leq 1e^{-03}$ with E-value < 10 were considered as statistically significant for the 5kb upstream regions.

2.2.2.5 Analysis of DNase I hypersensitive site in SESN3 upstream region

DNase I hypersensitive sites (DHS) are open chromatin region of DNA, sensitive to DNase I cleavage. It is believed that, the occurrence of DHS, notably in the promoter region (84) is an indicator of potential binding site for transcription factor. I extracted the available DHS data in various human cell lines and mouse (14.5 days and 8 week) liver from ENCODE project (51) and visualized them for upstream regions of *SESN3* genes in UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeUwDnase). The images generated from the browser were positioned according to the coordinate of the *SESN3* upstream region of block diagram and studied for active BMo.

2.2.2.6 Experimental validation of potential transcription factors

Human HEK293 cells were transfected with plasmid DNAs carrying coding sequences for control GFP (green fluorescent protein), human *FOXO3* and *SOX2* genes. The constructs also contained FLAG tag sequence on the N-terminus. After 48 hours of transfection, cells were processed for chromatin immunoprecipitation (ChIP) analysis for the predicted TF binding sequences as previously described (85). The sequences for the PCR primers are: *FOXO3* ChIP forward primer 5'-ACAAATCCTGGTACGCTGGA-3', reverse primer 5'– CAGGACTGTGCATTATGACATCA – 3'; *SOX2* ChIP forward primer 5'– CCAGTAGGCGATGCAAGTTA – 3', and reverse primer 5'– CTAGACGCCCGCAACCTG – 3'.

2.2.2.7 CRISPR/Cas9 gene knockout experiment

Human *FOXO3* and *SOX2* CRISPR/Cas9 single guide RNA (sgRNA) sequences were designed using an online program at crispr.mit.edu (Dr. Feng Zhang lab) for gene knockout. The selected two sgRNA sequences for the human *FOXO3* and *SOX2* genes are: 5'-CACTTCGAGCGGAGAGAGCG-3' (*FOXO3* sgRNA1), 5'-TCCACTTCGAGCGGAGAGAG-3' (*FOXO3* sgRNA2), 5'-TGGGCCGCTTGACGCGGTCC-3' (*SOX2* sgRNA1), and 5'-ATGGGCCGCTTGACGCGGTC-3' (*SOX2* sgRNA2). The DNA oligonucleotides were cloned into a lentiCRISPRv2 vector (a gift from Dr. Feng Zhang, Addgene plasmid #52961) as described previously (86, 87). To generate gene knockout stable cell lines, HEK293T cells were transfected with control *GFP*, *FOXO3*, or *SOX2* sgRNA plasmids. The transfected cells were selected using puromycin (1 μg/ml) for 7 days, and then maintained in the culture medium containing 0.5 μg/ml puromycin.

DNA constructs preparation

The coding sequences for *GFP*, human *FOXO3*, and *SOX2* genes were cloned into a pcDNA3 vector using PCR amplification and restriction digestion.

Cell culture and transfection

Human HEK293T and HepG2 cells were cultured in DMEM/high glucose medium containing 10% FBS. HEK293T cells were transfected with plasmid DNA using polyethylenimine and HepG2 cells were transfected using TurboFect reagent (Thermo Fisher Scientific).

mRNA analysis

Total RNAs were isolated from cultured cells using TRI Reagent (Sigma). mRNA levels for selected genes were analyzed by real-time PCR. Peptidylprolyl isomerase A (*PPIA*) was chosen as an internal control gene. Primer sequences are listed as follows: human *PPIA* forward primer: 5'- AGGTCCCAAAGACAGCAGAA-3', human *PPIA* reverse primer: 5’-GAAGTCACCACCCTGACACA-3’, human *SESN3* forward primer: 5’-GTACCAACTGCCGGAAAGTG-3’, and human *SESN3* reverse primer: 5’-CCACTGTGTTTGCTTGGACA.

2.2.2.8 Mapping protein interactions between the potential TFs

Eukaryotic TFs often regulate the expression of genes by forming protein complexes and several examples have been documented in the literature including that of *FOXOs* interacting with *SMAD3* (88) , *HNF4a* (89) etc to modulate the transcription of their target genes. Manually curated set of protein-protein interactions from the BioGRID database (57) were employed to map the physical associations between the predicted TFs from the Tomtom analysis for the 5kb upstream region. This not only allowed the construction of a protein interaction network between the predicted TFs but allowed the dissection of the major TFs based on their number of protein interactions in the network.

2.2.3 Results and Discussion

SESN3 has similar pattern of expression (RNA seq based expression GeneCards (90)) across most of the body fluids like blood, liver secretome, and multiple tissue types, indicating the consistent and universal transcriptional regulation of this gene. However, little is known about the factors and mechanisms controlling its expression. This study attempts to identify the cis-regulatory binding sites controlling *SESN3* and all possible regulatory proteins which may be involved in regulating the expression of *SESN3* gene at transcriptional level.

2.2.3.1 Identification of potential binding motifs by in silico phylogenetic footprinting in the regulatory regions of SESN3 across primates and rodents

Phylogenetic footprinting analysis facilitates the search for regions of conserved chromosomal fragments where the likelihood of transcription factor binding is high. These protein-binding sites, which are short fragments of DNA, often range from 6–30 bp in length (72), (91). A robust set of binding sites and corresponding TFs controlling the *SESN3* gene were identified by performing motif discovery based on phylogenetic alignments of orthologous sequences from a diverse set of primates and rodents using the human *SESN3* gene as a reference (see Materials and Methods). *In silico* phylogenetic footprinting (72), was applied for identifying the best conserved motifs in those orthologous regions (37). Briefly, 5kb upstream sequences of *SESN3* gene for human and its orthologs species (Figure 7) were analyzed by MEME(82), to identify the potential binding sites conserved across the species. I used the gene start as the reference to obtain the 5kb upstream. Based on the alignments, PWMs representing each of the 20 most significant BMo enriched across the analyzed sequences were identified. It was observed that most of the established binding motif PWMs in publicly available databases ranges in length between 4 bp to 30 bp (See Materials and Methods) therefore, the discovered motifs in current study would be able to capture most of these recognition sequences, including large co-complex TF binding sites or palindrome motifs, if they are present in the *SESN3* upstream.

2.2.3.2 Distribution of binding motifs for SESN3 across species

Genes of many eukaryotes display a more complex architecture of associated regulatory elements, including cis-promoter elements with binding sites for basal transcription factors, and distal /trans elements with host specific transcription factors binding sites (63). Several elegant studies suggests the importance of not only knowing the binding location of a eukaryotic TF (67) but also the complex combinatorial interplay between them(65-68). Therefore, the identified conserved novel motif sites were firstly screened across multiple speciesas shown in Figure 7. These binding motifs were quite different from each other; as indicated by the Pearson correlation coefficient values (data not shown). Relative positions of the discovered binding sites in 5kb upstream sequences across the species, organized by phylogenetic distance along with the combined significance of motif co-occurrence were shown as a block diagram (Figure 7A). The conservation of motifs was observed high in the region between -1 and -2.5 kb of the *SESN3* gene promoter.

DNase I hypersensitive sites (DHSs) are DNase I enzyme sensitive regions of chromatin, where chromatin has less condensed structure due to chromatin remodeling for facilitating transcriptional activation and other downstream events (3). I used the DHS data available for human cell lines and mouse liver (14.5 days and 8 weeks), generated from University of Washington as part of the ENCODE project (92). This analysis strongly suggested several predicted motifs (Figure 7B-C) in 5 kb upstream region of the *SESN3* genes to be active and open for transcription factor binding, especially within 1 kb of the gene promoter.

2.2.3.3 Prediction and validation of transcriptional regulatory apparatus targeting discovered motifs of SESN3 upstream region

I downloaded the motif databases viz. JASPAR CORE 2014, TRANSFAC, UniPROBE mouse and Jolma 2013 (See Materials and Methods) separately and then combined all together to perform the motif comparison analysis using Tomtom with proper filtering criteria (p-value ≤ $1e^{-03}$ and E-value <10).

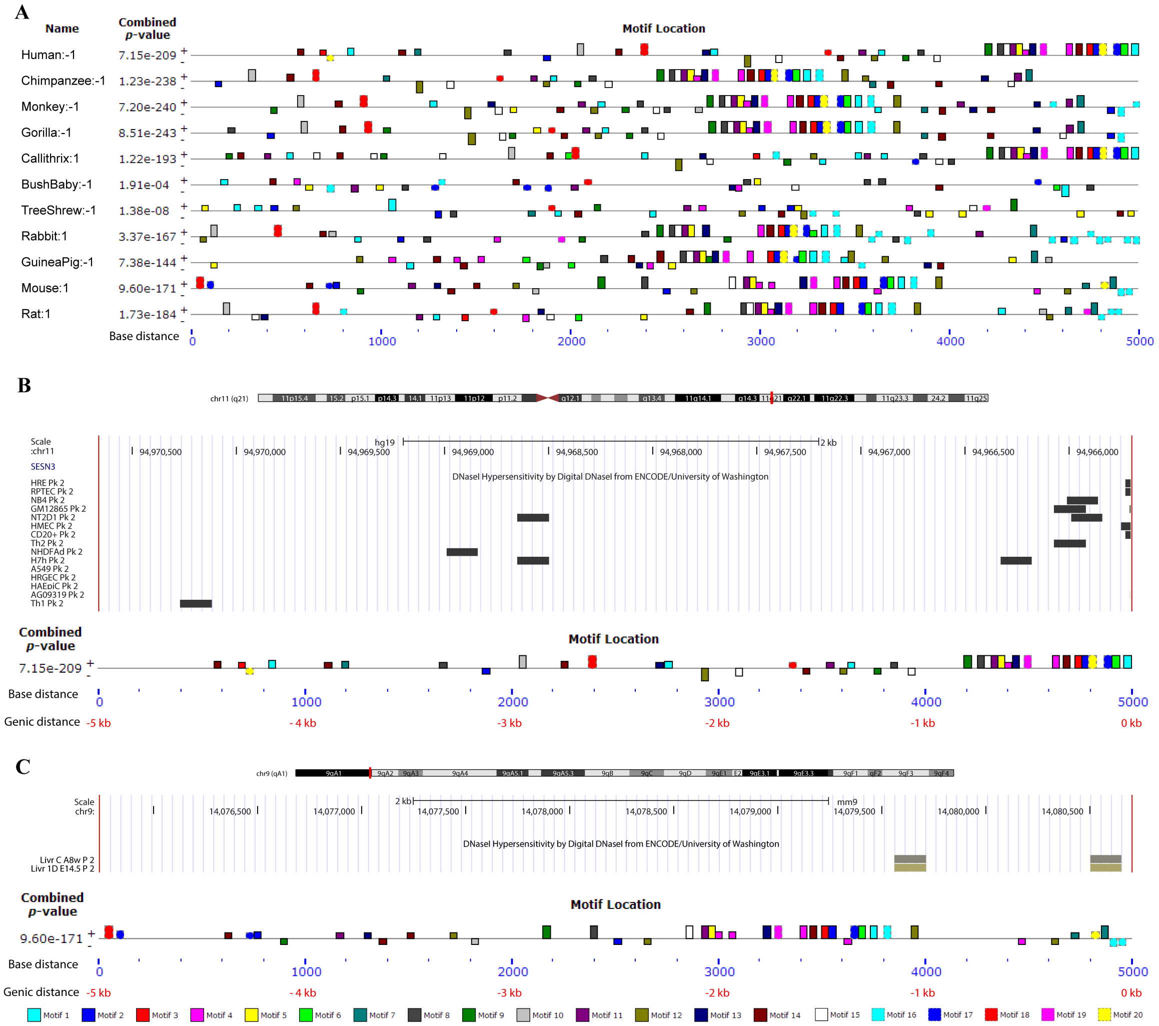


Figure 7: Block diagram showing occurrence of conserved motifs. (A) Location of twenty motifs identified and their distribution in 5 kb upstream sequences across human-*SESN3* & its other primate/rodent orthologous species were shown in the block diagram. The combined best matches of a sequence to a group of motifs were shown by combined p value. Sequence strand specified as "+" (input sequence was read from left to right) and "-" (input sequence was read on its complementary strand from right to left) with respect to the occurrence of motifs. Coordinates of each motif across species is shown as a sequence scale (from left to right, in blue) below the diagram. DNase I hypersensitive region was shown in 5kb upstream region of *SESN3* in (B) human cell lines and (C) mouse liver (8 week adult and 14.5 days embryo) using ENCODE project, represented by UCSC browser visualization tool. An overlap of DHS signal was found and shown as dark band over respective motifs in block diagram. The two coordinates on x-axis represents the *5kb upstream regions* as base distance (in blue) and genic distance (with respect to gene start site, in red) of *SESN3* gene.

High confidence set of TFs predicted to regulate the expression of SESN3 via Tomtom (49) included *FOXOs*, *SMADs*, *SOXs*, *HNF4A*, and *TCFs* (Figure 8A). Binding motifs which corresponded to high confidence TFs overlapping with DHS signals viz. *SOX2* and *FOXO3* were validated using ChIP-PCR approach in HEK293 cells (See Materials and Methods). *SOX2* and *FOXO3* transcription factors were found to exhibit significantly enriched binding to the predicted location in the human *SESN3* promoter region compared to a negative control *GFP* (Green Fluorescent Protein) (Figure 8E-F). Thus, this validation confirms the active BMos discovered for *FOXO3* and *SOX2* in the promoter region of the human *SESN3* gene. To further verify the functional relevance of these TFs in the regulation of the *SESN3* gene, overexpression and knockout of *FOXO3* and *SOX2* were also performed in human cell lines. The overexpression of *FOXO3* or *SOX2* were found significantly activating the *SESN3* gene in human HepG2 cells (Figure 8G). However, they did not significantly affect the *SESN3* gene expression in human HEK293 cells (93) suggesting that there might be cell-type-specific effects. Nevertheless, knockout of either *FOXO3* or *SOX2* downregulated the *SESN3* gene expression (Figure 8H).

There are 5 different isoforms of *SESN3*. Therefore, it is possible to have alternative regulatory elements in the first intron of the gene. In addition to the previous analysis, an *in silico* phylogenetic foot printing with 3 kb upstream and 2 kb instream query sequence of the primates and rodents were also performed for motif discovery and potential TFs binding to these new motifs. The new analysis might not produce the same set of motifs similar to the previously identified consensus sequences because the sequence search spaces are different, however, the motifs which overlap fully or partially with common DHS signals to the previous analysis were projected to produce reproducible results. A set of 20 overrepresented consensus motifs (E-value < e^{-44}) were identified among which, motifs overlapping with the DHS signals, and their corresponding potential binding transcription factors were considered for downstream analysis (93). Nearly, 64% of the previously detected TFs (whose binding motifs were supported with DHS) were reproducibly detected in the new analysis including *SOXs, FOXOs, SMADs, TCFs, HAP1, LEF1, GATA1, POU3F4, POU5F1, EKLF* and *TFAP4*. Hence, inclusion of instream region increased the coverage of predicted TFs in this analysis corresponding to the newly identified motifs.

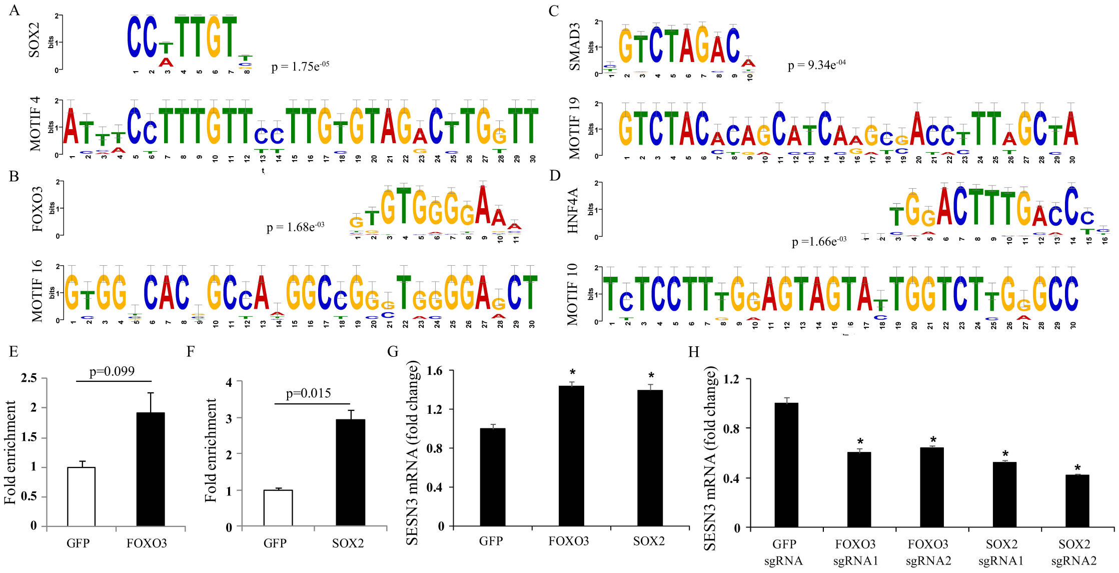


Figure 8: Tomtom analysis results for conserved motifs and experimental validation. (A-D) Transcription factors predicted for 20 consensus sequences (as query motif) by Tomtom analysis. Selected set of DHS overlapped motif aligning with their TF's PWM (top) and query motif (bottom) with binding specificity mentioned by p-values. (E-F) Validation of *FOXO3* and *SOX2* binding to predicted BMo location in *SESN3* upstream region by ChIP analysis. (G) Overexpression of *FOXO3* and *SOX2* activated the *SESN3* gene expression in human HepG2 hepatoma cells. (H) Knockout of *FOXO3* or *SOX2* using CRISPR/Cas9 approach downregulated the *SESN3* gene in human HEK293 cells. (* p<0.05).

Further, in order to prioritize these predicted TFs and to know potential protein complexes that might be responsible for regulation, I integrated the currently available human protein interaction network from the BioGRID (83) to construct a network of physical associations between TFs predicted to be binding to the *SESN3* gene regulatory regions (see Materials and Methods). This resulted in a network of 67 TFs with 125 associations among them, with TFs like *SMAD3, HDAC2, TCF3, SMAD2, CEBPA, SOX2, SMAD1* and *TAL1* exhibiting high degree of associations. Such physically interacting TF-TF network could provide potential co-complex interactions contributing to the regulation of *SESN3* gene. To identify high confidence list of TFs, this network was further dissected to include only the TFs which were predicted to bind the BMos with a high confidence ($p<e^{-03}$) from Tomtom analysis and their corresponding motifs overlapping with DHS signals thereby resulting in a subset of TF-TF interactions which are likely to control *SESN3* promoter. The resulting network of 30 nodes with 60 interactions is shown in Figure 9. The hubs of this TF-TF interaction network included *SMAD3*, *TCF3, SMAD2, HDAC2, SOX2, TAL1* and *TCF12*. *FOXOs*, which have been documented to regulate the *SESN3* gene transcription (94). For instance - Motif 4 identified in this analysis was predicted to be bound by *SMAD3* (Figure 8, $p = 9.34e^{-04}$, E-value = 3.76) and efficient controlled by forming a hub with most other high confident TFs as is evident from the interaction network analysis.

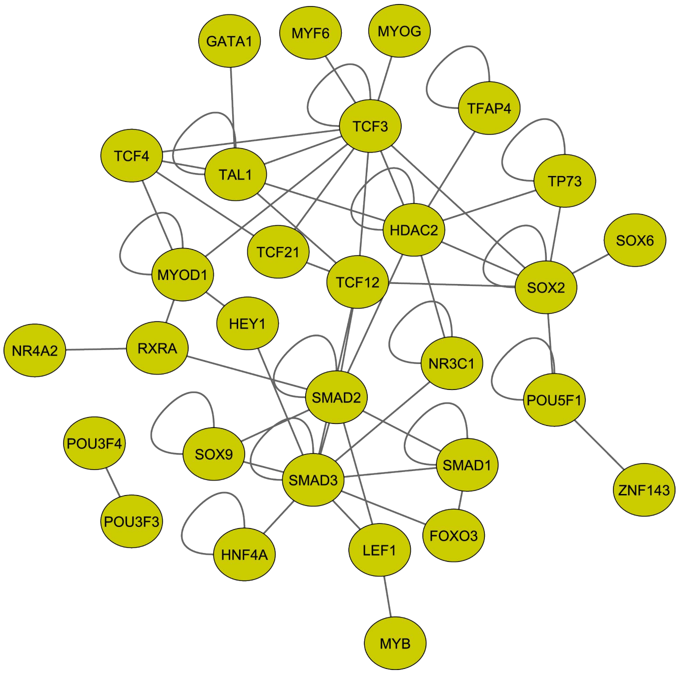


Figure 9: Interaction network of high confidence transcription factors. Protein interaction network between TFs constructed for high confident ($p \leq 0.001$, E-value < 10) transcription factors using BioGRID database with TFs belongs to DHS signal overlapped BMo were shown.

2.2.4 Conclusion

This study was among the first efforts to identify transcription factor binding sites in the *SESN3* gene promoter using an unbiased computational approach. A set of high confidence set of TFs correspond to novel BMos were identified and the hubs of TF-TF interaction network that include *SMADs*, *SOXs* and *TCFs*. *FOXOs* were obtained. These TFs were documented to regulate the *SESN3* gene transcription (94), also found to interact with *SMAD3* in this study, suggesting their interplay to combinatorically control *SESN3*. Some of them including *FOXO3* and *SOX2* were experimentally validated for their binding affinity in identified BMos using ChIP-PCR technique. These findings can form a roadmap to further understand the regulation mechanism of the *SESN3* gene.

CHAPTER 3

TO INVESTIGATE THE COMPLETE TRANSCRIPTOME ARCHITECTURE OF DEVELOPING MOUSE EYE

3.1 Express: A database of transcriptome profiles encompassing known and novel transcripts across multiple development stages in eye tissues

3.1.1 Introduction

The eye is a complex sensory organ that consists of an anterior segment that comprises of the cornea, iris, lens, ciliary body and anterior sclera, and a posterior segment that comprises of the retina, choroid and the optic nerve. Eye development is coordinated by a complex regulatory program that involves a myriad of signaling, transcriptional and post-transcriptional events (95-98). With the advancement of sequencing technologies (such as Next Generation Sequencing (NGS)) and its broad application on a genome wide scale (99-101), it is possible to explore the mechanisms governing the developmental "oculome" (96). Indeed, over the past decade, several studies have reported on the transcriptome of specific eye tissues at various development stages (96, 102-106).

Transcriptome studies reported on the lens for various developmental stages (105, 107) and retina (108-114) were mostly limited to comparative gene expression analysis, by restricting to known or annotated genes. However, the complete transcriptome and various isoforms in the context of developmental stages in tissues of eye are not fully characterized. In this study, a comprehensive and user-friendly platform termed "*Express*" was established, which enables the investigation of the transcriptomic profiles in mouse lens and retina tissues across various development stages. *Express* provides a one-stop portal for investigating gene expression at the resolution of individual transcripts encoded by not just the annotated coding and non-coding genes, but importantly also many novel gene loci in the mouse genome. *Express* facilitates this by allowing users to view the transcript level expression profiles of a gene across multiple developmental stages as heatmaps and simultaneously enables the visualization of the genomic location of the transcripts in an embedded genome browser. Users can view and download the various visualizations as well as the underlying data to facilitate rational design of experiments to study transcript structure, expression and splicing alterations across different developmental stages.

3.1.2 Materials and Methods

To obtain a comprehensive understanding of the transcriptome during development in lens and retinal tissues in mouse eye, publicly available RNA-seq datasets corresponding to the raw RNA sequence reads of mouse eye subcomponents were collected from different developmental stages (Appendix 1 and 2). Briefly, these datasets were aligned to the mouse reference genome, quantified for expression levels of known and novel transcripts followed by the normalization of the expression levels across samples. Resulting raw and normalized expression levels were then organized into a database using My Structured Query Language (MySQL). PHP: Hypertext Preprocessor (PHP) backend Application Program Interface (API) helps to query the database and a user-friendly frontend enables the visualization of the query results as heatmap and browser views across development stages as shown in Figure 10.

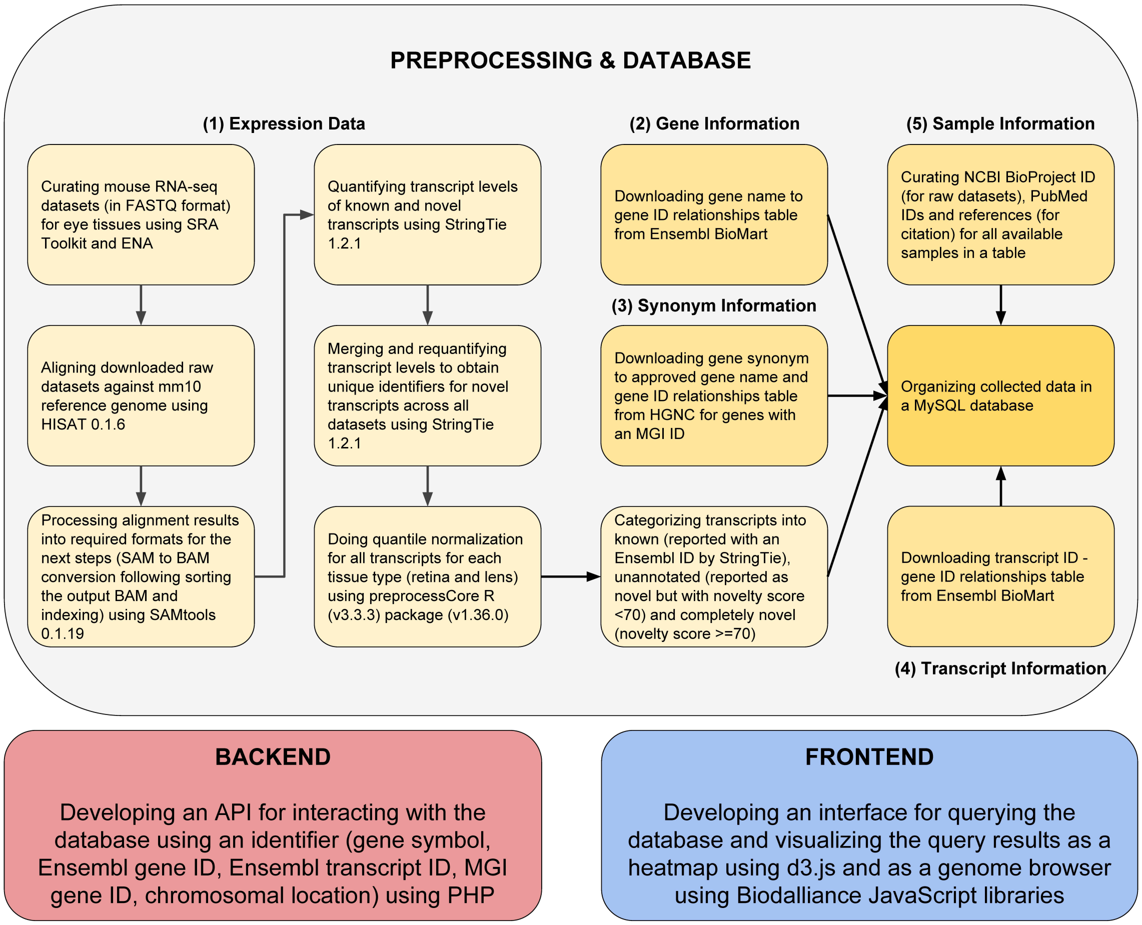

PREPROCESSING & DATABASE
(1) Expression Data
Curating mouse RNA-seq datasets (in FASTQ format) for eye tissues using SRA Toolkit and ENA
Aligning downloaded raw datasets against mm10 reference genome using HISAT 0.1.6
Processing alignment results into required formats for the next steps (SAM to BAM conversion following sorting the output BAM and indexing) using SAMtools 0.1.19
Quantifying transcript levels of known and novel transcripts using StringTie 1.2.1
Merging and requantifying transcript levels to obtain unique identifiers for novel transcripts across all datasets using StringTie 1.2.1
Doing quantile normalization for all transcripts for each tissue type (retina and lens) using preprocessCore R (v3.3.3) package (v1.36.0)
(2) Gene Information
Downloading gene name to gene ID relationships table from Ensembl BioMart
(3) Synonym Information
Downloading gene synonym to approved gene name and gene ID relationships table from HGNC for genes with an MGI ID
Categorizing transcripts into known (reported with an Ensembl ID by StringTie), unannotated (reported as novel but with novelty score <70) and completely novel (novelty score >=70)
(5) Sample Information
Curating NCBI BioProject ID (for raw datasets), PubMed IDs and references (for citation) for all available samples in a table
Organizing collected data in a MySQL database
Downloading transcript ID - gene ID relationships table from Ensembl BioMart
(4) Transcript Information
BACKEND
Developing an API for interacting with the database using an identifier (gene symbol, Ensembl gene ID, Ensembl transcript ID, MGI gene ID, chromosomal location) using PHP
FRONTEND
Developing an interface for querying the database and visualizing the query results as a heatmap using d3.js and as a genome browser using Biodalliance JavaScript libraries

Figure 10: Overview of the transcriptome profiling and database construction for Express. Transcriptomes of mouse lens and retina spanning several development stages (with biological replicates) were collected from published sources listed in Appendix 1 and 2. Curated RNA sequence data was quality filtered using FASTX Toolkit. High quality raw sequence reads were processed and aligned to mouse reference genome mm10 using HISAT and outputs were collected as SAM files. Post-processing (i.e. conversion of SAM to sorted Binary Alignment Map (BAM)) of aligned reads was accomplished using SAMTools. Aligned and post-processed RNA-seq BAM files associated with each developmental stage were utilized for identifying and quantifying the expression levels of known and novel transcripts across respective development stages of tissue subtypes using StringTie. Quantile normalization was performed for samples per tissue type using preprocess R package. The novel transcripts reported by StringTie were categorized into unannotated (novelty score < 70) and completely novel transcripts (novelty score >= 70). These normalized expression levels of known, unannotated and completely novel transcripts were organized into a table. Gene information mapping gene names to gene IDs was downloaded from Ensembl BioMart following Hugo Gene Nomenclature Committee (HGNC). Sample information was manually curated for samples and NCBI BioProject ID, PubMed ID and study reference were obtained per sample. These collected data were then organized into a My Structured Query Language (MySQL) database.

3.1.2.1 Data collection and processing

The raw RNA-seq reads of multiple development stages (each with its biological replicate) of mouse eye were collected from Gene Expression Omnibus (GEO) (115) and European Nucleotide Archive (ENA) (116). Appendix 1 and 2 show the relevant source of the RNA-seq datasets along with several metrics for lens and retina respectively, resulting from the alignment of the reads to the mouse reference genome (mm10). The single end datasets were downloaded in FASTQ format (A text-based format for storing both the nucleotide sequence and its corresponding quality scores) using the Sequence Read Archive (SRA) Toolkit (fastq-dump command), and the paired end datasets were directly downloaded from ENA (European Nucleotide Archive). The quality of the sequence reads were ensured using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) with a minimum of Phred quality score 20 for each sample.

An in-house NGS data analysis pipeline was employed for this study. Briefly, I used Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT, version 0.1.6) (117) for aligning short reads from RNA-seq experiments onto reference genome. HISAT (with default parameters) can rapidly align the quality filtered reads collected from different sources (Appendix 1 and 2) against the mouse reference genome mm10. SAM (Sequence Alignment/Map) files obtained from HISAT were post-processed using SAMtools (version 0.1.19) (118, 119) for converting SAM to BAM (Binary Alignment/Map) followed by sorting the output BAM files, and finally these BAM files were indexed using SAMtools. The sorted BAM files obtained after post-processing were used to quantify the expression levels of known and novel transcripts across development stages.

Transcript quantification and discovery from the aligned RNA-seq datasets was accomplished using StringTie (version 1.2.1) (120). StringTie is a novel network flow algorithm based on a fast and highly efficient assembler to quantify the transcripts of each genomic locus considering all possible multiple splice events. In addition to annotated transcripts, it can also provide the information of possible novel transcripts in each sample. The transcript level expression data for each sample quantified using StringTie were stored as GTFs (Gene Transfer Format files) providing expression levels for both known and novel transcripts. All the GTFs obtained for each sample were grouped and provided as an input for StringTie “merge” mode along with mouse reference genome (mm10) to obtain a reference annotation file (in GTF) including novel transcripts. Next, the reference merged GTF was used in re-running StringTie with the sorted BAM files for the corresponding samples, to obtain GTFs per sample having the same transcript identifier for a given novel transcript across all the samples.

The known transcripts were defined as the transcripts that were annotated as reference mouse transcripts in the Ensembl database (121). In contrast, novel transcripts were defined as the transcripts that were exclusively predicted by StringTie with little or no overlap with existing mouse transcript annotations in mm10. The length of the discovered transcript were examined onto the annotated reference transcript coordinates and a novelty score for each novel transcript was calculated by using the below formula,

$$\text{Novelty Score} = \left(1 - \frac{\text{length of overlapping region}}{\text{full length of novel transcript}}\right) \times 100$$

The novel transcripts having a novelty score (NS) ≥70 were considered as completely novel and the novel transcripts having novelty score <70 were considered as unannotated transcripts. Since sequencing or processing artifacts at various steps of the transcript quantification analysis could potentially contribute to high number of transcript isoforms, the transcripts were classified into three categories namely a) known transcripts annotated in Ensembl database (https://www.ensembl.org/Mus_musculus/Info/Index) b) completely novel transcripts i.e., transcripts which exhibit a novelty score of at least 70 and c) the remaining transcripts were classified as unannotated transcripts and excluded from all the downstream analysis. A quantification matrix was generated for both the lens and retinal transcriptomes with respect to different development stages by extracting the TPM (Transcripts Per Million reads sequenced) values from StringTie outputs.

3.1.2.2 Normalization of transcript expression levels across samples in a tissue

Although RNA-seq samples originating from the same laboratory are unlikely to have significant technical variation among the replicates and developmental stages, there could still be variations arising due to factors like tissue preparation, RNA extraction and sequencing depth differences. In the analyzed datasets for both lens and retina, RNA sequencing datasets originating from multiple labs and protocols were analyzed. Hence, in addition to providing the default option of raw expression levels of a transcript across developmental stages, a widely adopted quantile normalization method was performed using the preprocessCore package (122) in R and the resulting normalized expression data was used for showing the expression heatmaps in *Express*. Quantile normalization is a global adjustment method that assumes the statistical distribution of each sample under study is the same (123). Normalization is achieved by forcing the observed distributions to be the same and the average distribution, obtained by taking the average of each quantile across samples, is used as the reference. Its application on both microarray and RNA sequencing data has consistently shown its superior performance compared to other competing methods (123, 124). Hence, this normalization on RNA-seq expression profile matrices was used across developmental stages in lens and retina respectively. Raw or normalized expression levels of replicates of a developmental stage were averaged for display purposes on *Express.* In addition to the quantile normalization, to exhibit only high quality relevant transcripts, the end user has the option to select only highly expressed transcripts for visualization. This is facilitated by including a selection filter which allows the visualization of the expression levels for only those transcripts of a gene which have at least a certain level of expression observed in at least one of the developmental stages shown.

3.1.2.3 Database construction and implementation

In order to build the *Express* database of transcriptome profiles encompassing known and novel transcripts across multiple development stages in eye tissues in mouse, several steps were employed. These steps are illustrated in Figure 10. Briefly, as described in the above sections, the aligned, quantified and then normalized datasets organized as final matrices for each tissue type were stored into an SQL table. *Express* stores both the raw as well as the quantile normalized expression levels of transcripts in Transcripts Per Million Reads (TPM) units. Moreover, the sample metadata information is manually curated with NCBI BioProject ID, PubMed ID and a reference for citing the corresponding dataset. Also, the table containing Ensembl gene ID, MGI (Mouse Genome Informatics) gene ID and chromosomal location for all genes in mouse genome was downloaded from Ensembl BioMart and the table containing gene synonym, approved gene name and Ensembl gene ID was downloaded from HGNC (HUGO Gene Nomenclature Committee) for genes that are linked to an MGI gene ID. Similarly, the transcript ID - gene ID relationships table were obtained from Ensembl BioMart for linking gene information to the expression data.

3.1.2.4 User interface and access

Backend

A PHP: Hypertext Preprocessor (PHP) Application Programming Interface (API) was developed for interacting with the database using a query (e.g. gene symbol, Ensembl gene ID, MGI gene ID, Ensembl transcript ID or chromosomal location) for the user given TPM cutoff and tissue type. Upon sending the query, TPM cutoff and tissue type to the API, the query type is identified, and the corresponding quantile normalized transcript level expression data is retrieved from the database. Next, the expression values are normalized between 0 and 1 per transcript and the final data is returned in JSON (JavaScript Object Notation) format to be visualized by the frontend. The backend PHP API can also be used programmatically to obtain expression data, which is documented on documentation page of *Express* (http://www.iupui.edu/~sysbio/express/docs.html).

Frontend

The frontend interacts with the user to accept input (a tissue type, a TPM cutoff, value type and a query) and a visualization of the retrieved data from the MySQL database is provided. The structure of queried transcripts in a genome browser developed using Biodalliance JavaScript library (http://www.biodalliance.org) was shown. The mouse transcript structures were obtained from GENCODE version M7 (GRCm38.p4) in

BigBed format (A binary file format, created by conversion from a Browser Extensible Data format file) available on http://www.gencodegenes.org/mouse_biodalliance.html. To this BigBed file, the structures of novel transcripts discovered by StringTie were added from the analysis. The default identifiers obtained from StringTie were renamed to include corresponding tissue type in the identifier for easy understanding in the genome browser. To modify the GENCODE transcript annotation, firstly, the BigBed file was converted into BED file, the structures of novel transcripts were added and then converted back to BigBed format using UCSC utilities (125). Also, the expression data per transcript across multiple developmental stages is shown as a heatmap developed by using d3.js JavaScript library (https://d3js.org). The heatmap is sorted by transcript groups (as introduced in the section “Data collection and preprocessing”) as known transcripts, completely novel transcripts and unannotated transcripts. The transcripts in each group are also sorted by the averaged expression value for all developmental stages for keeping highly expressed transcripts at the top. The front end provides two select boxes for choosing an available TPM cutoff (0, 1, 2, and 5), and the tissue type and a textbox for entering the query. The frontend interface allows the user to choose a minimum expression cutoff for a transcript, which enables the display of only those transcripts resulting from search exhibiting this minimum expression level cutoff in at least one developmental stage. The default cutoff is set to 5 TPM. The value type select box can also be used to query for raw expression values or quantile normalized expression values. After search is performed, the results are shown as a heatmap along with a genome browser to view the transcript structure. The heatmap and browser view can be toggled using the button on the right-hand side of the navigation bar. Also, using the Export dropdown menu, it is possible to export heatmap view and browser view in SVG (Scalable Vector Graphics) format and heatmap data in TSV (Tab Separated Values).

3.1.2.5 Experimental validation of the RNA-seq identified transcripts for lens and retinal expressed genes

The University of Delaware animal facility hosted all the mice used in these experiments, which were performed following the guidelines defined in the Association for Research in Vision and Ophthalmology (ARVO) statement for the use of animals in ophthalmic and vision research. C57Bl/6 mouse lenses were microdissected at three stages, namely, embryonic day (E) 15.5, post-natal day (P) 0 and P10. Retina was dissected from four stages, namely P10, P20, P30 and P48. The day of detection of vaginal plug was defined as E0.5. Each of three biological replicates at E15.5 comprised of six lenses, and at P0 and P10 comprised of two lenses. Each of the biological replicates for retinal expression comprised of 2 retinas from P10, P20, P30 and P48. Total RNA was extracted from lenses using RNeasy Mini kit (Qiagen Inc, Valencia, CA) and cDNA was synthesized using Bio-Rad iScript^{TM} cDNA Synthesis Kit (Bio-Rad Laboratories, Hercules, CA), for use as a template in quantitative PCR (RT-qPCR) analysis. Forward and reverse primers were designed on the longest isoform of the transcript on exonic sequence flanking an intronic region such that the product sizes were < 300 bp. RT-qPCR was performed using Power SYBR Green PCR Master Mix (Invitrogen life technology, Grand Island, NY). Several house-keeping genes namely, *Actb*, *B2m* (Beta 2-microglobulin), and *Hprt* were used for normalization (126-130). Fold-change differences between target gene expression compared to specific housekeeping gene expression was estimated using the ΔΔCt method. The first comparison of gene expression in the ΔΔCt method was performed independently with several housekeeping genes. The second comparison was calculated based on expression at E15.5 (lens samples), and at P10 (retina samples). Statistical significance was calculated using two-way ANOVA as described (131).

3.1.3 Results and Discussion

3.1.3.1 Overview of *Express* database

Express is a database of transcriptome profiles encompassing known and novel transcripts across multiple development stages in mouse eye tissues. Several steps involved in preprocessing, post-processing, quantification and normalization of collected data followed by its organization in *Express* are illustrated in Figure 10 (see Materials and Methods). *Express* contains 81779 distinct transcripts for mouse lens and 178367 distinct transcripts for mouse retinal samples. Novel transcripts are defined as those that are not annotated in the reference genome annotation (see Materials and Methods). The proportions of the known and completely novel transcripts for each developmental stage at 5 TPM threshold in lens and retina are shown in Figure 11A and 11B, respectively. In the following sections, the composition of the datasets and functionality of the database were illustrated as well as the validations of several genes in lens and retinal tissues were presented, to demonstrate the utility of *Express* for studying eye development.

3.1.3.2 Analysis of lens and retinal RNA-seq data for building *Express*

Express contains gene and transcript level expression data obtained from 21 lens and 35 retinal RNA-seq mouse samples as shown in Appendix 1 and 2 respectively. Lens samples include developmental stages from E15 to P9 with alignment rates ranging from 86% to 94% (See Appendix 1). The retinal samples include developmental stages from P2 to P90 with majority of them exhibiting a high overall alignment rate varying from 80% to 97% (See Appendix 2). To control the technical variation in expression levels between samples, a quantile normalization of all the samples was performed in a given tissue type (See Materials and Methods). Both raw as well as normalized expression levels in Transcripts Per Million (TPM) reads sequenced units, are stored in the database and are available to download from the *Express* website.

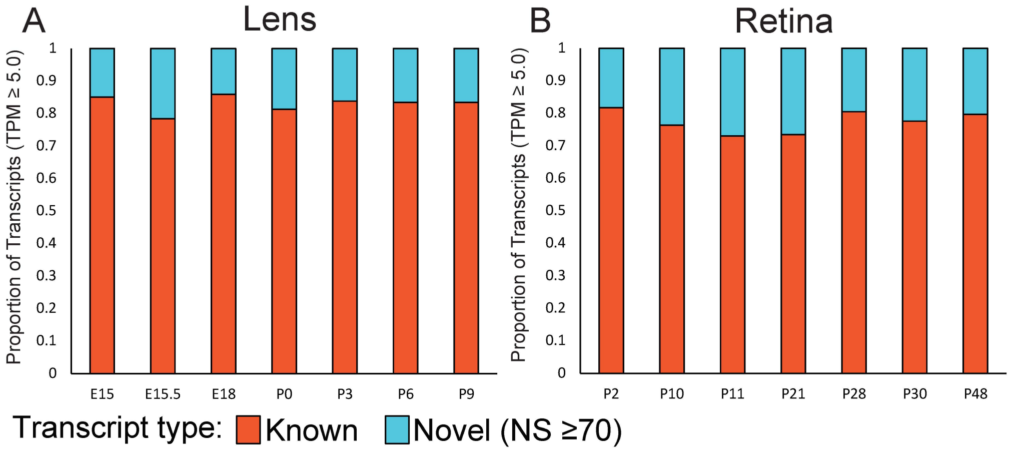


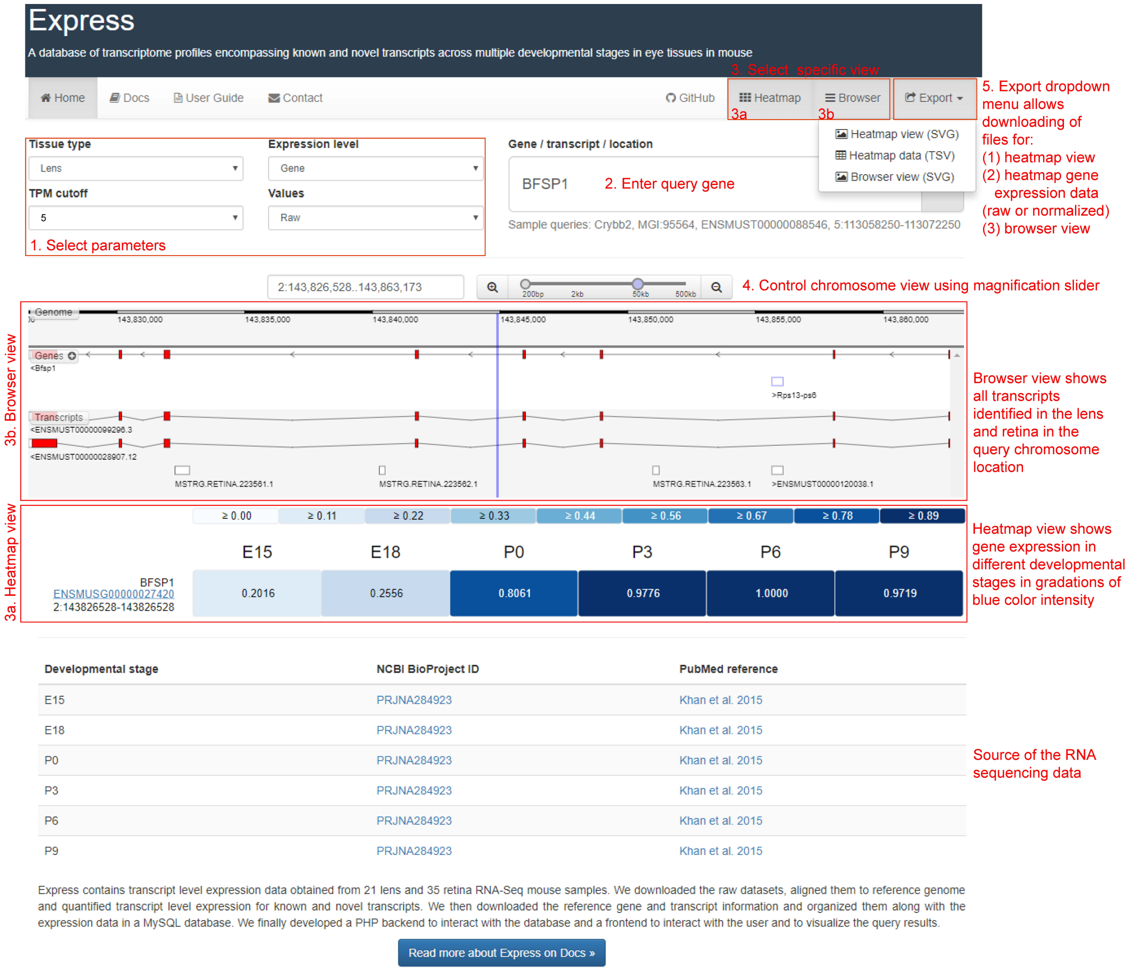
Figure 11: Histograms showing the proportion of known and completely novel transcripts across developmental stages at >5 TPM (A) for lens samples from E15 to P9 and (B) for retinal samples from P2 to P90. Multiple datasets associated with a given developmental stage are merged to facilitate the ease of comparison across stages.

3.1.3.3 User guide for exploring Express database

To retrieve gene expression data from *Express*, the following features have been added to the web interface. Step-by-step instructions for using *Express* are also available as a User Manual (see Figure 12 and the web interface of *Express* following the webpage -http://www.iupui.edu/~sysbio/express/user-guide.html).

The parameters to investigate the expression of a gene are (a) tissue type, namely lens, retina and lens cell subtype; (b) expression level, namely gene or transcript (splice isoform) level; (c) TPM (transcripts per million) cutoff of 0, 1, 2 and 5 and tpm values which could be raw or values after quantile normalization.

User can query a gene name, ENSEMBL ID or chromosome location to investigate gene expression in selected tissue type. The output can be viewed in (a) heatmap or (b) browser view using toggle buttons on the top right side of the web interface. Heatmap view shows gene expression at different developmental stages with color index as gradations of blue color intensity, denoting higher intensity for high gene expression compared to other developmental stages investigated in this study. The browser view shows all the genes and transcripts expressed in lens and retina in the query chromosomal location. Unannotated genes are displayed as *MSTRG.XXXX.XXXXX.X.* The chromosomal window on the browser view can be increased or decreased using the magnification slider provided on the top right of the browser view panel. The heatmap view and the browser view can be downloaded as high-resolution images using a dropdown export menu provided on the top right-hand side of the web interface. Further sources for the RNA seq-data used for the analysis in this study are provided at the bottom of the web interface.



Express

A database of transcriptome profiles encompassing known and novel transcripts across multiple developmental stages in eye tissues in mouse

Tissue type: Lens

Expression level: Gene

TPM cutoff: 5

Values: Raw

Gene / transcript / location: BFSP1

Sample queries: Crybb2, MGI:95564, ENSMUST00000088546, 5:113058250-113072250

1. Select parameters

2. Enter query gene

3. Select specific view

4. Control chromosome view using magnification slider

5. Export dropdown menu allows downloading of files for:
(1) heatmap view
(2) heatmap gene expression data (raw or normalized)
(3) browser view

Browser view shows all transcripts identified in the lens and retina in the query chromosome location

≥ 0.00	≥ 0.11	≥ 0.22	≥ 0.33	≥ 0.44	≥ 0.56	≥ 0.67	≥ 0.78	≥ 0.89

	E15	E18	P0	P3	P6	P9
BFSP1 ENSMUSG00000027420 2:143826528-143826528	0.2016	0.2556	0.8061	0.9776	1.0000	0.9719

Heatmap view shows gene expression in different developmental stages in gradations of blue color intensity

Developmental stage	NCBI BioProject ID	PubMed reference
E15	PRJNA284923	Khan et al. 2015
E18	PRJNA284923	Khan et al. 2015
P0	PRJNA284923	Khan et al. 2015
P3	PRJNA284923	Khan et al. 2015
P6	PRJNA284923	Khan et al. 2015
P9	PRJNA284923	Khan et al. 2015

Source of the RNA sequencing data

Express contains transcript level expression data obtained from 21 lens and 35 retina RNA-Seq mouse samples. We downloaded the raw datasets, aligned them to reference genome and quantified transcript level expression for known and novel transcripts. We then downloaded the reference gene and transcript information and organized them along with the expression data in a MySQL database. We finally developed a PHP backend to interact with the database and a frontend to interact with the user and to visualize the query results.

Read more about Express on Docs »

Figure 12: User guide for employing Express to investigate eye gene expression is highlighted in panel 3A. 1) User selects parameters, 2) Enters query gene or chromosomal region, 3) Selects view options, 3a. with heatmap view can visualize gene expression in various developmental stages, 3b. with Browser view can visualizes different transcripts, 4) Uses magnification slider to controls chromosomal range, 5) Can use the Export dropdown menu to download heatmap view, raw or normalized gene expression data or browser view.

As shown in Figure 12, for instance – on investigating the gene expression profile of *Bfsp1* in the lens at the gene level, at a threshold of 5 TPM for raw expression level, an output is generated with both heatmap and browser view. In the browser view, all genes expressed in the lens and retina at the chromosomal location as *Bfsp1* can be visualized. In the heatmap view, relative expression of *Bfsp1* at various developmental stages is shown. The expression of splice forms of *Bfsp1* can be compared at different developmental stages using the transcript level option i.e. while the expression of isoform ENSMUST00000099296 increases with development, the expression of ENSMUST00000028907 is highest at P0.

3.1.3.4 Development of *Express* as a user-friendly tool

Express provides transcript level expression data for mouse lens and retina across different developmental stages for known and novel transcripts as identified by StringTie. The mouse developmental stages are expressed as embryonic (E) or post-natal (P) followed by a number that indicates the number of days after fertilization or birth, respectively (*e.g.* E18 corresponds to an embryo dissected 18 days after the vaginal plug was observed, while P0 corresponds to the day of birth). A summary of eye developmental stages for ready comparison of ocular morphological changes with *Express* data stages is shown in Figure 13.

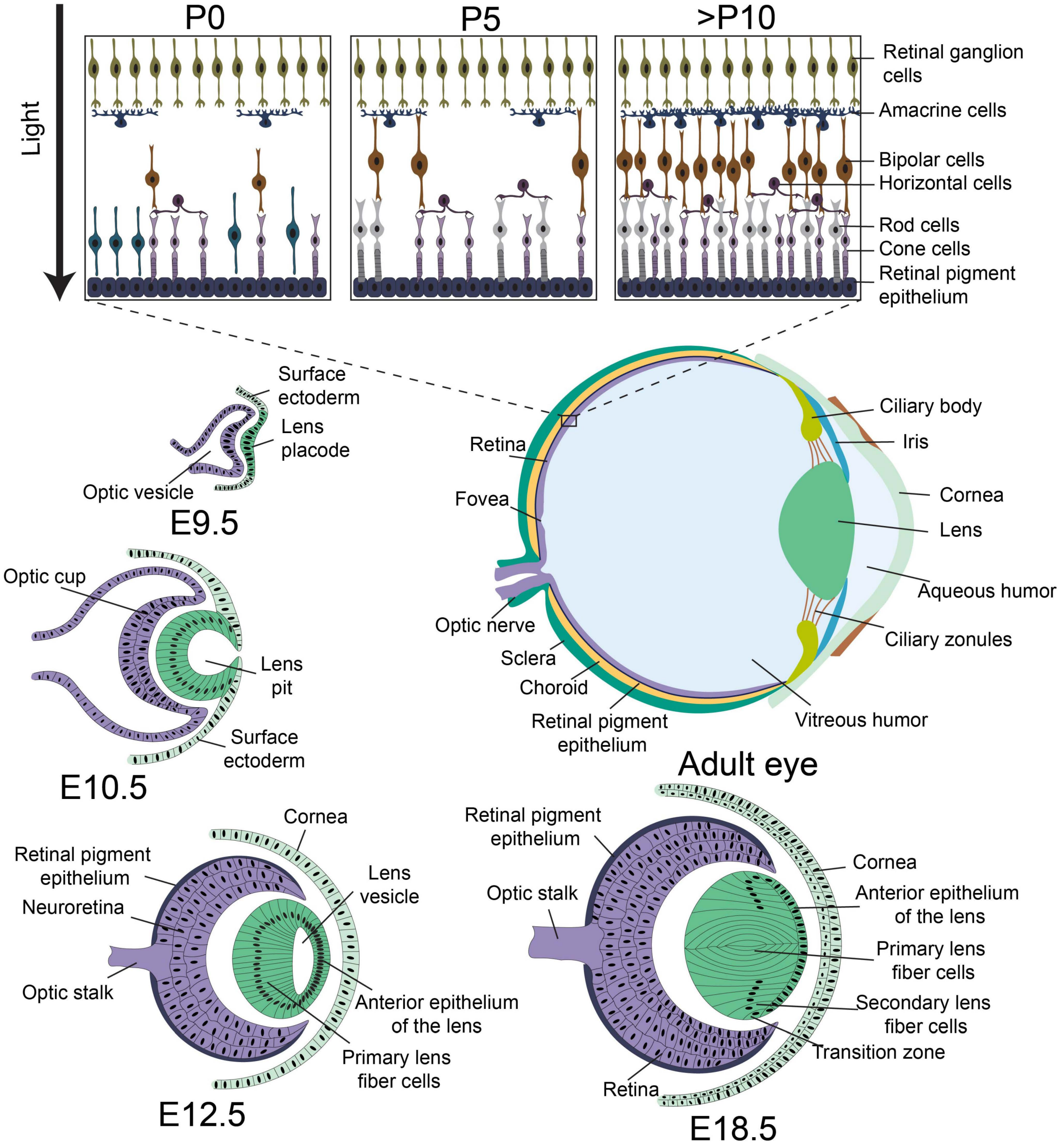


Figure 13: Overview of the mouse eye development and user interface. In the initial stages of eye development, the optic vesicle interacts with the overlying non-neural surface ectoderm at embryonic day (E) 9.5 in mouse and induces its thickening to form the lens placode. Subsequently at E10.5 the optic vesicle and the lens placode interact to develop into the optic cup and the lens pit, respectively. The lens pit closes to the form the lens vesicle and the overlying ectoderm contributes to the corneal epithelium. The posterior cells of the lens vesicle differentiate to form the primary lens fiber cells while cells of the anterior epithelium of the lens divide to form new epithelial cells that migrate towards the transition zone. Cells at the transition zone exit the cell cycle and terminally differentiate to form the secondary fiber cells. Further, the fiber cells migrate towards the center of the lens, and as they terminally differentiate, undergo organelle degradation,

resulting in an organelle free zone in the center of the lens by E18.5. Further development and differentiation events lead to the formation of the adult eye where the anterior region consists of the cornea, iris, cilliary body and cilliary zonules. The posterior of the lens consists of the retina, retinal pigment epithelium, choroid and sclera. A more detailed diagram of the retina shows that it is composed of several distinct cell types, including the retinal ganglion cells, amacrine cells, bipolar cells, horizontal cells and the rod and cone photoreceptors. Retinal ganglion cells and cone cells are differentiated and functional by E18.5 and by postnatal day (P) 5, amacrine cells, bipolar cells, horizontal cells and rod cells are formed. By P10, all the neuronal cells in the retina have completely connected synaptic junctions. Rod and cone cells synapse with horizontal cells for communicating with other photoreceptors and with bipolar cells, which further synapse with amacrine cells. The amacrine cells in turn synapse with the retinal ganglion cells.

Processed RNA-seq data is available for 7 developmental stages of the lens (E15, E15.5, E18, P0, P3, P6, and P9) and 11 development stages of the retina (P2, P10, P11, P21, P28, P30 P40, P48, P50, P60 and P90). In express, users can also search for cell-type specific expression profiles where available. For instance, a representation for lens dataset such as P0:E and P0:F stands for the epithelial and fiber compartments in lens. The fraction of transcripts for each developmental stage for lens and retinal samples is shown in Figure 11. Although majority of the lens developmental stages exhibit ~17% of completely novel transcripts, the proportion of completely novel transcripts in retina were found to be significantly higher and varying with expression threshold. Observed fraction of completely novel transcripts was found to be <25% across majority of the retina stages when transcripts were filtered to include only those expressed greater than 5 TPM in retinal samples (see Figure 11). The number of retina-expressed transcripts that are identified to be novel in this study is comparable to that previously reported in the human retina (132), and therefore supports the finding that retinal cells potentially express a large number of uncharacterized transcripts. In *Express*, users can filter to view only those transcripts resulting from a search that satisfy one of the four levels of confidence in expression levels – 1) transcripts exhibiting a non-zero expression level in TPM in at least one developmental stage, 2) transcripts with at least 1 TPM in at least one developmental stage 3) transcripts with at least 2 TPM in at least one developmental stage and 4) transcripts expressed with at least 5 TPM in at least one developmental stage (default threshold).

At 5 TPM cut-off, the lens samples were found to exhibit ~16% completely novel transcripts across stages. In contrast, the retinal samples were found to comprise of ~22%

completely novel transcripts. When lower expression thresholds were used the fraction of completely novel transcripts significantly increased in retinal samples. It is speculated that the high number of novel transcripts in retinal samples is likely due to the several distinct types of cells in the retina. Indeed, the total number of transcripts identified in mouse retina in this study are very similar to the numbers reported in human retinal samples (132).

3.1.3.5 Validation of transcript-expression in lens and retina

Several genes and their corresponding transcripts that were found to significantly altered across the developmental stages in lens and retina were identified. The expression pattern of these genes as well as other established genes were verified as a representative set of very significantly altering transcripts across stages to evaluate expression levels reported in *Express*. In particular, the expression profile of the selected transcripts (in the form of a heatmap) were downloaded from *Express* for each tissue subtype and their levels were experimentally validated for multiple development stages using RT-qPCR (see Materials and Methods). In lens, the expression of *Pax6*, *Elavl4* and *Rbm5* was validated (Figure 14A). *Pax6* (Paired box 6) is a transcription factor essential for eye development in mice and humans. Mutations in *Pax6* have been linked to congenital cataract, aniridia and anophthalmia in humans (133) and haplo-insuffciency of *Pax6* in mice results in small eyes (Sey) in mice (134, 135). RT-qPCR shows that *Pax6* expression is elevated in early postnatal stages in agreement with *Express* (Figure 14A). *Elavl4* (ELAV (Embryonic Lethal, Abnormal Vision, Drosophila)-like 4 (Hu antigen D) belongs to ELAV protein family and is expressed in the mouse lens and frog retina (136, 137). Elevated expression of *Elavl4* in the mouse lens increases the expression level of its targets (*GAP43* and *CamKIIα*), which is a similar outcome to its overexpression in brain tissue (137). The expression of *Elavl4* in mouse lens was found high during embryonic stages and gradually reduced in postnatal stages (Figure 14A), as predicted by the transcriptome datasets in *Express*. *Rbm5* (RNA binding motif protein 5) belongs to the Rbm protein family and is associated with lung cancer (138-141). *Rbm5* is observed to be highly expressed at embryonic stages and repressed during early postnatal stages (Figure 14A). While this is the first report of *Rbm5* expression in the lens, another member of the Rbm family, *Rbm24* is expressed in the vertebrate eye and its deficiency in zebrafish causes microphthalmia (106, 142).

Further, the expression of *Lhx2*, *Pabpc1, Tia1* and *Tubb2b* were also validated in the retina (Figure 14B). *Lhx2* (LIM homeobox 2) encodes an eye field transcription factor that is expressed from the earliest stages of optic development. *LHX2* mutations in human as well as its knockout in mice causes anophthalmia (143, 144). As indicated by *Express*, RT-qPCR show that *Lhx2* expression reduces in the retina in late postnatal stages (Figure 14B).

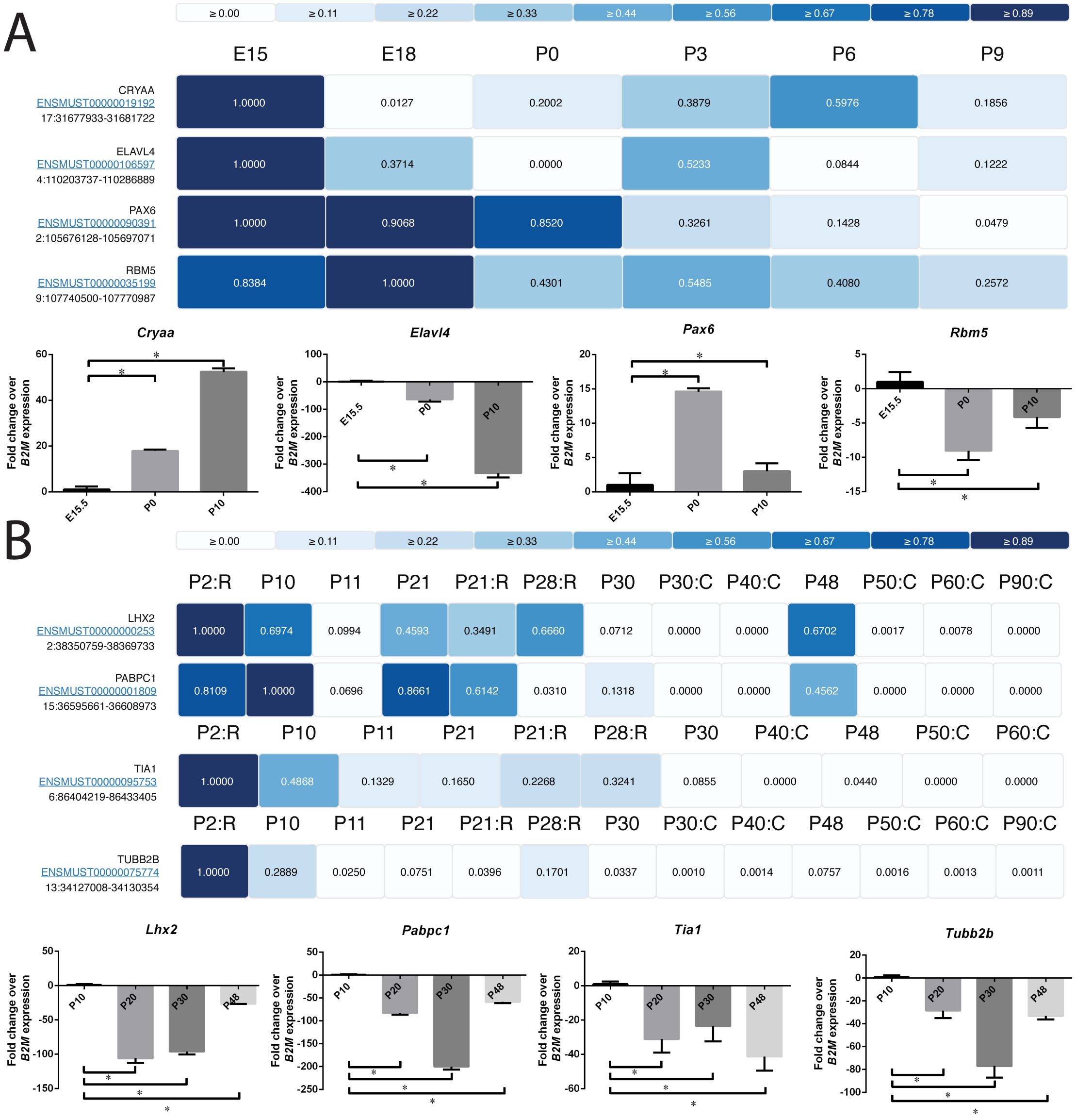


Figure 14: Heatmaps showing the expression profiles of selected transcripts in multiple development stages of mouse eye tissues. Expression data were normalized by the maximum expression level of a given transcript across stages and visualized as heatmap in Express. Expression profile of selected transcripts for (A) lens and (B) retina were downloaded from Express and shown in form of heatmap. In retina datasets some development stages are marked as “C” for cone and “R” for rod cells and unmarked development stages represent whole tissue. The marked datasets are derived from the given cell type. Their expression profile was also verified for multiple development stages using qPCR with *B2M* as housekeeping control and shown as additional panels.

Pabpc1 (Poly A-binding protein, cytoplasmic 1) binds to the poly A tail of mRNA and modulates its susceptibility to cap-mediated mRNA decay (145). RT-qPCR shows that *Pabpc1* is expressed highly at early postnatal stages and its expression reduces significantly at later developmental stages (Figure 14B) until P30 when its expression increases again. Tia1 (T-Cell-Restricted Intracellular Antigen-1) promotes the recruitment of U1 snRNP to splice sites and is implicated in lymphoma and leukemia (146-148), which is also expressed in the mouse lens (149). *Tia1* expression was found gradually decreased in the retina with age (Figure 14B). Tubb2b (Tubulin, Beta 2B Class IIb) is a component of microtubules. *Tubb2b* mutations result in congenital fibrosis of extraocular muscles (CFOEM), which leads to ptosis (drooping eyelids) and restricted eye movements in humans (150). RT-qPCR confirms that *Tubb2b* expression is high at P10 and reduces sharply at P30 before increasing again at P48 (Figure 14B) as predicted by *Express*. The levels of the control genes were also verified and compared across time points for reference.

Express was investigated to compare with the established expression pattern for the gamma-Crystallin family of genes. A previous study describes the expression of different *Cryg* family transcripts at the mouse stages E16.5, P1, P10, P20, P30, P40, P80, P120, and P180 (Goring et al. 1992). *Cryg* gene expression was compared for the stages in *Express* that are closest in developmental time to the stages in the Goring et al. study. Specifically, *Cryg* expression in *Express* for E15, P0 and P9 that are close to the stages E16.5, P1 and P10 in the Goring et al. study, was compared. Using raw expression and TPM cut-off of 5, a good agreement between the *Express* and previous findings for the general trends of the *Cryg* genes was observed, namely for *Cryge*, *Crygf*, *Crygb*, *Crygc* and *Crygd* (Figure 15). *Cryga* showed a slight deviation from the Goring et al. study in that it did not exhibit a slight reduction at P0 prior to being high at P10 (although it exhibits general agreement with the previous study in that the expression of *Cryga* was higher at P9 compared to E15). Therefore, these findings offer further support that gene expression data in *Express* reflects the experimentally validated and established gene expression patterns in the lens.

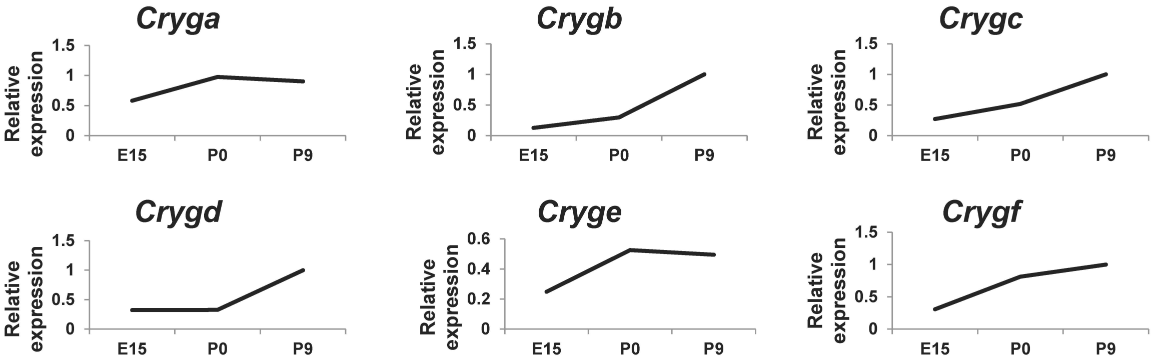


Figure 15: Gene expression analysis of Cryg family of genes. Raw Expression profiles of Cryg genes in mouse lens were downloaded from the Express database to investigate if they are in agreement with previously described patterns (reference: Goring et al.,1992). Expression values for specific Cryg genes in the stages closest to the developmental time points in the previous study are plotted (E15 in Express in lieu of E16.5 in Goring et al. 1992, P0 in lieu of P1, P9 in lieu of P10). The general expression patterns for the Cryg genes correlate well between Express and the previous findings. Specifically, Crygb, Crygc, Crygd and Crygf expression increases as development progresses and is highest at P9, while Cryge expression elevates at P0 and beyond. The only minor deviation is exhibited by Cryga whose expression increases from E15.5 through P9 in Express, instead of the slight decrease at P1 prior to increasing again at P10 as previously described.

3.1.4 Conclusion

A number of studies in the past have focused on studying the expression landscape of genes using microarrays across developmental stages (151-153) in mouse eye development and specialized databases (106, 154) have been built. However, the current understanding of the transcript structure, expression and their splicing alterations is incomplete. Here, *Express* is presented as the first large-scale transcriptomic resource based on eye tissue RNA-seq data to provide a user-friendly portal for studying and visualizing the expression levels of both the known and novel transcript isoforms across developmental stages in mouse eye tissues. Further, several transcripts were validated using RT-qPCR across multiple developmental stages in mouse lens and retinal tissues to confirm that the *Express*-quantified levels of transcripts are in agreement with the detected expression levels from RNA-seq quantification pipeline employed in this study. Several transcripts encoding RNA-binding proteins were found to be highly expressed in embryonic development and down-regulated in post-natal stages suggesting a complex post-transcriptional control of gene expression in early eye development.

The analysis suggests that retinal samples exhibit a significant number of novel transcripts comparable to a recent analysis of human retinal transcriptomes (132). It can be speculated that these novel RNA transcripts may reflect cell type specific functions. Hence resources like *Express* can not only further elaborate the understanding of tissue-specific developmental transcriptome but can also serve to improve gene annotations in mouse.

Express can be a useful resource for prioritization of candidate genes from exome sequencing analysis for patients with ocular defects as well as for providing a functional and developmental context to investigate the significance of differentially expressed genes in mouse mutants with eye defects.

3.2 Transcriptome analysis of developing lens reveals abundance of novel transcripts and extensive splicing alterations

3.2.1 Introduction

The past decade has seen a surge in transcriptome-level studies for specific developmental stages of the eye and its tissue sub-types (102, 155). The development of the eye involves a complex and highly orchestrated regulatory program with several specification and differentiation processes (96, 97). The lens is a transparent tissue that focuses light on the retina (156). It originates from the surface ectoderm early in embryogenesis and is composed of two cell types, namely the anteriorly located epithelial cells and the posteriorly located fiber cells (98, 157). During development and throughout the life of the animal, epithelial cells differentiate into fiber cells that elongate and migrate towards the center of the lens, while degrading their organelles, including nucleus.

Greater than 94% of multi-exonic genes in the human genome are alternatively spliced (158). Further, alternative splicing is an essential and highly controlled post-transcriptional regulatory mechanism which provides transcriptomic and proteomic diversity in eukaryotic organisms (159). Due to the extensive prevalence of splicing events in higher eukaryotes, various transcriptomic datasets across developmental stages have been previously explored in multiple model organisms to study the structure and composition of protein-coding and non-coding genes (160-163). These RNA-Seq based studies revealed more accurate and comprehensive set of known and novel genes for downstream functional and comparative analysis.

Previous studies report that ocular tissues such as the retina can exhibit highly diverse transcript profiles with hundreds of novel transcripts, likely contributed by the ensemble of multiple cell types abundant in retina (102, 132). However, few RNA-Seq based studies have been conducted so far for investigating the lens transcriptome (164, 165) especially over different developmental stages (105, 107). Further, these studies have used only known or annotated genes in their analysis. Thus, to date the complete lens transcriptome and the various isoforms expressed in the developing lens has not been fully characterized. In this study, I investigated the transcriptomic alterations and splicing events from publicly available lens RNA-Seq data and have constructed a comprehensive molecular portrait of known as well as novel transcript isoforms in the mouse lens across developmental stages.

3.2.2 Materials and Methods

To obtain a comprehensive understanding of the transcriptome and splicing alterations across various stages of lens development, I re-investigated the processed RNA-seq datasets of mouse lens from different developmental stages documented in Appendix 1. This study utilized the processed dataset resulted from previous study (166) where in house NGS data processing pipeline was used as illustrated in the workflow (Figure 16). The sorted binary alignment files (sorted-BAM) obtained after post-processing were employed for further data processing i.e. quantification of expression levels of transcripts and splicing analysis.

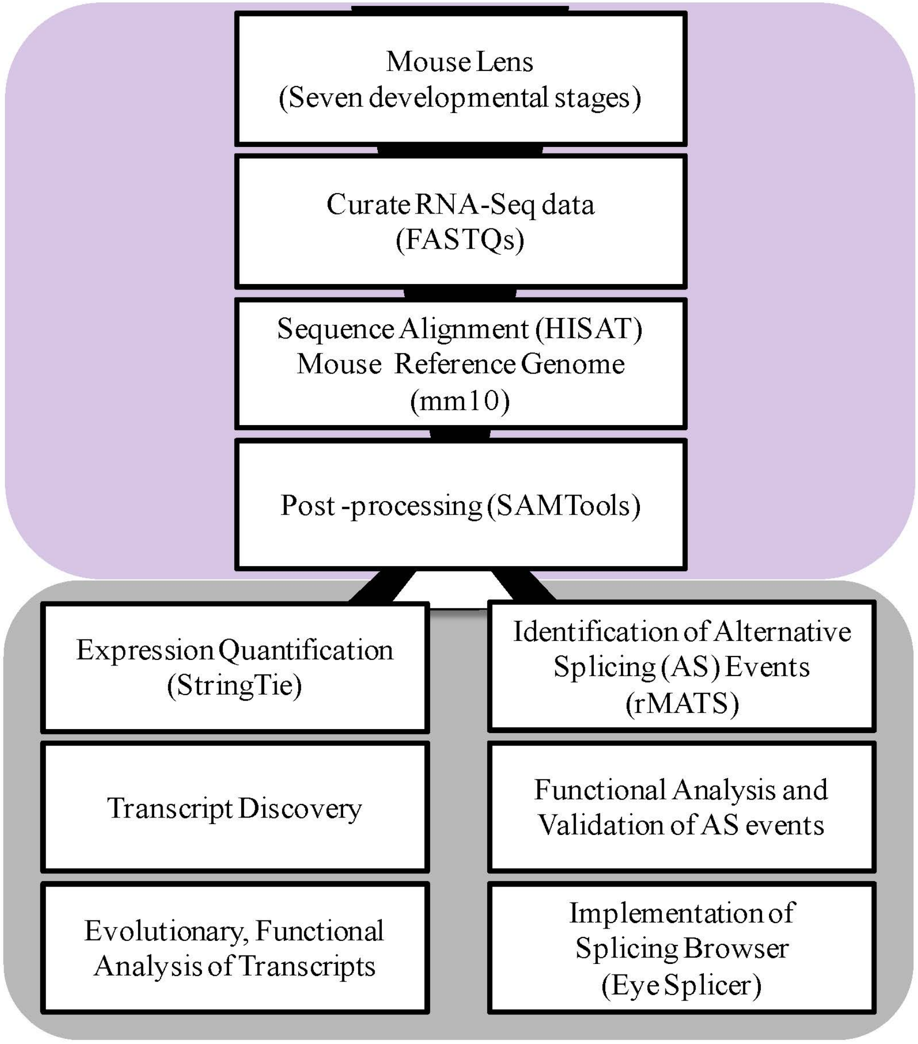


Figure 16: Overview of the transcriptome analysis across developmental stages in mouse lens. Transcriptomes of mouse lens spanning seven developmental stages (three embryonic; E15, E15.5, E18 and four postnatal; P0, P3, P6, P9 stages with biological replicates) were collected from published sources. Curated RNA sequence data was quality filtered using FASTX toolkit. High quality raw sequence reads were further processed using in house NGS data processing pipeline (as described previously) Processed dataset were utilized for two purposes. Firstly, for identifying and quantifying the expression levels of known and novel transcripts across seven developmental stages using StringTie, followed by an evolutionary and functional analysis to uncover high confident completely novel transcripts in developing lens. Secondly, the processed bam files were also employed for the identification of alternative splicing events using rMATS (replicate Multivariate Analysis of Transcript Splicing) (167) followed by functional analysis of genes belonging to the enriched splice events. Finally, the results of the most prominent splicing events namely skipped exon and retained intron events are also made available through Eye splicer, a web based splicing browser showing developmentally altered splicing events in mouse lens.

3.2.2.1 Transcript identification and quantification from the aligned RNA-seq datasets

StringTie (version 1.2.1) (120) was used for identification and quantification of transcripts from the aligned RNA-Seq reads (BAM files). StringTie uses a novel network flow algorithm for fast and highly efficient assembly and quantitate the transcripts of each genomic locus considering all possible multiple splice events. In addition to annotated transcripts, it can also provide the information of possible novel transcripts in each sample. Transcript level expression data quantified using StringTie were stored in GTF (Gene Transfer Format) providing expression levels for both known as well as novel transcripts against mouse reference genome (mm10 - Mus_musculus.GRCm38.84.gtf). All the GTFs previously obtained for each sample were grouped and provided as an input for stringtie "merge" mode along with mouse reference genome (mm10 - Mus_musculus.GRCm38.84.gtf). The merged GTF thus obtained was then utilized as reference annotation file in re-running StringTie with the sorted-BAM for the corresponding samples. As a result, I obtained a matrix of expression levels for 90689 transcripts (68166 annotated and 22523 novel transcripts) in the mouse genome. Known transcripts are defined as the transcripts whose genomic co-ordinates and annotations completely overlapped with those reported in Ensembl database (121) for the mouse genome. In contrast, novel transcripts were defined as the transcripts that were exclusively predicted by StringTie and hence could overlap partially with already annotated exonic regions in the mouse genome. A quantification matrix was generated for lens transcriptome with respect to different developmental stages extracting the TPM (transcripts per million) values from StringTie outputs. This matrix was utilized for downstream analysis.

3.2.2.2 Defining and investigating the novel transcripts across developmental stages

I calculated the proportion of known and novel transcripts for each RNA-seq sample with an expression threshold of TPM > 1.0 and averaged the values for corresponding replicates from each developmental stage. The obtained proportions were represented as a bar graph for each developmental stage. Similarly, the proportion of known and novel transcripts with varying expression thresholds (TPM > 0.5, > 2 and > 5) was calculated and represented as bar graphs to study the reproducibility of the observed trends.

To investigate the discovered novel transcripts for their extent of novelty with respect to the known transcript architectures documented in the mouse reference genome mm10, I mapped the length of the discovered transcript to annotated reference transcript coordinates and calculated a novelty score for each novel transcript by using the below formula,

$$\text{Novelty Score} = \left(1 - \frac{\text{length overlapping region}}{\text{full length of novel transcript}}\right) \times\ 100$$

The distribution of novelty score of novel transcripts in each developmental stage was examined and represented it as a density plot. I performed K–S (Kolmogorov–Smirnov) test to investigate for statistically significant differences in the novelty score distributions between any pair of developmental stages. Based on prior calculations and distribution of novelty scores, the novel transcripts were characterized into two groups; partially novel transcripts (PNTs, novelty score < 70%) and completely novel transcripts (CNTs, novelty score ≥ 70%). I analyzed the expression levels of transcripts across all stages for each transcript group - known, partially annotated novel and completely novel transcripts and performed Wilcoxon rank sum test to study the distribution of expression levels between transcript groups for each developmental stage separately. These results were represented as box plots in supplementary material.

This study also investigates the distribution of the number of exons and length of the transcripts for known, partially novel and completely novel transcripts. K–S (Kolmogorov-Smirnov) test was performed to evaluate whether length distributions of transcripts significantly differ. Likewise, exon counts were also compared for these three categories of transcripts.

3.2.2.3 RT-PCR analysis of CNTs

To validate the expression levels of novel transcripts discovered from RNA-Seq analysis, total RNA was extracted using a RNeasy Mini kit (Qiagen Inc, Valencia, CA) from microdissected C57Bl/6 mouse lenses at three stages, namely, embryonic day (E) 15.5, and post-natal day (P)0 and P10. Each of the three biological replicates at E15.5 comprised of six lenses, and at P0 and P10 comprised of two lenses. RNA was treated with RNase free DNase (Qiagen Inc #79254, Valencia, CA). cDNA was synthesized from 200 ng of total RNA, representing three biological replicates at each developmental stage using Bio-Rad iScript^{TM} cDNA Synthesis Kit (Bio-Rad Laboratories, Hercules, CA), and was used as a template in PCR analysis. Primers were designed for the exonic regions of these four CNTs. The PCR products were run on 1% agarose gel. Presence of specific bands at the expected size were indicative of transcript expression in the lens.

3.2.2.4 Phylogenetic conservation of mouse lens transcriptome

Although some reports indicate that mouse lens is likely to have a diverse transcriptome, the evolutionary significance of the transcriptome is poorly understood. Hence to address this, I investigated the evolutionary conservation of the identified transcripts. Multiple sequence alignment of genomic loci across several genomes provides a comprehensive snapshot of the evolutionary conservation, which can act as a proxy for functional preservation of a selected region (168). For instance, protein coding genomic loci were documented to be highly conserved across the genome than non-functional genomic loci (169). I applied this technique to conjecture and identify novel transcripts which could be functionality active across large phylogenetic distances. I downloaded the phastCons scores (170) from the UCSC Genome Browser for the complete mouse genome. PhastCons score employed in this study provides an estimate of the individual nucleotide level conservation, calculated based on multiple sequence alignment of 46 vertebrate genomes with respect to mouse reference genome mm10. It ranges from 0-1 with higher the score higher is the conservation of the individual nucleotide across the genomes. For this study, I utilized the available nucleotide resolution conservation score data for mm10 and calculated the phastCons score for each exon of the novel transcripts by averaging the per-base scores and then computed a representative conservation score for each transcript as the mean phastCons score of the exons representing the novel transcript. Final scores were analyzed for known (annotated) transcripts, PNTs and CNTs to compare their relative extents of conservation.

Since Gene Ontology (GO) based functional enrichment analysis can provide important clues about the functions and molecular processes predominantly associated with novel transcripts, I analyzed the Partially Novel Transcripts (PNTs) that shared majority (>70%) of their genomic region with known/annotated transcript containing genes to understand the likely functions associated with them. This involved filtering the PNTs with phastCons score (> 0.8) to first identify highly conserved transcripts and using the resulting set of genes associated with these PNTs for downstream functional analysis. Functional enrichment analysis was performed with p-value threshold $< 10^{-10}$ for collected genes using Cytoscape (171)-ClueGO (172) plugin and was represented as a clustered GO network. Significant clustering of genes, color coded by annotation group, based on enriched GO biological processes were highlighted in these representations. Transcripts belonging to the completely novel class share less than 30% of their genomic region with known transcripts. This study is based on the hypothesis that completely novel transcripts with high conservation and expressed in at least one developmental stage could be active with uncharacterized function. Hence, I filtered the transcripts based on phastCons score (> 0.8) and analyzed their expression pattern. Expression profiles normalized by their maximum expression level across stages for these highly conserved completely novel transcripts were hierarchically clustered using Cluster 3.0 (52) and visualized as a heatmap using Java Treeview (53). Representative hierarchically clustered panels of transcripts expressed in only one specific developmental stage and in all developmental stages were shown separately. Novelty Score (NS) and phastCons Score (PS) indices for transcripts were shown as an additional scale bar in each heatmap. A subset of broadly expressed, highly conserved and 100% novel transcripts were selected for experimental validation and discussed in the results section.

3.2.2.5 Analysis of differential alternative splicing

RNA-Seq data provides an opportunity to detect differential alternative splicing events across conditions. The two replicates of RNA-seq for each developmental stage of mouse lens tissue were merged and investigated with rMATS (replicate Multivariate Analysis of Transcript Splicing) (167) to identify differential alternative splicing (AS) events. rMATS provides a computational framework to identify all possible splicing events which are altered between two samples, by inspecting the status of exons/introns as they are included or excluded resulting from alternative splicing. I used sorted BAM (Binary Alignment/Map) files, obtained from aligning the raw RNA-seq datasets against the mouse reference genome using HISAT as discussed above, as input to rMATS by pairing with their corresponding replicates from each developmental stage. This allowed the pair-wise comparison of developmental stages for alterations in various splicing events. Since rMATS requires all input datasets to have the same read length, the dataset from E15.5 which had a different read length compared to others was excluded. Also, the GFF (General Feature Format) file downloaded from Ensembl (version 82, September 2015) (173) were provided as input to rMATS and the default thresholds were used for remaining options. Briefly, rMATS enabled the analysis of the inclusion/exclusion of target exons/introns contributing to different types of alternative splicing events, namely skipped exon (SE), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), mutually exclusive exons (MXE) and retained intron (RI), across any pair of developmental stages with replicates. An AS event is quantified based on the difference in the level of inclusion of an exon which is defined as the splice index or Percentage Splicing Index ($\psi\ score$) between two samples or conditions and ranges between 0 and 1. PSI represents the inclusion/exclusion of an exon for a transcript isoform considering all alternate possible isoforms. Reads aligning to the alternative exon or to its junctions with adjacent constitutive exons provide support for the inclusion isoform, whereas reads aligning to the junction between the adjacent constitutive exons support the exclusion isoform; the relative read density of these two sets forms the standard estimate of ψ.

Significant differences in the values of ψ for an exon, between a pair of conditions compared to a null distribution indicate its differential abundance. rMATS code was executed for all pairs of six developmental stages (E15, E18, P0, P3, P6 and P9) and generated a summary table with the number of different alternative splicing events that were detected below 1% FDR threshold (Table 3). Since skipped exon and retained intron events were the most abundant, I collected these events from raw rMATS outputs specifically those which are supported by reads that span splicing junctions and reads on target below 1% FDR. Functional enrichment analysis of genes belonging to these splicing events was performed using ClueGO (172).

Summary of the number of high confident Alternative Splicing (AS) events detected using rMATS pipeline (FDR <0.01) across developmental stages with replicates.

AS Event	E15 vs E18	E15 vs P0	E15 vs P3	E15 vs P6	E15 vs P9	E18 vs P0	E18 vs P3	E18 vs P6	E18 vs P9	P0 vs P3	P0 vs P6	P0 vs P9	P3 vs P6	P3 vs P9	P6 vs P9
SE	7	52	40	55	55	120	75	123	87	10	9	14	3	5	0
MXE	3	6	18	8	12	6	16	4	10	14	2	3	10	14	0
RI	62	46	25	26	37	73	29	31	34	6	2	9	0	1	2
A5SS	1	9	0	4	0	7	0	2	0	1	1	2	0	0	0
A3SS	3	4	3	3	3	12	5	7	5	1	2	0	1	0	1

Table 3: Identification of alternative splicing events using rMATS. Abbreviations used in the table stand for the following types of splicing events and definitions: SE- Skipped Exon, MXE- Mutually Exclusive Exon, RI- Retained Intron, A5SS- Alternative 5' Splice Site, A3SS- Alternative 3' Splice Site, PSI- Percent Spliced Index, FDR- False Discovery Rate.

3.2.2.6 Experimental validation of the skipped exons

To confirm splicing events during lens development, genes were selected based on their potential relevance to lens biology and which were predicted with less than 5% FDR in splicing analysis. For alternative splicing analysis, primers were designed on exons flanking the alternatively spliced exon (skipped exon) on either side. Total RNA from E15.5, P0 and P10 C57Bl/6 mouse lens was collected as described above. RNA was treated with RNase free DNase (Qiagen Inc #79254, Valencia, CA). 200ng of lens total RNA was used as template for cDNA synthesis using *in vitro* reverse transcription kit as described earlier and cDNA was used as a template for PCR reactions. The different splice isoforms were identified based on size differences of PCR products separated by 1% agarose gel electrophoresis. Further, the PCR products obtained using RNA from P0 lens were analyzed by Sanger sequencing. The different splice isoform DNA bands from the P0 lens samples were excised from the gel and subjected to DNA purification using Wizard® SV Gel and PCR Clean-Up System (Promega #A9281, Madison, WI). DNA isolated from specific splice isoforms was sequenced by Sanger sequencing method.

3.2.2.7 Development of a splicing browser for studying splicing alterations across developmental stages

The abundant AS events that were detected in this study namely skipped exons and retained introns, were made available for visualization via Eye Splicer (http://www.iupui.edu/~sysbio/eye-splicer/), an interactive web-based splicing browser for studying splicing alterations in mouse lens. Eye Splicer is built using the JavaScript library from Biodalliance (http://www.biodalliance.org). As Biodalliance requires BED (Browser Extensible Data) or BigBed formatted input files, these tables were preprocessed into BED formatted text files and generated the corresponding BigBed files, which are the compressed version of BED files and hence suitable for the web using the UCSC tools(125). Eye Splicer has a simple interface with the lists of genes that have exons alternatively spliced below 1% FDR for skipped exons and retained introns, shown on the left menu and an interactive genome browser on the right which allows the visualization of the exons of interest upon selection from the gene lists or upon search using its text field that supports coordinate based search or gene name / Ensembl ID based search. Any viewable section of the splicing browser can be exported using the Export button as SVG (scalable vector graphics). Eye Splicer is freely available on http://www.iupui.edu/~sysbio/eye-splicer/ and can be accessed without any login requirement.

3.2.3 Results and Discussion

Although mouse lens transcriptome profiling has been the focus of few studies in recent years (105, 107, 164, 165), an understanding of the complete repertoire of expressed transcripts and their splicing alterations during lens development is far from complete. In this study, the transcriptomic alterations and alternative splicing events were investigated in mouse lens across developmental stages. Overview of the analysis pipeline is illustrated in Figure 16. In brief, the available RNA-Seq data for mouse lens across varying developmental stages were collected and the raw sequence reads were processed using HISAT (117) and StringTie (120). The processed and quantified data were formatted into expression matrices and were utilized for investigation of complete transcriptomic architecture, extent of transcript novelty, and their evolutionary conservation (see Materials and Methods). Additionally, I investigated the alternative splicing events using rMATS (167) followed by an extensive functional analysis of the genes associated with enriched splicing event types. The most prominent splicing event types namely skipped exon and retained intron events were made available through Eye splicer (http://www.iupui.edu/~sysbio/eye-splicer/), a web based splicing browser showing developmentally altered splicing events in mouse lens.

3.2.3.1 Overview of the dataset and construction of the developmental transcriptomes in lens

The processed RNA sequencing data was collected to facilitate downstream analysis (See Materials and Methods, Figure 16 and Appendix 1). Overall, datasets exhibited a good read quality (Phred score > 20) and a high fraction of read alignment to the reference genome (alignment score ≥ 93%) using HISAT.

Since previous reports studying the eye transcriptomes indicated diverse transcriptomic architecture (96), the goal was to investigate whether such diversity exists in different developmental stages of lens. For this purpose, the expression of transcripts and corresponding exons were quantified using StringTie and a matrix of expression levels for 90689 transcripts (68166 annotated and 22523 novel transcripts) were constructed. The analysis indicated the existence of ~25% novel transcripts in the developmental mouse lens transcriptome. In order to further investigate the extent of the novel transcripts in each developmental stage, the proportion of known and novel transcripts (with TPM > 1.0) was analyzed across different developmental stages (Figure 17A). The results show that in each of the developmental stages of mouse lens there are about ~ 35 - 50% of novel transcripts. Such variations in the distribution of known versus novel transcripts with respect to different developmental stages was found to be consistent despite filtering for different TPM thresholds (i.e. > 0.5, > 2.0, and > 5.0). These observations support the presence of a diverse transcriptome with thousands of novel transcripts being expressed in various lens developmental stages as well as the predominance of complex transcriptional and post-transcriptional regulatory mechanisms in embryonic and post-natal stages during mouse lens formation.

3.2.3.2 Embryonic stages exhibit the highest extent of novelty for the newly discovered transcripts with a significant decrease in post-natal stages

To further investigate whether the expression of these novel transcripts differs between stages, novelty score of a transcript was calculated to measure the differences in the extent of novelty across stages using KS (Kolmogorov–Smirnov) test. Novelty score of a transcript is defined as the percentage of non-overlapping novel transcript length to the reference annotated transcriptome (Figure 17B). In the embryonic stages, each pair of neighboring developmental stages were found to be significantly different in their distribution of novelty scores for the novel transcripts (p-value ≤ 0.005) and this pattern was observed until birth (P0). In general, the novelty score distributions of the novel transcripts for embryonic stages were observed to be significantly higher compared to those seen in post-natal stages (median novelty score: 10.89 *vs* 9.04, p=1.06e-12, KS-test, Figure 17B).

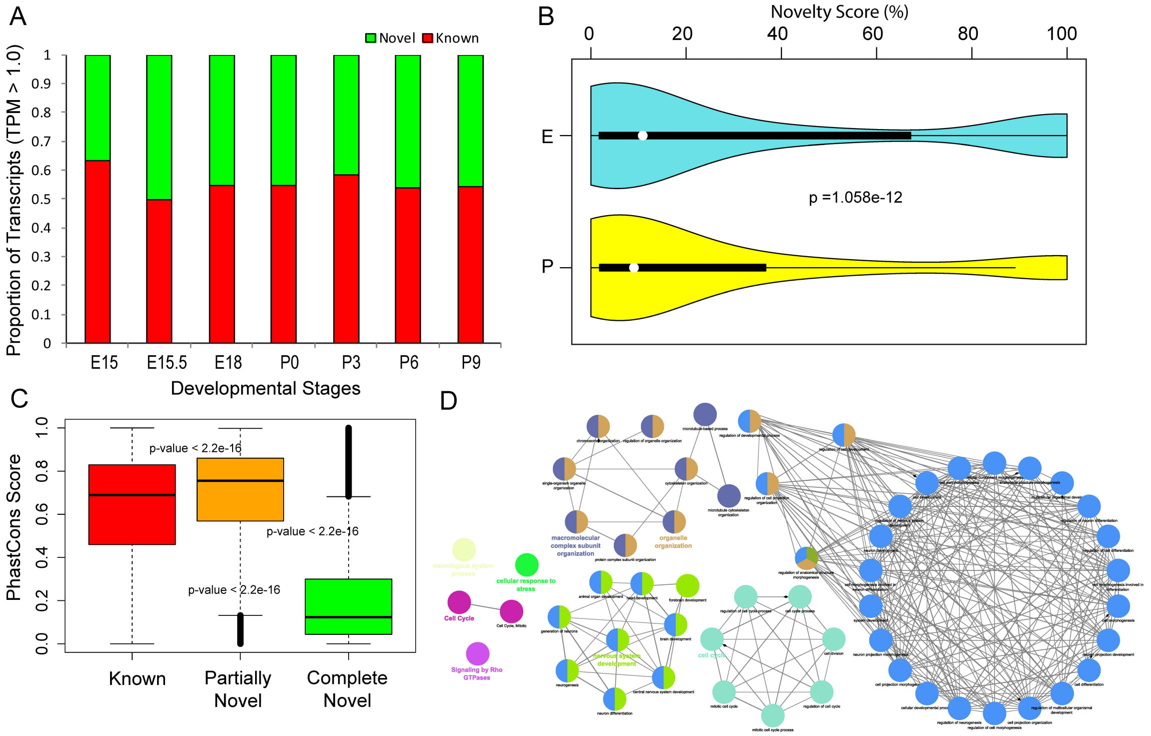


Figure 17: (A) Histogram showing the proportion of known and novel transcripts identified across various lens developmental stages in mouse. Only transcripts exhibiting an expression higher than 1 TPM (Transcripts Per Million reads sequenced) are considered in this plot. (B) Violin plot showing the distributions of novelty scores of identified transcripts, expressed in embryonic and postnatal stages. Novelty score of the transcripts expressed (with TPM > 5.0) at least in one stage were employed to generate two violin plots corresponding to the embryonic (E15, E15.5, E18) and postnatal (P0, P3, P6, P9) stages respectively. Differences in the distribution of novelty scores between embryonic and post-natal stages were compared using Kolmogorov–Smirnov test. Median novelty score for E and P were 10.89 and 9.043 respectively. (C) This panel shows the distribution of PhastCons scores (nucleotide level conservation), reflecting the extent of conservation for known, partially novel (novelty score <70%) and completely novel (novelty score ≥ 70%) transcripts identified across developmental stages in lens. Each pair of these transcript classes were found to be significantly different in their extent of conservation (p < 2.2e-16, Wilcoxon rank sum test) with median conservation scores 0.67, 0.76, and 0.13 for known, partially novel and completely novel transcript groups respectively. (D) Gene ontology enrichment analysis for genes corresponding to the high confidence partially novel transcripts (PS >0.76). Functional grouping of the GO-terms based on GO hierarchy was represented as clustered GO-network using the Cytoscape(171)-ClueGO(172) plugin. Significant clustering (p < 1e-10) of genes (color coded by functional annotation group they belong to) based on enriched GO-biological processes, with size of the nodes indicating the level of significant association of genes per GO-term, were shown.

3.2.3.3 Significant fraction of the partially novel transcripts in lens were found to be highly conserved across vertebrates and associated with neural system development, structural morphogenesis, protein localization, cell division and differentiation processes

In this study, a total of 22523 novel transcripts (~25% of total transcripts) were identified in mouse lens (18) along with their novelty score and expression levels. As discussed above, differences were observed in the distribution of novelty scores of transcripts between embryonic and postnatal developmental stages. Hence, the novel transcripts were further classified based on their novelty score (See Materials and Methods). The novel transcripts were categorized into two groups; Partially Novel Transcripts (PNTs, novelty score < 70%, 13207 transcripts) and Completely Novel Transcripts (CNTs, novelty score ≥ 70%, 9316 transcripts).

To investigate and compare the extent of conservation of known and novel transcripts, phastCons scores was used from UCSC Genome Browser, which provide a nucleotide level conservation score across 46 vertebrate genomes, facilitating a measure to quantify conservation for mouse genomic loci (see Materials and Methods). I calculated the phastCons score distributions for each group of transcripts; known transcripts, PNTs and CNTs (Materials and Methods section, Figure 17C). A significant difference in phastCons score distributions was observed among these groups (median for known transcripts = 0.67, median for PNTs = 0.76, and median for CNTs = 0.13; Wilcoxon rank sum test, p-value < 2.2e-16). The score distribution indicates that PNTs exhibit higher conservation patterns than already known transcripts while their patterns were less comparable to CNTs. These observations suggest that since lens tissue and corresponding cell line transcriptomes have been poorly or rarely studied by genome annotation consortiums like ENCODE (92) or FANTOM (174), it is possible that hundreds of transcripts specific to lens may have been rarely documented in genomic/transcriptomic resources. However, integrative analyses and databases based on next generation RNA-sequencing datasets specific to such overlooked tissues, would be able to capture such missing transcript isoforms or poorly annotated genes, suggesting the need for such focused studies. In contrast, most of the CNTs were found to be poorly conserved based on phastCons score profiles. Interestingly, a few of the CNTs were found as outliers in the box plot exhibiting extremely high conservation (Figure 17C, CNTs, above third quartile), which met the median phastCons threshold of both known and CNTs, and hence are likely to be active but functionally uncharacterized for biological processes.

To understand whether particular functions and processes are over-represented as gene ontology (GO) categories for these novel transcripts, I performed functional enrichment analysis of the PNTs by using the annotations of the corresponding mouse genes with which they overlap partially. To generate a high confident set of evolutionary conserved novel transcripts with annotated information, I filtered the PNTs with phastCons score > 0.8 and obtained a set of 3982 genes satisfying these criteria. I performed functional enrichment analysis of these genes with corrected p-value (Bonferroni correction) threshold < 10^{-10} using ClueGO (172). ClueGO is a Cytoscape plugin which enables the functional grouping of GO terms or gene sets to represent the enriched functional themes as networks. There was significant clustering of genes into 26 thematic groups based on enriched GO terms using ClueGO (172). Specific biological processes and associated modules are highlighted in Figure 17D. Results from this study demonstrate that ‘alternative mRNA splicing via spliceosome’, ‘mRNA metabolism process’, ‘ubiquitin mediated proteolysis’, ‘nervous system development’, ‘neurological system process’, ‘organelle organization’, ‘cell cycle’, ‘protein localization’ etc were over-represented in PNTs (Figure 17D). For instance, group 19 (i.e. nervous system development) was found to be significantly enriched (adjusted p-value = 9.93e-32) with 841 genes i.e. ~30% of the genes annotated with neurogenesis, neuron differentiation and nervous system developmental processes. These observations clearly reveal the role of several poorly characterized transcripts associated with nervous system development, RNA metabolism, cell cycle, organelle and chromatin organization, regulation of anatomical structure morphogenesis and cell differentiation, during lens development.

3.2.3.4 Majority of the complete novel transcripts are widely expressed across developmental stages albeit exhibiting significantly lower expression, conservation and length compared to partially novel transcripts

The expression level of a transcript across biological replicates was averaged in each developmental stage in order to compare the distribution of expression levels for known transcripts, PNTs and CNTs. I included the subset of transcripts in each class which were found to be expressed in all seven stages which resulted in 23121 known transcripts, 4531 PNTs and 4027 CNTs. Interestingly, all the three transcript classes was observed to exhibit significantly different expression profiles for each developmental stage (Wilcoxon rank sum test, p-value < 0.001), with known and PNTs exhibiting significantly higher expression compared to CNTs. This analysis also revealed that PNTs are highly expressed than known transcripts (Wilcoxon rank sum test, p-value < 0.001). These observations are similar to the conservation pattern of PNTs being higher than other transcript groups. These results indicate that PNTs are significantly more expressed than CNTs across all developmental stages and are often more expressed than even annotated transcripts suggesting that these PNTs are likely functional in lens development.

I also investigated the transcript structure of different transcript classes by comparing the number of exons and length distributions. Significant difference were observed in the distribution of exonic composition for PNTs and known transcripts ($p < 2.2e-16$, Kolmogorov–Smirnov test), with majority of the PNTs being multiexonic (>3 exons). In particular, about 20% of the PNTs were found to have more than 20 exons and were enriched in genes associated with several processes including 'microtubule cytoskeleton organization', 'cell cycle', 'nervous system development', 'cell projection morphogenesis', 'embryo development', 'focal adhesion' and 'chromatin remodeling'. In contrast, I observed that ~ 90% of CNTs were single or bi-exonic with a small fraction of them exhibiting multiexonic structure (Appendix 3A). I also investigated the length for the three groups of transcripts and found significantly (p-value < 2.2e-16) varying distribution of lengths (Appendix 3B). The known transcripts exhibited an expected distribution of transcript length as previously described (175) with an abundance of transcripts having length between ~10^2 bp and ~10^4 bp. However, among the novel transcript groups; PNTs exhibited a distribution more similar to that of known transcripts when compared to CNTs. In particular, most of the PNTs had length ranging from 10^2 – 10^7 bp with abundance of transcripts having length in the range of 10^5-10^6 bp. In contrast, majority of the CNTs ranging in length from 10^2 to 10^5 bp dominated by relatively shorter length (100-1000 bp) transcripts. Indeed, studies from GENCODE consortium

(39) observed that human long noncoding RNAs (lncRNAs) are typically encoded as single or biexonic transcripts with significantly lower exome lengths compared to annotated protein coding transcripts, suggesting that several of the CNTs detected in this study are likely to be noncoding RNAs.

Although transcripts belonging to the CNT class were generally poorly conserved compared to the other two groups (Figure 17C), nevertheless a small fraction (~8.6%) of CNTs exhibited high conservation with phastCons scores greater than 0.76. In order to further interrogate the activity of these ~8.6% completely novel transcripts, I analyzed their expression profile across developmental stages. Further they were filtered to obtain a set of CNTs with a phastCons score > 0.8 and expressed in at least one developmental stage, after excluding RNA-seq samples from E15.5 which originate from a different study in order to avoid any potential batch effect. I found a total of 647 CNTs that exhibited varying levels of expression across developmental stages (18). Figure 18A shows a clustering snapshot of the distribution of these expression profiles across stages with expression levels of a transcript normalized by its maximum level across developmental stages (Materials and Methods). I analyzed the expression profiles of CNTs based on hierarchical clustering to identify representative panels of transcripts expressed in only one specific developmental stage (Figure 18B) and in all developmental stages analyzed (Figure 18C). The results indicate that ~10% of the CNTs (phastCons score > 0.8) were expressed in specific developmental stages as shown in Figure 18B.

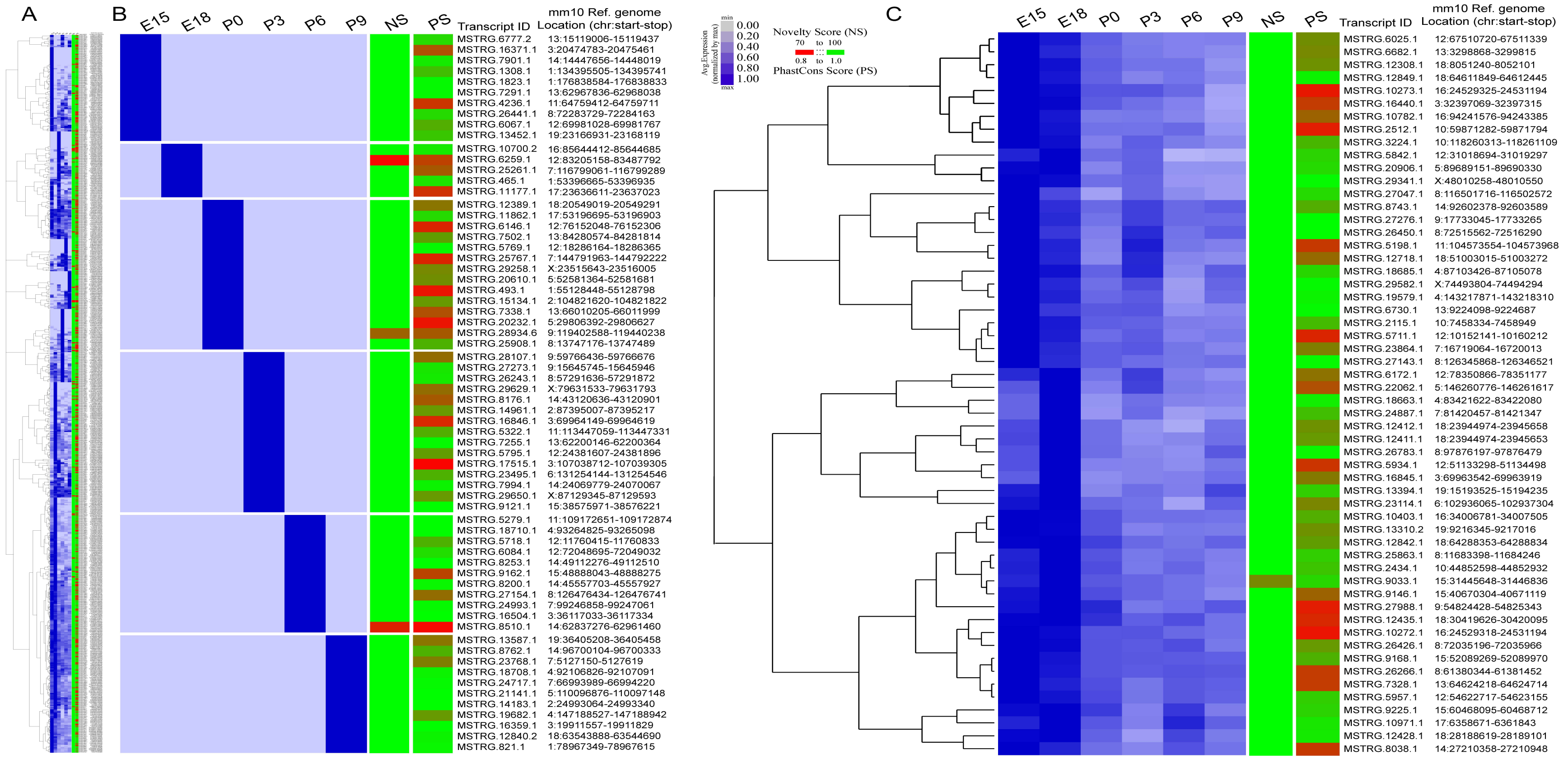


Figure 18: Completely novel transcripts (CNTs) with high conservation score (phastCons Score > 0.8) and expressed in atleast one developmental stage are shown across the panels. Expression profiles are normalized by the maximum expression level of a given transcript across stages and hierarchically clustered using Cluster 3.0 and visualized as a heatmap using Java Treeview. Samples from E15.5 that came from a different study than the rest of the samples were excluded from this expression analysis in order to avoid the batch effect. Heat maps showing the expression profiles of (A) 647 completely novel (novelty score ≥ 70%) transcripts hierarchically clustered with representative transcript groups expressed (B) in only one specific developmental stage and (C) in all the developmental stages. Novelty score (NS) and phastCons score (PS) indices for transcripts are also shown in as an additional scale bar in each heat map.

In contrast, ~47% of the transcripts were found to be expressed across all developmental stages, with a selected set of hierarchically clustered CNTs following this trend shown in Figure 18C. This suggests that a small fraction of CNTs with uncharacterized function could be potentially regulating stage specific developmental processes while majority of the CNTs could have broader functional roles across stages albeit uncharacterized.

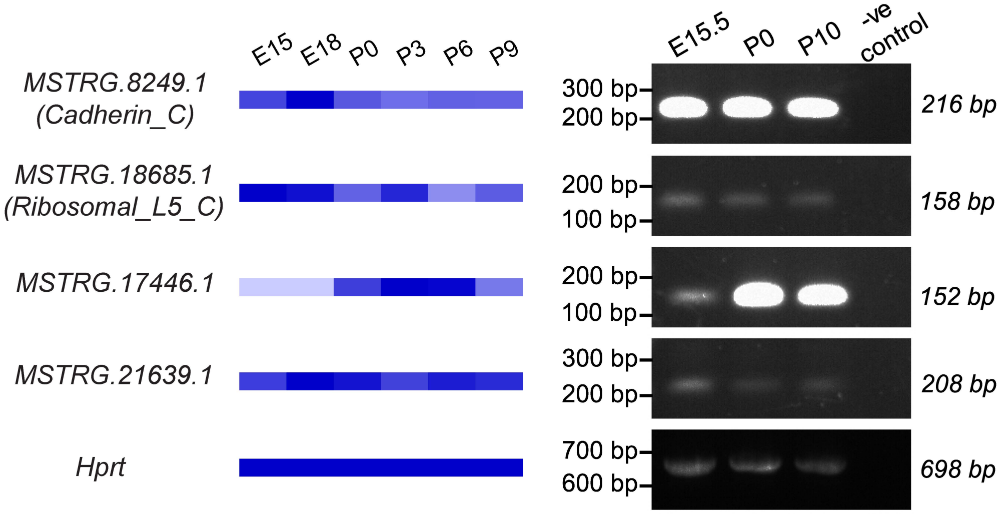


Figure 19: RT-PCR analysis validates expression of two CNTs with a predicted ORF (*MSTRG.8249.1* and *MSTRG.18685.1*) and two CNTs with no known ORF (*MSTRG.17446.1* and *MSTRG.21639.1*) in E15.5, P0 and P10 lenses. Note that *MSTRG.17446.1* is undetected in this analysis at stage E15.5. *Hprt* represents a loading control. Negative control is included for all CNTs tested where the RT-PCR reaction was performed using the same primers as for the CNTs but without any cDNA.

For these 647 CNTs, that were 100% novel and exhibited a phastCons score > 0.8, ORF prediction analysis was performed using an ad hoc Python script that detects both canonical and non-canonical start codons in six open reading frames. It was observed that 202 of them encode for ORFs with 121 of them exhibiting at least one hit using HMMSCAN(176) against Pfam, suggesting that at least 18% of the CNTs are likely to encode for functional domains. Further, four of the CNTs (shown in Figure 19) were validated by RT-PCR in three different developmental stages of lens, among which two transcripts were predicted to encode for ORFs (Figure 19). All the four completely novel transcripts were found to be expressed in P0 and P10 stages. As predicted from the transcriptomic analysis, the MSTRG.17446.1 transcript was not detected at E15.5. These results further validate the stage-specific expression of CNTs shown in Figure 19.

3.2.3.5 Splicing analysis reveals abundance of skipped exons and retained intron events across developmental stages

Alternative splicing is an important molecular mechanism which contributes to the transcriptomic diversity in higher eukaryotes (177). Increasing evidence supports the role of splicing and post-transcriptional regulatory alterations in development (178) and disease (179-182), in addition to their prominent role in generating multiple transcripts and protein isoforms in normal cells.

Since significant differences in the distribution of novelty scores was observed for novel transcripts between the embryonic and post-natal stages in mouse lens, it was speculated that alternative splicing could contribute to these differences. In addition to contributing to transcript isoforms, splicing events can also contribute to differential regulation of the gene products across developmental stages by controlling the abundance of the required isoform. Hence, I employed rMATS (167), a framework for detecting splicing alterations from next generation RNA-sequencing datasets, to investigate such key events for molecular diversity across developmental stages (see Materials and Methods). Table 3 shows the number of high confident Alternative Splicing (AS) events detected using rMATS pipeline (FDR <0.01) across every pair of developmental stages with replicates. Table 3 includes the number of detected AS events reported to be significant by rMATS, for the five types of events namely skipped exon (SE), alternative 5’ splice site (A5SS), alternative 3’ splice site (A3SS), mutually exclusive exons (MXE) and retained intron (RI). These results clearly indicate an abundance of SE and RI events compared to the other types during lens development.

3.2.3.6 Skipped exon events are the most abundant splicing events during lens development and are associated with differentiation, development and cytoskeletal regulatory pathways

Skipped exons are one of the most prevalent alternative splicing events in higher eukaryotes (183). In these events, the splicing machinery can ‘skip over’ an exon by splicing it, thereby masking its contribution in the final RNA or protein product. I obtained 418 significant (FDR < 1%) exon skipping events corresponding to 266 exons observed in 399 transcripts from 213 genes across various developmental stages (Appendix 4). Functional enrichment analysis of the genes associated with skipped exonic events was conducted using ClueGO (172) (Materials and Methods). Several significant (adjusted p-value < 2.05e-04, Figure 20A) groups of functional processes were found to be enriched including ‘mRNA processing’, ‘microtubule-based process’, ‘splicing factor NOVA regulated synpatic proteins’, ‘regulation of intrinsic apoptotic signaling pathway’, ‘lens development in camera-type eye’, ‘protein polymerization’, ‘tight junction’, ‘positive regulation of developmental growth’ and ‘striated muscle cell differentiation’. These observations indicate the prevalence of skipped exonic events in several differentiation and developmental processes via post-transcriptional regulation. For instance, 6 genes significantly (adjusted p-value = 8.30e-05) associated with the term ‘lens development were found in camera-type eye’. The genes that belong to this functional theme include *Cdk4* (Cyclin-Dependent Kinase 4), *Cryba1* (Crystallin, Beta A1), *Lim2* (Lens Intrinsic Membrane Protein 2), *Meis1* (Meis Homeobox 1), *Pax6* (Paired Box 6), and *Smarca4* (SWI/SNF Related, Matrix Associated, Actin Dependent Regulator of Chromatin, Subfamily A, Member 4), which contributes to ~8% of genes annotated with lens developmental processes.

Paired Box 6 (*Pax6*) is a transcription factor encoded by 14 exonic gene *Pax6*. This gene has previously been documented as a key regulator for sensory developmental processes (184, 185) and lens regeneration (186). The result indicate that a particular exon, ENSMUSE00001311933 (*Pax6*) was included in all developmental stages except P9 with high Percent Splicing Index (PSI) values ranging between 0.93 and 0.99. Similarly, I found that ENSMUSE00000736151 (Cyclin-Dependent Kinase 4, *Cdk4*) is differentially included in E18 (PSI value = 0.964) *versus* P0 (PSI value = 0.8025) (FDR < 1%) and ENSMUSE00000691476 (Crystallin, Beta A4, *Cryba4*) is included all developmental stages except P0 with high PSI values ranging between 0.97 and 0.99, suggesting its importance in lens development (FDR < 1%).

3.2.3.7 Several skipped exonic events during lens development could be verified by RT-PCR and Sanger sequencing

The expression of alternate isoforms of *Pax6* and *Cdk4* was validated by RT-PCR and Sanger sequencing across developmental stages. Both *Pax6* and *Cdk4* follow the predicted trend (Figure 20B). For example, the ENSMUSE00001311933 exon of *Pax6* is expressed at stages E15.5 and P0, while its expression is undetected at P10. *Cdk4* exon ENSMUSE00000736151 is expressed at all three stages, E15.5, P0 and P10. Further, the skipped exonic events detected in four other genes (*Banf1*, *Cryaa*, *Eif4g2*, *Rbm5*) were validated, that have been detected at an FDR <5% (Figure 20B). Additional validation of these splicing events in P0 lens using Sanger sequencing independently confirmed the findings (Materials and Methods). Mutations in *Cryaa* have been previously shown to cause cataracts in humans and mice (187, 188). *Eif4g2* and *Rbm5* encode for RNA binding proteins and *Banf1* encodes a DNA binding protein. While the function of these genes has not been characterized in the lens, they exhibit high expression in the lens tissue. Interestingly, another Rbm family protein, Rbm24, is expressed highly in vertebrate lens development (106) and its deficiency in Zebrafish causes microphthalmia (142). Taken together, all these five genes have alternatively spliced isoforms that were differentially expressed across lens developmental stages. For example, the ENSMUSE00000145472 exon of *Banf1* is skipped at P10. Further, the ENSMUSE00000352893 exon of *Cryaa* is not highly expressed at any of the lens developmental stages tested, suggesting no potential function of the ENSMUST00000019192 transcript during late embryonic and early postnatal stages of mouse lens development. The isoform of *Eif4g2* containing exon ENSMUSE00000203223 is expressed in all developmental stages, while the alternate isoform without the exon is not as highly expressed. *Rbm5* has a distinct expression pattern during lens development. While the *Rbm5* isoform including the ENSMUSE00001225318 exon is expressed at all stages, the isoform with skipped ENSMUSE00001225318 exon is expressed highly only at P0. This suggests a potential function for the ENSMUSE00001225318 exon at early perinatal stages. Together, the RT-PCR validation analysis suggests that alternatively spliced isoforms of genes expressed in the lens are also differentially expressed at different developmental stages.

This indicates that certain isoforms of genes function specifically during embryonic or postnatal development, indicating the significant contribution of post-transcriptional regulation to the functional diversity of the isoforms.

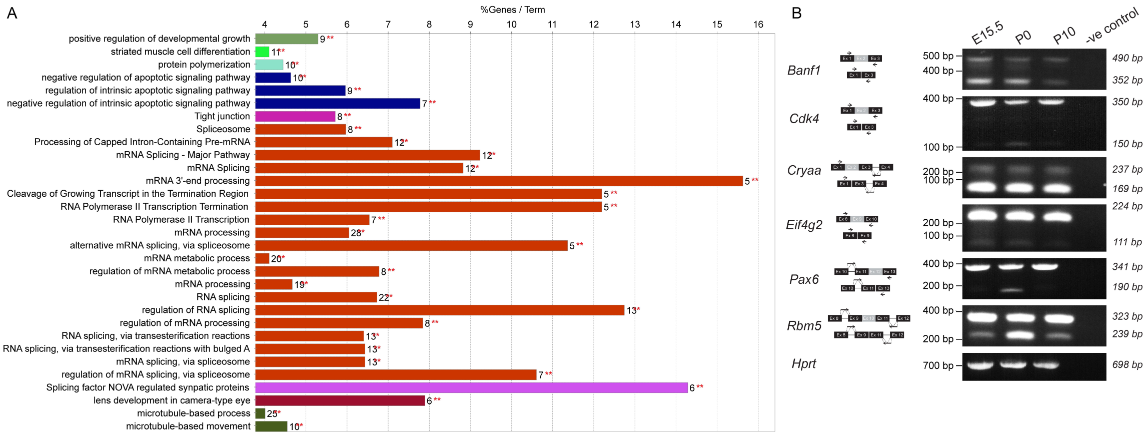


Figure 20: Functional analysis and validation of the high confident exon skipping events discovered across lens developmental states. (A) Functional enrichment analysis of genes associated with high confidence (FDR 1%) skipped exon events identified using rMATS(167) pipeline in atleast one pairwise comparison of developmental stages. For each biological process per group (color coded), the % genes per GO term with number of query genes (** in red) in the analysis is shown in histogram (B) Experimental validation by RT-PCR analysis of a selected set of high confident skipped exonic events reveals that selected mRNA isoforms with skipped events are more abundant during embryonic and perinatal stages. The schematic of the expected products are shown next to the gene. For validation, primers (arrows) were designed on the exons (black box) flanking the alternatively spliced exon (grey box). For all the genes, band with higher molecular weight is the isoform including the alternatively spliced exon and band with lower molecular weight is the isoform with the skipped exon. *Hprt* represents a loading control. Negative control is included for all isoforms tested where the RT-PCR reaction was performed using the same primers as for the isoforms but without any cDNA.

3.2.3.8 Genes associated with retained intronic events are enriched for developmental check point, cellular response to stress and RNA-splicing regulators

Retained intron (RI) is an important but less characterized AS mechanism. It causes retention of intronic region that may or may not also include some exonic regions during splicing. It is commonly suggested that, most of the transcripts exhibiting RI, could open a new targeting motif for small interfering RNA (siRNA) at RI loci, thus are degraded by nonsense-mediated decay (9). However, recent studies indicate that intron-retaining mRNAs are likely to have a more conserved role in development and numerous diseases (189). The splicing analysis indicated that retained intron events are the second most abundant alternative splicing events after skipped exon events (Table 3). I obtained 193 significant (FDR < 1%) intron retention events corresponding to 178 exons observed in 192 transcripts from 168 genes across various developmental stages (Appendix 4). Functional enrichment analysis of the genes which exhibited retained intronic events at 1% FDR threshold clearly revealed an enrichment for genes annotated with significant groups (p-value < 2.25e-04) such as 'RNA splicing', 'M Phase', 'cellular responses to stress', 'autodegradation of *Cdh1* by *Cdh1:APC/C*', 'regulation of RNA splicing', 'snRNP assembly', 'response to epidermal growth factor' and 'mitophagy' suggesting that the genes whose regulation is controlled by intron retention appear to be associated with developmental check points or stress related. For instance, several genes (*Anapc2, Anapc5, Cdk4, Ehmt2, Ensa, H3f3b, Id1, Mcm7, Ncapg, Nup35, Pole, Ppp1cc, Psmc4, Psmd11, Psmd4, Rps27a, Tpr and Trp53*) were found to be associated with cell cycle [M-Phase], which were found to be exhibiting retained introns in various developmental stages. Similarly, genes associated with 'autodegradation of Cdh1 by Cdh1: APC/C' were significantly enriched (p-value = 1.87e-07) with 15 genes (*Anapc2, Anapc5, Atg4b, Becn1, Cdk4, Ehmt2, H3f3b, Id1, Map1lc3b, Psmc4, Psmd11, Psmd4, Rps27a, Trp53, Wipi2*) contributing to 6% of the genes associated with *Cdh1* mediated proteolysis/ degradation of mitotic proteins. *Cdh1* (epithelial cadherin) is an important protein which controls the mitotic arrest with G1-phase elongation in neurogenesis(190).

3.2.3.9 Eye Splicer: an interactive web-based genome browser for visualizing alternative splicing events across lens developmental stages

To facilitate easy access to the discovered splicing events across lens developmental stages, an interactive web-based genome browser Eye Splicer (accessible via http://www.iupui.edu/~sysbio/eye-splicer/) was set up, powered by Biodalliance JavaScript library that enables visualizing skipped exon and retained intron events across developmental stages as tracks. After collecting the inclusion levels from rMATS, I converted these into BED formatted text files, which were further converted into BigBed files to make them suitable for loading into Eye Splicer (see Methods).

3.2.4 Conclusion

In this study, transcriptomic alterations and splicing events were investigated during lens formation (i.e. across different developmental stages; E15, E15.5, E18, P0, P3, P6 and P9), and constructed a molecular portrait of known and novel transcript isoforms in the mouse lens. Approximately 25% of the total transcripts were classified into partially and completely novel transcript types (PNTs and CNTs) based on their extent of overlap with current annotations, that uncovered the properties of these transcript sub-types. I found that the extent of novelty of transcripts decreased significantly in post-natal lens stages compared to embryonic stages, suggesting the presence of several uncharacterized novel transcript forms expressed during early lens development. PNTs were found to exhibit significantly higher conservation as well as expression levels compared to both completely novel and known transcripts, across the developmental stages studied here. Functional analysis of PNTs suggested the prominent role of several processes such as neural system development, structural morphogenesis, protein localization, cell division and differentiation, important for lens development. Notably, majority of the CNTs were widely expressed across developmental stages albeit exhibiting significantly lower expression, conservation and length compared to partially novel transcripts. The expression of several of these CNTs across lens developmental stages was also confirmed.

Functional analysis of the genes exhibiting the most abundant alternative splicing events, namely skipped exon and retained intron events. Several genes such as *Banf1, Cdk4, Cryaa, Eif4g2, Pax6* and *Rbm5* that are associated with lens development were found to exhibit skipped exonic events. The expression of different isoforms as well as novel genes in developing mouse lens were validated by qRT-PCR. Further, a splicing browser 'Eye Splicer' was developed to access and view developmentally altered splicing events in mouse lens. Together, this in-depth analysis provides a high-resolution architecture of the mouse lens transcriptome and provides a one-stop portal for furthering the understanding of splicing alterations during lens development.

CHAPTER 4

SYSTEMATIC TISSUE-SPECIFIC ANNOTATION OF FUNCTIONAL BINDING SITES OF RBPS IN THE HUMAN GENOME

4.1 Seten: a tool for systematic identification and comparison of processes, phenotypes, and diseases associated with RNA-binding proteins from condition-specific CLIP-seq profiles

4.1.1 Introduction

Genes are transcribed into RNAs and get matured through several layers of post-transcriptional regulation processes. These processes as well as transport, degradation and translation of the RNAs are mediated by RNA-binding proteins (RBPs) (191, 192). In cells, RNA is found to be assembled with RBPs and other proteins forming ribonucleoprotein complexes (RNPs) (193). For example, the SR protein SF2/ASF acts from alternative splicing to translation of an RNA (194). Moreover, some heterogeneous nuclear ribonucleoproteins (hnRNPs) are known to participate in RNA splicing, 3'-end processing, transcriptional regulation, and immunoglobulin gene recombination (195). Understanding these dynamic post-transcriptional regulatory networks requires the study of interactions between RNAs and RBPs. For this purpose, crosslinking and immunoprecipitation (CLIP) and related experimental protocols have been devised. All CLIP protocols involve RNA-RBP ultraviolet (UV) crosslinking followed by immunoprecipitation against an RBP of interest (196). There are several CLIP protocols: CLIP-seq, PAR-CLIP, HITS-CLIP, and iCLIP. CLIP-seq protocol involves sequencing the cDNA library created from the RNA which is previously purified by proteinase digestion after UV crosslinking and immunoprecipitation (197). For instance, PAR-CLIP (photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation) is a modified CLIP-seq technology that involves the use of photoreactive ribonucleoside analogs. These analogs can be ultraviolet crosslinked to interacting RBPs and are modified upon crosslinking. Hence, they can be used to separate RNAs bound by the RBP of interest from the background unbound RNAs (198). HITS-CLIP (high throughput sequencing of RNA isolated by crosslinking and immunoprecipitation) is another CLIP protocol that overcomes the limitation in the low number of tags by yielding more number of tags for the same cost (199). iCLIP (individual-nucleotide resolution UV crosslinking and immunoprecipitation) is yet another CLIP protocol that provides genome scale, high resolution and specificity method to enable analysis of cDNAs that are truncated at the RNA-RBP crosslink sites (200). Several computational methods have been developed for peak detection indicating the extent of binding from the data produced by these protocols. A common first step in all these frameworks before the peak detection is to map all the reads to the genome/transcriptome using algorithms such as Bowtie, RMAP, Novoalign (http://www.novocraft.com/products/novoalign/) and TopHat (201-204). After mapping, cluster detection is performed, where a read belongs to a cluster if it overlaps with at least one nucleotide with another read from the cluster. At this step in order to filter noise, reads with a length greater than a determined threshold and clusters with a minimum number of unique reads can be selected for peak detection. The most common approach for peak detection is to analyze clusters distribution profiles by improving the signal to noise ratio, and hence removing background and false positives. The softwares that use this strategy include WavClusteR, PARalyzer, Piranha, PIPE-CLIP, and dCLIP (205-210).

Although these tools are available for post-processing CLIP-seq data, there is no specific tool to either perform an enrichment analysis on such datasets nor to compare them for functional or phenotypic differences. Perhaps, the only tool which can perform gene set enrichment analysis for ChIP-seq datasets and could be configured for CLIP-seq datasets is ChIP-Enrich (211). Although, ChIP-seq and CLIP-seq protocols are fundamentally different at several levels including approaches used for cross-linking, reagents used for sequencing library preparation, efficiency of crosslinking as well as the peak calling algorithms employed, ChIP-Enrich provides an option to perform enrichment analysis of CLIP-seq processed outputs. The principle of an enrichment analysis is to associate gene sets (i.e. groups of relevant genes; e.g. processes, phenotypes or diseases) with a given study by using the fact that the co-functioning genes should have a higher potential to be detected by the high-throughput technologies (e.g. CLIP protocols). Such an approach can make the analysis of large gene lists move from an individual gene-oriented view to a relevant gene group-based analysis (212). Huang et al. (212) categorize the available enrichment analysis methods into three groups. First, singular enrichment analysis (SEA) group, enrichment p-value in these methods is calculated on each gene set from the pre-selected interesting gene list utilizing Fisher's exact, Chi-square, or binomial statistical methods. In the second group, gene set enrichment analysis (GSEA) methods, complete set of genes (without pre-selection) and corresponding experimental values are given, and they utilize Kolmogorov–Smirnov-like, t-Test, permutation, or z-score statistical methods. The last group is modular enrichment methods, which are similar to SEA but hierarchy among gene sets or genes are considered into enrichment p-value calculation by utilizing Kappa statistics and

Czekanowski-Dice Pearson's correlation (212). While these methods are available for functional analysis or functional enrichment of genes from microarray and RNA-seq with some efforts specific to RIP-chip data (213), no methods are available which can consider the binding affinity or scores of an RBPs binding potential on an RNA from CLIP-seq protocols to identify/perform an enrichment analysis using both functional and gene set enrichment approaches. Since it is increasingly being appreciated and an array of new technologies such as RBP Bind-n-Seq (214, 215) and DO-RIP-seq (216) are being developed to study the binding affinities of RBPs on target sites, it becomes important to leverage the signal strength of binding from CLIP-seq profiles for downstream functional analysis. Seten can do so by assuming that the binding score resulting from a peak calling method is a proxy for the extent of regulatory control of the RBP on the target transcript. The primary foundation of Seten (http://www.iupui.edu/~sysbio/seten/) is to identify and compare processes, phenotypes and diseases associated with RNA-binding proteins from condition-specific CLIP-seq profiles, given binding profile datasets provided as BED (Browser Extensible Data) files. Seten comes with a web interface (WI) developed in JavaScript and a command line interface (CLI) developed in Python. Seten WI provides an easy to use frontend without the need for installation and a better visualization and comparison of the enrichment results. Seten CLI can analyze multiple datasets in a single command and both the interfaces can be configured with multiple options.

4.1.2 Materials and Methods

4.1.2.1 CLIP-seq datasets

To test Seten and construct a database of precomputed functional and gene set enrichment results, peak-detected datasets from CLIPdb and ENCODE projects (217,

218) were used. Human RBP datasets were downloaded along with peak calling scores from CLIPdb and multiple samples of an RBP were merged for a cell line, which resulted in 68 unique RBP-cell line pairs. Similarly, human RBP datasets along with their detected peaks were also downloaded from the ENCODE project in BigBed format and converted to BED format using UCSC BigBed tools (125). There are 138 unique RBP-cell line pairs after merging biological replicates of RBPs within a cell line in this dataset. Additionally, an iCLIP-based peak-detected dataset for a noncanonical RBP FASTKD2, including three replicates which were merged and analyzed as a single dataset (219). Biological replicates or the datasets were merged for the same RBP-cell line pairs by concatenating their corresponding BED files using Unix cat command in order to maximize the number of binding data available per RBP-cell line. In this study, the scores associated with a detected peak from a CLIP-seq experiment are also referred to as binding affinity scores of an RBP on the target RNA because they represent a proxy measure for the extent of binding on the transcript.

4.1.2.2 Gene set collections

Gene sets are groups of relevant genes that share the same pathway, function or phenotype. Gene set collections for fruit fly, human, mouse, rat, worm and yeast were manually downloaded and organized. The gene set collections obtained are pathway annotations (BioCarta, KEGG and Reactome), Gene Ontology annotations (biological process, molecular function, cellular compartment), Human Phenotype Ontology (HPO – human only) and MalaCards Disease Ontology (human only) (220-226). The number of gene sets in gene set collections and the availability of organisms are given in Table 4.

Gene set collection	Fruit fly	Human	Mouse	Rat	Worm	Yeast
BioCarta	NA	314	276	NA	NA	NA
KEGG	135	302	298	298	134	113
Reactome	1171	1972	1477	1441	1024	822
GO Biological Process	4104	11176	11014	11325	2867	2886
GO Molecular Function	2174	3944	4453	3803	1714	1909
GO Cellular Compartment	841	1508	1473	1465	737	757
Human Phenotype Ontology	NA	7268	NA	NA	NA	NA
Malacards Disease Ontology	NA	9815	NA	NA	NA	NA

Table 4: The number of gene sets in gene set collections and the availability of organisms. NA entries represent that the corresponding gene set collection is not available for that organism.

4.1.2.3 Obtaining distinct gene scores list

Binding sites from the input BED file are mapped onto their corresponding gene symbols using a mapping table downloaded from Ensembl for each available organism (121). After mapping is complete, in the case where multiple scores are available for a gene, multiple methods were provided to obtain a single score to represent that gene, which results in distinct set of genes and their corresponding scores representing the extent of binding by an RBP. The available methods are maximum, minimum, mean, median, and sum. Therefore, for instance, if the selected method is sum, then the final score given to the corresponding gene will be sum of all scores available for that gene. The default method selected is ‘maximum’.

4.1.2.4 Gene set association analysis

To apply gene set association analysis, a previously reported competitive method was applied to transcription factor binding datasets in order to test if an RBP preferentially targets to genes in a given gene set (227). This method finds the common genes between given RBP targets and genes in given gene set and compares the scores of common genes to the scores of randomly permutated genes from RBP targets by a competitive test where Mann Whitney U test is used to test if the median score of the common genes is significantly higher than that of randomly permutated genes (228). Options to set thresholds for maximum number of genes in a given gene set to allow more specific gene sets to be used (defaults to < 350) and minimum number of common genes between RBP targets and genes in given gene set (defaults to > 5) were provided. Additional option to control the number of permutations to perform (defaults to 1000) was also provided. At each permutation, the method checks if the p-value from Mann Whitney U test is significant using another option (defaults to < 0.05) and counts the significant tests. At the end, the final corrected p-value is computed as

$$max\,(1 - \frac{\#\,sign.\ tests}{\#\,total\ tests}, \frac{1}{\#\,total\ tests}).$$

Such corrected p-values resulting from gene set enrichment analysis are referred to as p-values for brevity.

4.1.2.5 Functional association analysis

Further, a functional association analysis was implemented using a two-sided Fisher’s exact test (FET) for traditional functional enrichment (229). A correction method is used to correct the p-values obtained from functional enrichment analysis (Fisher's exact test or FET). Currently, Seten's web interface has only one method, which is false discovery rate (FDR) or Benjamini-Hochberg(230). Seten's command line interface includes several other methods or correcting the resulting p-values. Note that such correction methods are only available for functional enrichment analysis as the gene set enrichment method employs a different correction approach as described above.

4.1.2.6 Processing CRISPR RNA-seq datasets of RBPs

Clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 is a recently developed system for engineering genomes, which has transformed the ability to manipulate genes in cell lines and animal models (231). In ENCODE project multiple RBPs have been screened using the CRISPR/Cas9 system followed by RNA-sequencing to better understand the downstream pathways impacted by the loss of function of an RBP. Hence, in order to generate a reference gold standard set of functional annotations which are effected by an individual RBP and as a means of benchmarking the quality of the annotations predicted by Seten and ChIP-Enrich (C-E) from CLIP-seq data, I obtained RNA-sequencing data from non-specific CRISPR control and those treated with gRNAs against three different RBPs, namely IGF2BP1, SRSF7 and PTBP1 in K562 cells (232). Since these RBPs had both eCLIP and CRISPR RNA-seq datasets available, they were ideal for performing a benchmarking analysis. This dataset comprised of eight non-specific CRISPR control RNA-seq datasets representing wildtype K562 cells and two replicate RNA-seq datasets each for the RBPs IGF2BP1, SRSF7 and PTBP1 where in gRNAs were used to deplete the functional form of RBPs. This enabled a quantitative differential expression analysis followed by gene set enrichment for various gene set collections using Seten, to develop a gold standard. In brief, all the available RNA-seq data for CRISPR control and knock-out data for multiple RBPs in K562 cell line was collected and processed the raw quality filtered (Phred Score > 30) sequence reads using HISAT (117) and StringTie (120) pipeline with default parameters, to generate gene expression levels in Transcripts Per Million (TPM) reads for all human annotated Ensembl genic features (121). The processed and gene expression quantified data were formatted into expression matrices and utilized for generating a reference set of functional annotations impacted by the respective RBPs as described below.

4.1.2.7 Generation of gold standard set of functional annotations using CRISPR RNA-seq datasets of RBPs

Gene expression matrices comprising of CRISPR control and knock-out for each RBP were used to compute a relative change in expression for each gene. Relative change in expression is defined as the ratio of the absolute change in the expression difference between the mean of replicates of control and knock-out respectively, divided by the mean expression level of the gene in the control RNA-seq datasets. By utilizing such a normalized relative change in expression of each gene across the entire genome for each combination of control and CRISPR knock-out datasets of an RBP, gene set enrichment analysis was performed using Seten for both reactome and GOBP gene set collections. This enabled the identification of gene sets enriched due to the loss of an RBP at a corrected p-value of 0.05 using the Seten's GSEA approach. Such gene sets have been defined in this study as the gold standard annotations for the RBP for the corresponding gene set collections. By utilizing these annotations, precision and recall values were computed for Seten and C-E to assess the performance of the tools. Precision was defined as the fraction of enriched gene sets from GSEA on the control vs CRISPR RNA-seq data for the respective RBPs, which overlapped with the gene sets from Seten's or C-E's GSEA on CLIP-seq data at the same corrected p-values threshold of 0.05. Likewise, recall was defined as the fraction of gene sets identified by Seten's or C-E's GSEA on CLIP-seq data which overlapped with the enriched gene sets from GSEA on the control vs CRISPR RNA-seq data for the respective RBPs, at the same corrected p-value thresholds. Similarly, precision and recall were also computed for the negative control bed files described below for Seten's gene set or functional enrichment methods, which enabled the calculation of F1-scores to assess the relative performance of the methods and options.

4.1.2.8 Evaluation of Seten's performance against negative control

In order to evaluate the performance of the tool, random bed files referred to as negative controls were generated, corresponding to each RBP's eCLIP dataset separately. I utilized bed tools (233) shuffle function with –chrom (to ensure that each chromosome is equally represented in random bed files), '– incl' (that keeps genomic features and assigns shuffled scores for peaks) and separately '– excl' (that excludes the genomic features and assigns random genome wide coordinates for each peak) parameters to generate two sets of arbitrary bed files. These two sets of negative control bed files were referred as 'Test Peaks with randomized peak score' and 'Test Peaks with randomized peak coordinates'. Next, the F1 score was computed as the harmonic mean of precision and recall, to measure the performance of Seten against gold standard functional annotations described in the previous section, for three types of bed files namely original eCLIP Peaks, Test Peaks with randomized peak score and Test Peaks with randomized peak coordinates. This enabled the assess to the relative impact on the performance, for different options and to benchmark the annotations predicted by Seten for each of these types of bed files against the GSEA results obtained from CRISPR RNA seq gold standard described above. The analysis was repeated for three different RBPs which had both eCLIP and CRISPR RNA seq data, namely IGF2BP1, PTBP1 and SRSF7 in K562 cell line. For test peak data, the analysis was repeated against 5 random bed files for each RBP and reported the average F1 scores.

4.1.2.9 Availability

Seten WI (Web Interface)

Seten WI server is accessible on http://www.iupui.edu/~sysbio/seten/. Its source code which can be used for initiating a local instance of Seten WI is available via the GitHub repository: https://github.com/gungorbudak/seten.

Seten CLI (Command Line Interface)

Seten CLI is a Python package and can be installed via the package manager or can be built from its source. Its GitHub repository has detailed information about installing and using Seten CLI: https://github.com/gungorbudak/seten-cli.

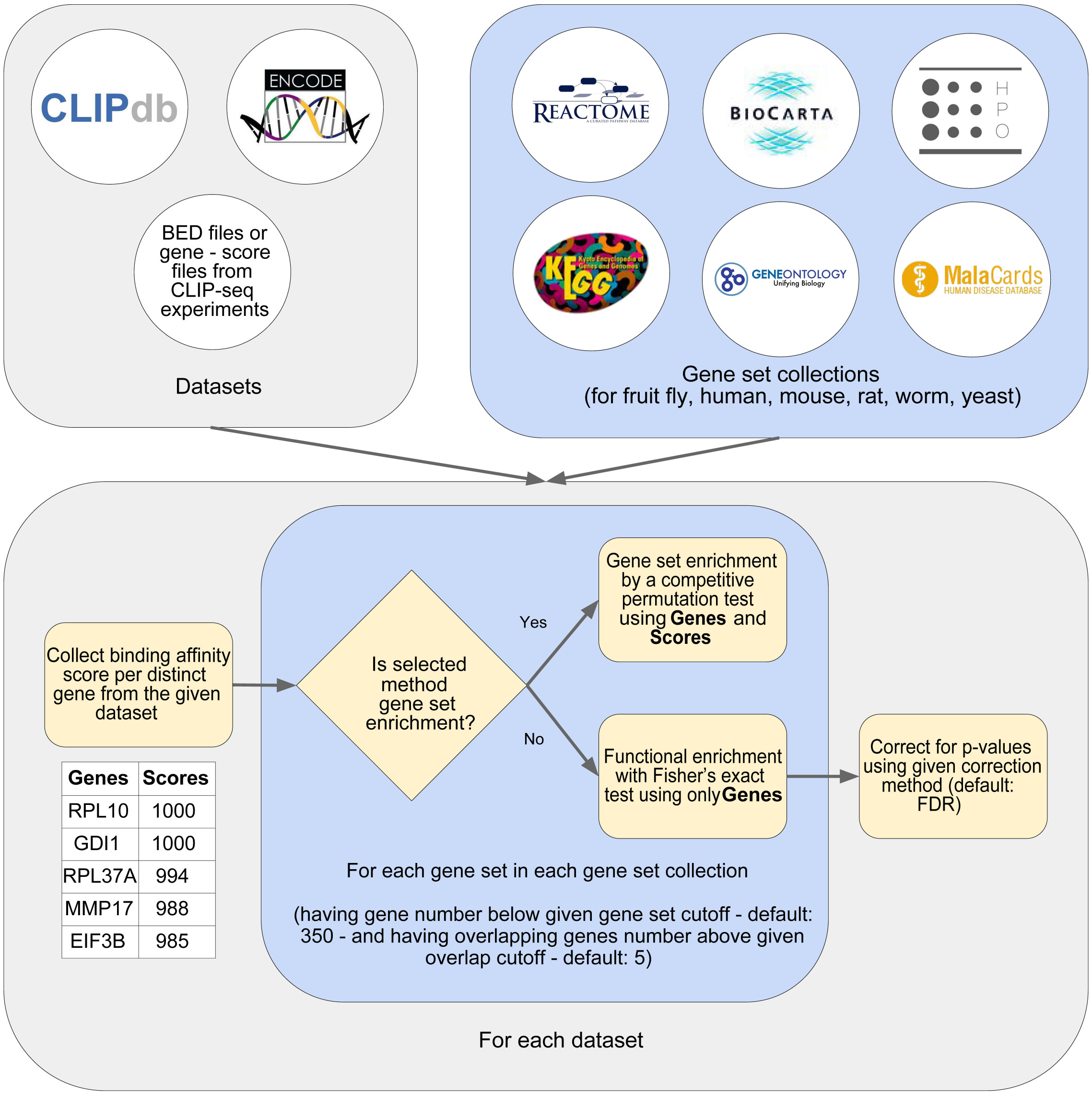
4.1.3 Results and Discussion

4.1.3.1 Overview of Seten

In Seten, for each input BED file which has at least the 5 columns namely chromosome, chromosome start, chromosome end, feature name and score associated with the binding of an RBP resulting from running a peak calling algorithm on a genome aligned CLIP-seq dataset, a Gene Set Enrichment Analysis (GSEA) is performed against the gene set collections selected by the user (Figure 21, see Materials and Methods). Both the web interface (WI) and command line interface (CLI) versions of Seten currently support the gene sets from BioCarta, KEGG, Reactome, Gene Ontology (GO) biological process, GO molecular function, GO cellular compartment, Human Phenotype Ontology (HPO) and MalaCards Disease Ontology for organisms including fruit fly, human, mouse, rat, worm and yeast with the CLI allowing the user to include additional gene set collections and organisms (Table 4). To facilitate easy access and navigation of existing CLIP-seq datasets, Seten WI includes results from precomputed functional analysis of 68 human RBPs obtained from CLIPdb as well as 138 human RBPs profiled in the ENCODE project. In addition to providing precomputed integrated functional analysis of dozens of CLIP-seq experiments, Seten WI also provides user friendly interface to compare the resulting annotations across experiments and RBPs as exportable bubble charts.

4.1.3.2 Seten's gene set enrichment outperforms functional enrichment, with peak scores contributing to the discovery of true annotations

In order to evaluate the performance of Seten, the gene set and functional enrichment implementations were employed in Seten and compared their performance against ‘negative control’ bed files generated using bedtools (233) for eCLiP peaks of RBPs (IGF2BP1, PTBP1, SRSF7 in K562 cell line) (See Materials and Methods). For each eCLIP dataset and their corresponding negative controls, the performance of the gene set and functional enrichment implementations were benchmarked against the annotations identified using CRISPR RNA-seq gold standard for the corresponding RBPs in K562 cell line (See Materials and Methods). Seten was run using default parameters (i.e. corrected p-value (< 0.01), gene set size (< 350), number of gene set hits per RBP (> 10)) for both the eCLIP and each negative beds separately. F1 score, which is the harmonic mean of precision and recall, was computed for respective Seten runs against CRISPR gold standard annotations (Materials and Methods). For random data, the process was repeated for five random negative controls for each RBP and report the average F1 score for each RBP, as shown in Figure 22.


CLIPdb
ENCODE
BED files or gene - score files from CLIP-seq experiments
Datasets
REACTOME
BIOCARTA
HPO
KEGG
GENEONTOLOGY Unifying Biology
MalaCards HUMAN DISEASE DATABASE
Gene set collections (for fruit fly, human, mouse, rat, worm, yeast)
Collect binding affinity score per distinct gene from the given dataset
Is selected method gene set enrichment?
Yes
No
Gene set enrichment by a competitive permutation test using Genes and Scores
Functional enrichment with Fisher's exact test using only Genes
Correct for p-values using given correction method (default: FDR)
For each gene set in each gene set collection
(having gene number below given gene set cutoff - default: 350 - and having overlapping genes number above given overlap cutoff - default: 5)
For each dataset

Genes	Scores
RPL10	1000
GDI1	1000
RPL37A	994
MMP17	988
EIF3B	985

Figure 21: Overview of Seten. Peak-detected datasets (in bed format) from RBP-specific CLIP-seq studies, CLIPdb and ENCODE projects (217, 218) are obtained as bed files to provide input to Seten. Several gene set collections were organized for multiple genomes including fruit fly, human, mouse, rat, worm and yeast. Currently included gene set collections comprise of pathway annotations (BioCarta, KEGG and Reactome), Gene Ontology annotations (biological process, molecular function, cellular compartment), Human Phenotype Ontology (HPO – human only) and MalaCards Disease Ontology (human only) (220-226). Scores associated with each gene from a BED file are employed for gene set enrichment analysis by organizing the scores according to the chosen scoring method. Scores mapped onto the genes are used to compute an enrichment using a competitive permutation test and corrected p-values from multiple testing are reported. In contrast, functional enrichment method only uses the associated genes and not the scores from BED files, for enrichment analysis using fisher's exact test and computes a false discovery rate.

The results show that gene set enrichment exhibits relatively higher performance than functional enrichment for both the Reactome and GO Biological Process annotations (Figure 22). F1-scores were also found to be significantly higher for eCLIP data compared to the negative controls resulting from randomized scores or genomic co-ordinates, for gene set enrichment method (Figure 22). In contrast, for functional enrichment, although results compared to randomized co-ordinates were higher, there was no significant difference in F1-scores compared to the randomized peak scores suggesting that while functional enrichment is not impacted by the scores, gene set enrichment implementation has a significant improvement due to the use of scores (Figure 22). Overall, although the maximum score was employed for each gene as the scoring method, the analysis demonstrates that Seten's gene set enrichment implementation is likely to outperform functional enrichment for inference of annotations from eCLIP profiles, by exploiting the scores which can act as proxy for the extent of binding.

4.1.3.3 Cell type-specific gene set associations can be identified by Seten

The results (Figure 22) also suggest that it is possible to not only identify the gene set associations of an RBP but RBPs profiled in different cell lines and conditions can be compared for one or more gene set collections. Such a feature is available in Seten WI for both precomputed CLIP-seq datasets as well as for user uploaded BED formatted CLIP results.

Seten can compare one or multiple gene set collections across conditions/cell lines of one or more RBPs to dynamically generate bubble charts for easy comparison of differences in the significance of associated gene sets. In both CLIPdb and ENCODE

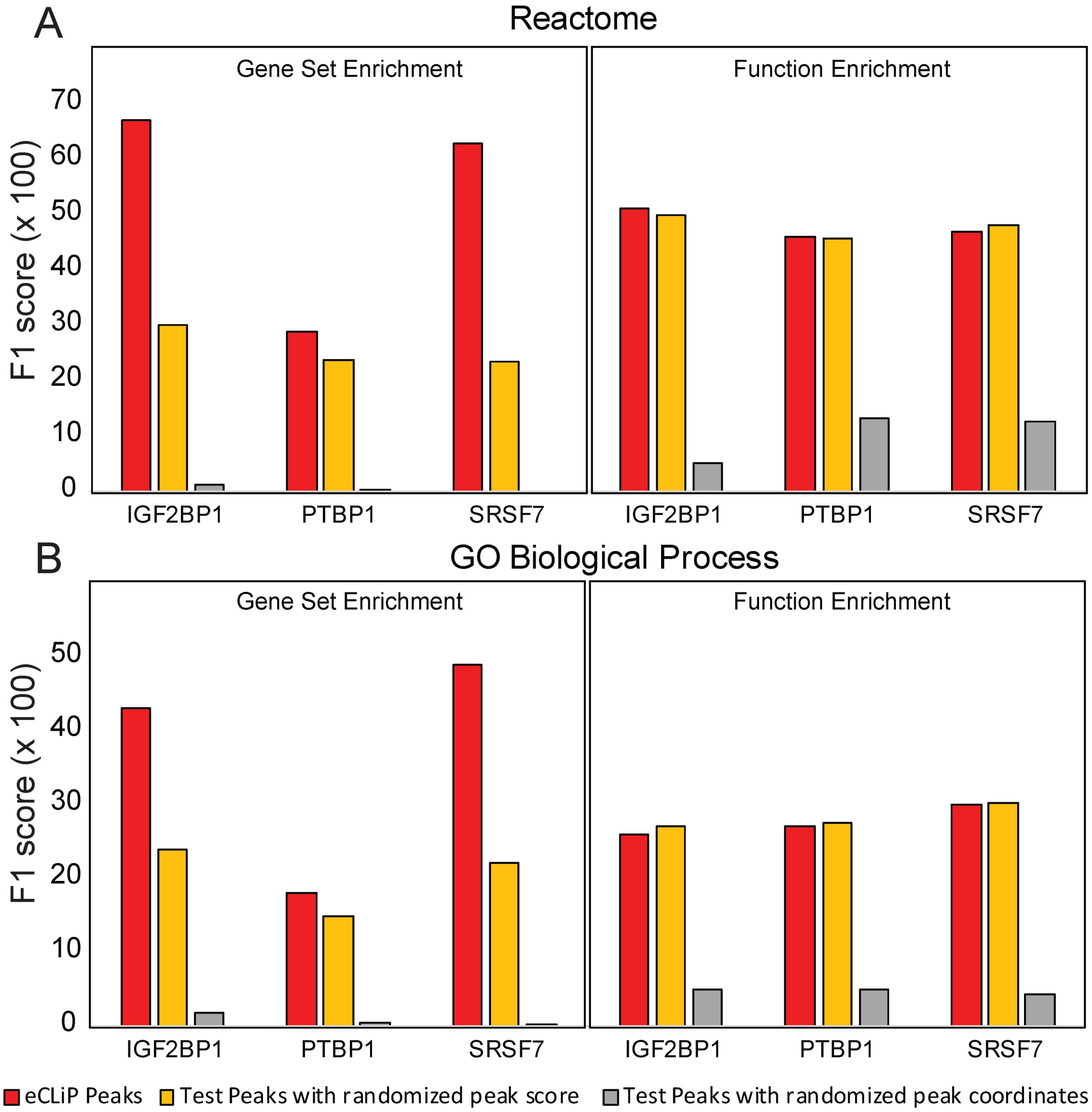


Figure 22: Comparison of Seten's gene set and functional enrichment methods against negative control. Histograms showing the performance comparison of Seten's Gene Set Enrichment Analysis (GSEA) and Functional Enrichment (FE) options along with their corresponding random bed files for each RBP, benchmarked against their CRISPR RNA-seq gold standard. F1 score, harmonic mean of precision and recall, represented on y-axis for each dataset/option was computed against CRISPR gold standard separately for (A) Reactome and (B) GO Biological Processes by running Seten using eCLiP peaks (in red) and 'negative control' peaks (Test Peaks with randomized peak score shown in orange, Test Peaks with randomized peak coordinates shown in grey). Negative control bed files for each RBP were generated using bed tools, as described in Materials and Methods.

datasets, some RBPs have CLIP data from different cell lines which allowed the usage of Seten for comparing these cell type-specific datasets.

The two FASTKD2 datasets discussed in the previous section also exhibit cell line specific differences as shown in Figure 23A. While the FASTKD2 / K562 (human bone marrow cell line having chronic myelogenous leukemia) gene set enrichment results show pure red-cell aplasia (p-value = 0.001) and diamond-blackfan anemia (p-value = 0.001), the other FASTKD2 / HEK293 does not exhibit these disease annotations.

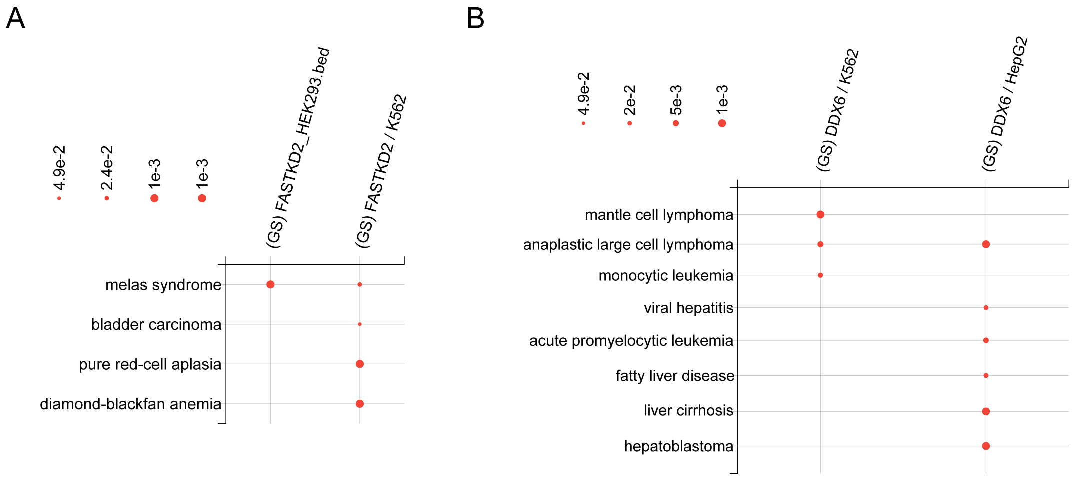


Figure 23: (A) The dynamically generated bubble chart from Seten WI, showing the comparison of significantly enriched MalaCards Disease Ontology terms for FASTKD2 in HEK293 and K562 cell lines. (B) The inset of a dynamically generated bubble chart from Seten WI, showing the comparison of significantly enriched MalaCards Disease Ontology terms for DDX6 in K562 and HepG2 cell lines. Only gene sets which had more than 5% of the total genes and exhibited a minimal p-value of 0.05 in one of the cell lines are included in this comparison. The radius of bubbles is computed as negative Log10(corresponding p-value).

DEAD-box helicase 6 (DDX6) is an RNA helicase found in P-bodies and stress granules and it functions in mRNA degradation and translation suppression (8). It has been shown to be contributing to lymphoma genesis by deregulation of BCL6 (B-Cell CLL/Lymphoma 6) in nodal marginal zone lymphoma (234). It has also been shown to be required for efficient hepatitis C virus replication (235). Figure 23B (a subset) shows a bubble chart comparing the significance scores for MalaCards Disease Ontology term associations for DDX6 / K562 (chronic myelogenous leukemia cell line) and DDX6 / HepG2 (hepatocellular carcinoma cell line). As is evident from the chart, while monocytic leukemia (p-value = 0.0110) is specific to K562 sample, fatty liver disease (p-value = 0.0140), liver cirrhosis (p-value = 0.001), and hepatoblastoma (p-value = 0.001) are specific to HepG2 sample. However, acute promyelocytic leukemia was also observed to be significant (p-value = 0.008) for DDX6 / HepG2. This might be seen because DDX6 activation has been observed in acute leukemia (236). Additionally, mantle cell lymphoma (p-value = 0.001 for DDX6 / K562) and anaplastic large cell lymphoma (p-value = 0.005 for DDX6 / K562 and 0.001 for DDX6 / HepG2) appear as significant hits. Moreover, viral hepatitis is one of the significant hits for DDX6 / HepG2 (p-value = 0.0150). These results suggest that Seten can be employed to study and navigate condition, cell line as well as tissue-specific variations in the gene set associations for RBPs, starting from CLIP-seq data.

4.1.3.4 Seten’s GO Biological Process and Reactome results agree with ChIP-Enrich gene set enrichment tool results

Since there are no existing tools to perform a gene set enrichment analysis on CLIP-seq datasets, in order to compare the results, ChIP-Enrich (C-E) gene set enrichment tool originally developed for ChIP-seq datasets was used by configuring its options to make it suitable for CLIP-seq datasets (211). Locus definition option was set to “Nearest gene” to assign all peaks to the nearest gene which is similar to implemented approach. The comparison was limited to GO Biological Process (GOBP) from Gene

Ontology and Reactome from pathway databases. The gene sets having more than 350 genes to be consistent with the default threshold in Seten were filtered out. Both tools for several datasets from ENCODE project were run. To compare the results, enriched gene sets using p-value threshold (corrected p-value < 0.05 in the respective tools) were filtered and then ranked them separately. Finally, the overlapping gene sets were taken between them and did a hypergeometric test to determine the significance of the overlap between the two approaches. First, the Alanyl-tRNA Synthetase (AARS – K562) GOBP and Reactome results were compared, which yielded a GOBP p-value of 6.15e-14 and a Reactome p-value of 5.44e-27 (Hypergeometric test) indicating a significant agreement of the discovered processes and pathways between the two methods (Figure 24A).

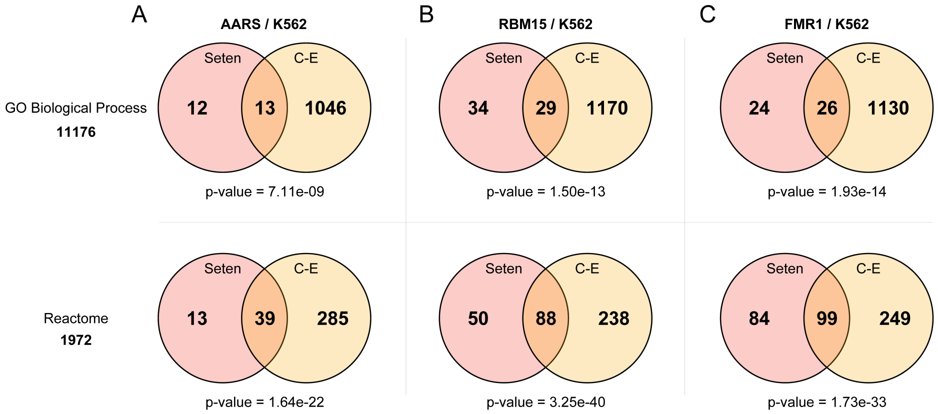


Figure 24: (A) The comparison of Seten and ChIP-Enrich using AARS – K562 dataset for GO Biological Process and Reactome gene set enrichment analysis results. (B) The comparison of Seten and ChIP-Enrich using RBM15 – K562 dataset for GO Biological Process and Reactome gene set enrichment analysis results. (C) The comparison of Seten and ChIP-Enrich using FMR1 – K562 dataset for GO Biological Process and Reactome gene set enrichment analysis results.

Next, the Putative RNA-binding protein 15 (RBM15 – K562) results for GOBP and Reactome gene set collections were compared to obtained a GOBP p-value of 6.42e-25 and a Reactome p-value of 2.70e-49 suggesting a significant overlap (Figure 24B).

Finally Fragile X Mental Retardation 1 (FMR1 – K562) GOBP and Reactome results were compared, which yielded a GOBP p-value of 2.59e-24 and a Reactome p-value of 6.90e-35 indicating the reproducibility of the enriched processes/pathways between the methods (Figure 24C).

4.1.3.5 Benchmarking of Seten and ChIP-Enrich against CRISPR RNA-seq reveals superior performance of Seten

Recent progress in utilizing CRISPR/Cas9 technologies for genome editing has enabled rapid sequencing-based profiling of genomic phenotypes (231). Although majority of the RBPs are known to be encoding for essential genes (237), ENCODE project has been successful in generating RNA-sequencing data of CRISPR/Cas9 based knock-outs of several RBPs including IGF2BP1, SRSF7 and PTBP1 in human K562 cell line (238). Hence, to generate a gold standard set of functional annotations impacted by these RBPs and to benchmark both Seten and ChIP-Enrich tools against this common reference set for which both eCLIP and CRISPR data are available, the CRISPR RNA-seq data was processed and organized as described in Materials and Methods. By utilizing the functional annotations obtained from gene set enrichment analysis of the relative gene expression changes from CRISPR control vs knock-out for each of these RBPs, as gold standard, the performance of both the tools was compared against this reference by computing precision and recall (see Materials and Methods). As shown in Figure 25, for each of these three RBPs, Seten was found to exhibit significantly higher precision for both Reactome and GO Biological Process annotations compared to that observed for ChIP-Enrich (C-E). Seten exhibited an average precision of 72% and 58% for Reactome and GOBP gene sets. In contrast, C-E was found to show an average precision of 42% and 8% respectively, indicating that Seten is more suitable for functional annotation of CLIP-seq data (Figure 25). Comparison of the average recall values between the tools indicated that, while Seten exhibited higher recall than C-E for Reactome (51% vs 47%), inverse trend was seen for GOBP annotations (32% vs 45%). A major contributor to the lower average recall of Seten is PTBP1, which was found to exhibit a relatively lower recall for both Reactome and GOBP annotations. In this context, it must be reminded that not all RBP's loss of binding events might result in corresponding changes in RNA expression levels of their targets - a major assumption in the calculation of recall. This could be due to a number of reasons such as A) redundancy in the functionality of RBPs where a paralogous RBP might complement the function of the mutated RBP, B) RNA levels might not be impacted but protein levels might be impacted C) quality of the binding site might be low or functional impact of the binding site might be minimal. Nevertheless, although the number of RBPs with both eCLIP and CRISPR data is currently limited, it is possible to conclude from this data that Seten achieves significantly higher precision and comparable recall as that of C-E.

It is important to note that currently there are very few high-resolution CRISPR datasets which stand orthogonal to CLIP-seq profiles. Also, since CRISPR screens are still in their infancy, it is unclear to what extent do they strictly identify only the direct effects of regulatory molecules like RBPs and not secondary off-target effects (231). Hence, additional orthogonal approaches to probe and measure the genome-wide impact due to the loss/gain of function of an RBP are needed to comprehensively understand, model and improve the functional annotations of RBPs using CLIP-seq profiles.

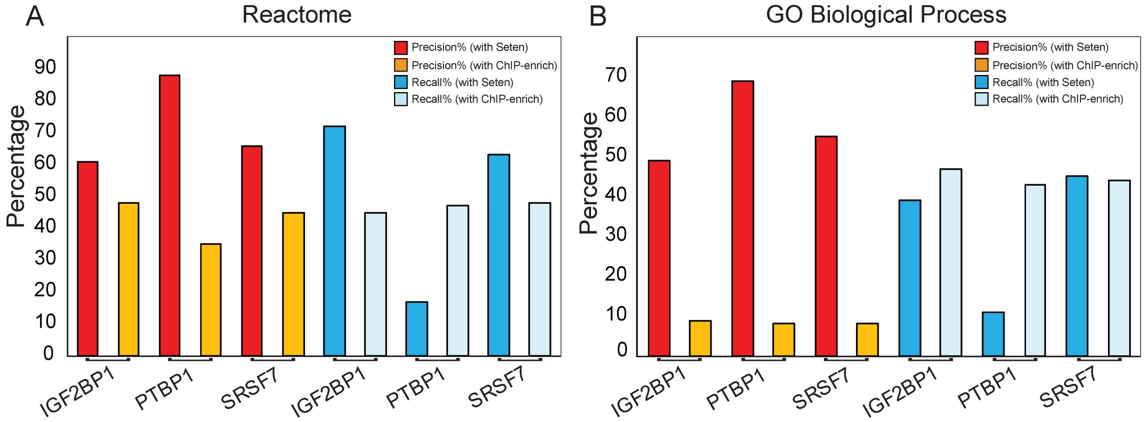


Figure 25: Benchmarking of predicted functional annotations from Seten and ChIP-Enrich against those identified from CRISPR based RNA-seq datasets of RNA-binding proteins in K562 cell line. Precision and recall plots for IGF2BP1, SRSF7 and PTBP1 using Seten and ChIP-Enrich for the gene set collections (A) Reactome and (B) GO Biological Process. In both cases, gene set enrichment approach as implemented in the respective tools was utilized to generate functional annotations from eCLIP-based profiles, to compare their relative performance. Seten was found to exhibit a significantly higher precision and comparable recall to that observed for ChIP-Enrich.

4.1.4 Conclusion

Seten is a computational framework that performs an enrichment analysis using the scores resulting from peak calling algorithms on CLIP-seq datasets. This tool can also perform a comparison of the identified processes and phenotypes across a set of profiled RBPs both within and across conditions or tissue types being studied. Thus, this study fills the gap in current understanding of the downstream biological context resulting from extensive rewiring in post-transcriptional networks. Seten is implemented as a web interface (WI) using JavaScript and a command line interface (CLI) using Python (http://www.iupui.edu/~sysbio/seten/). Seten WI provides exportable visualizations of results as bar charts and bubble charts (in SVG format) and requires no installation or dependency except for an up-to-date browser. Using Seten CLI, multiple datasets can be analyzed using a single command.

4.2 SliceIt: A genome-wide resource and visualization tool to design CRISPR/Cas9 screens for editing protein-RNA interaction sites in the human genome

4.2.1 Introduction

CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) is identified as a defense system that protects bacteria and archaea from mobile genetic elements (239-241). This RNA guided interference mechanism has been successfully employed in eukaryotic cells (both in vitro and in vivo) by Zhang (242) and Charpentier (243) groups. It is now established that sgRNAs (single guide RNAs) can be engineered to target a 17-20 bp stretch of DNA sequence preceding a protospacer adjacent motif (PAM) (244). CRISPR/Cas9 system has been crucial in multiple disciplines but is especially useful for understanding the gene function by manipulating precise genomic locations (244-246). This emerging technology enables the re-investigation of multilayered functional dependency of regulating molecules such as kinases, transcription factors (TFs), long non-coding RNAs (lncRNAs), and other protein coding gene groups, to provide specific perturbations and to study their comprehensive genome wide effects (245, 247-250).

RNA binding proteins (RBPs) post-transcriptionally regulate a variety of biological processes as described previously (192, 193, 251-253). Several crosslinking and immunoprecipitation (CLIP)-seq protocols have been developed over the years (254-256) to delineate the molecular interaction of RNA binding proteins (RBPs) and their target RNAs at single-nucleotide resolution in a cell. However, most of the millions of binding sites identified from these high throughout CLIP-seq studies do not have functional evidence for their contribution to the fate of the RNA molecule, except perhaps binding the RNA target, creating an ambiguity with little or no functional relevance of these interactions in cellular context (257, 258). It has also been argued that various CLIP-seq protocols don't agree with each other in recovering binding sites and often produce noisy signals resulting in a number of false positive binding sites (259). With the advent of genome modification via CRISPR/Cas9 systems (244-246), it is possible to investigate the functional connectivity between RBPs and RNAs, specifically with high resolution. CRISPR/Cas9 system can potentially be employed to investigate the functional aspect of localized RBP-RNA interactions in cells (Figure 26). This system, though originally developed to directly target and edit DNA, recent studies also report the use of variants of Cas9 for tracking RNA (260, 261). However, the system's efficiency to access/edit RNA is highly compromised with low signal to noise ratio and accuracy. In addition, editing RNA (which can't be repaired by cellular pathways) will not enable the use of expression of the target RNA molecule as a proxy to measure the functionality of the binding site. CRISPR/Cas9 system is precise and cheaper compared to other gene editing techniques. Apart from single target site perturbation, it can also be used to target multiple loci simultaneously with different sgRNAs and using a single Cas9 variant. The ability of using single Cas9 protein with multiple sgRNAs opens the doors for high throughput editing of target loci. Therefore, to understand the impact of RBP binding sites perturbation in human cells, CRISPR/Cas9 system is proposed to edit the DNA locations where RBP binds to RNA.

SliceIt (7) is the first comprehensive database of in silico sgRNA library to edit the currently known millions of protein-RNA interaction sites in the human genome. It stands as a one-stop portal for designing CRISPR/Cas9 screens for functional dissection of post-transcriptional regulatory networks. SliceIt enables the designing of multiple sgRNAs for each binding site based on user's desired location for editing the genome by defining specificity and efficiency thresholds, to facilitate uncovering their functional role in modulating post-transcriptional regulation of a transcript. Predicted sgRNAs are available to be visualized in a genome browser with additional layers of information such as SNPs and cis-exon expression across human tissue types. SliceIt also provides an option to download data for every search query in CSV or Excel file format. SliceIt uses Flask micro framework in the backend to efficiently parse results and to communicate with user interface and the databases. In order to handle multiple queries efficiently, dynamic implementation of multiprocessing is used to parallelize querying process from Elasticsearch cluster to effectively reduce query search time by several fold.

4.2.2 Materials and Methods

4.2.2.1 Data collection and processing

eCLIP (262) based binding profiles of scores of RBPs in (hepatocellular carcinoma (HepG2) and chronic myelogenous leukemia (K562) cell lines was obtained from the ENCODE project (70). In total, the included dataset comprised of 2.23 million unique binding sites for 68 RBPs in HepG2 and 2.38 million unique binding sites for 86 RBPs in K562 cell line. The genomic coordinates of these RBPs' eCLIP profiles (in .bed file format) was downloaded, parsed and used for prediction of sgRNAs localized to each binding site of RBPs. In addition to CLIP profiles, dbSNPs was also downloaded (263), GWAS catalog (16) and exon expression profiles for 53 human tissues from the GTEx project (264) (GTEx data was extracted by using recount workflow (265, 266)) and integrated with SliceIt, to provide comprehensive information for a genomic region of interest.

4.2.2.2 Prediction of sgRNAs around RBPs' binding sites

A typical sgRNA comprises of a 19-20 base long oligonucleotide that could be designed to target user defined homologous sequence on the host genome. In this study, CRISPR-DO was used (244) to design sgRNAs targeting all possible RBP binding sites from various CLIP experiments (262) in HepG2 and K562 cell lines from ENCODE (70). CRISPR-DO requires an input genomic region in bed format along with other essential metrics such as genome assembly and spacer length. The RBP binding site was customized coordinates obtained from the ENCODE project for human reference genome (hg38) and employed a flanking distance of ±50bp from the mid-point of the binding site (if BS length is <100 bp) and automated the standalone CRISPR-DO software (http://cistrome.org/crispr/source) to predict potential sgRNAs per binding site using ad-hoc scripts. This enabled the development of an in-house pipeline for sgRNA design to millions of RBP binding sites on a compute cluster.

4.2.2.3 Processing of CRISPR DO outputs

CRISPR-DO provides genomic location, 30 bp sequence (i.e. 20 bp sgRNA sequence + PAM + 7 bp flank sequence), sgRNA strand orientation, specificity score, efficiency score and other flagged annotations for all possible sgRNA predictions per genomic region of interest. Predicted sgRNAs and other information for each binding site were tagged with respective queried binding site (BS) coordinates and corresponding cell line as well as RBP information. These predictions were concatenated into a single file. The distance between the mid-point of the BS and PAM site was also computed, to include it as an additional column in the processed outputs.

4.2.2.4 Database construction and implementation

SliceIt is a database implemented in elasticsearch with a web interface that was developed using Bootstrap 4.0 software. The database is hosted at https://sliceit.soic.iupui.edu/. SliceIt currently allows users to search by gene name or Ensembl ID, coordinates and by RNA binding protein name at various efficiency and specificity thresholds.

Data Interface

SliceIt interface comprises of three different search options to query the data and the search result of all of these primary functions include (i) Retrieval of information on guide RNAs corresponding to the binding sites (ii) Retrieval of SNPs and GWAS information that fall within the region of interest (iii) exon expression levels across tissues that are in 500 bp proximity to region of interest and are defined as cis-exons in this study (iv) Visualization of data in IGV JS genome browser. When a user queries for a gene, SliceIt provides a list of various genes and their corresponding Ensembl IDs in the form of an auto suggest dropdown box. For each search the recommended and default cut-off for efficiency and specificity scores for selection of sgRNAs are 0.3 and 50 respectively. These numbers are recommended for selection of optimal guide RNA design by CRISPR-DO tool. Results retrieved in SliceIt are organized into the following sections

Annotations

The output page displays annotation information obtained from Ensembl for search based on gene name or coordinate range within a chromosome. These annotations cover the gene name, description, location coordinates, strand information along with a link to Ensembl database to access more information.

Genome Browser

SliceIt helps visualize the locations of binding sites and sgRNAs with the help of IGV JS genome browser embedded in the output page. By default, the tracks that are displayed include (i) hg38 reference genome (ii) GWAS (iii) SNP from dbSNP (iv) Binding sites in HepG2 cell line (v) Binding sites in K562 cell line (vi) Predicted sgRNAs. The users have an option to remove and add any of these tracks, change track color, name and height. Apart from the default tracks that are loaded for every search query, users can also load various other tracks by using drop down menus. These options for additional tracks include exon expression tracks for various tissues, HepG2 and K562 binding sites tracks for individual RBPs. SliceIt also has an additional functionality that allows users to add their own data in the form of a track in genome browser by using the "Add custom track from URL" section on the results page. This allows users to add indexed BED, BAM, Wig, Bigwig and BedGraph file formats that are hosted on an external server.

Data Tables

The retrieved raw data is displayed in "Data View" tab in a tabular form with various options such as CSV and Excel export, search filter by coordinates, efficiency or specificity and column sorting. For each binding site, SliceIt provides 5 different sgRNAs that are filtered based on the highest efficiency and specificity scores. The data view also has 3 other sub-tabs that provide data regarding SNPs, GWAS and Cis-exon expression.

Backend

In the backend, SliceIt runs on Python's Flask micro web framework to efficiently process the query, parse the data and return output. Predicted sgRNAs, dbSNPs, GWAS and Exon expression data is stored in elasticsearch (https://www.elastic.co/) that is hosted on an external cluster. For each search, the query input is passed to the backend via flask framework and SliceIt automatically designs various queries to efficiently retrieve data from elasticsearch and Ensembl.

4.2.2.5 sgRNA design for experimental validation

SliceIt serves as the first comprehensive predictive engine for designing sgRNAs to edit the currently known millions of protein-RNA interaction sites in the human genome that augment to conduct high-resolution binding site block/silencing experiments. Two sgRNAs (5'-TGAATCTCGCTCTGTTGCCC-3' for BS chr2:99157353-99157403 and 5'- GGTTGATCCCGAACACAGGA-3' for BS chr2:99159478-99159514) in LIPT1 (Lipoyltransferase 1) gene locus from SliceIt were designed.

Establishing the CRISPR Cas9 system

Lentiviral vector digestion, oligo annealing and cloning into digested vector: Lentiviral CRISPR v2 plasmid (a gift from Feng Zhang (Addgene plasmid # 52961 ; http://n2t.net/addgene:52961 ; RRID:Addgene_52961) was digested with BsmBI and dephosphorylated with alkaline phosphatae for 2 hrs at 37°C. Digested plasmid was purified from gel using QIAquick gel extraction kit as per manufacturer's instructions.

Oligos were phosphorylated and annealed using T4 PNK (NEB M0201S) enzyme and T4ligation buffer (ATP added) in a thermocycler using following parameters: 37°C for 30min, 95°C for 5 min and ramp down to 25°C at 5°C/min. Annealed oligos are diluted at 1:200 dilution into sterile water. Diluted oligos were ligated with digested plasmid using T4DNA ligase, incubated over night at 16°C. Lentiviral plasmid was transfected into Stbl3 bacteria (Invitrogen C7373-03) using heat shock transfection method. For each oligo, three clones were selected for plasmid isolation. Plasmids were isolated using GeneJET Plasmid Miniprep Kit (K0503) and sent for sanger sequencing.

Lentiviral packaging

One day before transfection 2.5x10^5 HEK293T cells were plated (6 well plate) in DMEM supplemented with 10% heat-inactivated fetal bovine serum (FBS). Cells were incubated at 37°C overnight to get a confluency of 70%. Transfection was carried out using polyethyleneimine (PEI) method with the ratio of PEI:pTarget:pVSVg:RRE:REV; 16:3:1: 2:2. In a sterile tube, total 3ug of DNA following the ratios was diluted to 200ul of serum-free DMEM. PEI (2ug/ul) based on a 2:1 ratio of PEI(ug): total DNA(ug) was added to diluted DNA. Mix was incubated for 15 min at room temperature. For each binding site, one oligo was transfected individually into the cells, to purturb the binding site. After 72 hrs, lentiviral particles were harvested and concentrated at 3,000g for 5min at 4°C. The supernatant is filtered through a 0.45um filtration on ice using synringe filter. HepG2 cells were freshly cultured in 6 well plate for 24hrs, followed by transduction using lentiviral concentrate. Cells were incubated for one week followed by which puromycin treatment was given for positive selection of transduced cells. GFP and plasmid insert was used as positive control. Fluorescence microscope was used to check transfection and transduction efficiency. After 1 week puromycin concentration was reduced, since the cell number was low. Cells were split after 50% confluency of wells. After 1 week of incubation, cells were harvested in two tubes. From one aliquot, RNA was isolated using Tri-reagent, cDNA generated and real-time PCR was run for analyzing the expression level of proximal exons. In order to validate gene modifications, DNA was isolated from the second aliquot and sent for sanger sequencing.

4.2.3 Results and Discussion

RBP driven post-transcriptional regulation likely depends on its binding efficiency to its target location (237, 267) (Figure 26A). This phenomenon is highly crucial for several key biological processes including in development (18, 95, 268, 269) and differentiation (251, 270-275). It can be studied by measuring the expression level of the target RNA or proximal/neighboring exon to that of the binding site of interest. I hypothesize that the perturbation of the RBPs' binding sites or its equivalent position on DNA can potentially promote dysregulated function of the post-transcriptional target RNA molecule, enabling the functional dissection of the millions of binding sites of RBPs being discovered by CLIP (276) and related technologies (254-256). Cas9 system has been extensively utilized to edit the genomic loci of interest (244, 277). However, it has not been used to systematically understand the impact of RBP binding site perturbation in human cell types. Thus, this "cause" and "effect" model of regulation was employed by perturbing equivalent binding site on DNA using Cas9 system, where dysregulation phenotype can be measured by expression analysis of the target RNA feature, such as inclusion or exclusion of exon. Briefly, the BS profiles for available RBPs from the ENCODE project was downloaded and preformatted the BS regions by flanking them up to 100 bp. CRISPR-DO was employed to design sgRNAs for each binding site. All predicted sgRNAs were deposited into a database called SliceIt (see Figure 26B and Materials and Methods)

SliceIt facilitates designing CRISPR/Cas9 screens in both low (as illustrated in Figure 26C) and high throughput modes, by enabling parametric flexibility for designing sgRNAs along with the ability to filter the binding sites for the presence of SNPs, GWAS hits and exon expression alterations across a wide range of human tissue types.

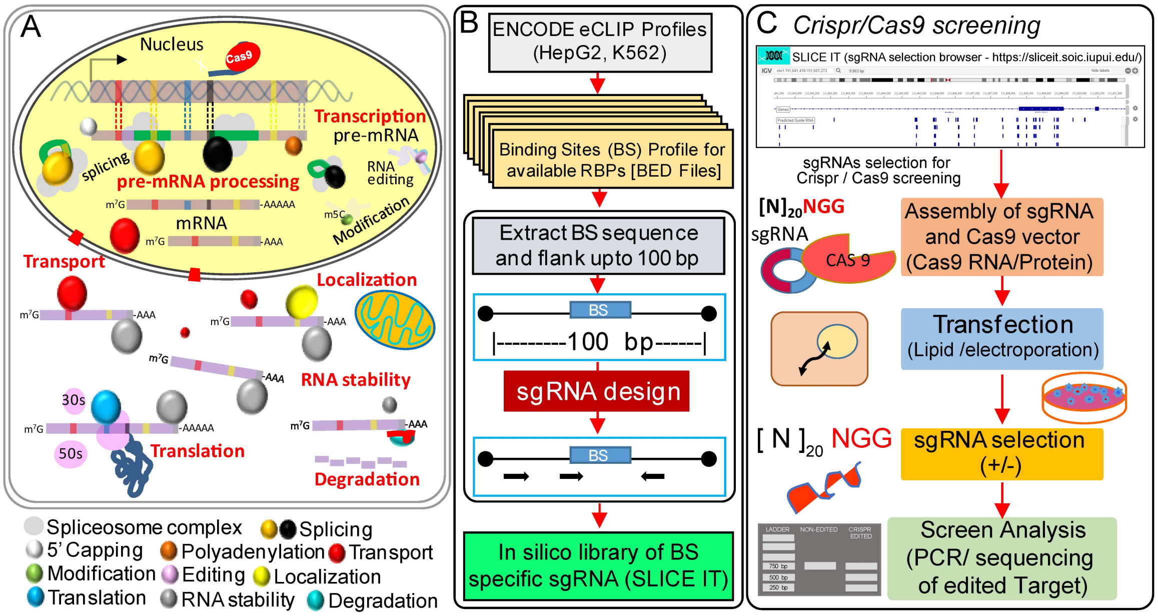


Figure 26: Strategy for validation of RBPs' binding sites. (A) A hypothesis driven CRISPR/Cas9 model for determining the effect of sgRNA on an exon proximal to targeted binding site (B) Construction of SliceIt, an in-silico guide RNA library for RBPs' binding sites (C) Proposed approach for designing CRISPR/Cas9 screening experiments based on the compendium of sgRNAs from SliceIt, for perturbation of binding sites.

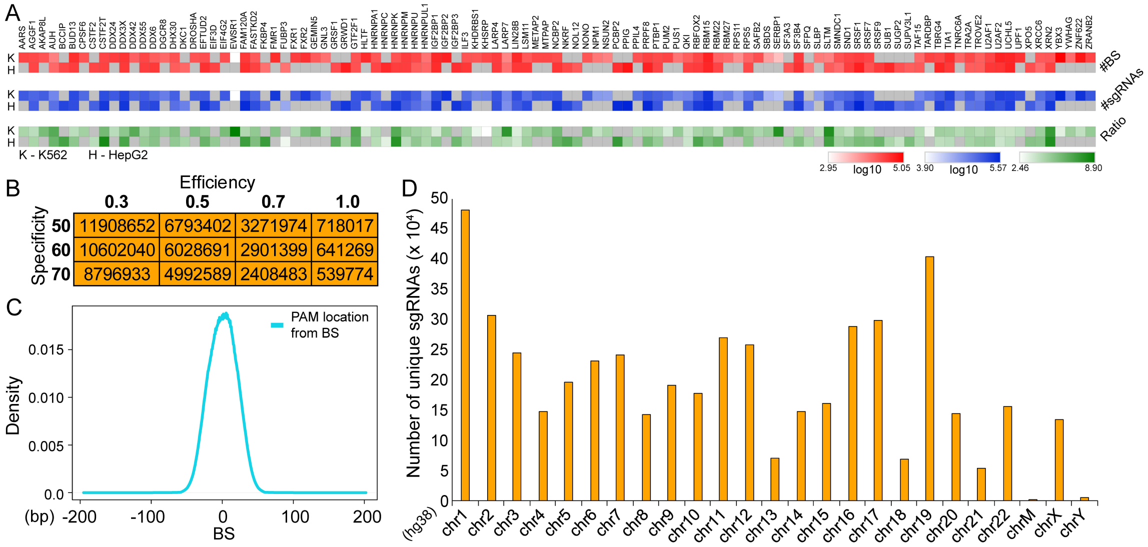
4.2.3.1 Overview of SliceIt

This study presents SliceIt (https://sliceit.soic.iupui.edu/), a database and visualization tool providing a comprehensive summary of in silico sgRNA (single guide

RNA) library, to facilitate rational design of CRISPR/Cas9 experiments in both low and high throughput fashion to perturb the protein-RNA interaction sites. CRISPR-DO (244) was used to design ~4.9 million unique sgRNAs targeting all possible RBP binding sites resulting from eCLIP experiments of scores of RBPs in HepG2 and K562 cell lines (from ENCODE) (see Materials and Methods). SliceIt provides a user-friendly environment, developed in highly advanced search engine framework called Elasticsearch. It is available in both table and genome browser views facilitating the easy navigation of RBP binding sites, sgRNAs, SNPs and GWAS hits, while querying for a gene, RBP or region of interest. It also provides exon expression profiles across 53 human tissues from the GTEx project (https://gtexportal.org/home/) (264), to examine locus specific expression changes proximal to the binding sites, to enable rational design of experiments in specific tissue/cell types. Users can also upload custom tracks in various file formats (in browser) to navigate additional genomic features in hg38 human genomic build. This custom track upload and navigation feature, in addition to the datasets already integrated into SliceIt, provide a functional context for user-generated datasets. All the binding site centric information is dynamically accessible via "search by gene", "search by coordinate" and "search by RBP" and readily available to download. SliceIt is the first comprehensive predictive engine for designing sgRNAs to edit the currently known millions of protein-RNA interaction sites in the human genome. It is a one-stop repertoire of guide RNA library and RBP binding sites along with several layers of functional information, to design high throughput CRISPR cas9 screens for studying the phenotypes and diseases associated with the binding sites of RBPs and to functionally dissect the post transcriptional regulatory networks.

4.2.3.2 Characteristics of sgRNA repertoire available in SliceIt

Approximately 4.6 million binding sites corresponding to 108 RBPs across the two cell lines (2.23 million unique sites in HepG2 and 2.38 million unique sites in K562) were downloaded from the ENCODE project (218) (Figure 27A). These binding sites were flanked to 100 bp (if <100 bp) and queried for sgRNA prediction using CRISPR-DO (see Materials and Methods). This resulted in a repository of ~4.9 million unique sgRNAs (3.73 million and 3.04 million unique sgRNAs for HepG2 and K562 cell lines respectively) predicted for ~4.6 million binding sites. Binding sites for each RBP and corresponding unique sgRNAs were log10 transformed and illustrated as a heatmap in Figure 27A. The ratio of number of sgRNAs and binding sites for each RBP were also calculated, representing an estimate of the average number of sgRNAs designed per binding site for each RBP in both the cell lines (Figure 27A). SliceIt comprises of a relatively unbiased collection of sgRNAs (2-8 sgRNAs per binding site) for the RBPs included in the database, making it an easily accessible and user-friendly web interface for designing the targeted post-transcriptional dysregulation experiments using Cas9 system. Additionally, it was observed that the total distribution of designed sgRNAs decreased with increasing efficiency as well as specificity (Figure 27B). Several CRISPR based experiments have shown that Cas9 directed double strand breaks occur mostly in close proximity to PAM region of targeted genomic loci (244). Thus, the positional occurrence of designed sgRNAs was investigated by calculating the distance of binding site from PAM. Most of the designed sgRNAs were proximal to the midpoint of the binding sites with an exponential decline as distance increased from the center of the binding site (Figure 27C). It is noteworthy to mention that the total number of designed sgRNAs and made available via SliceIt varied between chromosomes and were generally correlated with the size of the chromosome (Figure 27D).



Specificity	Efficiency 0.3	0.5	0.7	1.0
50	11908652	6793402	3271974	718017
60	10602040	6028691	2901399	641269
70	8796933	4992589	2408483	539774

Figure 27: (A) Heatmap showing the number of unique binding sites (log10 transformed - red) and sgRNAs (log10 transformed-blue) and the ratio of the number of sgRNAs and binding sites (green) for each RBP, representing an estimated average sgRNAs designed per binding site for each RBP in both the cell lines (H-HepG2 and K-K562). RBPs for which currently no binding site information is available from ENCODE project for either cell line are greyed out. (B) Distribution of the total number of designed sgRNAs available from SliceIt as a function of the predicted efficiency and specificity scores. (C) Density plot showing the distribution of distances between sgRNA's PAM location and the mid-point of the targeted binding site, for all the sgRNAs available from SliceIt. (D) Distribution of the absolute number of sgRNAs across human chromosomes present in SliceIt.

4.2.3.3 SliceIt database construction, visualization and accessibility

SliceIt is an efficient search engine, consisting of several layers of omics data including RBP binding site profiles, SNPs, GWAS and tissue-specific exon expression levels (GTEx) under the niche of Flask server module. The complete pipeline as illustrated in Figure 28A, details on how the flask server interacts with front-end, retrieves and parses data from elasticsearch for providing an output. Basically, this pipeline takes only a few seconds to search the query, process and render an output.

While most other alternative tools run an algorithm in the cloud to generate guide RNA predictions, SliceIt takes advantage of precomputed data to achieve a speed that is several fold faster than other comparable tools.

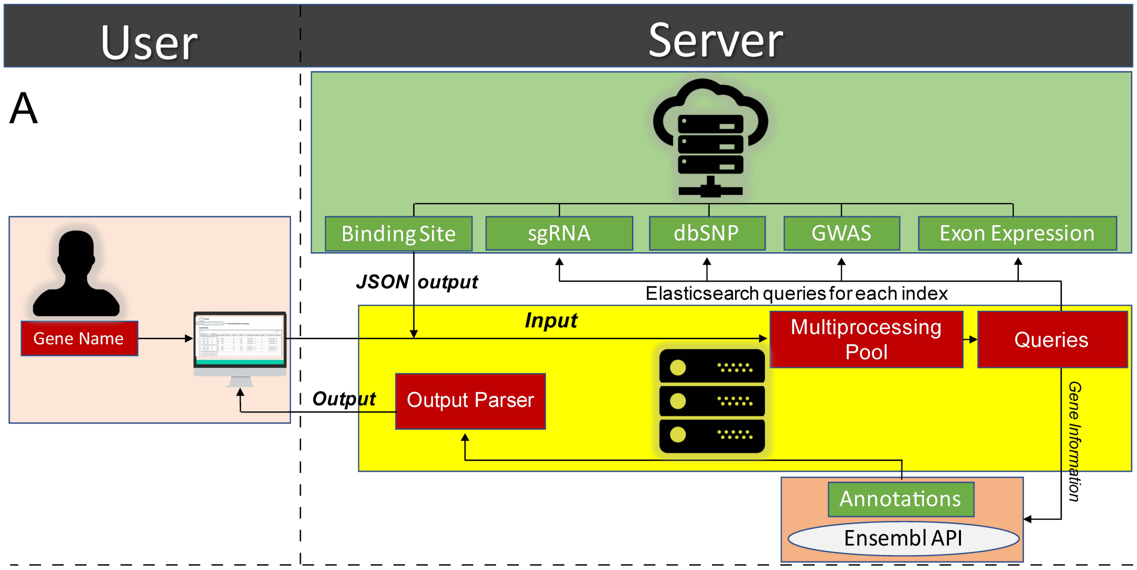
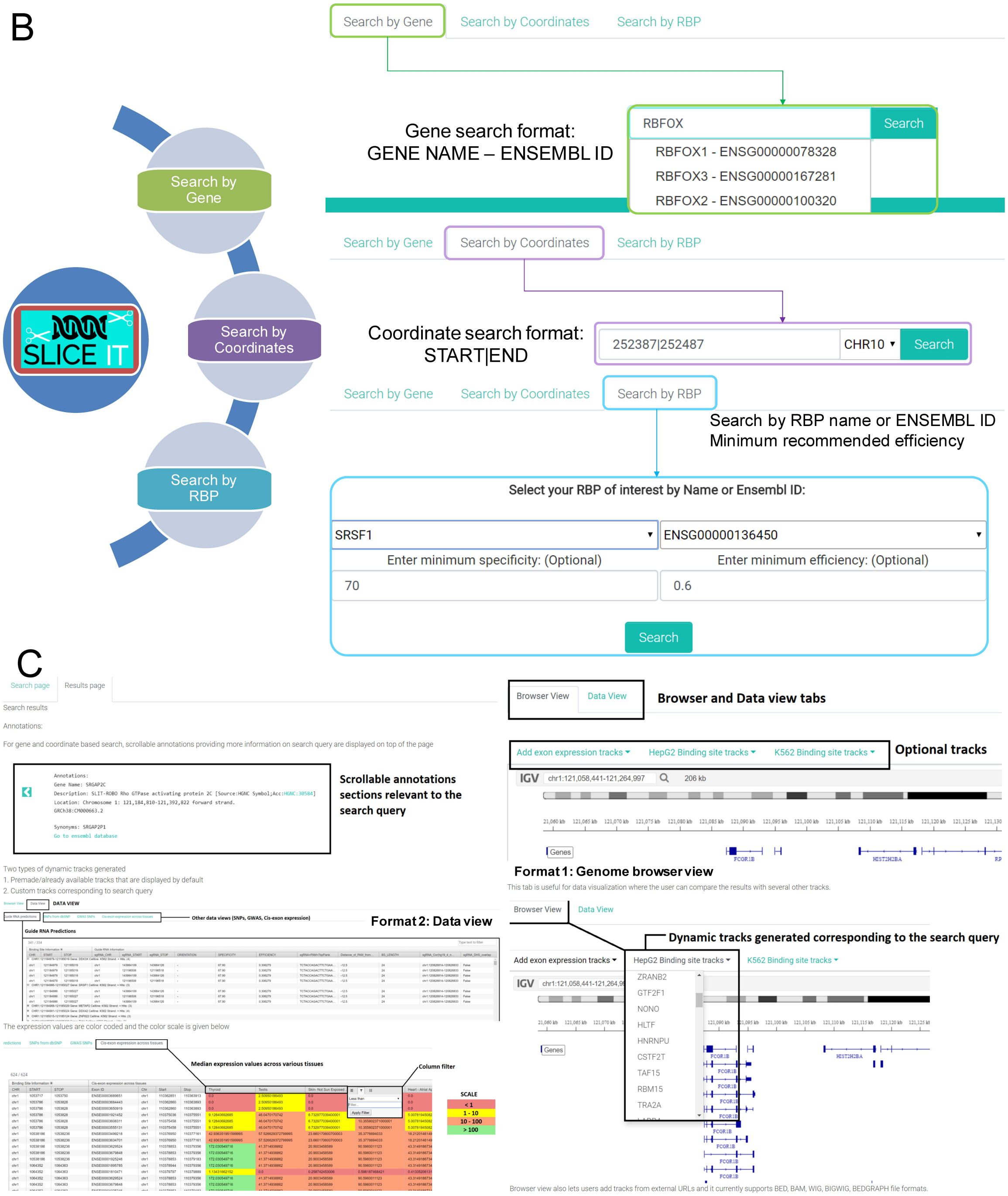


Figure 28: (A) Search processing pipeline showing the functionality of SliceIt's communication with frontend and elasticsearch cluster to parse and display query output.

SliceIt enables searching a transcribed region with RBP binding sites in the genome for designed sgRNAs. This is facilitated by allowing the user to search for gene region, genomic co-ordinates and for targets of an RBP (Figure 28B). For instance, a user could search for the sgRNAs that can be designed to edit the binding sites of a member of the RBFOX family of RBPs or search for a specific genomic region defined by chromosomal co-ordinates. Alternatively, a user can visualize the target binding sites of a selected RBP for which sgRNAs satisfy specific design thresholds. For each search query, SliceIt outputs query annotations, data visualization with IGV JS genome browser, SNPs in the region, GWAS SNPs in the region, predicted sgRNA and cis-exon expression information in tabular form as shown in Figure 28C.



Search by Gene | Search by Coordinates | Search by RBP

Gene search format:
GENE NAME – ENSEMBL ID

RBFOX | Search

RBFOX1 - ENSG00000078328
RBFOX3 - ENSG00000167281
RBFOX2 - ENSG00000100320

Search by Gene | Search by Coordinates | Search by RBP

Coordinate search format:
START|END

252387|252487 | CHR10 ▾ | Search

Search by Gene | Search by Coordinates | Search by RBP

Search by RBP name or ENSEMBL ID
Minimum recommended efficiency

Select your RBP of interest by Name or Ensembl ID:

SRSF1 ▾ | ENSG00000136450 ▾

Enter minimum specificity: (Optional): 70

Enter minimum efficiency: (Optional): 0.6

Search

Search page | Results page

Search results

Annotations:

For gene and coordinate based search, scrollable annotations providing more information on search query are displayed on top of the page

```
Annotations:
Gene Name: SRGAP2C
Description: SLIT-ROBO Rho GTPase activating protein 2C [Source:HGNC Symbol;Acc:HGNC:30584]
Location: Chromosome 1: 121,184,810-121,392,822 forward strand.
GRCh38:CM000663.2

Synonyms: SRGAP2P1
Go to ensembl database
```

Scrollable annotations sections relevant to the search query

Two types of dynamic tracks generated
1. Premade/already available tracks that are displayed by default
2. Custom tracks corresponding to search query

Browser View | Data View — **DATA VIEW**

Guide RNA predictions | SNPs from dbSNP | GWAS SNPs | Cis-exon expression across tissues — **Other data views (SNPs, GWAS, Cis-exon expression)**

Format 2: Data view

Guide RNA Predictions

341 / 334 — Type text to filter

Binding Site Information			Guide RNA Information										
CHR	START	STOP	sgRNA_CHR	sgRNA_START	sgRNA_STOP	ORIENTATION	SPECIFICITY	EFFICIENCY	sgRNA+PAM+7bpFlank	Distance_of_PAM_from...	BS_LENGTH	sgRNA_Co(hg19_4_n...	sgRNA_DHS_overlap
⊟ CHR1:121184979-121185019 Gene: DDX3X Cellline: K562 Strand: + Hits: (4)													
chr1	121184979	121185019	chr1	143984109	143984126	-	67.90	0.306279	TCTACCAGACTTCTGAA...	-12.5	24	chr1:120826814-120826833	False
chr1	121184979	121185019	chr1	121196506	121196518	-	67.90	0.306279	TCTACCAGACTTCTGAA...	-12.5	24	chr1:120826814-120826833	False
chr1	121184979	121185019	chr1	143984109	143984126	-	67.90	0.306279	TCTACCAGACTTCTGAA...	-12.5	24	chr1:120826814-120826833	False
chr1	121184979	121185019	chr1	121196506	121196518	-	67.90	0.306279	TCTACCAGACTTCTGAA...	-12.5	24	chr1:120826814-120826833	False
⊟ CHR1:121184986-121185027 Gene: SRSF1 Cellline: K562 Strand: + Hits: (3)													
chr1	121184986	121185027	chr1	143984109	143984126	-	67.90	0.306279	TCTACCAGACTTCTGAA...	-12.5	24	chr1:120826814-120826833	False
chr1	121184986	121185027	chr1	121196506	121196518	-	67.90	0.306279	TCTACCAGACTTCTGAA...	-12.5	24	chr1:120826814-120826833	False
chr1	121184986	121185027	chr1	143984109	143984126	-	67.90	0.306279	TCTACCAGACTTCTGAA...	-12.5	24	chr1:120826814-120826833	False
⊞ CHR1:121184988-121185029 Gene: METAP2 Cellline: K562 Strand: + Hits: (3)													
⊞ CHR1:121184991-121185024 Gene: DDX42 Cellline: K562 Strand: + Hits: (4)													
⊞ CHR1:121185015-121185124 Gene: ZNF622 Cellline: K562 Strand: + Hits: (3)													

The expression values are color coded and the color scale is given below

Guide RNA predictions | SNPs from dbSNP | GWAS SNPs | Cis-exon expression across tissues

Median expression values across various tissues

Column filter: Less than ▾ | Filter... | Apply Filter

624 / 624

Binding Site Information			Cis-exon expression across tissues								
CHR	START	STOP	Exon ID	Chr	Start	Stop	Thyroid	Testis	Skin- Not Sun Exposed		Heart - Atrial Ap
chr1	1053717	1053750	ENSE00003689651	chr1	110362851	110363913	0.0	2.50950186493	0.0		0.0
chr1	1053786	1053828	ENSE00003684443	chr1	110362860	110363893	0.0	2.50950186493	0.0		0.0
chr1	1053786	1053828	ENSE00003650919	chr1	110362860	110363893	0.0	2.50950186493	0.0		0.0
chr1	1053786	1053828	ENSE00001921452	chr1	110375036	110375551	9.12840692685	46.0470170742	6.732977039400001		5.00781945082
chr1	1053786	1053828	ENSE00003608311	chr1	110375458	110375551	9.12840692685	46.0470170742	6.732977039400001	10.355802371000001	5.00781945082
chr1	1053786	1053828	ENSE00003555131	chr1	110375458	110375551	9.12840692685	46.0470170742	6.732977039400001	10.355802371000001	5.00781945082
chr1	10538186	10538236	ENSE00003499218	chr1	110376950	110377161	42.936351951599995	57.526629372799995	23.660170600700003	35.3776694033	18.2120146149
chr1	10538186	10538236	ENSE00003634701	chr1	110376950	110377161	42.936351951599995	57.526629372799995	23.660170600700003	35.3776694033	18.2120146149
chr1	10538186	10538236	ENSE00003629524	chr1	110378853	110379356	172.030549716	41.3714938862	20.9003458589	90.5960011123	43.3149186734
chr1	10538186	10538236	ENSE00003679848	chr1	110378853	110379356	172.030549716	41.3714938862	20.9003458589	90.5960011123	43.3149186734
chr1	10538186	10538236	ENSE00001925246	chr1	110378853	110379183	172.030549716	41.3714938862	20.9003458589	90.5960011123	43.3149186734
chr1	1064352	1064363	ENSE00001895785	chr1	110378944	110379356	172.030549716	41.3714938862	20.9003458589	90.5960011123	43.3149186734
chr1	1064352	1064363	ENSE00001810471	chr1	110379797	110379889	1.13431662152	0.0	0.258742453006	0.596187468421	0.41335206131
chr1	1064352	1064363	ENSE00003629524	chr1	110378853	110379356	172.030549716	41.3714938862	20.9003458589	90.5960011123	43.3149186734
chr1	1064352	1064363	ENSE00003679848	chr1	110378853	110379356	172.030549716	41.3714938862	20.9003458589	90.5960011123	43.3149186734

SCALE
🟥 < 1
🟨 1 - 10
🟧 10 - 100
🟩 > 100

Browser View | Data View — **Browser and Data view tabs**

Add exon expression tracks ▾ | HepG2 Binding site tracks ▾ | K562 Binding site tracks ▾ — **Optional tracks**

IGV chr1:121,058,441-121,264,997 — 206 kb

Format 1: Genome browser view

This tab is useful for data visualization where the user can compare the results with several other tracks.

Browser View | Data View

Dynamic tracks generated corresponding to the search query

Add exon expression tracks ▾ | HepG2 Binding site tracks ▾ | K562 Binding site tracks ▾

ZRANB2
GTF2F1
NONO
HLTF
HNRNPU
CSTF2T
TAF15
RBM15
TRA2A

IGV chr1:121,058,441-121,264,99

Browser view also lets users add tracks from external URLs and it currently supports BED, BAM, WIG, BIGWIG, BEDGRAPH file formats.

Figure 28: (B) Screenshot of SliceIt showing three different search options currently available in the database i.e. Search by gene name or Ensembl ID, Search by coordinates and Search by RBP name at various efficiency and specificity thresholds. (C) Figure describing various components of a typical search result page from SliceIt. Highlighted sections include Annotations, Data View, Exon expression with color coding and Genome browser view.

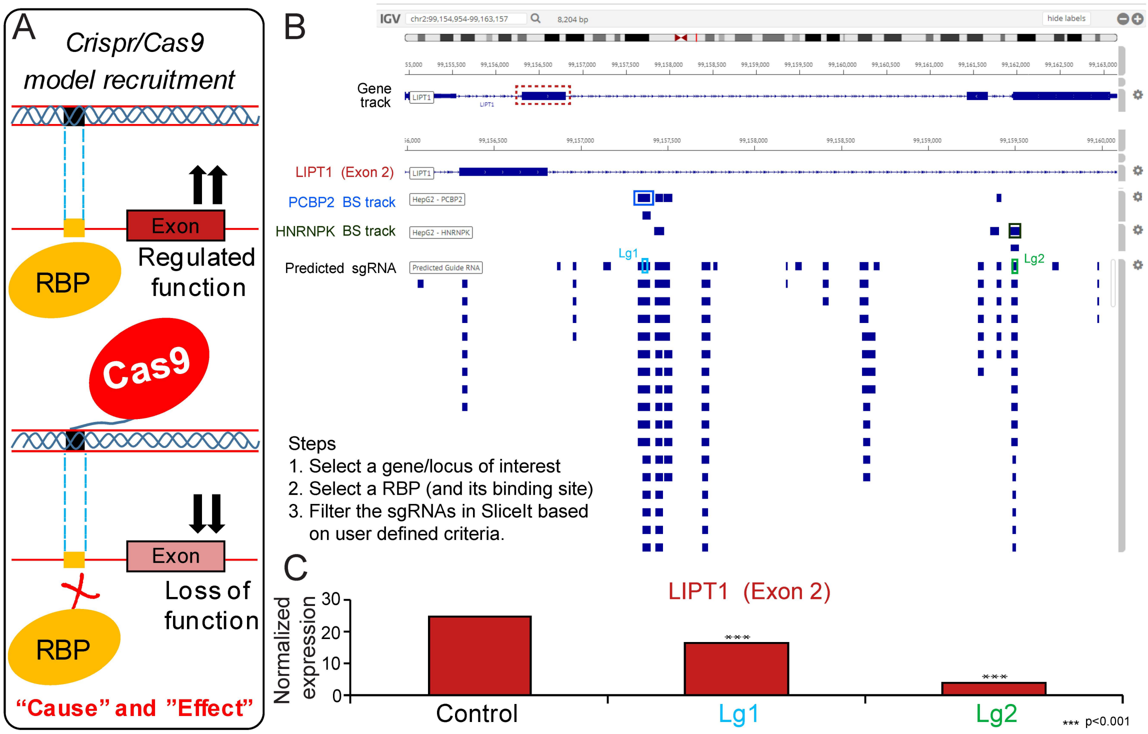
Cis-exon expression information across human tissues obtained from the GTEx project is displayed in Fragments Per Kilobase of transcript per Million mapped reads (FPKM) units and is color coded to indicate various ranges of expression levels (see Materials and Methods). An expression level of less than 1 FPKM is displayed in dark red, greater than 1 and less than 10 FPKMs in yellow, greater than 10 and less than 100 FPKMs in orange and an expression level higher than 100 FPKMs is displayed in green.

4.2.3.4 SliceIt aids in functional validation of RBP binding sites using CRISPR/Cas9 experiments

SliceIt is an integrative omics resource that facilitates systematic experimental designs for editing RNA binding sites and their functional dissection across human tissues. It provides a robust set of sgRNAs that could be used for systematic perturbation of RBP' binding sites occurring in a genomic loci or gene of interest. Several other omics datasets were also integrated into SliceIt that can potentially help the users to define their criteria for RBP binding site centric Cas9 experiments.

In order to validate the designed sgRNAs reported in SliceIt, the "cause" and "effect" model of regulation was employed by perturbing equivalent binding site on DNA using Cas9 genome editing system and study the impact of the perturbation on proximal exon expression levels (Figure 29A). SliceIt was used to extract the list of sgRNAs targeting each binding site under consideration (with ± 50 bp flank sequence). sgRNAs which have efficiency > 0.70 and specificity > 70 % with minimal distance between on-site PAM (Protospacer Adjacent Motif) and center of the binding site were selected. The binding sites in 'Browser View' of SliceIt were also navigated to verify if the sites are accompanied by at least one GWAS or dbSNP annotations. SliceIt also enables the exon level expression profiles. Hence, it was also confirmed if the exon proximal to the binding site is expressed > 1 FPKM in primary tissue samples from GTEx project, corresponding to the cell line model being used. For this particular study, SliceIt was used to design potential sgRNAs for perturbing the binding sites and selected two sgRNAs for editing two different binding sites on LIPT1 (Lipoyltransferase 1). LIPT1 encodes for an acyl group transferase, involved in lipoic acid metabolism (278, 279) and glycine degradation (280). LIPT1 genomic loci was navigated and queried for two RBPs independently for respective binding sites (and predicted sgRNAs) as shown in Figure 29B (see Materials and Methods). The second exon of LIPT1 was focused and investigated the impact of RBP binding sites (BS1 and BS2) targeted by respective sgRNAs (Lg1 and Lg2) designed by SliceIt (Figure 29B). LIPT1 was confirmed to be significantly expressed in primary liver tissue samples, since the cell line model is HepG2 cell line. The main objective was to validate if these sgRNAs are likely to perturb the binding sites in HepG2 cells, with a high efficiency as predicted by SliceIt. Perturbation can result in increase or decrease in proximal exon expression levels, since it depends on whether the binding site can enhance or repress the activity of the exon i.e, binding site can be a splicing enhancer or repressor. SgRNA plasmid library constructs were confirmed using Sanger sequencing (Appendix 5). Plasmid transfection and lentiviral transduction efficiency was measured as GFP signals on fluorescence microscopy. GFP signals detected, confirmed 80-100% transfection efficiency in HEK293T cells and 75-90% transduction efficiency in HepG2 cells. For gene LIPT1, primers were designed to estimate the abundance of the second exon to validate the effect of two proximal binding sites' upon perturbation, using two different sgRNAs (labeled as Lg1 and Lg2 in Figure

29B). Upon normalizing with housekeeping gene PPIA, qPCR results showed significant (p<0.001) decrease in the exon expression levels compared to wild type HepG2 exon expression levels once transduced with Lg1 and Lg2 lentiviruses, respectively (Figure 29C). Hence, qPCR results confirmed that the designed sgRNAs targeted the binding sites, resulting in significant reduction of the proximal exon expression levels as a result of the perturbation of the binding sites. Interestingly, sgRNA - Lg2, designed to target the distal binding site of exon 2 exhibited higher reduction of the exon level than the sgRNA – Lg1 designed to perturb the proximal binding site.



Steps
1. Select a gene/locus of interest
2. Select a RBP (and its binding site)
3. Filter the sgRNAs in SliceIt based on user defined criteria.

Figure 29: Functional validation of RBP binding sites using CRISPR/Cas9 experiments in human cell lines. (A) A hypothesis driven CRISPR/Cas9 model for determining the effect of sgRNA on an exon proximal to targeted binding site (B) A genome browser view of SliceIt illustrating the genomic loci of LIPT1 gene, the second exon, RBP centric query for binding sites (BS1 and BS2 color coded boxes) targeted by respective sgRNAs (Lg1 and Lg2, color coded boxes) designed by SliceIt (C) Binding site perturbation experiments using sgRNAs; Lg1 and Lg2. Exon expression levels (normalized) proximal to binding sites were quantified by qPCR (*** = p<0.001).

4.2.4 Conclusion

SliceIt is a comprehensive resource and visualization platform that enables the users to systematically design experiments to study the impact of a binding site of RBP on a particular RNA target. Hence, it can help the users in dissecting the role of binding sites in A) modulating splicing, stability and localization of RNA B) controlling the protein isoform levels, across a multitude of tissue types and cell lines by facilitating the generation of high quality custom set of sgRNAs for the well-established CRIPSR/Cas9 genome editing system. It is a one-stop repertoire that enables the design of small scale experiments to study a specific binding site's role in modulating post-transcriptional regulation or medium range studies such as RBP centric functional screens or genome-scale CRISPR/Cas9 screens to edit the protein-RNA interaction networks. SliceIt also enhances the applicability of CRISPR/Cas9 system by focusing on binding sites on lncRNAs that can be perturbed for studying their contribution in downstream post-transcriptional control. Additionally, the “custom track” feature of SliceIt enables the users to re-purpose the compendium of SliceIt according to their choice. For instance, users can modulate the regulome for a novel RBP of interest and study the results in the context of existing protein-RNA interaction maps and expression profiles across tissue types.

4.3 PEEK: Prioritization of RBP-Binding sites using Expression and Evolutionary Constraints

4.3.1 Introduction

Eukaryotic cells encode for thousands of RNA-binding proteins (RBPs) that associate physically with a unique set of RNA targets (281). They bind to RNA at specific recognition sites (282), and play a crucial role in post-transcriptional regulation of the gene products (191). RBPs have been implicated in various genetic diseases in humans, such as cancers, neurological diseases and metabolic disorders (283). Identification of RBP binding sites (BS) on mRNA clarifies the molecular function of RBPs. Various CLIP (Cross-linking Immunoprecipitation) protocols have been developed to uncover the targets of RBPs. Such protocols predominantly involve irradiation of cells using ultraviolet light for crosslinking RNAs to the interacting RBPs (284), followed by immunoprecipitation of the protein of interest bound to RNAs, purification of RNA by proteinase digestion and generation of a cDNA library for the purified RNA (285).

Several variants of CLIP protocols have been established so far (198-200, 286) and documented millions of binding sites (BS) on RNA (70, 217, 287). Although several studies have investigated hundreds of RBPs and their reported BS (288-292), however, little or no evidence is available for their functional impact on post-transcriptional targets (193, 259, 288). Also, there are existing CLIP-seq peak calling algorithms such as CLIPper (https://github.com/YeoLab/clipper), PARalyzer (206) and Piranha (207) that delineate the regions in the transcriptome that are significantly associated with RBP-binding. However, these methods rely on identification of statistically significant BS based on the enrichment of reads aligned to the reference genome. To date, computational methods to identify the biologically functional binding sites in a tissue specific manner have had limited success. Hence, elucidating the functionally relevant RBP binding sites across various tissues is critically important for unraveling the role of RBP in physiology and disease.

RNA-binding proteins are known to be conserved across a wide range of species (192, 293), therefore, investigating the extent of conservation of BS provides a significant insight into the targets which exhibit high or low conservation across species. Canonically, genomic elements with higher extent of conservation across species contribute to more significantly converged biological function (294-297). Genomic elements such as genes, transcripts or exons demonstrate their functionality based on their level of expression in a tissue specific manner (100, 298, 299). Therefore, extent of conservation of BS combined with expression level of proximal exons was inferred to be a powerful method to annotate the BS with their biological relevance across tissues. Hence, it is imperative to develop a robust algorithm to dissect the functional protein RNA interactions that could elucidate the molecular mechanism involved in maintaining the functional diversity in transcriptome. In this study, first the documented BS (70, 217, 287) was characterized and decoded its impact on exome. Further, an algorithm was developed that integrates the exon expression and evolution constrains of genomic elements across multiple species around these BS.

4.3.2 Materials and Methods

4.3.2.1 Data collection

The BS profile of several RBPs was downloaded from ENCODE (99) and CLIPdb (217), originally in bed (Browser Extensible Data) format. These downloaded files were parsed and pre-formatted for downstream integrative analysis. Similarly, Gene Expression Omnibus (115) platform was used to download the unprocessed RNA sequencing data (fastq files) of different tissues across 10 species from multiple studies (298, 300-302) (Appendix 6). The RNA seq data (raw fastq) was also downloaded for 369 Liver Hepatocellular Carcinoma (LIHC) patients, 448 Kidney Renal Clear Cell Carcinoma (KIRC) patients and 154 Glioblastoma (GBM) patients from The Cancer Genome Atlas (TCGA) (303). In addition to that, the MAF blocks in MAF [Multiple Alignment File] files were also downloaded from UCSC genome browser (304). These MAFs provide the information of whether a particular region ('block') of the human genome is conserved across 46 species.

4.3.2.2 Alignment and Quantification of RNA-Seq data

Raw RNA sequence files (in fastq format) downloaded from multiple resources (as described previously) were first examined for quality assurance using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and preprocessed for high quality reads using FASTx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). High quality sequence reads (Phred score≥20) were aligned onto respective reference genome (Appendix 7) using HISAT (117) with default parameters. These aligned reads were carried out for post processing steps using samtools and converted into sorted.bam format. These sorted bam files were processed for transcript assembly and quantification using StringTie (120). Obtained expression profiles were formatted into separate matrices for each data type.

4.3.2.3 Annotation of binding sites and expression profiling of proximal exons in LIHC

The eCLIP (286) based BS profile of 56 RBPs (in HepG2 cell line obtained from ENCODE project) was downloaded. For this study, binding sites having p<0.001 as documented in ENCODE were considered (99). Bedtools (233) were used to annotate these binding sites onto genomic elements obtained from Ensembl (305) biomart. Based on their locus specific genomic positioning, the BS were characterized into three bins; "exonic", "intronic" and "junctional" binding. Certainly, if a binding site is occupied entirely by an exonic region, it was referred to as exonic; otherwise, the binding site was referred to as intronic. In case, a genomic locus was annotated with multiple exons, the longest exon was considered for downstream analysis. If certain regions of the binding site were occupied in both exon and intron, the binding was referred to as junctional binding.

The exon expression profile was extracted from constructed matrix of 369 LIHC patients (TCGA) as described previously. Next, the impact of annotated binding sites on expression of local exon (i.e. binding site on exon locus or closest to) was investigated by comparing with neighbor or random exons within the gene boundary.

4.3.2.4 Construction of BCM (Binary Conservation Matrix) and EEM (Exon Expression Matrix) matrices using MAF blocks

MAF blocks downloaded from UCSC genome browser (https://genome.ucsc.edu/) contains the conservation profile of genomic elements across 46 species. RBPs' BS profile (hg19 build) obtained from ENCODE and CLIPdb were parsed from the MAFs using ad hoc python script. For each BS, the equivalent conservation block coordinates were extracted and annotated in binary format as "0" for its absence and "1" for its presence in 10 species. This step is repeated for all the BS included in current analysis. The binary format records for all the binding sites of an RBP across 10 species were concatenated into Binary Conservation Matrix (BCM).

Similarly, genomic coordinates of human (hg19) exons were obtained from Ensembl (305) and inspected in MAF blocks using ad hoc python script to estimate the equivalent exonic coordinates for 10 species. Human exonic coordinates and its equivalent conserved exons for these species (genomic positions) were concatenated to generate Exon Conservation Matrix (ECM). Each genomic element in ECM was parsed to obtain the expression profile of a given tissue for respective species (generated previously). This step was repeated to obtain the expression profile of human exons and its equivalent exon in 10 species per tissue type. Finally, the cross species expression profile was converted into Exon Expression Matrix (EEM) for a given tissue type. Same step was followed for all other tissues. Please note, CrossMap (306) was employed to liftover the MAFs, and other genomic coordinates to Ensembl build (hg38) wherever applicable.

4.3.2.5 Prioritization of RBP-Binding sites using Expression and Evolutionary Constraints (PEEK)

This study proposes to model the relationship between exon levels and the evolutionary conservation pattern of BS that elucidate the impact of BS on the expression levels of associated exons in a tissue-specific manner. To do so, a modified version of matrix eQTL engine (307) was implemented. In a typical eQTL analysis, which involves millions of association tests, it is common to treat each genotype variable as categorical and model its effect on gene expression with ANOVA (Analysis of Variance). It is known that ANOVA model can be viewed as linear regression and F-test can be employed as a statistical test to significantly speed up the association calculations when large matrices are involved (307). Hence, matrix eQTL engine was modified for matlab due to its efficient Basic Linear Algebra Subroutine (BLAS), to facilitate rapid binding site exon expression level association calculations and to identify functional binding sites. Here F-test statistic is defined as $\frac{(n-k-1)r^2}{k(1-r^2)}$ where k is the number of regressors, n the vector size and r the sample correlation.

Thus, ANOVA based Prioritization of RBP-Binding sites using Expression and Evolutionary Constraints (PEEK) was accomplished by correlating the EEM and BCM using the modified version of R package Matrixeqtl (307). A binding site was considered to control an exon if the distance between them was ≤ 5kb, equivalent to the cis-eQTL analysis commonly employed in association studies. Hence, a list of BS was obtained and exons that were significantly associated to each other and referred them as ‘prioritized’ or ‘functional’ BS and associated exons.

4.3.2.6 Annotation of PEEK binding sites and expression profiling of proximal exons in cancer

AnnotatePeaks.pl script from Homer (308) was used to annotate the PEEK prioritized binding sites in genomic boundary of the known genes. The expression profile of associated neighbor exons were extracted from exon expression matrices of 369 Liver Hepatocellular Carcinoma (LIHC), 154 Glioblastoma (GBM) and 448 Kidney Renal Clear Cell Carcinoma (KIRC) patients from TCGA project (8) (as described previously). Expression profile of associated exon was compared with genomic locus specific random exons to measure the impact of PEEK binding sites in cancer.

4.3.2.7 Validation of peek prioritized binding sites using Crispr cas9 system

SliceIt was developed (7) as a part of this thesis, that consist of a massive collection of sgRNAs designed around protein-RNA interaction sites in human to conduct high-resolution binding site perturbation experiments. SliceIt was used to design the sgRNAs targeting the genomic loci in close proximity to selected PEEK prioritized binding sites in FUS (FUS RNA Binding Protein), RBCK1 (RANBP2-Type and C3HC4-Type Zinc Finger Containing 1) and RNPEPL1 (Arginyl Aminopeptidase Like 1) genes. Further, Crispr Cas9 system in HepG2 cells was established as described previously (7). Briefly, Lentiviral transfection method was used to perturb RBP binding sites. Lentiviral CRISPR plasmids (Addgene plasmid # 52961) were cloned with oligos and enriched in Hek293T cells for transducing HepG2 cells. This tool was utilized for genomic perturbation experiments using selected sgRNAs targeting equivalent genomic loci of BS. The expression level of proximal exon(s) was estimated using qPCR to verify the impact of functional BS.

4.3.3 Results and discussion

RNA-binding proteins (RBPs) are involved in a variety of post transcription regulation processes by directly interact with a diverse set of RNA species to designate their function (7). Several databases such as ENCODE provides millions of binding sites (BS) of ~100 RBPs based on UV cross-linking and immunoprecipitation (CLIP) experiments. Indeed, CLIP and its variant high throughput protocols(254-256) have established a peer platform to understand the RBP driven post transcriptional regulation (7), the data generated often compromises with signal noise and functional ambiguity (257, 258). Also, these protocols provide little or no relevance on tissue specific post transcriptional regulatory circuits. A robust algorithm was developed to dissect the functional protein RNA interactions that could elucidate the molecular mechanism involved in maintaining the functional diversity in transcriptome. First, the RBPs' binding sites obtained from ENCODE were characterized and decoded its impact on exome. PEEK, an integrated approach was developed which facilitates tissue-specific evolutionary annotation of binding sites of RBPs by mining hundreds of RNA-Seq datasets spanning 10 vertebrate species and 4 tissues (298, 300-302, 309). Several of functional binding sites obtained from PEEK were verified by Crispr/ Cas9 system.

4.3.3.1 Majority of the binding sites of RBPs influence the proximal exons

To investigate the relationship between the occurrence of RBP-binding site and the expression of a nearby exon, eCLIP profiles of 56 RBPs from HepG2 cell line (ENCODE) were employed. The binding event were classified as exonic, intronic or junctional binding as shown in Figure 30A. To match the cell type, RNA-Seq data for 372 Liver Hepatocellular Carcinoma (LIHC) patients was obtained from TCGA project (8) and the expression levels of all annotated exons in the human genome (305) were obtained using HISAT (117) and StringTie (120) pipeline (see Materials and Methods). Upon considering all binding events of RBP falling on the annotated intronic, exonic or junctional regions of the human genome, it was observed that a majority (62%) of the binding sites were found to occur on exonic features, while 10% were found to occur on junctions (Figure 30B). Exons on which exonic binding occurred on 'Itself', exhibited a higher expression compared to preceding and subsequent exons (i.e. "Before" and "After") as shown in Figure 30D. Likewise, RBP binding events at the 'junction' of exons exhibited higher expression than the neighboring exons (Figure 30C and 30F). In contrast, the intronic binding was found to result in lower expression of the proximal exons compared to distal random exons from the same gene (Figure 30E). In general, the analysis demonstrated that full or partial exonic binding of an RBP results in its increased expression while intronic binding contributes to lowering the expression of neighboring exons. Therefore, this data represents a relationship between binding site occurrence and the expression of proximal exons. Further, the study emphasized that irrespective of the directionality of exon levels, binding sites confer an impact on proximal exon. This could be identified by determining the association between binding site and exon expression across species in an evolutionary context.

4.3.3.2 Overview of PEEK framework

The BS profile of 135 RBPs was obtained from ENCODE (99) and CLIPdb (217), and annotated the genomic coordinates with the human genome annotation file. Similarly, coordinates of human exons were obtained from Ensembl (305) biomart (Figure 31A). Each binding site and exon was mapped to the corresponding MAF block (obtained from UCSC genome browser (304)) that could explain its conservation across 46 species (Materials and Methods). Eventually, a list of binding sites and exons, each mapped to their corresponding MAF blocks were generated. A Binary Conservation Matrix (BCM) was generated using the list of binding sites mapped to their corresponding MAF blocks, where the binding site coordinates, along with its presence (“1”) or absence (“0”) across the 10 species were documented. Similarly, the exon conservation matrix (ECM) was constructed using human exon coordinates and its equivalent genomic coordinates with 70% overlap of MAF blocks in other species. Therefore, ECM documented the human

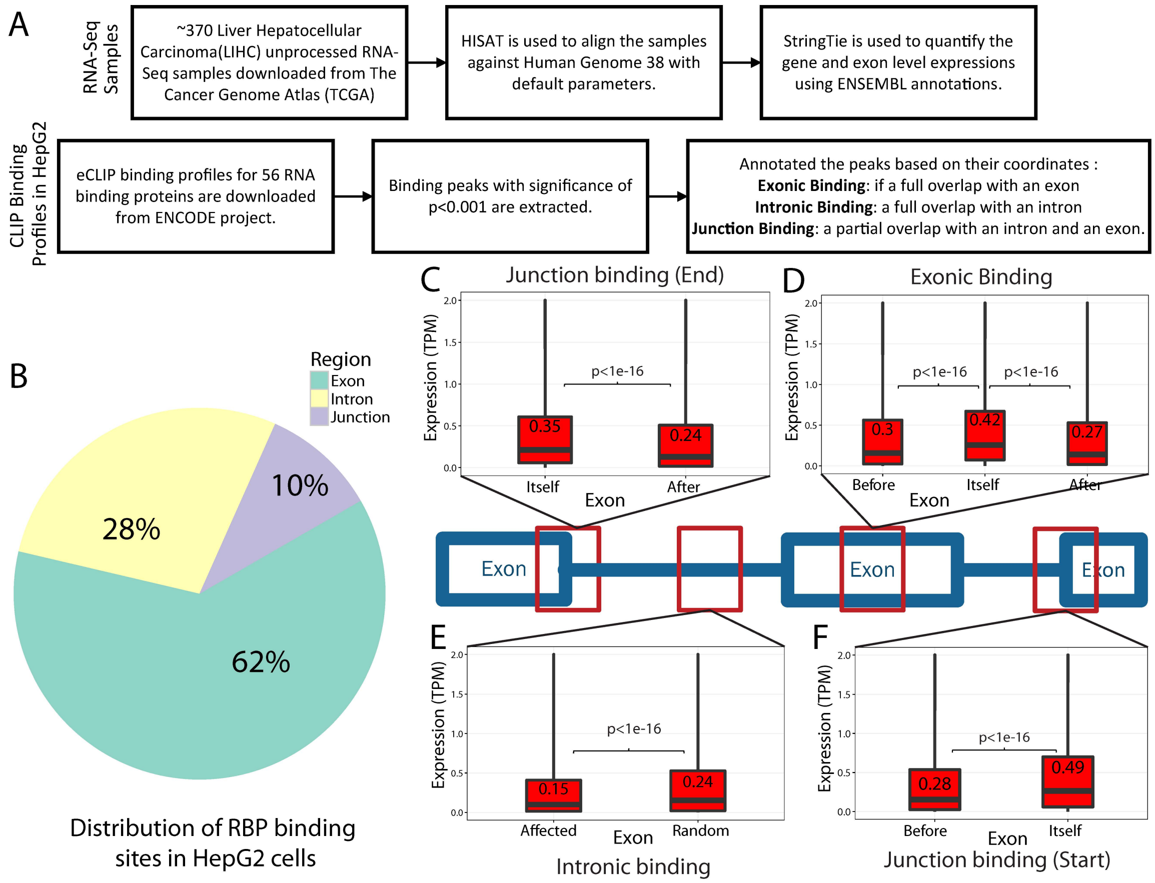
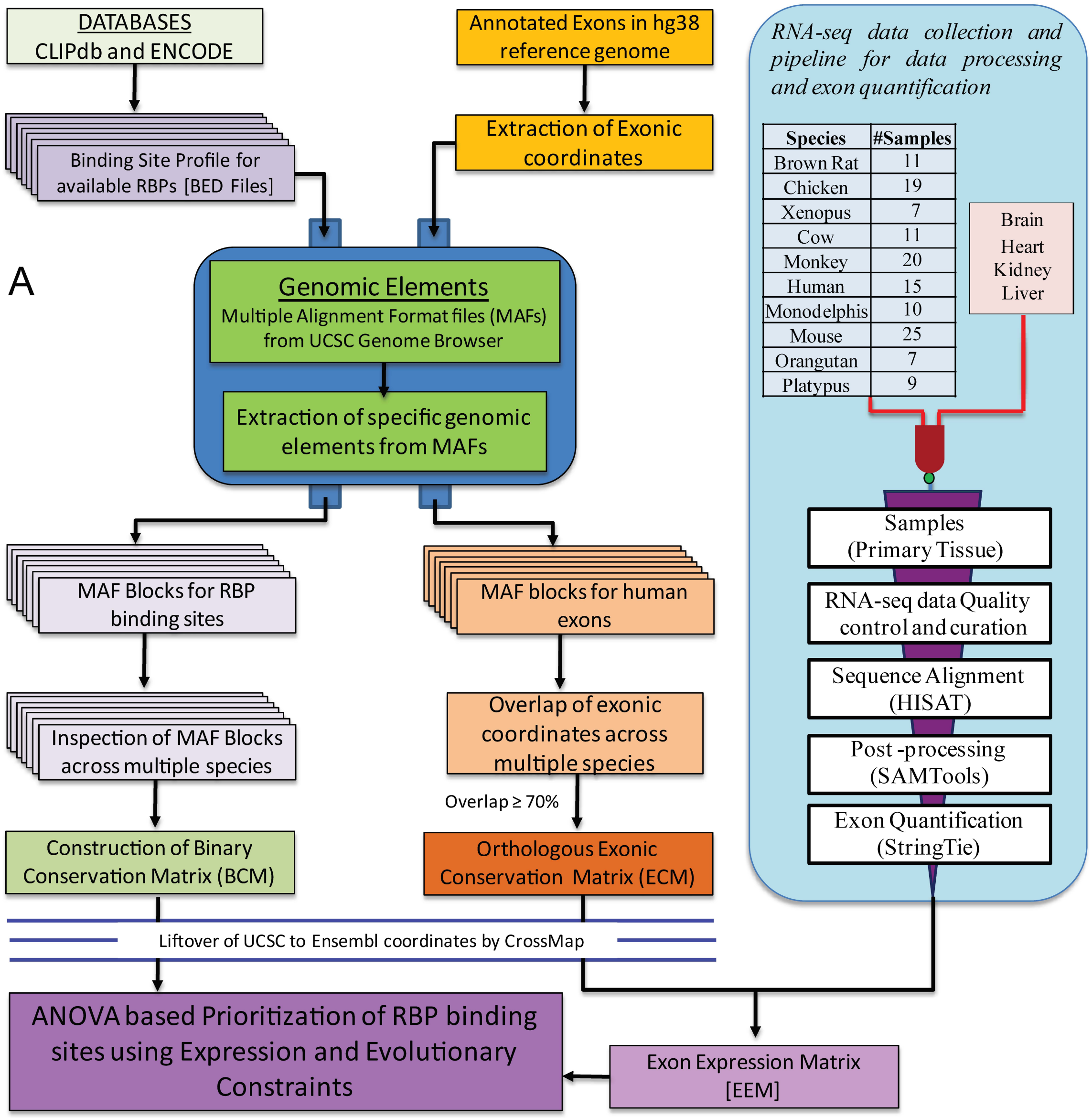


Figure 30: Analysis to demonstrate the relationship between binding site occurrence and expression of proximal exons. (A) The flowchart presents an overview of the steps carried out for processing the datasets employed in the analysis. All binding events of RBPs were classified into exonic, intronic and junction binding. (B) Pie chart representing the distribution of binding sites from eCLIP profile of 56 RBPs on intronic, exonic and junction regions. Box plots in (C) and (F) show a comparison of expression levels between junction bound (marked as 'Itself') and neighboring exons. (D) Boxplots showing expression levels of bound exon (marked as 'Itself') and neighboring exons. (E) Boxplots showing expression of neighborhood exons (marked as 'Affected') and random distal exons from the same gene for intronic binding events. This preliminary analysis confirmed that RBP binding sites can significantly influence the expression of only nearby exons, suggesting that functional binding sites can be identified using an association between binding site and exon expression across species in an evolutionary context.

exon ID, coordinates of the exons in humans and predicted the coordinates of exons in other species using the MAF blocks.

Following the construction of ECM and BCM, all coordinates belonging to an older genome version were updated to the Ensembl 84 version for all species using CrossMap (306). ECM was utilized in the construction of Exon Expression Matrix (EEM). To construct the exon for which the expression values of exons were required across 10 species and 4 vital tissues (brain, kidney, liver and heart). To obtain the expression values, a total of 134 RNA-Seq samples were processed for 10 species (Figure 31A). Quality filtered RNA-seq reads were aligned to the reference genomes using HISAT(117). SAMtools (118) was utilized to convert SAM files to sorted BAM files, and exon level expression quantification was performed using StringTie (120) (see Materials and Methods). Using the expression levels of exons across 10 species and 4 tissues, the Exon Expression Matrix (EEM) was constructed.

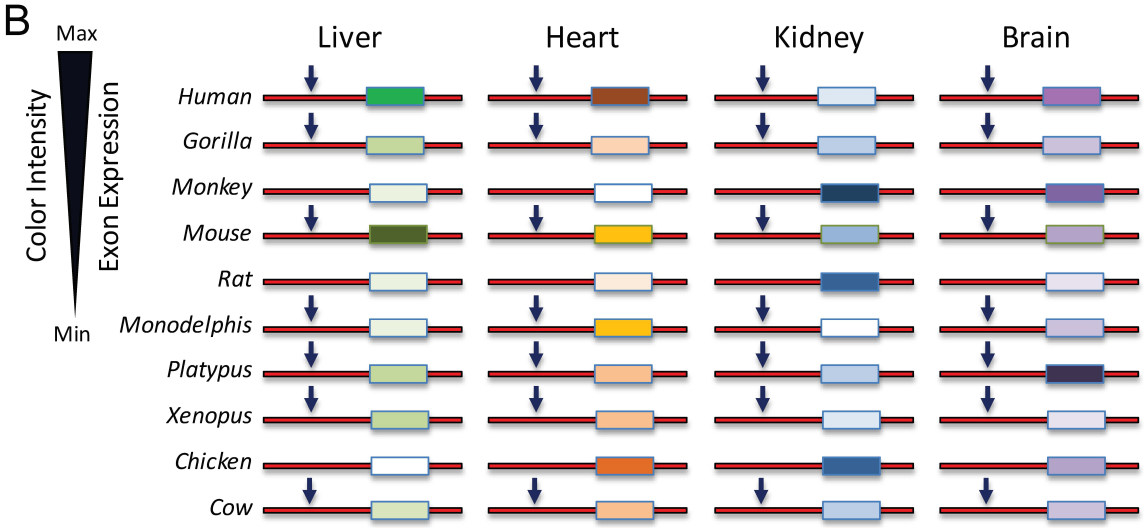
Lastly, ANOVA based prioritization of binding sites of RBPs was accomplished by correlating the EEM and BCM using the R package Matrixeqtl (307). A binding site was considered to control an exon if the distance between them was ≤ 5kb, analogous to the cis-eQTL analysis commonly employed in association studies (Figure 31B). A list of binding sites and exons that were significantly associated to each other were obtained. For this study, these binding sites and exons would henceforth be referred as ‘prioritized’ or ‘functional’ binding sites and exons.


DATABASES
CLIPdb and ENCODE
Binding Site Profile for available RBPs [BED Files]
Annotated Exons in hg38 reference genome
Extraction of Exonic coordinates
A
Genomic Elements
Multiple Alignment Format files (MAFs) from UCSC Genome Browser
Extraction of specific genomic elements from MAFs
MAF Blocks for RBP binding sites
Inspection of MAF Blocks across multiple species
Construction of Binary Conservation Matrix (BCM)
MAF blocks for human exons
Overlap of exonic coordinates across multiple species
Overlap ≥ 70%
Orthologous Exonic Conservation Matrix (ECM)
Liftover of UCSC to Ensembl coordinates by CrossMap
ANOVA based Prioritization of RBP binding sites using Expression and Evolutionary Constraints
Exon Expression Matrix [EEM]
RNA-seq data collection and pipeline for data processing and exon quantification
Brain
Heart
Kidney
Liver
Samples (Primary Tissue)
RNA-seq data Quality control and curation
Sequence Alignment (HISAT)
Post -processing (SAMTools)
Exon Quantification (StringTie)

Species	#Samples
Brown Rat	11
Chicken	19
Xenopus	7
Cow	11
Monkey	20
Human	15
Monodelphis	10
Mouse	25
Orangutan	7
Platypus	9

PEEK: Prioritization of RBP binding sites using Expression and Evolutionary Constraints

Figure 31: Overview of PEEK (A) A computational framework for prioritization of RBP binding sites using expression and evolutionary constraints (See methods).



(B) Influence of prioritized binding site over associated exon expression. Using the PEEK method, the prioritization of RBPs' binding sites in multiple tissues was carried out. The figure imitates the influence of prioritized binding site over the associated nearby (≤ 5000 bp) exon expression level in available tissue types across multiple species.

4.3.3.3 PEEK illustrates only a small fraction of binding sites of RBPs are functional across tissues

PEEK prioritized binding sites were investigated across the 4 body sites; brain, liver, kidney and heart. It was observed that about 2.4% (~184,000 binding sites for each tissue type) of the experimentally determined RBP binding sites obtained from CLIPdb and ENCODE were prioritized at FDR (310) ≤ 0.05 (Figure 32A). These functional binding sites were used in the subsequent downstream analyses. The percentage of binding sites that were prioritized, dropped to 0.5 percent at FDR< 0.01, implying that only a small fraction of the experimentally discovered sites is likely to have a 'functional' impact in a tissue specific manner at the post-transcriptional level. Interestingly, the brain harbored the highest number of prioritized binding sites at FDR ≤ 0.05, followed by kidney, liver and then heart. The fraction of prioritized binding sites observed in PEEK, further concorded with the level of structural and functional complexity of these vital organs with brain being the most complex organ (311).

4.3.3.4 Significant fraction of functional binding sites is tissue specific

This study indicated that among the total 339,576 prioritized binding sites identified across all the four tissues, about 10% are unique to each tissue and ~55% of them were prioritized in more than one tissue. This observation further indicate that majority of the functional binding sites could be active across multiple tissues (Figure 32B). Brain was found to have the highest number of uniquely prioritized binding sites among all tissues, suggesting a higher activity of RBPs in the brain as opposed to other tissues.

4.3.3.5 Majority of the functional binding sites are intronic

Introns are highly variable but evolutionary conserved genomic elements with relatively higher level of conservation near exons (295). These genomic elements have been extensively implicated in a variety of gene regulation mechanism (312-314). For instance - it contain several tissue-specific branch point sequence (BPS) at splice site that determine the fate of intron exclusion during RNA splicing in human (315). Recently, Sun L et al, have experimentally demonstrated that introns are highly structured than exons in vivo (316) and that significantly impacts the regulation of pre-mRNA splicing (317) and several other post transcriptional processes. This urges for an extensive investigation of highly orchestral regulatory mechanism in context to RBPs binding site and the neighboring exons. The study demonstrated that the genomic features associated with the functional binding sites as illustrated in Figure 32B clearly revealed an

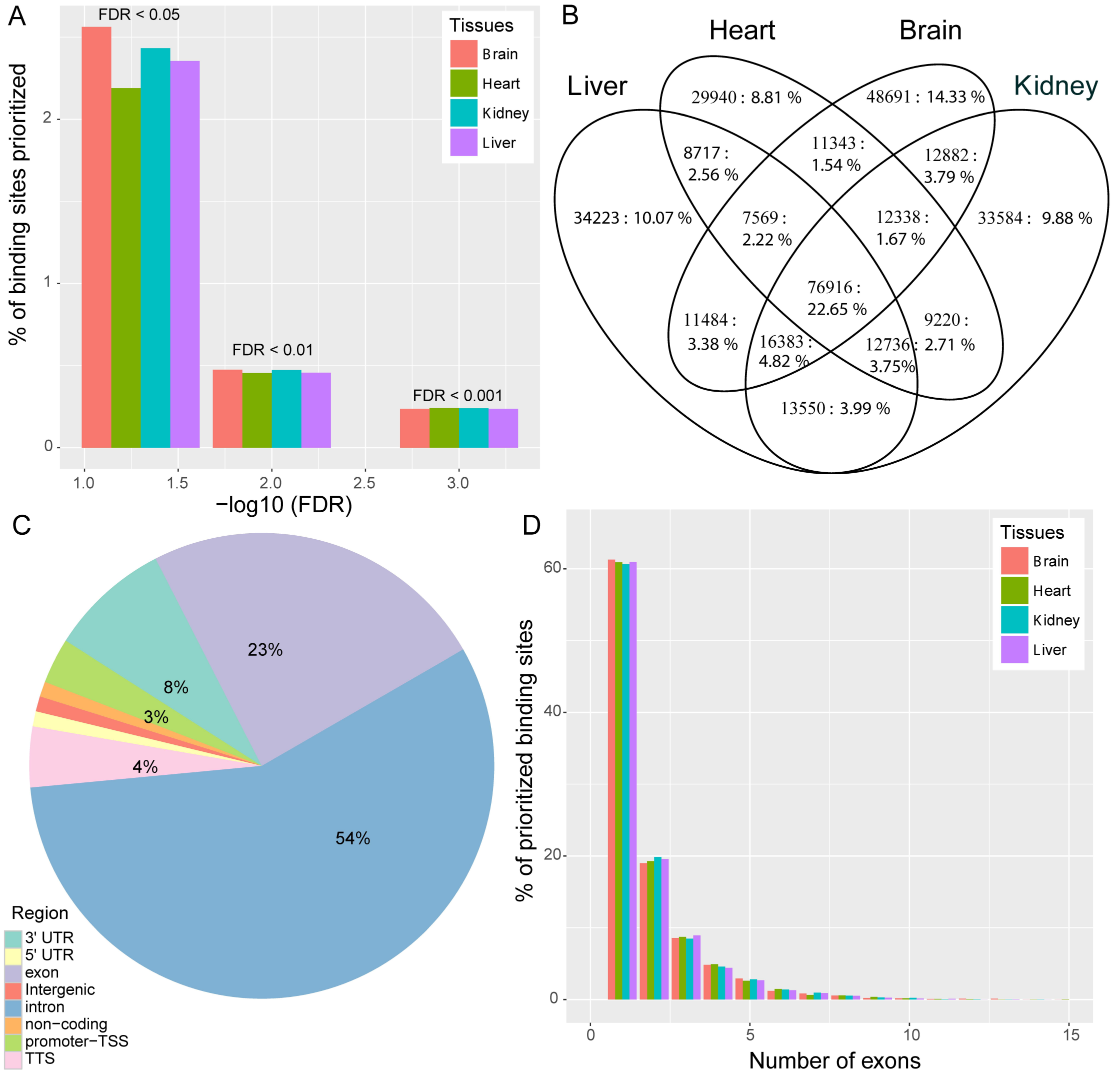


Figure 32: (A) Percentage of binding sites detected as functional at various FDR thresholds. 2.56% (224,000) out of a total of 8 million binding sites were found to be functional at FDR < 0.05. The percentage decreases to 0.5% and 0.25% as the FDR thresholds are lowered to 0.01 and 0.001 respectively. (B) Tissue wise segregation of functional binding sites. The Venn diagram depicts the number and percentage of binding sites that are functional across each tissue (FDR < 0.05), and those that are common among all tissues. 14% of the total prioritized binding sites are unique to brain and about 22% of the binding sites are common among all tissues. (C) A majority of the functional binding sites are intronic. The location of prioritize binding sites at the gene level was uncovered using HOMER (308). (D) Percentage of prioritized binding sites vs number of exons associated. The figure shows the percentage of prioritized binding sites that have been associated to exons. 60% of the prioritized binding sites are associated to 1 exon (FDR < 0.05), and about 20% of the binding sites are associated to 2 exons. Overall, 80% of the binding sites are associated to at most 2 exons.

enrichment for intronic regions (Odds Ratio = 2.38, p < 2.2e-16, Fisher's test) (Figure 32C). Likewise, an under-representation of functional binding sites was found in the exonic regions (Odds Ratio = 0.4, p < 2.2e-16, Fisher's test) suggesting that majority of functional binding sites are unlikely to be identified by traditional methods that employ coding sequence or structural information.

4.3.3.6 Most binding sites influence at most 2 proximal exons

The percentage of prioritized binding sites associated with number of prioritized exons identified using PEEK framework were calculated. Upon segregating these prioritized exons based on the number of binding sites they were associated with; it was revealed that about 80% of the prioritized binding sites influenced the expression of at most 2 exons across all four tissues (Figure 32D). Although the threshold of distance used in the association analysis could contribute to these fractions, it is possible to speculate that majority of the functional binding sites are likely to control at most two nearby exons. It was also observed that the distance between the functional binding site and its associated exon is fairly uniform within the 5kb threshold employed in the analysis (Appendix 8).

4.3.3.7 PEEK identified functional binding site influences the expression of proximal exons

Next, the influence of functional binding sites on prioritized exons was investigated in gene centric and transcriptome wide case studies. Firstly, SF3B1 (Splicing Factor 3b Subunit 1) was selected that plays an important role in regulation of several genes involved in cell cycle, RNA processing and telomere maintenance (318, 319). Canonically, SF3B1 has been characterized as a crucial splicing co-factor of SF3B complex, mainly involved in BPS (branchpoint sequence) recognition and modelling of the transcriptome (320). Several driver mutations in this gene have been implicated for anomalous RNA splicing in multiple cancers (321, 322). In this study, I investigated all the functional binding sites and associated exons in SF3B1 (chr2: 197388515-197435091:-1) identified by PEEK as shown in Figure 33A, where exons are color coded (intensity) as per the proportion of associated functional binding sites from PEEK in a tissue specific manner. This study identified a functional binding site targeted by TROVE2, which was found significantly associated (adjusted p-value$<10^{-4}$) with the expression of proximal exon, ENSE00000964873 (Figure 33B) in all four tissues. Interestingly, it was observed that, this exon showed ~200 folds decreased expression in human and ~20 fold decreased expression in mouse with respect to that in Platypus. The study speculates that the absence of associated regulatory site contributes to such atypically higher expression of this exon in Platypus (Figure 33B).

For transcriptome-wide case study, it was investigated whether prioritized exons associated to functional binding sites showed a different expression profile than that of random exons. The exon level expression was extracted from cancer patients of three types; Glioblastoma, Liver Hepatocellular Carcinoma and Kidney Renal clear cell carcinoma for brain, liver and kidney respectively (see Materials and Methods). The data of cancer patients showed that the expression level of prioritized exons was significantly higher than the expression level of random exons for each tissue type (Appendix 9). To confirm whether prioritized exons indeed exhibited variation in expression than random exons, the Median Absolute Deviation (MAD) values were computed for each prioritized exon across patients (Figure 33C). The variability in the expression levels of prioritized exons was significantly higher than the variability in the expression levels of random exons across cancer patients, and this observation was consistent across all tissues.

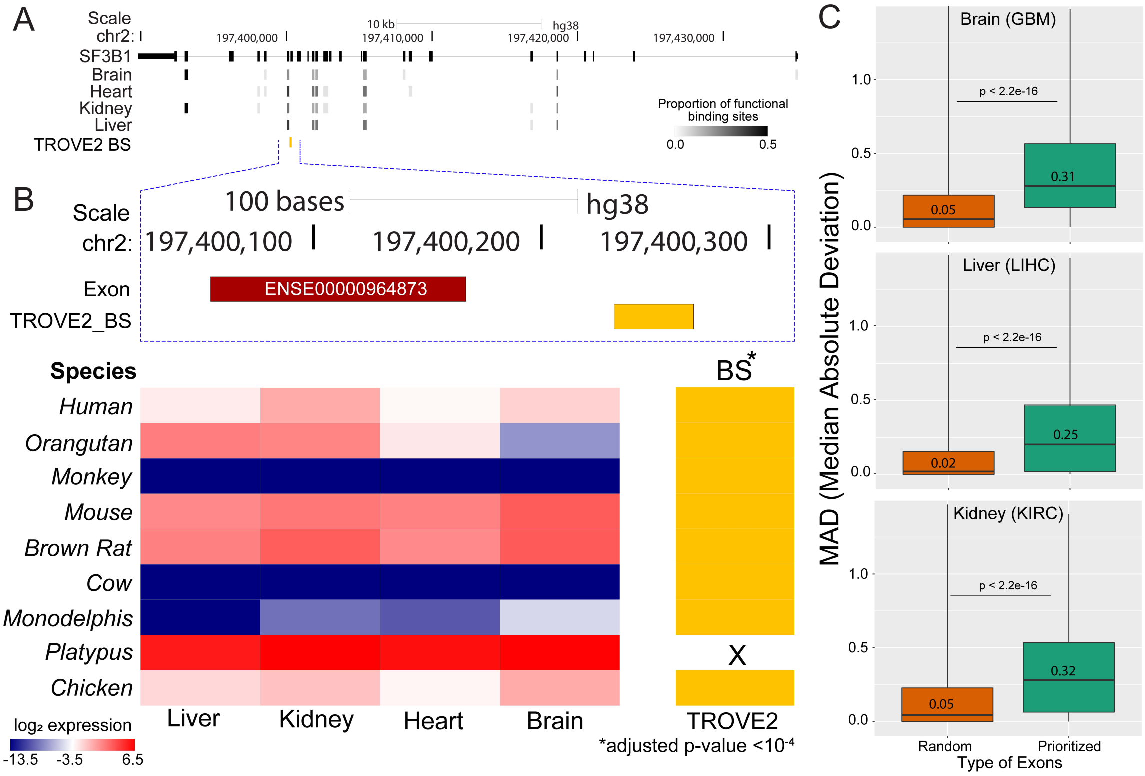


Figure 33: PEEK identified functional binding site influences the expression of proximal exons. (A) Genomic tracks showing the genomic boundary of SF3B1 where functionally important exons identified by PEEK were color coded (intensity) as per the proportion of associated functional binding sites in a tissue specific manner. (B) A functional binding site (in yellow color) targeted by TROVE2, was identified which was found significantly associated (adjusted p-value $<10^{-4}$) with the expression of proximal exon (ENSE00000964873) in human and equivalent in other species across all four tissues. Expression profile of associated exon is shown as heatmap. (C) The influence of functional binding sites on prioritized exons was investigated across cancer patients. Expression levels of exons in each tissue were obtained from cancer patients; Glioblastoma, Liver Hepatocellular Carcinoma and Kidney Renal clear cell carcinoma for brain, liver and kidney respectively. The box plots represent the distribution of median absolute deviation (MAD) values (i.e. variability in the expression level of each exon across different patients) of the prioritized exons vs. the random exons in different cancers.

4.3.3.8 Crispr cas9 system verifies PEEK prioritized binding sites

I employed Crispr Cas9 system for verification of functional binding sites identified by PEEK in HepG2 cell line. The experiment was strategically designed to validate the functional impact of prioritized binding sites by estimating the expression change of proximal exons upon perturbation of (a) single binding site (b) distance measured binding sites and (c) locus specific regulatory binding sites in respective genomic loci. For this analysis, I selected PEEK prioritized BS and associated proximal exons (at distance < 1.5 kb) in the genomic loci of three genes; RNPEPL1, FUS and RBCK1.

These genes were selected based on their association to a diverse functional mechanism in liver metabolism and diseases. For instance - RNPEPL1 encodes for an aminopeptidase which preferentially hydrolyzes an N-terminal methionine, citrulline or glutamine(323). I observed that high expression of RNPEPL1 could be prognostic ($p<0.00001$) for liver cancer in TCGA cohort (https://www.proteinatlas.org/ENSG00000142327-RNPEPL1/pathology/liver+cancer). Fus is an important sub-component of the heterogeneous nuclear ribonucleoprotein (hnRNP) complex (324). It plays a vital regulatory role in RNA metabolism (325) and other cellular processes such as DNA repair mechanism (326). This gene, in association with LATS1/2 activates Hippo pathway and hence inhibits the progression of hepatocellular carcinoma (HCC) (327). Similarly, RBCK1 encodes for E3 ubiquitin-protein ligase protein that transfers ubiquitin from E2-complex to its substrates (328). RBCK1 negatively regulates tumor necrosis factor(329) and has been associated with the pathogenesis of liver cancer (330).

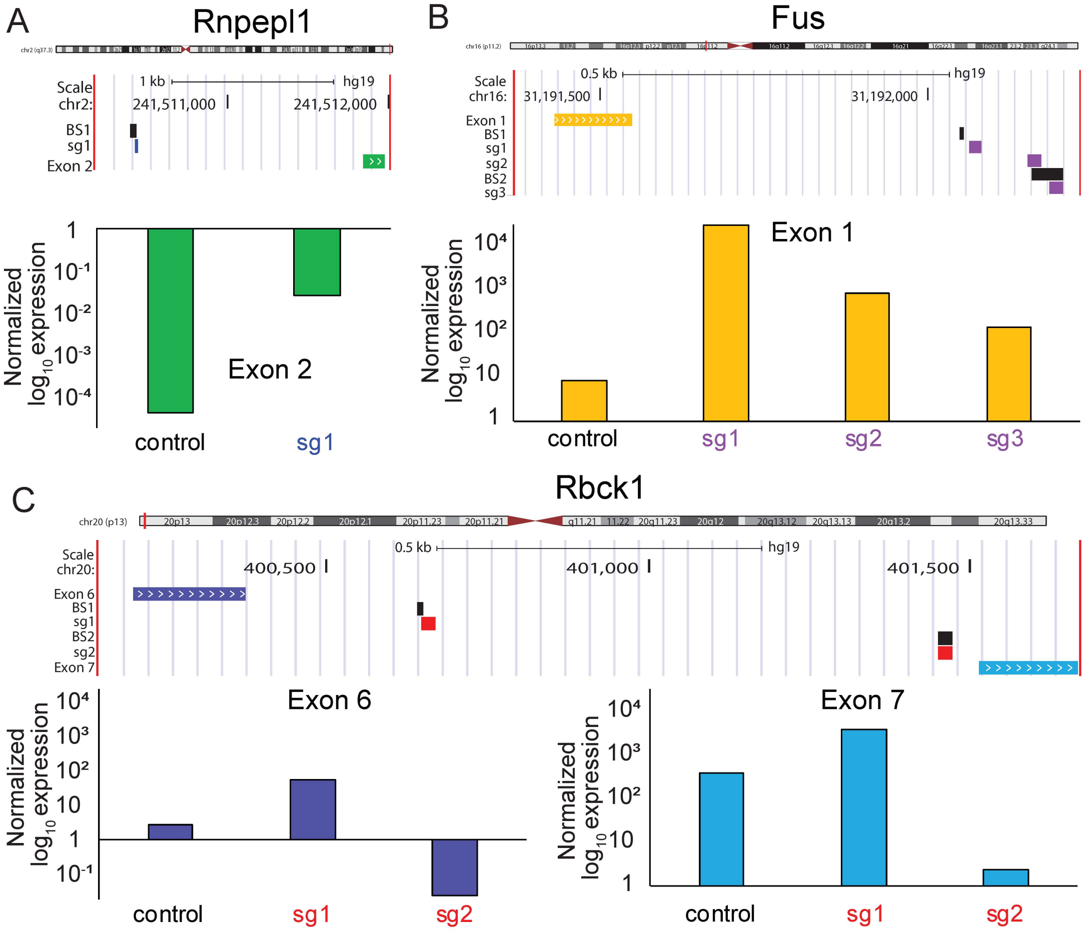


Figure 34: Validation of PEEK prioritized binding using Crispr cas9 system. Crispr Cas9 system was employed (see Materials and Methods) to validate the functional impact of PEEK prioritized binding sites by estimating the expression change of proximal exons (at distance < 1.5 kb) upon perturbation of (A) *single binding site* in RNPEPL1 (B) *distance measured binding sites* in FUS and (C) *locus specific regulatory binding sites* in RBCK1. For each gene, primers were designed to estimate the abundance of the exon associated to PEEK BS in respective genomic loci. The exon expression was measured (normalizing with housekeeping gene PPIA) in wild type and CRISPR edited cells through qRT-PCR.

I used SliceIt (7) to extract potential sgRNAs (efficiency > 0.70 and specificity > 70%) designed to target these PEEK BS with minimal distance between on-site PAM (Protospacer Adjacent Motif) and center of the BS. sgRNA guided CrisprCas9 system was established to perturb these selected RBP binding sites in HepG2 cells (Figure 34, Materials and Methods). For each gene, primers were designed to estimate the abundance of the exon associated to PEEK BS.

The exon expression (normalizing with housekeeping gene PPIA) was measured and analyzed in wild type and CRISPR edited cells through qRT-PCR. A significant increase ($p<0.001$) in expression of (a) RNPEPL1-Exon2 was observed upon perturbation of *single binding site* (Figure 34A) and (b) FUS-Exon1 upon perturbation of *distance measured binding sites*, with minimal increase in distant binding site (Figure 34B) with respect to wild type. In RBCK1, I found that the selected prioritized BS1 and BS2 upon perturbation, independently contributes to significant increase and decrease in expression of proximal exons respectively (Figure 34C) and hence confirms the (c) *locus specific regulatory binding site* associated to cellular process. Nonetheless, the qPCR results confirmed that PEEK prioritized binding sites, significantly affect the change in exon expression levels as a result of the perturbation of the binding sites. Interestingly, this study demonstrated several instances to further confirm the functional efficacy of PEEK prioritized binding sites where RBP binds and regulates in site specific manner (Figure 34).

4.3.4 Conclusion

Existing databases such as ENCODE and CLIPDB have documented millions of binding sites of more than 135 RBPs, a method for prioritizing binding sites of RBPs based on their biological significance is still lacking. PEEK was developed as a semi-automated and scalable computational framework, that prioritize the documented RBP-binding sites based on their biological importance in a tissue specific manner. This study identified that majority of the binding sites of RBPs influence the proximal exons.

Interestingly, only a small fraction i.e. 2.4% (184,000) of all binding sites of RBPs are functional across tissues (at 5% FDR) and significantly contribute to tissue specific binding. Majority of the functional binding sites were found to be intronic and could influence at most 2 proximal exons. The PEEK identified functional binding sites were annotated in multiple cancers and observed that the expression of exon proximal to these binding sites was significantly higher than random exons. Hence, this study presents a novel approach which facilitates a detailed insight of the functional binding site and proximal exon to dissect the regulatory mechanism underlying in multiple cancers.

CHAPTER 5

SUMMARY

Regulatory proteins such as TFs and RBPs are highly appreciated for complex interplay with the respective targeted genomic/transcriptomic elements via TRNs and PTRNs. The mechanistic understanding of these two regulation types require high resolution tissue-specific functional annotation of both the proteins as well as their target sites.

I reconstructed a high resolution roadmap of gene centric transcriptional regulation of two genes; Uromodulin in kidney (72) and Sestrin3 in liver cells (93) by implementing a novel in silico phylogenetic foot printing approach (37) on the upstream regulatory regions of a diverse set of individual gene orthologs. This analysis allowed the identification of a reliable set of binding motifs in the upstream regulatory regions and constructed a high confidence compendium of transcription factors involved in gene regulation processes. (Chapter 2)

This study elaborates the understanding of transcriptome profiling and temporal post transcriptional switching of isoform in developing mouse eye. For instance - Express(166) unifies various mouse lens and retina RNA-seq data and provides user-friendly visualization of the transcriptome to facilitate gene discovery in the eye. It serves as an effective portal for analyzing the pruned RNA-seq expression datasets presently collected for the lens and retina. It also allows a wild-type context for the detailed analysis of targeted gene-knockout mouse ocular defect models and facilitate the prioritization of candidate genes from RNA-seq data of eye disease patients.

Transcriptomic alterations and splicing events were also investigated during mouse lens formation using RNA-seq data from multiple developmental stages and constructed a molecular portrait of known and novel transcripts. This study elucidates that the extent of novelty of expressed transcripts decreases significantly in post-natal lens compared to embryonic stages. Also, examination of the splice isoforms revealed skipped exon and retained intron to be the most abundant alternative splicing events during lens development. Further, a splicing browser, Eye Splicer was developed (http://www.iupui.edu/~sysbio/eye-splicer/) to facilitate exploration of developmentally altered splicing events. This study improves the current knowledge of post-transcriptional regulatory networks during mouse lens development. (Chapter 3)

In this study, a computational framework for systematic tissue-specific annotation of functional binding sites of RBPs was developed in the human genome to uncover disease associated binding events and the PTRNs. Several tools such as Seten (253) and SliceIt (7) were also developed, that enables the user to annotate the condition-specific CLIP-seq profiles with relevant biological processes, phenotypes, and diseases associated with RBPs. In particular, SliceIt efficiently provide a multi-omics resource for designing Crispr Cas9 experiments to verify the functionality of these RBP binding profiles. (Chapter 4)

A computational framework, PEEK was developed that employ tissue-specific cross-species RNA-seq information from more than 100 samples encompassing 4 tissue (Kidney, Liver, Brain, Heart) and 10 species, to prioritize and evolutionarily annotate the binding sites of RBPs across tissues and validate several of these high confidence functional binding sites predicted to control the proximal exons in human cell lines. (Chapter 4)

5.1 Significance and Innovation

My thesis potentially contributes to the research community by providing methods, web interface and software which transforms the ability to build high-quality regulatory binding maps of RBPs and TF's in a tissue specific manner using RNA-seq datasets. For instance- the novel method '*in silico* phylogenetic foot printing' developed in this study illustrates a genome wide application to scrutinize the conserved TF binding motifs and delimits the identification of functionally important cis regulatory genomic loci and enhancers. This method could also help to understand the diaspora of regulatory genomic elements departed due to evolution. The method is scalable and could be used for deriving the novel potential therapeutic targets to synchronize the expression of causative genes in several disorders.

My thesis provides a broad spectrum of temporal and evolutionary dynamics of transcriptome and their regulation at transcriptional and post transcriptional level. The approach employed in identification and characterization of transcripts, including novel transcripts across developmental stages of mouse eye, made an appeal on improving the genomic annotations to further understand the complete transcriptomic architecture, especially in stage specific disorders, but not limited to eye.

My thesis provides several methods and web interfaces that further advance the ability to functionally annotate hundreds of RBPs and their RNA binding sites across tissues in the human genome. For instance – the novel method "PEEK" developed in this study, provides a semi-automated and scalable computational framework to annotate and identify functional binding sites of RBP in multiple tissue type across species and can efficiently delineate the functionally relevant binding sites from millions of CLIP binding sites. It enables the systematic annotation of these functional sites in vertebrate genomes. It provides the fundamental knowledge about the roles of RBPs in the context of disease phenotype, networks and pathways. Overall, this study is a significant piece of work which can accelerate the progress in molecular diagnostics and drug target identification.

5.2 Future work

In future, I aim to conduct a large scale pooled Crispr/Cas9 based screening experiment for system level verification of functional binding sites bound by RBPs. I will be designing ~20000 top ranked guide-RNAs (sgRNAs) using SliceIt (7) that specifically targets these binding sites. I will further develop a computational pipeline to analyze the Crispr/Cas9 perturbed RNA-seq readouts to observe any locus specific expression changes proximal to binding sites.

Taken together, this study offers a wide range of applications to the biomedical researchers by aiding the identification of crucial therapeutic targets that are significantly regulated by RBP’s in physiological as well as pathological conditions.

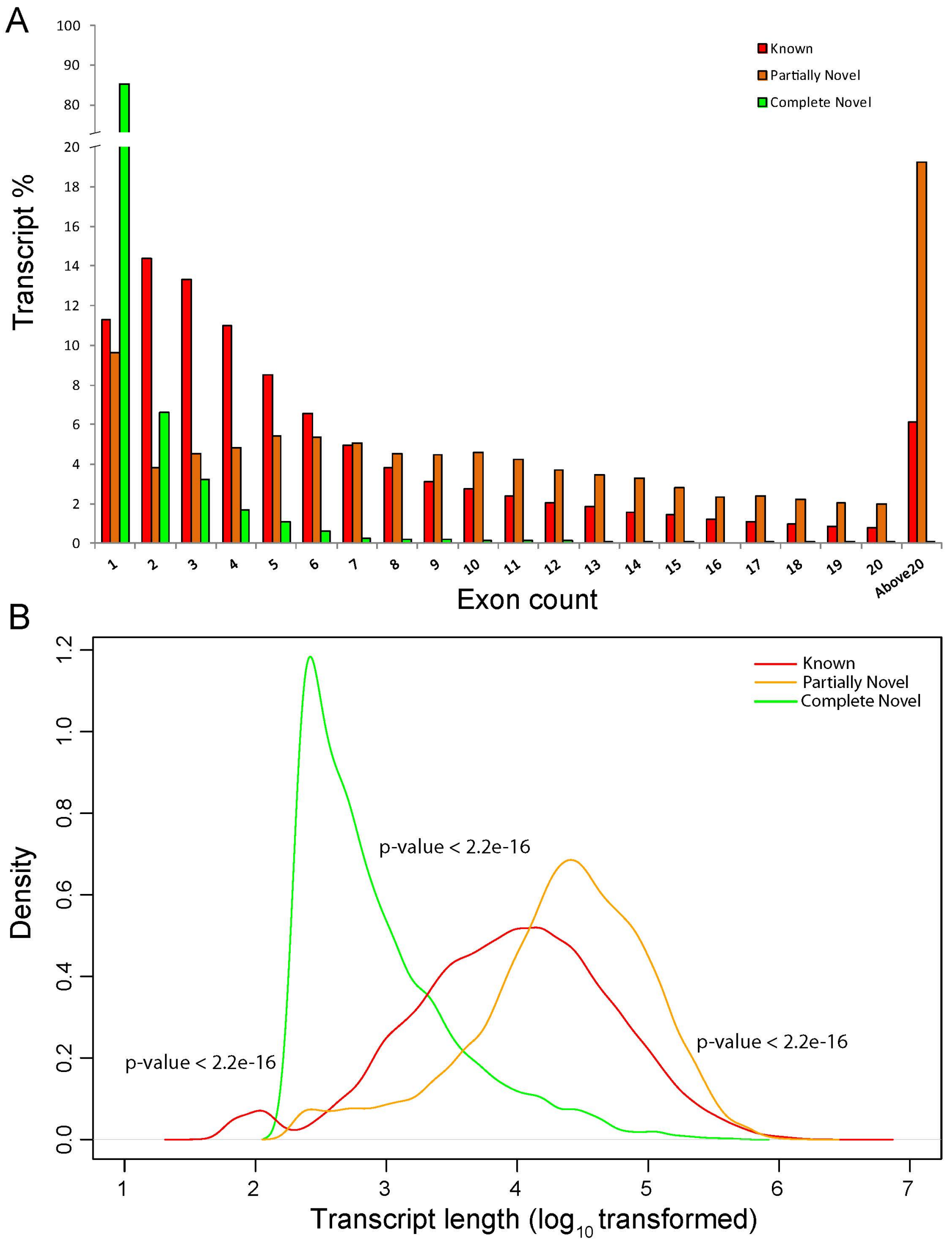
APPENDICES

#	SRA ID	PMID	Development Stage	Read Type	Read Length	Read Count	Base Count	Overall Alignment Rate (%)
1	SRR2039769	26225632	E15	PE	100	13772390	2754478000	94
2	SRR2039770	26225632	E15	PE	100	13542500	2708500000	95
3	SRR953395	24161570	E15.5	SE	52	48552190	2524713880	94
4	SRR953394	24161570	E15.5	SE	52	47574424	2473870048	94
5	SRR953393	24161570	E15.5	SE	52	42525381	2211319812	94
6	SRR2039771	26225632	E18	PE	100	17810970	3562194000	93
7	SRR2039772	26225632	E18	PE	100	18019388	3603877600	93
8	SRR1222595	25489224	P0	SE	51	33174286	1691888586	88
9	SRR1222596	25489224	P0	SE	51	29919226	1525880526	87
10	SRR1222672	25489224	P0	SE	51	29965660	1528248660	89
11	SRR1222673	25489224	P0	SE	51	28652759	1461290709	86
12	SRR1222674	25489224	P0	SE	51	30661663	1563744813	89
13	SRR1222675	25489224	P0	SE	51	24833352	1266500952	89
14	SRR2039773	26225632	P0	PE	100	17766309	3553261800	93
15	SRR2039774	26225632	P0	PE	100	14533000	2906600000	93
16	SRR2039775	26225632	P3	PE	100	15495833	3099166600	93
17	SRR2039776	26225632	P3	PE	100	13072393	2614478600	93
18	SRR2039777	26225632	P6	PE	100	16965754	3393150800	93
19	SRR2039778	26225632	P6	PE	100	17658286	3531657200	93
20	SRR2039779	26225632	P9	PE	100	18874309	3774861800	93
21	SRR2039780	26225632	P9	PE	100	13563853	2712770600	93

Appendix 1: RNA-seq samples for mouse lens.
The table shows SRA ID (Sequence Read Archive) for the sample, PMID (PubMed) for the study, developmental stage, read type, read length, read count, base count, and overall alignment rate using HISAT (Hierarchical Indexing for Spliced Alignment of Transcripts).

#	SRA ID	PMID	Development Stage	Read Type	Read Length	Read Count	Base Count	Overall Alignment Rate (%)
1	SRR1023063	24382353	P2	SE	76	29939891	2275431716	93
2	SRR1023064	24382353	P2	SE	76	36280541	2757321116	94
3	SRR1784052	26324254	P10	SE	50	27028041	1351402050	94
4	SRR1784053	26324254	P10	SE	50	24267068	1213353400	94
5	SRR1784054	26324254	P10	SE	50	28777255	1438862750	94
6	SRR1574329	25801704	P11	PE	90	108113500	19460430000	96
7	SRR1574330	25801704	P11	PE	90	106003809	19080685620	97
8	SRR1574333	25801704	P11	SE	90	176162120	8631943880	87
9	SRR1574334	25801704	P11	SE	90	176206610	8634123890	87
10	SRR1023073	24382353	P21	SE	76	39574659	3007674084	94
11	SRR1023074	24382353	P21	SE	76	38855951	2953052276	94
12	SRR1784070	26324254	P21	SE	50	34792825	1739641250	95
13	SRR1784071	26324254	P21	SE	50	31461693	1573084650	95
14	SRR1784072	26324254	P21	SE	50	37264811	1863240550	95
15	SRR1176996	24812086	P28	SE	50	45056397	2252819850	92
16	SRR1176997	24812086	P28	SE	50	51508183	2575409150	91
17	SRR1176998	24812086	P28	SE	50	52339450	2616972500	92
18	SRR1687694	25712131	P30	PE	100	16507045	3334423090	84
19	SRR1687695	25712131	P30	PE	100	16384187	3309605774	84
20	SRR1687696	25712131	P30	PE	100	14348324	2898361448	81
21	SRR1687697	25712131	P30	PE	100	14251738	2878851076	81
22	SRR1687698	25712131	P30	PE	100	25674252	5186198904	83
23	SRR1427139	25002228	P30	SE	51	44636149	2276443599	68
24	SRR1427140	25002228	P30	SE	51	40396217	2060207067	66
25	SRR1427141	25002228	P40	SE	51	52430453	2673953103	74
26	SRR1427142	25002228	P40	SE	51	46344921	2363590971	75
27	SRR1213798	25489233	P48	PE	90	6779886	1220379480	96
28	SRR1213799	25489233	P48	PE	90	6803332	1224599760	97
29	SRR1213800	25489233	P48	PE	90	6826280	1228730400	96
30	SRR1427143	25002228	P50	SE	51	42904725	2188140975	70
31	SRR1427144	25002228	P50	SE	51	41243569	2103422019	68
32	SRR1427145	25002228	P60	SE	51	56910039	2902411989	48
33	SRR1427146	25002228	P60	SE	51	46737505	2383612755	56
34	SRR1427147	25002228	P90	SE	51	40366735	2058703485	76
35	SRR1427148	25002228	P90	SE	51	45096977	2299945827	71

Appendix 2: RNA-seq samples for mouse retina.
The table shows SRA ID (Sequence Read Archive) for the sample, PMID (PubMed) for the study, developmental stage, read type, read length, read count, base count, and overall alignment rate using HISAT (Hierarchical Indexing for Spliced Alignment of Transcripts).



Appendix 3: Analysis of the genomic structure of identified transcripts.
(A) Histogram showing the distribution of the number of exons for known, partially novel and completely novel transcripts. (B) Kernel density distribution of transcript lengths (log10 transformed) for known, partially novel and completely novel. Statistical differences in the distributions of lengths were computed using the non-parametric Kolmogorov–Smirnov test on every pair of transcript types. Completely novel transcripts as a group were found to be significantly shorter than both partially novel and known transcripts. In contrast, partially novel transcripts were found to be longer than even the known transcripts.

(A) Selected high confident exon skipping events detected using rMATS pipeline (FDR <0.01) across developmental stages with replicates. Values across stages correspond to PSI values of the exons.

Exon ID	Coordinates (mm10)	strand	Gene Name	E15	E18	P0	P3	P6	P9
ENSMUSE00000334942	19:57051130-57051234	-	Ablim1	NA	0.13	0.61	0.729	0.66	0.612
ENSMUSE00000668725	3:148849766-148849804	-	Adgrl2	0.202	0.24	0.61	NA	0.57	0.611
ENSMUSE00001324776	1:82891460-82891507	+	Agfg1	0.232	0.18	0.41	0.532	0.48	0.489
ENSMUSE00001039657	18:6057517-6057591	-	Arhgap12	0.289	0.32	0.82	NA	NA	NA
ENSMUSE00000700987	2:10056770-10056806	-	Atp5c1	0.822	0.76	0.55	0.55	0.62	0.574
ENSMUSE00000230008	18:32426224-32426352	+	Bin1	NA	0.68	0.18	NA	NA	NA
ENSMUSE00000217920	9:70004306-70004341	+	Bnip2	NA	0.5	0.9	0.806	NA	0.845
ENSMUSE00000736151	10:127064202-127064453	+	Cdk4	0.959	0.96	0.8	NA	NA	NA
ENSMUSE00000691476	5:112251747-112251797	-	Cryba4	0.974	0.98	NA	0.994	0.99	0.99
ENSMUSE00000311733	14:47726471-47726554	+	Ktn1	NA	0.24	0.5	0.476	NA	0.517
ENSMUSE00000440236	6:93680789-93680877	-	Magi1	0.118	0.09	NA	NA	0.48	0.655
ENSMUSE00000667965	7:143518850-143518885	-	Nap1l4	0.477	0.42	0.29	0.228	0.25	0.245
ENSMUSE00001311933	2:105695306-105695456	+	Pax6	0.995	1	0.94	0.993	1	NA
ENSMUSE00000317905	15:93452117-93452173	+	Pphln1	0.151	0.07	0.57	0.356	0.5	0.639
ENSMUSE00000635082	9:86790056-86790139	-	Snap91	NA	0.85	0.27	0.248	0.2	NA
ENSMUSE00001196118	11:80393084-80393176	+	Zfp207	0.368	NA	0.67	NA	0.61	0.593

(B) Selected high confident intron retention events detected using rMATS pipeline (FDR <0.01) across developmental stages with replicates. Values across stages correspond to PSI values of the exons.

Exon ID	Coordinates (mm10)	strand	Gene Name	E15	E18	P0	P3	P6	P9
ENSMUSE00000784872	3:103174340-103177419	+	Bcas2	0.039	0.0685	0.026	0.035	0.0225	0.031
ENSMUSE00000787504	11:101295534-101296316	-	Becn1	NA	0.1185	0.0635	0.062	0.051	0.044
ENSMUSE00000643467	2:91013238-91019497	+	Celf1	0.3755	0.709	0.612	0.8395	0.8565	0.846
ENSMUSE00001342001	1:165338188-165340023	-	Dcaf6	0.082	0.1555	0.0585	0.0635	0.052	0.0535
ENSMUSE00000842895	11:106782469-106784018	-	Ddx5	0.1375	0.1835	0.2265	0.3185	0.302	0.3625
ENSMUSE00000492954	3:95628541-95632102	+	Ensa	0.458	0.6615	0.4515	0.4335	0.4195	NA
ENSMUSE00001357022	3:152213977-152215630	+	Fubp1	0.0095	0.0535	0.058	0.0585	0.032	0.0515
ENSMUSE00000857219	1:161038225-161038539	+	Gas5	0.2265	0.289	0.0485	0.068	0.1195	0.099
ENSMUSE00001326780	7:31134414-31135739	-	Gramd1a	0.4465	0.7265	0.28	0.4175	0.319	0.307
ENSMUSE00000765273	11:50379468-50379964	+	Hnrnph1	0.093	NA	0.2775	0.1575	0.181	0.165
ENSMUSE00000756514	X:95947770-95950446	-	Las1l	0.0135	0.034	0.0545	0.041	0.043	0.0575
ENSMUSE00000764755	X:94537676-94538065	-	Maged1	0.064	0.1	0.0455	0.0525	0.045	0.0545
ENSMUSE00000777868	5:21743379-21746090	+	Pmpcb	0.0105	0.026	0.0865	0.033	0.0485	0.0665
ENSMUSE00000740654	X:8143848-8144679	-	Rbm3	0.0675	0.0985	0.0195	0.0405	0.044	0.0695
ENSMUSE00001332031	1:55014483-55016490	-	Sf3b1	0.1365	0.178	0.2265	0.264	0.284	0.282

Appendix 4: Identification of alternative splicing events using rMATS (replicate Multivariate Analysis of Transcript Splicing).

Abbreviations used in the table stand for the following types of splicing events and definitions: SE- Skipped Exon, MXE- Mutually Exclusive Exon, RI- Retained Intron, A5SS- Alternative 5’ Splice Site, A3SS- Alternative 3’ Splice Site, PSI- Percent Spliced Index, FDR- False Discovery Rate.

>Lg1

```
CGGATTTTTCTTGGCTTTATATATCTTGTGGAAGGACGAAACACCGCTGGTAGGGGAGTCAAGAGAGTTTTAGAGCTAGA
AATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTGAATTCGCTAGC
TAGGTCTTGAAAGGAGTGGGAATTGGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCCCACAGTCCCCGAGAAGTT
GGGGGGAGGGGTCGGCAATTGATCCGGTGCCTAGAGAAGGTGGCGCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGGC
TCCGCCTTTTTCCCGAGGGTGGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAACGGGTTT
GCCGCCAGAACACAGGACCGGTTCTAGAGCGCTGCCACCATGGACAAGAAGTACAGCATCGGCCTGGACATCGGCACCAA
CTCTGTGGGCTGGGCCGTGATCACCGACGAGTACAAGGTGCCCAGCAAGAAATTCAAGGTGCTGGGCAACACCGACCGGC
ACAGCATCAAGAAGAACCTGATCGGAGCCCTGCTGTTCGACAGCGGCGAAACAGCCGAGGCCACCCGGCTGAAGAGAACC
GCCAGAAGAAGATACACCAGACGGAAGAACCGGATCTGCTATCTGCAAGAGATCTTCAGCAACGAGATGGCCAAGGTGGA
CGACAGCTTCTTCCACAGACTGGAAGAGTCCTTCCTGGTGGAAGAGGATAAGAAGCACGAGCGGCACCCCATCTTCGGCA
ACATCGTGGACGAGTGGCCTACCACGAGAAGTACCCCACCATCTACCACCTGAGAAAGAAACTGGTGGACAGCACCGACA
GGCCGACCTGCGGCTGATCTATCTGGCCCTGGCCCACATGATCAAGTTCCGGGGCCACTTCCTGATCGAGGGCGACCTGA
ACCCCGACAACAGCGACGTGGACAGCTGTTCATCCAGCTGGTGCAGACCTACAACCAGCTGTTCGAGGAAAACCCCATCA
ACGCAGCGGCGTGGACGCCAGCCATCCTGTCTGCAGACTGAGCAGAGCAGACGCTGAAATCTGATCGCCAGCTGCCGGCG
AGAAAAAAAATGGGCCTGTTCGG
```

>Lg2

```
TTCCGAATTTCCTGGCTTTATATATCTTGTGGAAGGACGAAACACCGAGTATGGATTAAATAAAGGAGTTTTAGAGCTAG
AAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTGAATTCGCTAG
CTAGGTCTTGAAAGGAGTGGGAATTGGCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCCCACAGTCCCCGAGAAGT
TGGGGGGAGGGGTCGGCAATTGATCCGGTGCCTAGAGAAGGTGGCGCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGG
CTCCGCCTTTTTCCCGAGGGTGGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTGAACGTTCTTTTTCGCAACGGGTT
TGCCGCCAGAACACAGGACCGGTTCTAGAGCGCTGCCACCATGGACAAGAAGTACAGCATCGGCCTGGACATCGGCACCA
ACTCTGTGGGCTGGGCCGTGATCACCGACGAGTACAAGGTGCCCAGCAAGAAATTCAAGGTGCTGGGCAACACCGACCGG
CACAGCATCAAGAAGAACCTGATCGGAGCCCTGCTGTTCGACAGCGGCGAAACAGCCGAGGCCACCCGGCTGAAGAGAAC
CGCCAGAAGAAGATACACCAGACGGAAGAACCGGATCTGCTATCTGCAAGAGATCTTCAGCAACGAGATGGCCAAGGTGG
ACGACAGCTTCTTCCACAGACTGGAAGAGTCCTTCCTGGTGGAAGAGGATAAGAAGCACGAGCGGCACCCCATCTTCGGC
AACATCGTGGACGAGTGGCCTACCACGAGAAGTACCCCACCATCTACCACCTGAGAAAGAAACTGGTGGACAGCACCGAC
AAGCCGACCTGCGGCTGATCTATCTGGCCCTGGCCCACATGATCAAGTTCCGGGGCCACTTCCTGATCGAGGGCGACCTG
AACCCCGACAACAGCGACGTGGACAAGCTGTTCATCCAGCTGGTGCAGACCTACAACCAGCTGTTCGAGAAAACCCCATC
ACGCAGCGCGTGACGCAAGCATCCTGTCTGCAGACTGAGCAAGAGCAGACGCTGGAAAATTCTGATCGCCCAGCTGCCGG
CCGAGAG
```

List of sgRNAs (yellow highlighted)
Lg1 CTGGTAGGGGAGTCAAGAGA
Lg2 AGTATGGATTAAATAAAGGA

Appendix 5: Sanger sequencing data confirming the sgRNA plasmid library constructs

Sno	SRA_ID	Tissue	Species	Read_Type	reads	overall alignment rate
1	SRR306838	Brain	Human	SE	24513415	72.56%
2	SRR306839	Brain	Human	SE	18850030	60.39%
3	SRR306841	Brain	Human	SE	24325223	65.94%
4	SRR306847	Heart	Human	SE	24128204	68.50%
5	SRR306848	Heart	Human	SE	12451849	49.80%
6	SRR306849	Heart	Human	SE	18444502	68.33%
7	SRR306850	Heart	Human	SE	25197713	61.81%
8	SRR306851	Kidney	Human	SE	22493518	71.56%
9	SRR306852	Kidney	Human	SE	20684752	69.08%
10	SRR306853	Kidney	Human	SE	31386619	61.33%
11	SRR306854	Liver	Human	SE	16391552	52.98%
12	SRR306855	Liver	Human	SE	26755509	75.26%
13	SRR306856	Liver	Human	SE	23866499	73.16%
14	SRR649360	Brain	Human	SE	19508676	62.00%
15	SRR649361	Brain	Human	SE	38402147	85.80%
16	SRR594475	Heart	Cow	PE	117554231	89.96%
17	SRR594476	Kidney	Cow	PE	115720336	92.87%
18	SRR594477	Liver	Cow	PE	103019718	92.36%
19	SRR594482	Brain	Cow	PE	28445194	95.12%
20	SRR594484	Heart	Cow	PE	21185451	95.24%
21	SRR594485	Kidney	Cow	PE	27567792	95.64%
22	SRR594486	Liver	Cow	PE	26021524	96.59%
23	SRR594491	Brain	Cow	PE	33628113	93.42%
24	SRR594493	Heart	Cow	PE	37897106	94.26%
25	SRR594494	Kidney	Cow	PE	22535945	96.00%
26	SRR594495	Liver	Cow	PE	29192793	95.67%
27	SRR306711	Brain	Chicken	SE	17557038	52.07%
28	SRR306714	Heart	Chicken	SE	23004385	59.95%
29	SRR306715	Heart	Chicken	SE	21117498	56.12%
30	SRR306716	Kidney	Chicken	SE	23021153	64.21%
31	SRR306717	Kidney	Chicken	SE	22796688	52.18%
32	SRR306718	Liver	Chicken	SE	30245926	62.10%
33	SRR306719	Liver	Chicken	SE	8595446	58.15%
34	SRR306720	Liver	Chicken	SE	22542615	64.52%
35	SRR594500	Brain	Chicken	PE	117728780	88.40%
36	SRR594502	Heart	Chicken	PE	107873185	81.63%
37	SRR594503	Kidney	Chicken	PE	117005256	87.92%
38	SRR594504	Liver	Chicken	PE	111950293	86.33%
39	SRR594509	Brain	Chicken	PE	32266164	88.49%
40	SRR594511	Heart	Chicken	PE	45037438	92.10%
41	SRR594513	Liver	Chicken	PE	25468656	93.84%
42	SRR594520	Heart	Chicken	PE	27260601	91.57%
43	SRR594521	Kidney	Chicken	PE	34660070	85.96%
44	SRR594522	Liver	Chicken	PE	18978066	44.47%
45	SRR649385	Brain	Chicken	SE	35609107	81.63%
46	SRR306778	Brain	Rhesus_Monkey	SE	22554234	61.07%
47	SRR306779	Brain	Rhesus_Monkey	PE	21461283	62.27%
48	SRR306783	Heart	Rhesus_Monkey	SE	20815484	62.38%
49	SRR306785	Kidney	Rhesus_Monkey	SE	24115366	49.15%
50	SRR306786	Liver	Rhesus_Monkey	SE	21711196	64.22%
51	SRR306787	Liver	Rhesus_Monkey	SE	9393115	65.77%
52	SRR306788	Liver	Rhesus_Monkey	SE	22831536	70.62%
53	SRR594446	Brain	Rhesus_Monkey	PE	35066763	92.30%
54	SRR594448	Heart	Rhesus_Monkey	PE	35248042	93.95%
55	SRR594449	Kidney	Rhesus_Monkey	PE	31891747	92.23%
56	SRR594450	Liver	Rhesus_Monkey	PE	28555788	93.49%
57	SRR594455	Brain	Rhesus_Monkey	PE	107669551	89.74%
58	SRR594457	Heart	Rhesus_Monkey	PE	109193093	89.73%
59	SRR594458	Kidney	Rhesus_Monkey	PE	108637672	90.04%
60	SRR594459	Liver	Rhesus_Monkey	PE	113094939	90.75%
61	SRR594464	Brain	Rhesus_Monkey	PE	26487487	91.83%

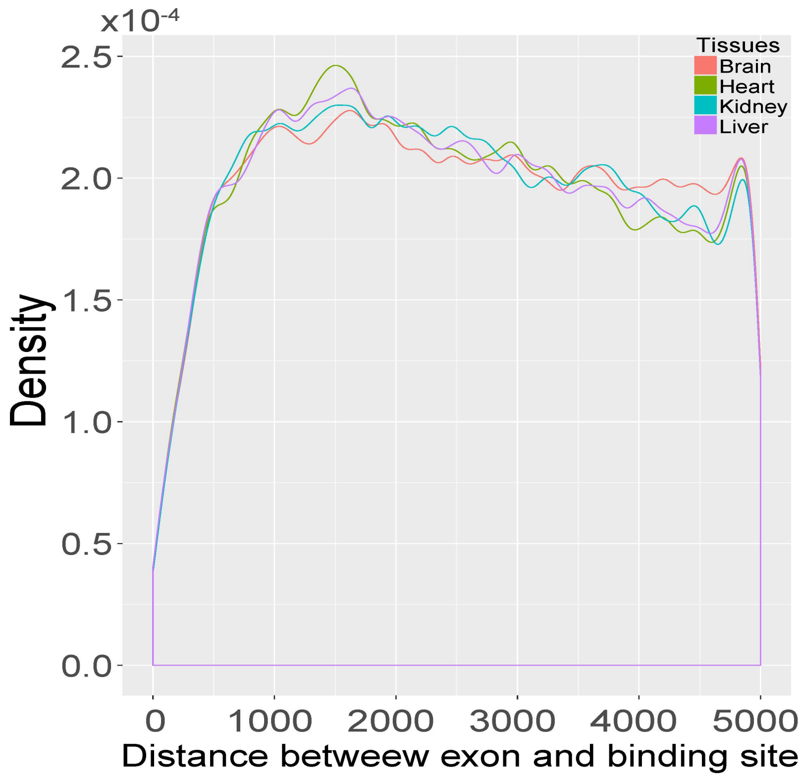
62	SRR594466	Heart	Rhesus_Monkey	PE	36101619	93.98%
63	SRR594467	Kidney	Rhesus_Monkey	PE	40389069	93.78%
64	SRR594468	Liver	Rhesus_Monkey	PE	26700682	93.91%
65	SRR649368	Brain	Rhesus_Monkey	SE	31270389	86.38%
66	SRR306743	Brain	Monodelphis	SE	47574556	57.88%
67	SRR306744	Brain	Monodelphis	SE	22273667	40.64%
68	SRR306748	Heart	Monodelphis	SE	17738032	53.39%
69	SRR306749	Heart	Monodelphis	SE	16630475	46.84%
70	SRR306750	Heart	Monodelphis	SE	32689495	58.92%
71	SRR306751	Kidney	Monodelphis	SE	21754688	56.60%
72	SRR306752	Kidney	Monodelphis	SE	14679277	54.69%
73	SRR306753	Liver	Monodelphis	SE	20729602	53.03%
74	SRR306754	Liver	Monodelphis	SE	19002730	43.46%
75	SRR649376	Brain	Monodelphis	SE	40483416	69.32%
76	SRR306758	Brain	Mouse	SE	18882745	74.76%
77	SRR306759	Brain	Mouse	SE	20757817	60.46%
78	SRR306760	Brain	Mouse	SE	17770683	77.14%
79	SRR306761	Brain	Mouse	SE	18759557	72.08%
80	SRR306762	Brain	Mouse	SE	19726026	51.82%
81	SRR306766	Heart	Mouse	SE	44668984	53.46%
82	SRR306767	Heart	Mouse	SE	24493681	63.84%
83	SRR306768	Heart	Mouse	SE	25903961	66.68%
84	SRR306770	Kidney	Mouse	SE	23639764	67.74%
85	SRR306771	Kidney	Mouse	SE	29158234	61.12%
86	SRR306772	Liver	Mouse	SE	48306727	40.84%
87	SRR306773	Liver	Mouse	SE	18444416	71.24%
88	SRR306774	Liver	Mouse	SE	34010208	57.28%
89	SRR594393	Brain	Mouse	PE	87264604	95.15%
90	SRR594395	Heart	Mouse	PE	35175982	95.88%
91	SRR594396	Kidney	Mouse	PE	119274786	95.77%
92	SRR594397	Liver	Mouse	PE	116292478	86.80%
93	SRR594402	Brain	Mouse	PE	118824353	91.56%
94	SRR594404	Kidney	Mouse	PE	118885190	95.62%
95	SRR594405	Liver	Mouse	PE	134045721	89.94%
96	SRR594410	Brain	Mouse	PE	32511234	93.98%
97	SRR594412	Heart	Mouse	PE	15968605	94.06%
98	SRR594413	Kidney	Mouse	PE	29821800	95.43%
99	SRR594414	Liver	Mouse	PE	34824609	91.64%
100	SRR649371	Brain	Mouse	SE	40407034	87.72%
101	SRR306725	Brain	Platypus	SE	24343340	54.26%
102	SRR306726	Brain	Platypus	SE	9306487	53.58%
103	SRR306727	Brain	Platypus	SE	13655446	55.80%
104	SRR306730	Heart	Platypus	SE	18807127	46.58%
105	SRR306731	Heart	Platypus	SE	16876055	45.19%
106	SRR306732	Kidney	Platypus	SE	17863241	43.08%
107	SRR306734	Kidney	Platypus	SE	16081270	40.75%
108	SRR306736	Liver	Platypus	SE	21513648	44.41%
109	SRR649381	Brain	Platypus	SE	30611890	73.37%
110	SRR306793	Heart	Orangutan	SE	20807820	66.71%
111	SRR306794	Kidney	Orangutan	SE	36798263	63.06%
112	SRR306795	Liver	Orangutan	SE	31482282	54.31%
113	SRR306796	Brain	Orangutan	SE	30547227	57.35%
114	SRR306797	Heart	Orangutan	SE	30043284	61.77%
115	SRR306798	Kidney	Orangutan	SE	21355541	74.67%
116	SRR306799	Liver	Orangutan	SE	35683453	69.25%
117	SRR594421	Heart	Brown_Rat	PE	35802852	91.96%
118	SRR594422	Kidney	Brown_Rat	PE	114089612	95.77%
119	SRR594423	Liver	Brown_Rat	PE	26181362	95.00%
120	SRR594428	Brain	Brown_Rat	PE	96368839	86.26%
121	SRR594430	Heart	Brown_Rat	PE	67008998	91.04%
122	SRR594431	Kidney	Brown_Rat	PE	116656722	94.12%
123	SRR594432	Liver	Brown_Rat	PE	131658529	93.98%

124	SRR594437	Brain	Brown_Rat	PE	32262802	94.05%
125	SRR594439	Heart	Brown_Rat	PE	25869367	91.99%
126	SRR594440	Kidney	Brown_Rat	PE	40114787	94.45%
127	SRR594441	Liver	Brown_Rat	PE	42043440	93.95%
128	SRR649392	Brain	Xenopus	SE	36987876	77.80%
129	SRR649393	Heart	Xenopus	SE	36468389	78.89%
130	SRR649394	Heart	Xenopus	SE	34826248	84.87%
131	SRR649395	Kidney	Xenopus	SE	36696254	78.81%
132	SRR649396	Kidney	Xenopus	SE	38280443	83.55%
133	SRR649397	Liver	Xenopus	SE	36658397	81.81%
134	SRR649398	Liver	Xenopus	SE	37331954	83.21%

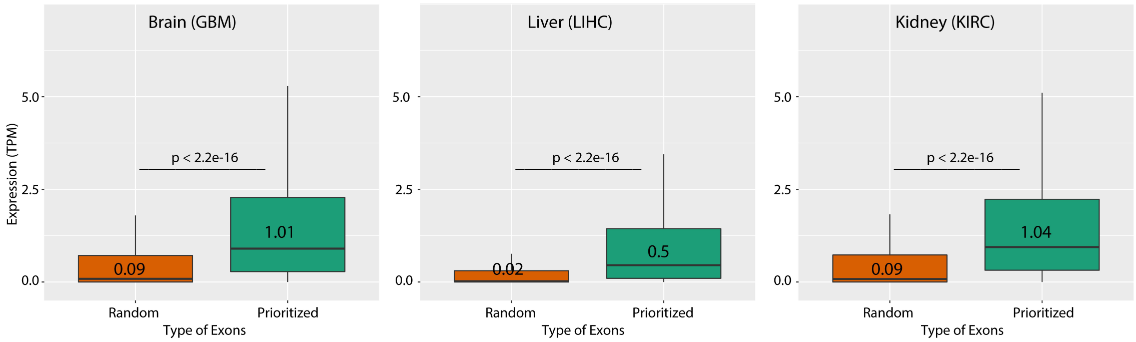
Appendix 6: List of samples used in this study and related metadata

Species	Biological name	Reference [Ensembl database]
Brown Rat	Rattus norvegicus	REFERENCE/Brown_Rat/Rnor_6.0.84/Rnor_6.0.84
Chicken	Gallus gallus	REFERENCE/Chicken/Galga14.84/Galga14.84
Cow	Bos taurus	REFERENCE/Cow/Bos_taurus.UMD3.1.84/Bos_taurus.UMD3.1.84
Human	Homo sapiens	REFERENCE/Human/h38.84/h38.84
Opossum	Monodelphis domestica	REFERENCE/Monodelphis/BROADO5.84/BROADO5.84
Mouse	Mus musculus	REFERENCE/Mouse/m38.84/m38.84
Orangutan	Pongo pygmaeus abelii	REFERENCE/Orangutan/PPYG2.84/PPYG2.84
Platypus	Ornithorhynchus anatinus	REFERENCE/Platypus/OANA5.84/OANA5.84
Rhesus Monkey	Macaca mulatta	REFERENCE/Rhesus_Monkey/MMUL_1.84/MMUL_1.84
Xenopus	Xenopus tropicalis	REFERENCE/Xenopus/JGI_4.2.84/JGI_4.2.84

Appendix 7: List of species and reference genomes used for alignment and downstream data analysis.



Appendix 8: Distance between binding site and associated exon - after setting a distance threshold of 5000 bp between binding sites of RBPs and the associated exons (FDR < 0.05), the distance is less than or equal to 3000 bp in most of the exon-binding site associations.



Appendix 9: Expression of prioritized exons vs. random exons.

The box plots represent the distribution of expression of prioritized exons from the PEEK pipeline vs. the expression of equal number of random exons for the tissues brain, liver and kidney. The expression levels of exons in each tissue were obtained from cancer patients, the cancer types being Glioblastoma, Liver Hepatocellular Carcinoma and Kidney Renal clear cell carcinoma for brain, liver and kidney respectively. Using the data for ~400 patients of each cancer type, the expression level of prioritized exons was found significantly higher than the expression level of random exons for each tissue type.

REFERENCES

1. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012;150(6):1274-86. Epub 2012/09/11. doi: 10.1016/j.cell.2012.04.040. PubMed PMID: 22959076; PMCID: 3679407.
2. Janga SC, Mittal N. Construction, structure and dynamics of post-transcriptional regulatory network directed by RNA-binding proteins. Advances in experimental medicine and biology. 2011;722:103-17. doi: 10.1007/978-1-4614-0332-6_7. PubMed PMID: 21915785.
3. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75-82. Epub 2012/09/08. doi: 10.1038/nature11232. PubMed PMID: 22955617.
4. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J. DNA-binding specificities of human transcription factors. Cell. 2013;152(1-2):327-39. doi: 10.1016/j.cell.2012.12.009. PubMed PMID: 23332764.
5. Shatkin AJ, Manley JL. The ends of the affair: capping and polyadenylation. Nature structural biology. 2000;7(10):838-42. doi: 10.1038/79583. PubMed PMID: 11017188.
6. Jonkhout N, Tran J, Smith MA, Schonrock N, Mattick JS, Novoa EM. The RNA modification landscape in human disease. Rna. 2017;23(12):1754-69. Epub 2017/09/01. doi: 10.1261/rna.063503.117. PubMed PMID: 28855326; PMCID: PMC5688997.
7. Vemuri S, Srivastava R, Mir Q, Hashemikhabir S, Dong XC, Janga SC. SliceIt: A genome-wide resource and visualization tool to design CRISPR/Cas9 screens for editing protein-RNA interaction sites in the human genome. Methods (San Diego, Calif). 2019. doi: 10.1016/j.ymeth.2019.09.004. PubMed PMID: 31494246.
8. Wang Y, Arribas-Layton M, Chen Y, Lykke-Andersen J, Sen GL. DDX6 Orchestrates Mammalian Progenitor Function through the mRNA Degradation and Translation Pathways. Mol Cell. 2015;60(1):118-30. doi: 10.1016/j.molcel.2015.08.014. PubMed PMID: 26412305; PMCID: PMC4592480.
9. Zhao Y, Lin J, Xu B, Hu S, Zhang X, Wu L. MicroRNA-mediated repression of nonsense mRNAs. eLife. 2014;3:e03032. doi: 10.7554/eLife.03032. PubMed PMID: 25107276; PMCID: 4359369.
10. Brown AS, Mohanty BK, Howe PH. Computational Identification of Post Translational Modification Regulated RNA Binding Protein Motifs. PLoS One.

2015;10(9):e0137696. Epub 2015/09/15. doi: 10.1371/journal.pone.0137696. PubMed PMID: 26368004; PMCID: PMC4569568.

11. Maniatis T. On the road from classical to modern molecular biology. Nat Med. 2012;18(10):1499-502. Epub 2012/10/09. doi: 10.1038/nm.2931. PubMed PMID: 23042362.

12. Lieberman J. Unveiling the RNA World. N Engl J Med. 2018;379(13):1278-80. Epub 2018/09/12. doi: 10.1056/NEJMcibr1808725. PubMed PMID: 30200806.

13. Nerenz RD, Lefferts J. Our Genome's "Dark Matter" Is the Next Frontier in Molecular Diagnostics. Clin Chem. 2017;63(3):792-3. Epub 2017/03/01. doi: 10.1373/clinchem.2016.268607. PubMed PMID: 28242839.

14. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19(6):329-46. Epub 2018/03/31. doi: 10.1038/s41576-018-0003-4. PubMed PMID: 29599501.

15. The long view on sequencing. Nat Biotechnol. 2018;36(4):287. Epub 2018/04/06. doi: 10.1038/nbt.4125. PubMed PMID: 29621212.

16. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, Pendlington ZM, Welter D, Burdett T, Hindorff L, Flicek P, Cunningham F, Parkinson H. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-D901. doi: 10.1093/nar/gkw1133. PubMed PMID: 27899670; PMCID: PMC5210590.

17. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113-20. Epub 2013/09/28. doi: 10.1038/ng.2764. PubMed PMID: 24071849; PMCID: PMC3919969.

18. Srivastava R, Budak G, Dash S, Lachke SA, Janga SC. Transcriptome analysis of developing lens reveals abundance of novel transcripts and extensive splicing alterations. Sci Rep. 2017;7(1):11572. Epub 2017/09/16. doi: 10.1038/s41598-017-10615-4. PubMed PMID: 28912564; PMCID: PMC5599659.

19. Gronostajski RM, Guaneri J, Lee DH, Gallo SM. The NFI-Regulome Database: A tool for annotation and analysis of control regions of genes regulated by Nuclear Factor I transcription factors. J Clin Bioinforma. 2011;1(1):4. Epub 2011/09/03. doi: 10.1186/2043-9113-1-4. PubMed PMID: 21884625; PMCID: PMC3143897.

20. Lu Z, Chang HY. Decoding the RNA structurome. Curr Opin Struct Biol. 2016;36:142-8. Epub 2016/03/01. doi: 10.1016/j.sbi.2016.01.007. PubMed PMID: 26923056; PMCID: PMC4785074.

21. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M, Nakanishi T, Teruya K, Satou K, Hirano T. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell. 2017;30(3):149-61. Epub 2017/04/02. doi: 10.1007/s13577-017-0168-8. PubMed PMID: 28364362; PMCID: PMC5486853.

22. Feng Y, Zhang Y, Ying C, Wang D, Du C. Nanopore-based fourth-generation DNA sequencing technology. Genomics Proteomics Bioinformatics. 2015;13(1):4-16. Epub 2015/03/07. doi: 10.1016/j.gpb.2015.01.009. PubMed PMID: 25743089; PMCID: PMC4411503.

23. Mueller A, Fischer K, Suluku R, Hoenen T. Sequencing of mRNA from Whole Blood using Nanopore Sequencing. J Vis Exp. 2019(148). Epub 2019/06/18. doi: 10.3791/59377. PubMed PMID: 31205309.
24. Santambrogio S, Cattaneo A, Bernascone I, Schwend T, Jovine L, Bachi A, Rampoldi L. Urinary uromodulin carries an intact ZP domain generated by a conserved C-terminal proteolytic cleavage. Biochemical and biophysical research communications. 2008;370(3):410-3. Epub 2008/04/01. doi: 10.1016/j.bbrc.2008.03.099. PubMed PMID: 18375198.
25. Rampoldi L, Scolari F, Amoroso A, Ghiggeri G, Devuyst O. The rediscovery of uromodulin (Tamm-Horsfall protein): from tubulointerstitial nephropathy to chronic kidney disease. Kidney international. 2011;80(4):338-47. Epub 2011/06/10. doi: 10.1038/ki.2011.134. PubMed PMID: 21654721.
26. El-Achkar TM, Wu XR. Uromodulin in kidney injury: an instigator, bystander, or protector? Am J Kidney Dis. 2012;59(3):452-61. Epub 2012/01/27. doi: S0272-6386(11)01698-2 [pii]

10.1053/j.ajkd.2011.10.054. PubMed PMID: 22277744; PMCID: 3288726.
27. Mutig K, Kahl T, Saritas T, Godes M, Persson P, Bates J, Raffi H, Rampoldi L, Uchida S, Hille C, Dosche C, Kumar S, Castaneda-Bueno M, Gamba G, Bachmann S. Activation of the bumetanide-sensitive Na+,K+,2Cl- cotransporter (NKCC2) is facilitated by Tamm-Horsfall protein in a chloride-sensitive manner. The Journal of biological chemistry. 2011;286(34):30200-10. Epub 2011/07/09. doi: 10.1074/jbc.M111.222968. PubMed PMID: 21737451; PMCID: 3191059.
28. El-Achkar TM, McCracken R, Rauchman M, Heitmeier MR, Al-Aly Z, Dagher PC, Wu XR. Tamm-Horsfall protein-deficient thick ascending limbs promote injury to neighboring S3 segments in an MIP-2-dependent mechanism. Am J Physiol Renal Physiol. 2011;300(4):F999-F1007. Epub 2011/01/14. doi: ajprenal.00621.2010 [pii]

10.1152/ajprenal.00621.2010. PubMed PMID: 21228114.
29. Mo L, Huang HY, Zhu XH, Shapiro E, Hasty DL, Wu XR. Tamm-Horsfall protein is a critical renal defense factor protecting against calcium oxalate crystal formation. Kidney international. 2004;66(3):1159-66. Epub 2004/08/26. doi: 10.1111/j.1523-1755.2004.00867.x. PubMed PMID: 15327412.
30. Mo L, Zhu XH, Huang HY, Shapiro E, Hasty DL, Wu XR. Ablation of the Tamm-Horsfall protein gene increases susceptibility of mice to bladder colonization by type 1-fimbriated Escherichia coli. Am J Physiol Renal Physiol. 2004;286(4):F795-802. Epub 2003/12/11. doi: 10.1152/ajprenal.00357.2003. PubMed PMID: 14665435.
31. Kottgen A, Hwang SJ, Larson MG, Van Eyk JE, Fu Q, Benjamin EJ, Dehghan A, Glazer NL, Kao WH, Harris TB, Gudnason V, Shlipak MG, Yang Q, Coresh J, Levy D, Fox CS. Uromodulin levels associate with a common UMOD variant and risk for incident CKD. J Am Soc Nephrol. 2010;21(2):337-44. Epub 2009/12/05. doi: ASN.2009070725 [pii]

10.1681/ASN.2009070725. PubMed PMID: 19959715; PMCID: 2834540.
32. Lhotta K. Uromodulin and chronic kidney disease. Kidney & blood pressure research. 2010;33(5):393-8. Epub 2010/10/16. doi: 10.1159/000320681. PubMed PMID: 20948228.

33. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome biology. 2012;13(9):R48. Epub 2012/09/07. doi: 10.1186/gb-2012-13-9-r48. PubMed PMID: 22950945; PMCID: 3491392.

34. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, Fields S, Stamatoyannopoulos JA. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature methods. 2009;6(4):283-9. Epub 2009/03/24. doi: 10.1038/nmeth.1313. PubMed PMID: 19305407; PMCID: 2668528.

35. Klepper K, Drablos F. MotifLab: a tools and data integration workbench for motif discovery and regulatory sequence analysis. BMC bioinformatics. 2013;14:9. Epub 2013/01/18. doi: 10.1186/1471-2105-14-9. PubMed PMID: 23323883; PMCID: 3556059.

36. Wang J, Lu J, Gu G, Liu Y. In vitro DNA-binding profile of transcription factors: methods and new insights. The Journal of endocrinology. 2011;210(1):15-27. Epub 2011/03/11. doi: 10.1530/JOE-11-0010. PubMed PMID: 21389103.

37. Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome research. 2002;12(5):739-48. Epub 2002/05/09. doi: 10.1101/gr.6902. PubMed PMID: 11997340; PMCID: 186562.

38. Rustici G, Kolesnikov N, Brandizi M, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Ison J, Keays M, Kurbatova N, Malone J, Mani R, Mupo A, Pedro Pereira R, Pilicheva E, Rung J, Sharma A, Tang YA, Ternent T, Tikhonov A, Welter D, Williams E, Brazma A, Parkinson H, Sarkans U. ArrayExpress update--trends in database growth and links to data analysis tools. Nucleic Acids Res. 2013;41(Database issue):D987-90. Epub 2012/11/30. doi: 10.1093/nar/gks1174. PubMed PMID: 23193272; PMCID: 3531147.

39. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigo R. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research. 2012;22(9):1775-89. Epub 2012/09/08. doi: 10.1101/gr.132159.111. PubMed PMID: 22955988; PMCID: 3431493.

40. Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics. 2003;19(2):295-6. Epub 2003/01/23. PubMed PMID: 12538257.

41. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. Journal of molecular biology. 1988;203(2):439-55. Epub 1988/09/20. PubMed PMID: 3199442.

42. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T,

Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E. Ensembl 2007. Nucleic Acids Res. 2007;35(Database issue):D610-7. Epub 2006/12/07. doi: 10.1093/nar/gkl996. PubMed PMID: 17148474; PMCID: 1761443.

43. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004;5(4):276-87. Epub 2004/05/08. doi: 10.1038/nrg1315. PubMed PMID: 15131651.

44. Chen DH, Chang AY, Liao BY, Yeang CH. Functional characterization of motif sequences under purifying selection. Nucleic Acids Res. 2013;41(4):2105-20. Epub 2013/01/11. doi: 10.1093/nar/gks1456. PubMed PMID: 23303791; PMCID: 3575792.

45. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37(Web Server issue):W202-8. Epub 2009/05/22. doi: 10.1093/nar/gkp335. PubMed PMID: 19458158; PMCID: 2703892.

46. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology. 1994;2:28-36. Epub 1994/01/01. PubMed PMID: 7584402.

47. Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, Valen E, Yusuf D, Lenhard B, Wasserman WW, Sandelin A. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Res. 2010;38(Database issue):D105-10. Epub 2009/11/13. doi: 10.1093/nar/gkp950. PubMed PMID: 19906716; PMCID: 2808906.

48. Robasky K, Bulyk ML. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. Nucleic Acids Res. 2011;39(Database issue):D124-8. Epub 2010/11/03. doi: 10.1093/nar/gkq992. PubMed PMID: 21037262; PMCID: 3013812.

49. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome biology. 2007;8(2):R24. Epub 2007/02/28. doi: 10.1186/gb-2007-8-2-r24. PubMed PMID: 17324271; PMCID: 1852410.

50. He HH, Meyer CA, Hu SS, Chen MW, Zang C, Liu Y, Rao PK, Fei T, Xu H, Long H, Liu XS, Brown M. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nature methods. 2013. doi: 10.1038/nmeth.2762. PubMed PMID: 24317252.

51. Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology. 2011;9(4):e1001046. doi: 10.1371/journal.pbio.1001046. PubMed PMID: 21526222; PMCID: 3079585.

52. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics. 2004;20(9):1453-4. Epub 2004/02/12. doi: 10.1093/bioinformatics/bth078. PubMed PMID: 14871861.

53. Saldanha AJ. Java Treeview--extensible visualization of microarray data. Bioinformatics. 2004;20(17):3246-8. Epub 2004/06/08. doi: 10.1093/bioinformatics/bth349. PubMed PMID: 15180930.

54. Traylor A, Hock T, Hill-Kapturczak N. Specificity protein 1 and Smad-dependent regulation of human heme oxygenase-1 gene by transforming growth factor-beta1 in

renal epithelial cells. Am J Physiol Renal Physiol. 2007;293(3):F885-94. Epub 2007/06/15. doi: 10.1152/ajprenal.00519.2006. PubMed PMID: 17567933.

55. Shie JL, Chen ZY, Fu M, Pestell RG, Tseng CC. Gut-enriched Kruppel-like factor represses cyclin D1 promoter activity through Sp1 motif. Nucleic Acids Res. 2000;28(15):2969-76. Epub 2000/07/25. PubMed PMID: 10908361; PMCID: 102679.

56. Cheng YH, Handwerger S. A placenta-specific enhancer of the human syncytin gene. Biology of reproduction. 2005;73(3):500-9. Epub 2005/05/13. doi: 10.1095/biolreprod.105.039941. PubMed PMID: 15888734.

57. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M. The BioGRID interaction database: 2013 update. Nucleic Acids Res. 2013;41(Database issue):D816-23. Epub 2012/12/04. doi: 10.1093/nar/gks1158. PubMed PMID: 23203989; PMCID: 3531226.

58. Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, Adato A, Peter I, Khen M, Atarot T, Groner Y, Lancet D. Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. Nucleic Acids Res. 2003;31(1):142-6. Epub 2003/01/10. PubMed PMID: 12519968; PMCID: 165497.

59. Vyletal P, Bleyer AJ, Kmoch S. Uromodulin biology and pathophysiology--an update. Kidney & blood pressure research. 2010;33(6):456-75. Epub 2010/11/27. doi: 10.1159/000321013. PubMed PMID: 21109754.

60. Scolari F, Caridi G, Rampoldi L, Tardanico R, Izzi C, Pirulli D, Amoroso A, Casari G, Ghiggeri GM. Uromodulin storage diseases: clinical aspects and mechanisms. Am J Kidney Dis. 2004;44(6):987-99. Epub 2004/11/24. doi: S0272638604012557 [pii]. PubMed PMID: 15558519.

61. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16-23. Epub 2000/05/17. PubMed PMID: 10812473.

62. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome research. 2004;14(6):1188-90. Epub 2004/06/03. doi: 10.1101/gr.849004. PubMed PMID: 15173120; PMCID: 419797.

63. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003;424(6945):147-51. Epub 2003/07/11. doi: 10.1038/nature01763. PubMed PMID: 12853946.

64. Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan Z, Schilstra MJ, Clarke PJ, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H. A genomic regulatory network for development. Science. 2002;295(5560):1669-78. Epub 2002/03/02. doi: 10.1126/science.1069883. PubMed PMID: 11872831.

65. Wathelet MG, Lin CH, Parekh BS, Ronco LV, Howley PM, Maniatis T. Virus infection induces the assembly of coordinately activated transcription factors on the IFN-beta enhancer in vivo. Mol Cell. 1998;1(4):507-18. Epub 1998/07/14. PubMed PMID: 9660935.

66. Britten RJ, Davidson EH. Gene regulation for higher cells: a theory. Science. 1969;165(3891):349-57. Epub 1969/07/25. PubMed PMID: 5789433.

67. Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome biology. 2003;5(1):201. Epub 2004/01/08. doi: 10.1186/gb-2003-5-1-201. PubMed PMID: 14709165; PMCID: 395725.

68. Kim J, Choi M, Kim JR, Jin H, Kim VN, Cho KH. The co-regulation mechanism of transcription factors in the human gene regulatory network. Nucleic Acids Res. 2012;40(18):8849-61. Epub 2012/07/17. doi: 10.1093/nar/gks664. PubMed PMID: 22798495; PMCID: 3467061.

69. Buske FA, Boden M, Bauer DC, Bailey TL. Assigning roles to DNA regulatory motifs using comparative genomics. Bioinformatics. 2010;26(7):860-6. Epub 2010/02/12. doi: 10.1093/bioinformatics/btq049. PubMed PMID: 20147307; PMCID: 2844991.

70. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, Rowe LD, Dreszer TR, Roe G, Podduturi NR, Tanaka F, Hong EL, Cherry JM. ENCODE data at the ENCODE portal. Nucleic Acids Res. 2016;44(D1):D726-32. doi: 10.1093/nar/gkv1160. PubMed PMID: 26527727; PMCID: PMC4702836.

71. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, Maurano MT, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, Stamatoyannopoulos JA. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489(7414):83-90. Epub 2012/09/08. doi: 10.1038/nature11212. PubMed PMID: 22955618.

72. Srivastava R, Micanovic R, El-Achkar TM, Janga SC. An intricate network of conserved DNA upstream motifs and associated transcription factors regulate the expression of uromodulin gene. The Journal of urology. 2014;192(3):981-9. doi: 10.1016/j.juro.2014.02.095. PubMed PMID: 24594405.

73. Raviscioni M, Gu P, Sattar M, Cooney AJ, Lichtarge O. Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. Journal of molecular biology. 2005;350(3):402-15. Epub 2005/06/11. doi: 10.1016/j.jmb.2005.04.054. PubMed PMID: 15946684.

74. Lee JH, Budanov AV, Karin M. Sestrins orchestrate cellular metabolism to attenuate aging. Cell metabolism. 2013;18(6):792-801. doi: 10.1016/j.cmet.2013.08.018. PubMed PMID: 24055102; PMCID: 3858445.

75. Budanov AV, Lee JH, Karin M. Stressin' Sestrins take an aging fight. EMBO molecular medicine. 2010;2(10):388-400. doi: 10.1002/emmm.201000097. PubMed PMID: 20878915; PMCID: 3166214.

76. Dong XC. The potential of sestrins as therapeutic targets for diabetes. Expert opinion on therapeutic targets. 2015;19(8):1011-5. doi: 10.1517/14728222.2015.1044976. PubMed PMID: 25944222; PMCID: 4504765.

77. Budanov AV, Sablina AA, Feinstein E, Koonin EV, Chumakov PM. Regeneration of peroxiredoxins by p53-regulated sestrins, homologs of bacterial AhpD. Science. 2004;304(5670):596-600. doi: 10.1126/science.1095569. PubMed PMID: 15105503.

78. Bae SH, Sung SH, Oh SY, Lim JM, Lee SK, Park YN, Lee HE, Kang D, Rhee SG. Sestrins activate Nrf2 by promoting p62-dependent autophagic degradation of Keap1 and prevent oxidative liver damage. Cell metabolism. 2013;17(1):73-84. doi: 10.1016/j.cmet.2012.12.002. PubMed PMID: 23274085.

79. Kang X, Petyaykina K, Tao R, Xiong X, Dong XC, Liangpunsakul S. The inhibitory effect of ethanol on Sestrin3 in the pathogenesis of ethanol-induced liver injury. American journal of physiology Gastrointestinal and liver physiology. 2014;307(1):G58-65. doi: 10.1152/ajpgi.00373.2013. PubMed PMID: 24833709; PMCID: 4080163.

80. Lee JH, Budanov AV, Talukdar S, Park EJ, Park HL, Park HW, Bandyopadhyay G, Li N, Aghajan M, Jang I, Wolfe AM, Perkins GA, Ellisman MH, Bier E, Scadeng M, Foretz M, Viollet B, Olefsky J, Karin M. Maintenance of metabolic homeostasis by Sestrin2 and Sestrin3. Cell metabolism. 2012;16(3):311-21. doi: 10.1016/j.cmet.2012.08.004. PubMed PMID: 22958918; PMCID: 3687365.

81. Tao R, Xiong X, Liangpunsakul S, Dong XC. Sestrin 3 protein enhances hepatic insulin sensitivity by direct activation of the mTORC2-Akt signaling. Diabetes. 2015;64(4):1211-23. doi: 10.2337/db14-0539. PubMed PMID: 25377878; PMCID: 4375082.

82. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015;43(W1):W39-49. doi: 10.1093/nar/gkv416. PubMed PMID: 25953851; PMCID: 4489269.

83. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, Reguly T, Nixon J, Ramage L, Winter A, Sellam A, Chang C, Hirschman J, Theesfeld C, Rust J, Livstone MS, Dolinski K, Tyers M. The BioGRID interaction database: 2015 update. Nucleic Acids Res. 2015;43(Database issue):D470-8. doi: 10.1093/nar/gku1204. PubMed PMID: 25428363; PMCID: 4383984.

84. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28(4):593-4. doi: 10.1093/bioinformatics/btr708. PubMed PMID: 22199392; PMCID: 3278762.

85. Xiong X, Tao R, DePinho RA, Dong XC. The autophagy-related gene 14 (Atg14) is regulated by forkhead box O transcription factors and circadian rhythms and plays a critical role in hepatic autophagy and lipid metabolism. The Journal of biological chemistry. 2012;287(46):39107-14. doi: 10.1074/jbc.M112.412569. PubMed PMID: 22992773; PMCID: 3493951.

86. Sanjana NE, Shalem O, Zhang F. Improved vectors and genome-wide libraries for CRISPR screening. Nature methods. 2014;11(8):783-4. Epub 2014/07/31. doi: 10.1038/nmeth.3047. PubMed PMID: 25075903; PMCID: PMC4486245.

87. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, Heckl D, Ebert BL, Root DE, Doench JG, Zhang F. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science. 2014;343(6166):84-7. doi: 10.1126/science.1247005. PubMed PMID: 24336571; PMCID: PMC4089965.

88. Bollinger LM, Witczak CA, Houmard JA, Brault JJ. SMAD3 augments FoxO3-induced MuRF-1 promoter activity in a DNA-binding-dependent manner. American journal of physiology Cell physiology. 2014;307(3):C278-87. doi: 10.1152/ajpcell.00391.2013. PubMed PMID: 24920680; PMCID: 4121583.

89. Ganjam GK, Dimova EY, Unterman TG, Kietzmann T. FoxO1 and HNF-4 are involved in regulation of hepatic glucokinase gene expression by resveratrol. The Journal of biological chemistry. 2009;284(45):30783-97. doi: 10.1074/jbc.M109.045260. PubMed PMID: 19740748; PMCID: 2781477.

90. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D. GeneCards Version 3: the human gene integrator. Database : the journal of biological databases and curation. 2010;2010:baq020. doi: 10.1093/database/baq020. PubMed PMID: 20689021; PMCID: 2938269.

91. Li L. GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. Journal of computational biology : a journal of computational molecular cell biology. 2009;16(2):317-29. doi: 10.1089/cmb.2008.16TT. PubMed PMID: 19193149; PMCID: 2756050.

92. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74. doi: 10.1038/nature11247. PubMed PMID: 22955616; PMCID: 3439153.

93. Srivastava R, Zhang Y, Xiong X, Zhang X, Pan X, Dong XC, Liangpunsakul S, Janga SC. Prediction and Validation of Transcription Factors Modulating the Expression of Sestrin3 Gene Using an Integrated Computational and Experimental Approach. PLoS One. 2016;11(7):e0160228. doi: 10.1371/journal.pone.0160228. PubMed PMID: 27466818; PMCID: PMC4965051.

94. Hagenbuchner J, Ausserlechner MJ. Mitochondria and FOXO3: breath or die. Frontiers in physiology. 2013;4:147. doi: 10.3389/fphys.2013.00147. PubMed PMID: 23801966; PMCID: 3687139.

95. Dash S, Siddam AD, Barnum CE, Janga SC, Lachke SA. RNA-binding proteins in eye development and disease: implication of conserved RNA granule components. Wiley Interdiscip Rev RNA. 2016. doi: 10.1002/wrna.1355. PubMed PMID: 27133484.

96. Lachke SA, Maas RL. Building the developmental oculome: systems biology in vertebrate eye development and disease. Wiley interdisciplinary reviews Systems biology and medicine. 2010;2(3):305-23. doi: 10.1002/wsbm.59. PubMed PMID: 20836031.

97. Zagozewski JL, Zhang Q, Eisenstat DD. Genetic regulation of vertebrate eye development. Clinical genetics. 2014;86(5):453-60. doi: 10.1111/cge.12493. PubMed PMID: 25174583.

98. Cvekl A, Ashery-Padan R. The cellular and molecular mechanisms of vertebrate lens development. Development. 2014;141(23):4432-47. doi: 10.1242/dev.107953. PubMed PMID: 25406393; PMCID: PMC4302924.

99. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636-40. doi: 10.1126/science.1105136. PubMed PMID: 15499007.

100. Consortium G. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015;348(6235):648-60. doi: 10.1126/science.1262110. PubMed PMID: 25954001.

101. Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. Nat Med. 2011;17(3):297-303. doi: 10.1038/nm.2323. PubMed PMID: 21383744.

102. Tian L, Kazmierkiewicz KL, Bowman AS, Li M, Curcio CA, Stambolian DE. Transcriptome of the human retina, retinal pigmented epithelium and choroid. Genomics. 2015;105(5-6):253-64. doi: 10.1016/j.ygeno.2015.01.008. PubMed PMID: 25645700; PMCID: 4404213.
103. Anand D, Lachke SA. Systems biology of lens development: A paradigm for disease gene discovery in the eye. Experimental eye research. 2017;156:22-33. doi: 10.1016/j.exer.2016.03.010. PubMed PMID: 26992779; PMCID: 5026553.
104. Chaitankar V, Karakulah G, Ratnapriya R, Giuste FO, Brooks MJ, Swaroop A. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. Progress in retinal and eye research. 2016. doi: 10.1016/j.preteyeres.2016.06.001. PubMed PMID: 27297499.
105. Khan SY, Hackett SF, Lee MC, Pourmand N, Talbot CC, Jr., Riazuddin SA. Transcriptome Profiling of Developing Murine Lens Through RNA Sequencing. Investigative ophthalmology & visual science. 2015;56(8):4919-26. doi: 10.1167/iovs.14-16253. PubMed PMID: 26225632; PMCID: 4525677.
106. Lachke SA, Ho JW, Kryukov GV, O'Connell DJ, Aboukhalil A, Bulyk ML, Park PJ, Maas RL. iSyTE: integrated Systems Tool for Eye gene discovery. Investigative ophthalmology & visual science. 2012;53(3):1617-27. doi: 10.1167/iovs.11-8839. PubMed PMID: 22323457; PMCID: PMC3339920.
107. Khan SY, Hackett SF, Riazuddin SA. Non-coding RNA Profiling of Developing Murine Lens. Experimental eye research. 2016. doi: 10.1016/j.exer.2016.01.010. PubMed PMID: 26808486.
108. Andzelm MM, Cherry TJ, Harmin DA, Boeke AC, Lee C, Hemberg M, Pawlyk B, Malik AN, Flavell SW, Sandberg MA, Raviola E, Greenberg ME. MEF2D drives photoreceptor development through a genome-wide competition for tissue-specific enhancers. Neuron. 2015;86(1):247-63. doi: 10.1016/j.neuron.2015.02.038. PubMed PMID: 25801704; PMCID: 4393375.
109. Busskamp V, Krol J, Nelidova D, Daum J, Szikra T, Tsuda B, Juttner J, Farrow K, Scherf BG, Alvarez CP, Genoud C, Sothilingam V, Tanimoto N, Stadler M, Seeliger M, Stoffel M, Filipowicz W, Roska B. miRNAs 182 and 183 are necessary to maintain adult cone photoreceptor outer segments and visual function. Neuron. 2014;83(3):586-600. doi: 10.1016/j.neuron.2014.06.020. PubMed PMID: 25002228.
110. Roger JE, Hiriyanna A, Gotoh N, Hao H, Cheng DF, Ratnapriya R, Kautzmann MA, Chang B, Swaroop A. OTX2 loss causes rod differentiation defect in CRX-associated congenital blindness. The Journal of clinical investigation. 2014;124(2):631-43. doi: 10.1172/JCI72722. PubMed PMID: 24382353; PMCID: 3904630.
111. Ruzycki PA, Tran NM, Kefalov VJ, Kolesnikov AV, Chen S. Graded gene expression changes determine phenotype severity in mouse models of CRX-associated retinopathies. Genome biology. 2015;16:171. doi: 10.1186/s13059-015-0732-z. PubMed PMID: 26324254; PMCID: 4556057.
112. Sundermeier TR, Zhang N, Vinberg F, Mustafi D, Kohno H, Golczak M, Bai X, Maeda A, Kefalov VJ, Palczewski K. DICER1 is essential for survival of postmitotic rod photoreceptor cells in mice. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2014;28(8):3780-91. doi: 10.1096/fj.14-254292. PubMed PMID: 24812086; PMCID: 4101655.

113. Uren PJ, Lee JT, Doroudchi MM, Smith AD, Horsager A. A profile of transcriptomic changes in the rd10 mouse model of retinitis pigmentosa. Molecular vision. 2014;20:1612-28. PubMed PMID: 25489233; PMCID: 4235044.
114. Zhang N, Tsybovsky Y, Kolesnikov AV, Rozanowska M, Swider M, Schwartz SB, Stone EM, Palczewska G, Maeda A, Kefalov VJ, Jacobson SG, Cideciyan AV, Palczewski K. Protein misfolding and the pathogenesis of ABCA4-associated retinal degenerations. Human molecular genetics. 2015;24(11):3220-37. doi: 10.1093/hmg/ddv073. PubMed PMID: 25712131; PMCID: 4424957.
115. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-5. doi: 10.1093/nar/gks1193. PubMed PMID: 23193258; PMCID: PMC3531084.
116. Gibson R, Alako B, Amid C, Cerdeno-Tarraga A, Cleland I, Goodgame N, Ten Hoopen P, Jayathilaka S, Kay S, Leinonen R, Liu X, Pallreddy S, Pakseresht N, Rajan J, Rossello M, Silvester N, Smirnov D, Toribio AL, Vaughan D, Zalunin V, Cochrane G. Biocuration of functional annotation at the European nucleotide archive. Nucleic Acids Res. 2016;44(D1):D58-66. doi: 10.1093/nar/gkv1311. PubMed PMID: 26615190; PMCID: PMC4702917.
117. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nature methods. 2015;12(4):357-60. doi: 10.1038/nmeth.3317. PubMed PMID: 25751142.
118. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. PubMed PMID: 19505943; PMCID: 2723002.
119. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2008;36(Database issue):D13-21. doi: 10.1093/nar/gkm1000. PubMed PMID: 18045790; PMCID: 2238880.
120. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290-5. doi: 10.1038/nbt.3122. PubMed PMID: 25690850; PMCID: 4643835.
121. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Nuhn M, Parker A, Patricio M, Pignatelli M, Rahtz M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Birney E, Harrow J, Muffato M, Perry E, Ruffier M, Spudich G, Trevanion SJ, Cunningham F, Aken BL, Zerbino DR, Flicek P. Ensembl 2016. Nucleic Acids Res. 2016;44(D1):D710-6. doi: 10.1093/nar/gkv1157. PubMed PMID: 26687719.

122. Bolstad B. Probe Level Quantile Normalization of High Density Oligonucleotide Array Data. Unpublished manuscript http://bmbolstadcom/stuff/qnormpdf. 2001.
123. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003;19(2):185-93. PubMed PMID: 12538238.
124. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloe D, Le Gall C, Schaeffer B, Le Crom S, Guedj M, Jaffrezic F, French StatOmique C. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14(6):671-83. doi: 10.1093/bib/bbs046. PubMed PMID: 22988256.
125. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26(17):2204-7. doi: 10.1093/bioinformatics/btq351. PubMed PMID: 20639541.
126. Cavalheiro GR, Matos-Rodrigues GE, Gomes AL, Rodrigues PM, Martins RA. c-Myc regulates cell proliferation during lens development. PLoS One. 2014;9(2):e87182. doi: 10.1371/journal.pone.0087182. PubMed PMID: 24503550; PMCID: PMC3913586.
127. He S, Limi S, McGreal RS, Xie Q, Brennan LA, Kantorow WL, Kokavec J, Majumdar R, Hou H, Jr., Edelmann W, Liu W, Ashery-Padan R, Zavadil J, Kantorow M, Skoultchi AI, Stopka T, Cvekl A. Chromatin remodeling enzyme Snf2h regulates embryonic lens differentiation and denucleation. Development. 2016;143(11):1937-47. doi: 10.1242/dev.135285. PubMed PMID: 27246713; PMCID: PMC4920164.
128. Mamuya FA, Wang Y, Roop VH, Scheiblin DA, Zajac JC, Duncan MK. The roles of alphaV integrins in lens EMT and posterior capsular opacification. J Cell Mol Med. 2014;18(4):656-70. doi: 10.1111/jcmm.12213. PubMed PMID: 24495224; PMCID: PMC4000117.
129. Shaham O, Gueta K, Mor E, Oren-Giladi P, Grinberg D, Xie Q, Cvekl A, Shomron N, Davis N, Keydar-Prizant M, Raviv S, Pasmanik-Chor M, Bell RE, Levy C, Avellino R, Banfi S, Conte I, Ashery-Padan R. Pax6 regulates gene expression in the vertebrate lens through miR-204. PLoS Genet. 2013;9(3):e1003357. doi: 10.1371/journal.pgen.1003357. PubMed PMID: 23516376; PMCID: PMC3597499.
130. Wigle JT, Chowdhury K, Gruss P, Oliver G. Prox1 function is crucial for mouse lens-fibre elongation. Nat Genet. 1999;21(3):318-22. doi: 10.1038/6844. PubMed PMID: 10080188.
131. Bookout AL, Mangelsdorf DJ. Quantitative real-time PCR protocol for analysis of nuclear receptor signaling pathways. Nucl Recept Signal. 2003;1:e012. doi: 10.1621/nrs.01012. PubMed PMID: 16604184; PMCID: PMC1402222.
132. Farkas MH, Grant GR, White JA, Sousa ME, Consugar MB, Pierce EA. Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes. BMC genomics. 2013;14:486. doi: 10.1186/1471-2164-14-486. PubMed PMID: 23865674; PMCID: 3924432.
133. Glaser T, Jepeal L, Edwards JG, Young SR, Favor J, Maas RL. PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. Nat Genet. 1994;7(4):463-71. doi: 10.1038/ng0894-463. PubMed PMID: 7951315.

134. Hogan BL, Horsburgh G, Cohen J, Hetherington CM, Fisher G, Lyon MF. Small eyes (Sey): a homozygous lethal mutation on chromosome 2 which affects the differentiation of both lens and nasal placodes in the mouse. Journal of embryology and experimental morphology. 1986;97:95-110. PubMed PMID: 3794606.
135. Hill RE, Favor J, Hogan BL, Ton CC, Saunders GF, Hanson IM, Prosser J, Jordan T, Hastie ND, van Heyningen V. Mouse small eye results from mutations in a paired-like homeobox-containing gene. Nature. 1991;354(6354):522-5. doi: 10.1038/354522a0. PubMed PMID: 1684639.
136. Amato MA, Boy S, Arnault E, Girard M, Della Puppa A, Sharif A, Perron M. Comparison of the expression patterns of five neural RNA binding proteins in the Xenopus retina. The Journal of comparative neurology. 2005;481(4):331-9. doi: 10.1002/cne.20387. PubMed PMID: 15593335.
137. Bitel CL, Perrone-Bizzozero NI, Frederikse PH. HuB/C/D, nPTB, REST4, and miR-124 regulators of neuronal cell identity are also utilized in the lens. Molecular vision. 2010;16:2301-16. PubMed PMID: 21139978; PMCID: 2994760.
138. Li G, Yi S, Yang F, Zhou Y, Ji Q, Cai J, Mei Y. Identification of mutant genes with high-frequency, high-risk, and high-expression in lung adenocarcinoma. Thoracic cancer. 2014;5(3):211-8. doi: 10.1111/1759-7714.12080. PubMed PMID: 26767003; PMCID: 4704307.
139. Yang C, Sun C, Liang X, Xie S, Huang J, Li D. Integrative analysis of microRNA and mRNA expression profiles in non-small-cell lung cancer. Cancer gene therapy. 2016;23(4):90-7. doi: 10.1038/cgt.2016.5. PubMed PMID: 26964645.
140. Su Z, Yin J, Zhao L, Li R, Liang H, Zhang J, Wang K. Lentiviral vector-mediated RBM5 overexpression downregulates EGFR expression in human non-small cell lung cancer cells. World journal of surgical oncology. 2014;12:367. doi: 10.1186/1477-7819-12-367. PubMed PMID: 25441176; PMCID: 4289049.
141. Shao C, Zhao L, Wang K, Xu W, Zhang J, Yang B. The tumor suppressor gene RBM5 inhibits lung adenocarcinoma cell growth and induces apoptosis. World journal of surgical oncology. 2012;10:160. doi: 10.1186/1477-7819-10-160. PubMed PMID: 22866867; PMCID: 3502321.
142. Maragh S, Miller RA, Bessling SL, Wang G, Hook PW, McCallion AS. Rbm24a and Rbm24b are required for normal somitogenesis. PLoS One. 2014;9(8):e105460. doi: 10.1371/journal.pone.0105460. PubMed PMID: 25170925; PMCID: 4149414.
143. Porter FD, Drago J, Xu Y, Cheema SS, Wassif C, Huang SP, Lee E, Grinberg A, Massalas JS, Bodine D, Alt F, Westphal H. Lhx2, a LIM homeobox gene, is required for eye, forebrain, and definitive erythrocyte development. Development. 1997;124(15):2935-44. PubMed PMID: 9247336.
144. Desmaison A, Vigouroux A, Rieubland C, Peres C, Calvas P, Chassaing N. Mutations in the LHX2 gene are not a frequent cause of micro/anophthalmia. Molecular vision. 2010;16:2847-9. PubMed PMID: 21203406; PMCID: 3012651.
145. Walters RW, Bradrick SS, Gromeier M. Poly(A)-binding protein modulates mRNA susceptibility to cap-dependent miRNA-mediated repression. Rna. 2010;16(1):239-50. doi: 10.1261/rna.1795410. PubMed PMID: 19934229; PMCID: 2802033.
146. Milne K, Kobel M, Kalloger SE, Barnes RO, Gao D, Gilks CB, Watson PH, Nelson BH. Systematic analysis of immune infiltrates in high-grade serous ovarian

cancer reveals CD20, FoxP3 and TIA-1 as positive prognostic factors. PLoS One. 2009;4(7):e6412. doi: 10.1371/journal.pone.0006412. PubMed PMID: 19641607; PMCID: 2712762.

147. Koreishi AF, Saenz AJ, Persky DO, Cui H, Moskowitz A, Moskowitz CH, Teruya-Feldstein J. The role of cytotoxic and regulatory T cells in relapsed/refractory Hodgkin lymphoma. Applied immunohistochemistry & molecular morphology : AIMM / official publication of the Society for Applied Immunohistochemistry. 2010;18(3):206-11. doi: 10.1097/PAI.0b013e3181c7138b. PubMed PMID: 20065852; PMCID: 3260943.

148. Forch P, Puig O, Martinez C, Seraphin B, Valcarcel J. The splicing regulator TIA-1 interacts with U1-C to promote U1 snRNP recruitment to 5' splice sites. The EMBO journal. 2002;21(24):6882-92. PubMed PMID: 12486009; PMCID: 139089.

149. Lachke SA, Alkuraya FS, Kneeland SC, Ohn T, Aboukhalil A, Howell GR, Saadi I, Cavallesco R, Yue Y, Tsai AC, Nair KS, Cosma MI, Smith RS, Hodges E, Alfadhli SM, Al-Hajeri A, Shamseldin HE, Behbehani A, Hannon GJ, Bulyk ML, Drack AV, Anderson PJ, John SW, Maas RL. Mutations in the RNA granule component TDRD7 cause cataract and glaucoma. Science. 2011;331(6024):1571-6. doi: 10.1126/science.1195970. PubMed PMID: 21436445; PMCID: 3279122.

150. Cederquist GY, Luchniak A, Tischfield MA, Peeva M, Song Y, Menezes MP, Chan WM, Andrews C, Chew S, Jamieson RV, Gomes L, Flaherty M, Grant PE, Gupta ML, Jr., Engle EC. An inherited TUBB2B mutation alters a kinesin-binding site and causes polymicrogyria, CFEOM and axon dysinnervation. Human molecular genetics. 2012;21(26):5484-99. doi: 10.1093/hmg/dds393. PubMed PMID: 23001566; PMCID: 3516133.

151. Zhang SS, Xu X, Liu MG, Zhao H, Soares MB, Barnstable CJ, Fu XY. A biphasic pattern of gene expression during mouse retina development. BMC developmental biology. 2006;6:48. doi: 10.1186/1471-213X-6-48. PubMed PMID: 17044933; PMCID: 1633734.

152. Farjo R, Yu J, Othman MI, Yoshida S, Sheth S, Glaser T, Baehr W, Swaroop A. Mouse eye gene microarrays for investigating ocular development and disease. Vision research. 2002;42(4):463-70. PubMed PMID: 11853762.

153. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH, Yung R, Asch E, Ohno-Machado L, Wong WH, Cepko CL. Genomic analysis of mouse retinal development. PLoS biology. 2004;2(9):E247. doi: 10.1371/journal.pbio.0020247. PubMed PMID: 15226823; PMCID: 439783.

154. King R, Lu L, Williams RW, Geisert EE. Transcriptome networks in the mouse retina: An exon level BXD RI database. Molecular vision. 2015;21:1235-51. PubMed PMID: 26604663; PMCID: 4626778.

155. Anand D, Lachke SA. Systems biology of lens development: A paradigm for disease gene discovery in the eye. Experimental eye research. 2016. doi: 10.1016/j.exer.2016.03.010. PubMed PMID: 26992779; PMCID: 5026553.

156. Sharma KK, Santhoshkumar P. Lens aging: effects of crystallins. Biochimica et biophysica acta. 2009;1790(10):1095-108. doi: 10.1016/j.bbagen.2009.05.008. PubMed PMID: 19463898; PMCID: 2743770.

157. Cvekl A, Duncan MK. Genetic and epigenetic mechanisms of gene regulation during lens development. Progress in retinal and eye research. 2007;26(6):555-97. doi: 10.1016/j.preteyeres.2007.07.002. PubMed PMID: 17905638; PMCID: 2136409.

158. Stevens M, Oltean S. Alternative Splicing in CKD. J Am Soc Nephrol. 2016. doi: 10.1681/ASN.2015080908. PubMed PMID: 26763787.
159. Christinat Y, Moret BM. Inferring transcript phylogenies. BMC bioinformatics. 2012;13 Suppl 9:S1. PubMed PMID: 22831154; PMCID: 3372451.
160. Fernandez-Valverde SL, Calcino AD, Degnan BM. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge Amphimedon queenslandica. BMC genomics. 2015;16:387. doi: 10.1186/s12864-015-1588-z. PubMed PMID: 25975661; PMCID: 4432959.
161. Calahorro F, Holden-Dye L, O'Connor V. Analysis of splice variants for the C. elegans orthologue of human neuroligin reveals a developmentally regulated transcript. Gene expression patterns : GEP. 2015;17(2):69-78. doi: 10.1016/j.gep.2015.02.002. PubMed PMID: 25726726.
162. Zimmermann C, Stevant I, Borel C, Conne B, Pitetti JL, Calvel P, Kaessmann H, Jegou B, Chalmel F, Nef S. Research resource: the dynamic transcriptional profile of sertoli cells during the progression of spermatogenesis. Molecular endocrinology. 2015;29(4):627-42. doi: 10.1210/me.2014-1356. PubMed PMID: 25710594.
163. Chen L, Kostadima M, Martens JH, Canu G, Garcia SP, Turro E, Downes K, Macaulay IC, Bielczyk-Maczynska E, Coe S, Farrow S, Poudel P, Burden F, Jansen SB, Astle WJ, Attwood A, Bariana T, de Bono B, Breschi A, Chambers JC, Consortium B, Choudry FA, Clarke L, Coupland P, van der Ent M, Erber WN, Jansen JH, Favier R, Fenech ME, Foad N, Freson K, van Geet C, Gomez K, Guigo R, Hampshire D, Kelly AM, Kerstens HH, Kooner JS, Laffan M, Lentaigne C, Labalette C, Martin T, Meacham S, Mumford A, Nurnberg S, Palumbo E, van der Reijden BA, Richardson D, Sammut SJ, Slodkowicz G, Tamuri AU, Vasquez L, Voss K, Watt S, Westbury S, Flicek P, Loos R, Goldman N, Bertone P, Read RJ, Richardson S, Cvejic A, Soranzo N, Ouwehand WH, Stunnenberg HG, Frontini M, Rendon A. Transcriptional diversity during lineage commitment of human blood progenitors. Science. 2014;345(6204):1251033. doi: 10.1126/science.1251033. PubMed PMID: 25258084; PMCID: 4254742.
164. Manthey AL, Terrell AM, Lachke SA, Polson SW, Duncan MK. Development of novel filtering criteria to analyze RNA-sequencing data obtained from the murine ocular lens during embryogenesis. Genomics data. 2014;2:369-74. doi: 10.1016/j.gdata.2014.10.015. PubMed PMID: 25478318; PMCID: 4248573.
165. Hoang TV, Kumar PK, Sutharzan S, Tsonis PA, Liang C, Robinson ML. Comparative transcriptome analysis of epithelial and fiber cells in newborn mouse lenses with RNA sequencing. Molecular vision. 2014;20:1491-517. Epub 2014/12/10. PubMed PMID: 25489224; PMCID: PMC4225139.
166. Budak G, Dash S, Srivastava R, Lachke SA, Janga SC. Express: A database of transcriptome profiles encompassing known and novel transcripts across multiple development stages in eye tissues. Experimental eye research. 2018;168:57-68. Epub 2018/01/18. doi: 10.1016/j.exer.2018.01.009. PubMed PMID: 29337142; PMCID: PMC5826895.
167. Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences of the United States of America. 2014;111(51):E5593-601. doi: 10.1073/pnas.1419161111. PubMed PMID: 25480548; PMCID: 4280593.

168. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome research. 2005;15(8):1051-60. doi: 10.1101/gr.3642605. PubMed PMID: 16024817; PMCID: 1182217.
169. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. Distinguishing protein-coding and noncoding genes in the human genome. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(49):19428-33. doi: 10.1073/pnas.0709013104. PubMed PMID: 18040051; PMCID: 2148306.
170. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 2013;41(Database issue):D64-9. doi: 10.1093/nar/gks1048. PubMed PMID: 23155063; PMCID: 3531082.
171. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27(3):431-2. doi: 10.1093/bioinformatics/btq675. PubMed PMID: 21149340; PMCID: 3031041.
172. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091-3. doi: 10.1093/bioinformatics/btp101. PubMed PMID: 19237447; PMCID: 2666812.
173. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Kahari AK, Keenan S, Martin FJ, Maurel T, McLaren W, Murphy DN, Nag R, Overduin B, Parker A, Patricio M, Perry E, Pignatelli M, Riat HS, Sheppard D, Taylor K, Thormann A, Vullo A, Wilder SP, Zadissa A, Aken BL, Birney E, Harrow J, Kinsella R, Muffato M, Ruffier M, Searle SM, Spudich G, Trevanion SJ, Yates A, Zerbino DR, Flicek P. Ensembl 2015. Nucleic Acids Res. 2015;43(Database issue):D662-9. doi: 10.1093/nar/gku1010. PubMed PMID: 25352552; PMCID: 4383879.
174. Lizio M, Harshbarger J, Abugessaisa I, Noguchi S, Kondo A, Severin J, Mungall C, Arenillas D, Mathelier A, Medvedeva YA, Lennartsson A, Drablos F, Ramilowski JA, Rackham O, Gough J, Andersson R, Sandelin A, Ienasescu H, Ono H, Bono H, Hayashizaki Y, Carninci P, Forrest AR, Kasukawa T, Kawaji H. Update of the FANTOM web resource: high resolution transcriptome of diverse cell types in mammals. Nucleic Acids Res. 2017;45(D1):D737-D43. doi: 10.1093/nar/gkw995. PubMed PMID: 27794045; PMCID: 5210666.
175. Hutchins AP, Poulain S, Fujii H, Miranda-Saavedra D. Discovery and characterization of new transcripts from RNA-seq data in mouse CD4(+) T cells. Genomics. 2012;100(5):303-13. doi: 10.1016/j.ygeno.2012.07.014. PubMed PMID: 22884873.

176. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR. HMMER web server: 2015 update. Nucleic Acids Res. 2015;43(W1):W30-8. doi: 10.1093/nar/gkv397. PubMed PMID: 25943547; PMCID: PMC4489315.
177. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11(5):345-55. doi: 10.1038/nrg2776. PubMed PMID: 20376054.
178. Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. Nucleic Acids Res. 2016;44(2):838-51. doi: 10.1093/nar/gkv1168. PubMed PMID: 26531823; PMCID: PMC4737145.
179. Kechavarzi B, Janga SC. Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome biology. 2014;15(1):R14. doi: 10.1186/gb-2014-15-1-r14. PubMed PMID: 24410894; PMCID: 4053825.
180. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. Genome Med. 2015;7(1):45. doi: 10.1186/s13073-015-0168-9. PubMed PMID: 26113877; PMCID: PMC4480902.
181. Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. Nat Commun. 2014;5:5274. doi: 10.1038/ncomms6274. PubMed PMID: 25409906; PMCID: PMC4467577.
182. Hollander D, Donyo M, Atias N, Mekahel K, Melamed Z, Yannai S, Lev-Maor G, Shilo A, Schwartz S, Barshack I, Sharan R, Ast G. A network-based analysis of colon cancer splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from ELK1. Genome research. 2016. doi: 10.1101/gr.193169.115. PubMed PMID: 26860615.
183. Kim E, Goren A, Ast G. Alternative splicing: current perspectives. BioEssays : news and reviews in molecular, cellular and developmental biology. 2008;30(1):38-47. doi: 10.1002/bies.20692. PubMed PMID: 18081010.
184. van Heyningen V, Williamson KA. PAX6 in sensory development. Human molecular genetics. 2002;11(10):1161-7. PubMed PMID: 12015275.
185. Yan Q, Gong L, Deng M, Zhang L, Sun S, Liu J, Ma H, Yuan D, Chen PC, Hu X, Liu J, Qin J, Xiao L, Huang XQ, Zhang J, Li DW. Sumoylation activates the transcriptional activity of Pax-6, an important transcription factor for eye and brain development. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(49):21034-9. doi: 10.1073/pnas.1007866107. PubMed PMID: 21084637; PMCID: 3000302.
186. Madhavan M, Haynes TL, Frisch NC, Call MK, Minich CM, Tsonis PA, Del Rio-Tsonis K. The role of Pax-6 in lens regeneration. Proceedings of the National Academy of Sciences of the United States of America. 2006;103(40):14848-53. doi: 10.1073/pnas.0601949103. PubMed PMID: 17003134; PMCID: 1595439.
187. Litt M, Kramer P, LaMorticella DM, Murphey W, Lovrien EW, Weleber RG. Autosomal dominant congenital cataract associated with a missense mutation in the human alpha crystallin gene CRYAA. Human molecular genetics. 1998;7(3):471-4. PubMed PMID: 9467006.
188. Brady JP, Garland D, Duglas-Tabor Y, Robison WG, Jr., Groome A, Wawrousek EF. Targeted disruption of the mouse alpha A-crystallin gene induces cataract and

cytoplasmic inclusion bodies containing the small heat shock protein alpha B-crystallin. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(3):884-9. PubMed PMID: 9023351; PMCID: 19608.
189. Wong JJ, Au AY, Ritchie W, Rasko JE. Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. BioEssays : news and reviews in molecular, cellular and developmental biology. 2016;38(1):41-9. doi: 10.1002/bies.201500117. PubMed PMID: 26612485.
190. Delgado-Esteban M, Garcia-Higuera I, Maestre C, Moreno S, Almeida A. APC/C-Cdh1 coordinates neurogenesis and cortical size during development. Nat Commun. 2013;4:2879. doi: 10.1038/ncomms3879. PubMed PMID: 24301314.
191. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014;15(12):829-45. doi: 10.1038/nrg3813. PubMed PMID: 25365966.
192. Neelamraju Y, Hashemikhabir S, Janga SC. The human RBPome: from genes and proteins to human disease. Journal of proteomics. 2015;127(Pt A):61-70. doi: 10.1016/j.jprot.2015.04.031. PubMed PMID: 25982388.
193. Janga SC. From specific to global analysis of posttranscriptional regulation in eukaryotes: posttranscriptional regulatory networks. Briefings in functional genomics. 2012;11(6):505-21. doi: 10.1093/bfgp/els046. PubMed PMID: 23124862.
194. Kim MY, Hur J, Jeong S. Emerging roles of RNA and RNA-binding protein network in cancer cells. BMB reports. 2009;42(3):125-30. PubMed PMID: 19335997.
195. Chaudhury A, Chander P, Howe PH. Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on hnRNP E1's multifunctional regulatory roles. Rna. 2010;16(8):1449-62. doi: 10.1261/rna.2254110. PubMed PMID: 20584894.
196. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. Science. 2003;302(5648):1212-5. doi: 10.1126/science.1090095. PubMed PMID: 14615540.
197. Konig J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2011;13(2):77-83. doi: 10.1038/nrg3141. PubMed PMID: 22251872.
198. Spitzer J, Hafner M, Landthaler M, Ascano M, Farazi T, Wardle G, Nusbaum J, Khorshid M, Burger L, Zavolan M, Tuschl T. PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. Methods in enzymology. 2014;539:113-61. doi: 10.1016/b978-0-12-420120-0.00008-6. PubMed PMID: 24581442.
199. Darnell RB. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. Wiley Interdiscip Rev RNA. 2010;1(2):266-86. doi: 10.1002/wrna.31. PubMed PMID: 21935890; PMCID: 3222227.
200. Huppertz I, Attig J, D'Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, Tajnik M, Konig J, Ule J. iCLIP: protein-RNA interactions at nucleotide resolution. Methods (San Diego, Calif). 2014;65(3):274-87. doi: 10.1016/j.ymeth.2013.10.011. PubMed PMID: 24184352.
201. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. 2009;10(3):R25. doi: 10.1186/gb-2009-10-3-r25. PubMed PMID: 19261174.

202. Smith AD, Chung WY, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ. Updates to the RMAP short-read mapping software. Bioinformatics. 2009;25(21):2841-2. doi: 10.1093/bioinformatics/btp533. PubMed PMID: 19736251.
203. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11. doi: 10.1093/bioinformatics/btp120. PubMed PMID: 19289445.
204. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013;14(4):R36. doi: 10.1186/gb-2013-14-4-r36. PubMed PMID: 23618408.
205. Comoglio F, Sievers C, Paro R. Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. BMC bioinformatics. 2015;16:32. doi: 10.1186/s12859-015-0470-y. PubMed PMID: 25638391.
206. Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, Ohler U. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome biology. 2011;12(8):R79. doi: 10.1186/gb-2011-12-8-r79. PubMed PMID: 21851591; PMCID: 3302668.
207. Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, Hannon GJ, Sanford JR, Penalva LO, Smith AD. Site identification in high-throughput RNA-protein interaction data. Bioinformatics. 2012;28(23):3013-20. doi: 10.1093/bioinformatics/bts569. PubMed PMID: 23024010; PMCID: 3509493.
208. Chen B, Yun J, Kim MS, Mendell JT, Xie Y. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome biology. 2014;15(1):R18. doi: 10.1186/gb-2014-15-1-r18. PubMed PMID: 24451213.
209. Wang T, Xie Y, Xiao G. dCLIP: a computational approach for comparative CLIP-seq analyses. Genome biology. 2014;15(1):R11. doi: 10.1186/gb-2014-15-1-r11. PubMed PMID: 24398258.
210. Reyes-Herrera PH, Ficarra E. Computational Methods for CLIP-seq Data Processing. Bioinformatics and biology insights. 2014;8:199-207. doi: 10.4137/bbi.s16803. PubMed PMID: 25336930.
211. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, Scott LJ, Sartor MA. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Res. 2014;42(13):e105. doi: 10.1093/nar/gku463. PubMed PMID: 24878920; PMCID: PMC4117744.
212. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13. doi: 10.1093/nar/gkn923. PubMed PMID: 19033363.
213. Erhard F, Dolken L, Zimmer R. RIP-chip enrichment analysis. Bioinformatics. 2013;29(1):77-83. doi: 10.1093/bioinformatics/bts631. PubMed PMID: 23104891.
214. Lambert NJ, Robertson AD, Burge CB. RNA Bind-n-Seq: Measuring the Binding Affinity Landscape of RNA-Binding Proteins. Methods in enzymology. 2015;558:465-93. doi: 10.1016/bs.mie.2015.02.007. PubMed PMID: 26068750.
215. Lambert N, Robertson A, Jangi M, McGeary S, Sharp PA, Burge CB. RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol Cell. 2014;54(5):887-900. doi: 10.1016/j.molcel.2014.04.016. PubMed PMID: 24837674; PMCID: PMC4142047.

216. Nicholson CO, Friedersdorf MB, Keene JD. Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. Rna. 2016. doi: 10.1261/rna.058115.116. PubMed PMID: 27742911.
217. Yang YC, Di C, Hu B, Zhou M, Liu Y, Song N, Li Y, Umetsu J, Lu ZJ. CLIPdb: a CLIP-seq database for protein-RNA interactions. BMC genomics. 2015;16:51. doi: 10.1186/s12864-015-1273-2. PubMed PMID: 25652745; PMCID: 4326514.
218. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo GW. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nature methods. 2016. doi: 10.1038/nmeth.3810. PubMed PMID: 27018577.
219. Popow J, Alleaume AM, Curk T, Schwarzl T, Sauer S, Hentze MW. FASTKD2 is an RNA-binding protein required for mitochondrial RNA processing and translation. Rna. 2015;21(11):1873-84. doi: 10.1261/rna.052365.115. PubMed PMID: 26370583.
220. Nishimura D. BioCarta. Biotech Software & Internet Report. 2001;2(3):117-20. doi: 10.1089/152791601750294344.
221. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44(D1):D457-62. doi: 10.1093/nar/gkv1070. PubMed PMID: 26476454; PMCID: PMC4702792.
222. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30. PubMed PMID: 10592173; PMCID: PMC102409.
223. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, Jassal B, Jupe S, Korninger F, McKay S, Matthews L, May B, Milacic M, Rothfels K, Shamovsky V, Webber M, Weiser J, Williams M, Wu G, Stein L, Hermjakob H, D'Eustachio P. The Reactome pathway Knowledgebase. Nucleic Acids Res. 2016;44(D1):D481-7. doi: 10.1093/nar/gkv1351. PubMed PMID: 26656494; PMCID: PMC4702931.
224. Milacic M, Haw R, Rothfels K, Wu G, Croft D, Hermjakob H, D'Eustachio P, Stein L. Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers (Basel). 2012;4(4):1180-211. doi: 10.3390/cancers4041180. PubMed PMID: 24213504; PMCID: PMC3712731.
225. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jahn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park SM, Riggs ER, Scott RH, Sisodiya S, Van Vooren S, Wapner RJ, Wilkie AO, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BB, Washingthon NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014;42(Database issue):D966-74. doi: 10.1093/nar/gkt1026. PubMed PMID: 24217912.
226. Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein TI, Safran M, Lancet D. MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2014;47:1 24 1-1 19. doi: 10.1002/0471250953.bi0124s47. PubMed PMID: 25199789.

227. Patra P, Izawa T, Pena Castillo L. REPA: Applying Pathway Analysis to Genome-wide Transcription Factor Binding Data. Computational Biology and Bioinformatics, IEEE/ACM Transactions on. 2015;PP(99):1-. doi: 10.1109/TCBB.2015.2453948.
228. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other1947:50-60. doi: 10.1214/aoms/1177730491.
229. Fisher RA. On the Interpretation of X2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society. 1922;85(1):87-94. doi: 10.2307/2340521.
230. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57(1):289-300.
231. D'Agostino Y, D'Aniello S. Molecular basis, applications and challenges of CRISPR/Cas9: a continuously evolving tool for genome editing. Briefings in functional genomics. 2017. doi: 10.1093/bfgp/elw038. PubMed PMID: 28057617.
232. RNA-seq profiling of CRISPR/Cas9 based knock outs of RNA-binding proteins in human cell line K562 - https://www.encodeproject.org/search/?type=Experiment&assay_title=CRISPR+RNA-seq&replicates.library.biosample.life_stage=adultENCODE Project;2017.
233. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2. doi: 10.1093/bioinformatics/btq033. PubMed PMID: 20110278; PMCID: PMC2832824.
234. Stary S, Vinatzer U, Mullauer L, Raderer M, Birner P, Streubel B. t(11;14)(q23;q32) involving IGH and DDX6 in nodal marginal zone lymphoma. Genes Chromosomes Cancer. 2013;52(1):33-43. doi: 10.1002/gcc.22004. PubMed PMID: 22965301.
235. Jangra RK, Yi M, Lemon SM. DDX6 (Rck/p54) is required for efficient hepatitis C virus replication but not for internal ribosome entry site-directed translation. J Virol. 2010;84(13):6810-24. doi: 10.1128/JVI.00397-10. PubMed PMID: 20392846; PMCID: PMC2903299.
236. Poppe B, Vandesompele J, Schoch C, Lindvall C, Mrozek K, Bloomfield CD, Beverloo HB, Michaux L, Dastugue N, Herens C, Yigit N, De Paepe A, Hagemeijer A, Speleman F. Expression analyses identify MLL as a prominent target of 11q23 amplification and support an etiologic role for MLL gain of function in myeloid malignancies. Blood. 2004;103(1):229-35. doi: 10.1182/blood-2003-06-2163. PubMed PMID: 12946992.
237. 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. 2009;106(48):20300-5. doi: 10.1073/pnas.0906940106. PubMed PMID: 19918083; PMCID: 2777960.
238. RNA-seq profiling of CRISPR/Cas9 based knock outs of RNA-binding proteins in human cell line K562 - https://www.encodeproject.org/search/?type=Experiment&assay_title=CRISPR+RNA-seq&replicates.library.biosample.life_stage=adultENCODE Project 2017;2017.

239. Hille F, Richter H, Wong SP, Bratovic M, Ressel S, Charpentier E. The Biology of CRISPR-Cas: Backward and Forward. Cell. 2018;172(6):1239-59. doi: 10.1016/j.cell.2017.11.032. PubMed PMID: 29522745.
240. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science. 2008;322(5909):1843-5. doi: 10.1126/science.1165771. PubMed PMID: 19095942; PMCID: PMC2695655.
241. Zhang ZT, Jimenez-Bonilla P, Seo SO, Lu T, Jin YS, Blaschek HP, Wang Y. Bacterial Genome Editing with CRISPR-Cas9: Taking Clostridium beijerinckii as an Example. Methods Mol Biol. 2018;1772:297-325. doi: 10.1007/978-1-4939-7795-6_17. PubMed PMID: 29754236.
242. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015;16(5):299-311. Epub 2015/04/10. doi: 10.1038/nrg3899. PubMed PMID: 25854182; PMCID: PMC4503232.
243. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014;346(6213):1258096. Epub 2014/11/29. doi: 10.1126/science.1258096. PubMed PMID: 25430774.
244. Ma J, Koster J, Qin Q, Hu S, Li W, Chen C, Cao Q, Wang J, Mei S, Liu Q, Xu H, Liu XS. CRISPR-DO for genome-wide CRISPR design and optimization. Bioinformatics. 2016;32(21):3336-8. doi: 10.1093/bioinformatics/btw476. PubMed PMID: 27402906; PMCID: PMC6095119.
245. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, Adamson B, Norman TM, Lander ES, Weissman JS, Friedman N, Regev A. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell. 2016;167(7):1853-66 e17. Epub 2016/12/17. doi: 10.1016/j.cell.2016.11.038. PubMed PMID: 27984732; PMCID: PMC5181115.
246. Horlbeck MA, Gilbert LA, Villalta JE, Adamson B, Pak RA, Chen Y, Fields AP, Park CY, Corn JE, Kampmann M, Weissman JS. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. eLife. 2016;5. doi: 10.7554/eLife.19760. PubMed PMID: 27661255; PMCID: PMC5094855.
247. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, Attenello FJ, Villalta JE, Cho MY, Chen Y, Mandegar MA, Olvera MP, Gilbert LA, Conklin BR, Chang HY, Weissman JS, Lim DA. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science. 2017;355(6320). Epub 2016/12/17. doi: 10.1126/science.aah7111. PubMed PMID: 27980086; PMCID: PMC5394926.
248. Koike-Yusa H, Li Y, Tan EP, Velasco-Herrera Mdel C, Yusa K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol. 2014;32(3):267-73. Epub 2014/02/19. doi: 10.1038/nbt.2800. PubMed PMID: 24535568.
249. Birsoy K, Wang T, Chen WW, Freinkman E, Abu-Remaileh M, Sabatini DM. An Essential Role of the Mitochondrial Electron Transport Chain in Cell Proliferation Is to Enable Aspartate Synthesis. Cell. 2015;162(3):540-51. doi: 10.1016/j.cell.2015.07.016. PubMed PMID: 26232224; PMCID: PMC4522279.
250. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP,

Rissland OS, Durocher D, Angers S, Moffat J. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 2015;163(6):1515-26. Epub 2015/12/03. doi: 10.1016/j.cell.2015.11.015. PubMed PMID: 26627737.

251. de Rooij L, Chan DCH, Keyvani Chahi A, Hope KJ. Post-transcriptional regulation in hematopoiesis: RNA binding proteins take control (1). Biochem Cell Biol. 2019;97(1):10-20. Epub 2018/06/14. doi: 10.1139/bcb-2017-0310. PubMed PMID: 29898370.

252. Brinegar AE, Cooper TA. Roles for RNA-binding proteins in development and disease. Brain Res. 2016;1647:1-8. Epub 2016/03/15. doi: 10.1016/j.brainres.2016.02.050. PubMed PMID: 26972534; PMCID: PMC5003702.

253. Budak G, Srivastava R, Janga SC. Seten: a tool for systematic identification and comparison of processes, phenotypes, and diseases associated with RNA-binding proteins from condition-specific CLIP-seq profiles. Rna. 2017;23(6):836-46. Epub 2017/03/25. doi: 10.1261/rna.059089.116. PubMed PMID: 28336542; PMCID: PMC5435856.

254. Ramanathan M, Porter DF, Khavari PA. Methods to study RNA-protein interactions. Nature methods. 2019;16(3):225-34. Epub 2019/02/26. doi: 10.1038/s41592-019-0330-1. PubMed PMID: 30804549.

255. Lee FCY, Ule J. Advances in CLIP Technologies for Studies of Protein-RNA Interactions. Mol Cell. 2018;69(3):354-69. Epub 2018/02/06. doi: 10.1016/j.molcel.2018.01.005. PubMed PMID: 29395060.

256. Queiroz RML, Smith T, Villanueva E, Marti-Solano M, Monti M, Pizzinga M, Mirea DM, Ramakrishna M, Harvey RF, Dezi V, Thomas GH, Willis AE, Lilley KS. Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). Nat Biotechnol. 2019;37(2):169-78. Epub 2019/01/05. doi: 10.1038/s41587-018-0001-2. PubMed PMID: 30607034.

257. Wheeler EC, Van Nostrand EL, Yeo GW. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. Wiley Interdiscip Rev RNA. 2018;9(1). Epub 2017/08/31. doi: 10.1002/wrna.1436. PubMed PMID: 28853213; PMCID: PMC5739989.

258. Zhao Y, Zhang Y, Teng Y, Liu K, Liu Y, Li W, Wu L. SpyCLIP: an easy-to-use and high-throughput compatible CLIP platform for the characterization of protein-RNA interactions with high accuracy. Nucleic Acids Res. 2019;47(6):e33. Epub 2019/02/05. doi: 10.1093/nar/gkz049. PubMed PMID: 30715466; PMCID: PMC6451120.

259. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nature methods. 2011;8(7):559-64. doi: 10.1038/nmeth.1608. PubMed PMID: 21572407.

260. Wang F, Wang L, Zou X, Duan S, Li Z, Deng Z, Luo J, Lee SY, Chen S. Advances in CRISPR-Cas systems for RNA targeting, tracking and editing. Biotechnol Adv. 2019. doi: 10.1016/j.biotechadv.2019.03.016. PubMed PMID: 30926472.

261. Nelles DA, Fang MY, O'Connell MR, Xu JL, Markmiller SJ, Doudna JA, Yeo GW. Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell. 2016;165(2):488-96. Epub 2016/03/22. doi: 10.1016/j.cell.2016.02.054. PubMed PMID: 26997482; PMCID: PMC4826288.

262. Konig J, Zarnack K, Luscombe NM, Ule J. Protein-RNA interactions: new genomic technologies and perspectives. Nat Rev Genet. 2012;13(2):77-83. doi: 10.1038/nrg3141. PubMed PMID: 22251872.
263. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308-11. PubMed PMID: 11125122; PMCID: PMC29783.
264. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5. Epub 2013/05/30. doi: 10.1038/ng.2653. PubMed PMID: 23715323; PMCID: PMC4010069.
265. Collado-Torres L, Nellore A, Jaffe AE. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. F1000Res. 2017;6:1558. doi: 10.12688/f1000research.12223.1. PubMed PMID: 29043067; PMCID: PMC5621122.
266. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analysis using recount2. Nat Biotechnol. 2017;35(4):319-21. doi: 10.1038/nbt.3838. PubMed PMID: 28398307.
267. Theil K, Herzog M, Rajewsky N. Post-transcriptional Regulation by 3' UTRs Can Be Masked by Regulatory Elements in 5' UTRs. Cell Rep. 2018;22(12):3217-26. Epub 2018/03/22. doi: 10.1016/j.celrep.2018.02.094. PubMed PMID: 29562178.
268. Bryant CD, Yazdani N. RNA-binding proteins, neural development and the addictions. Genes Brain Behav. 2016;15(1):169-86. Epub 2015/12/09. doi: 10.1111/gbb.12273. PubMed PMID: 26643147; PMCID: PMC4944654.
269. Naito T, Tanaka H, Naoe Y, Taniuchi I. Transcriptional control of T-cell development. Int Immunol. 2011;23(11):661-8. Epub 2011/09/29. doi: 10.1093/intimm/dxr078. PubMed PMID: 21948191.
270. Chang X, Li B, Rao A. RNA-binding protein hnRNPLL regulates mRNA splicing and stability during B-cell to plasma-cell differentiation. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(15):E1888-97. Epub 2015/04/01. doi: 10.1073/pnas.1422490112. PubMed PMID: 25825742; PMCID: PMC4403190.
271. Wei YN, Hu HY, Xie GC, Fu N, Ning ZB, Zeng R, Khaitovich P. Transcript and protein expression decoupling reveals RNA binding proteins and miRNAs as potential modulators of human aging. Genome biology. 2015;16:41. doi: 10.1186/s13059-015-0608-2. PubMed PMID: 25853883; PMCID: 4375924.
272. Moore KS, von Lindern M. RNA Binding Proteins and Regulation of mRNA Translation in Erythropoiesis. Frontiers in physiology. 2018;9:910. Epub 2018/08/09. doi: 10.3389/fphys.2018.00910. PubMed PMID: 30087616; PMCID: PMC6066521.
273. Fassnacht C, Ciosk R. Cell Fate Maintenance and Reprogramming During the Oocyte-to-Embryo Transition. Results Probl Cell Differ. 2017;59:269-86. Epub 2017/03/02. doi: 10.1007/978-3-319-44820-6_10. PubMed PMID: 28247053.
274. Dou XM, Zhang XS. [RNA-binding protein PTB in spermatogenesis: Progress in studies]. Zhonghua Nan Ke Xue. 2016;22(9):856-60. Epub 2017/10/27. PubMed PMID: 29071887.
275. Newman R, McHugh J, Turner M. RNA binding proteins as regulators of immune cell biology. Clin Exp Immunol. 2016;183(1):37-49. Epub 2015/07/24. doi: 10.1111/cei.12684. PubMed PMID: 26201441; PMCID: PMC4687516.

276. Ule J, Jensen K, Mele A, Darnell RB. CLIP: a method for identifying protein-RNA interaction sites in living cells. Methods (San Diego, Calif). 2005;37(4):376-86. Epub 2005/11/30. doi: 10.1016/j.ymeth.2005.07.018. PubMed PMID: 16314267.

277. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34(2):184-91. Epub 2016/01/19. doi: 10.1038/nbt.3437. PubMed PMID: 26780180; PMCID: PMC4744125.

278. Soreze Y, Boutron A, Habarou F, Barnerias C, Nonnenmacher L, Delpech H, Mamoune A, Chretien D, Hubert L, Bole-Feysot C, Nitschke P, Correia I, Sardet C, Boddaert N, Hamel Y, Delahodde A, Ottolenghi C, de Lonlay P. Mutations in human lipoyltransferase gene LIPT1 cause a Leigh disease with secondary deficiency for pyruvate and alpha-ketoglutarate dehydrogenase. Orphanet J Rare Dis. 2013;8:192. Epub 2013/12/18. doi: 10.1186/1750-1172-8-192. PubMed PMID: 24341803; PMCID: PMC3905285.

279. Tort F, Ferrer-Cortes X, Thio M, Navarro-Sastre A, Matalonga L, Quintana E, Bujan N, Arias A, Garcia-Villoria J, Acquaviva C, Vianey-Saban C, Artuch R, Garcia-Cazorla A, Briones P, Ribes A. Mutations in the lipoyltransferase LIPT1 gene cause a fatal disease associated with a specific lipoylation defect of the 2-ketoacid dehydrogenase complexes. Human molecular genetics. 2014;23(7):1907-15. Epub 2013/11/22. doi: 10.1093/hmg/ddt585. PubMed PMID: 24256811.

280. Mayr JA, Feichtinger RG, Tort F, Ribes A, Sperl W. Lipoic acid biosynthesis defects. J Inherit Metab Dis. 2014;37(4):553-63. Epub 2014/04/30. doi: 10.1007/s10545-014-9705-8. PubMed PMID: 24777537.

281. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett. 2008;582(14):1977-86. doi: 10.1016/j.febslet.2008.03.004. PubMed PMID: 18342629; PMCID: PMC2858862.

282. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LO, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. A compendium of RNA-binding motifs for decoding gene regulation. Nature. 2013;499(7457):172-7. doi: 10.1038/nature12311. PubMed PMID: 23846655; PMCID: PMC3929597.

283. Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. Trends Genet. 2013;29(5):318-27. doi: 10.1016/j.tig.2013.01.004. PubMed PMID: 23415593.

284. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jr., Jungkamp AC, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010;141(1):129-41. doi: 10.1016/j.cell.2010.03.009. PubMed PMID: 20371350; PMCID: PMC2861495.

285. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature. 2008;456(7221):464-9. Epub

2008/11/04. doi: 10.1038/nature07488. PubMed PMID: 18978773; PMCID: PMC2597294.
286. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K, Stanton R, Rigo F, Guttman M, Yeo GW. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nature methods. 2016;13(6):508-14. doi: 10.1038/nmeth.3810. PubMed PMID: 27018577; PMCID: PMC4887338.
287. Zhu Y, Xu G, Yang YT, Xu Z, Chen X, Shi B, Xie D, Lu ZJ, Wang P. POSTAR2: deciphering the post-transcriptional regulatory logics. Nucleic Acids Res. 2019;47(D1):D203-D11. doi: 10.1093/nar/gky830. PubMed PMID: 30239819; PMCID: PMC6323971.
288. Tollervey JR, Curk T, Rogelj B, Briese M, Cereda M, Kayikci M, Konig J, Hortobagyi T, Nishimura AL, Zupunski V, Patani R, Chandran S, Rot G, Zupan B, Shaw CE, Ule J. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nat Neurosci. 2011;14(4):452-8. doi: 10.1038/nn.2778. PubMed PMID: 21358640; PMCID: PMC3108889.
289. Macias S, Plass M, Stajuda A, Michlewski G, Eyras E, Caceres JF. DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. Nature structural & molecular biology. 2012;19(8):760-6. doi: 10.1038/nsmb.2344. PubMed PMID: 22796965; PMCID: PMC3442229.
290. Sauliere J, Murigneux V, Wang Z, Marquenet E, Barbosa I, Le Tonqueze O, Audic Y, Paillard L, Roest Crollius H, Le Hir H. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nature structural & molecular biology. 2012;19(11):1124-31. doi: 10.1038/nsmb.2420. PubMed PMID: 23085716.
291. Hoell JI, Larsson E, Runge S, Nusbaum JD, Duggimpudi S, Farazi TA, Hafner M, Borkhardt A, Sander C, Tuschl T. RNA targets of wild-type and mutant FET family proteins. Nature structural & molecular biology. 2011;18(12):1428-31. doi: 10.1038/nsmb.2163. PubMed PMID: 22081015; PMCID: PMC3230689.
292. Ascano M, Jr., Mukherjee N, Bandaru P, Miller JB, Nusbaum JD, Corcoran DL, Langlois C, Munschauer M, Dewell S, Hafner M, Williams Z, Ohler U, Tuschl T. FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature. 2012;492(7429):382-6. doi: 10.1038/nature11737. PubMed PMID: 23235829; PMCID: PMC3528815.
293. Gerstberger S, Hafner M, Ascano M, Tuschl T. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. Advances in experimental medicine and biology. 2014;825:1-55. doi: 10.1007/978-1-4939-1221-6_1. PubMed PMID: 25201102; PMCID: PMC4180674.
294. Dermitzakis ET, Reymond A, Lyle R, Scamuffa N, Ucla C, Deutsch S, Stevenson BJ, Flegel V, Bucher P, Jongeneel CV, Antonarakis SE. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. Nature. 2002;420(6915):578-82. doi: 10.1038/nature01251. PubMed PMID: 12466853.
295. Hare MP, Palumbi SR. High intron sequence conservation across three mammalian orders suggests functional constraints. Mol Biol Evol. 2003;20(6):969-78. doi: 10.1093/molbev/msg111. PubMed PMID: 12716984.

296. Margulies EH, Blanchette M, Program NCS, Haussler D, Green ED. Identification and characterization of multi-species conserved sequences. Genome research. 2003;13(12):2507-18. doi: 10.1101/gr.1602203. PubMed PMID: 14656959; PMCID: PMC403793.

297. Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. Nat Rev Genet. 2004;5(6):456-65. doi: 10.1038/nrg1350. PubMed PMID: 15153998.

298. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grutzner F, Bergmann S, Nielsen R, Paabo S, Kaessmann H. The evolution of gene expression levels in mammalian organs. Nature. 2011;478(7369):343-8. doi: 10.1038/nature10532. PubMed PMID: 22012392.

299. Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, Buffenstein R, Lee SR, Chang KT, Rhee H, Kim JS, Yang KS, Gladyshev VN. Gene expression defines natural changes in mammalian lifespan. Aging Cell. 2015;14(3):352-65. doi: 10.1111/acel.12283. PubMed PMID: 25677554; PMCID: PMC4406664.

300. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338(6114):1593-9. doi: 10.1126/science.1228186. PubMed PMID: 23258891; PMCID: PMC3568499.

301. Fietz SA, Lachmann R, Brandl H, Kircher M, Samusik N, Schroder R, Lakshmanaperumal N, Henry I, Vogt J, Riehn A, Distler W, Nitsch R, Enard W, Paabo S, Huttner WB. 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. 2012;109(29):11836-41. doi: 10.1073/pnas.1209647109. PubMed PMID: 22753484; PMCID: PMC3406833.

302. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635-40. doi: 10.1038/nature12943. PubMed PMID: 24463510.

303. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77. Epub 2015/02/19. doi: 10.5114/wo.2014.47136. PubMed PMID: 25691825; PMCID: PMC4322527.

304. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]. 2009;Chapter 1:Unit1 4. doi: 10.1002/0471250953.bi0104s28. PubMed PMID: 19957273; PMCID: PMC2834533.

305. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, Gil L, Giron CG, Gordon L, Hourlier T, Hunt S, Johnson N, Juettemann T, Kahari AK, Keenan S, Kulesha E, Martin FJ, Maurel T, McLaren WM, Murphy DN, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ruffier M, Sheppard D, Taylor K, Thormann A, Trevanion SJ, Vullo A, Wilder SP, Wilson M, Zadissa A, Aken BL, Birney E, Cunningham F, Harrow J, Herrero J, Hubbard TJ, Kinsella R, Muffato M, Parker A, Spudich G, Yates A, Zerbino DR, Searle SM. Ensembl 2014. Nucleic Acids Res. 2014;42(Database issue):D749-55. doi: 10.1093/nar/gkt1196. PubMed PMID: 24316576; PMCID: 3964975.

306. Zhao H, Sun Z, Wang J, Huang H, Kocher JP, Wang L. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics. 2014;30(7):1006-7. doi: 10.1093/bioinformatics/btt730. PubMed PMID: 24351709; PMCID: PMC3967108.
307. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353-8. doi: 10.1093/bioinformatics/bts163. PubMed PMID: 22492648; PMCID: PMC3348564.
308. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576-89. doi: 10.1016/j.molcel.2010.05.004. PubMed PMID: 20513432; PMCID: PMC2898526.
309. Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Colak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338(6114):1587-93. doi: 10.1126/science.1230612. PubMed PMID: 23258890.
310. BENJAMINI Y, HOCHBERG Y. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing1995.
311. Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome biology. 2011;12(12):R120. Epub 2011/12/21. doi: 10.1186/gb-2011-12-12-r120. PubMed PMID: 22182830; PMCID: PMC3334615.
312. Shaul O. How introns enhance gene expression. Int J Biochem Cell Biol. 2017;91(Pt B):145-55. Epub 2017/07/05. doi: 10.1016/j.biocel.2017.06.016. PubMed PMID: 28673892.
313. Wang C, Szaro BG. Post-transcriptional regulation mediated by specific neurofilament introns in vivo. Journal of cell science. 2016;129(7):1500-11. Epub 2016/02/26. doi: 10.1242/jcs.185199. PubMed PMID: 26906423.
314. Li X, Liu S, Zhang L, Issaian A, Hill RC, Espinosa S, Shi S, Cui Y, Kappel K, Das R, Hansen KC, Zhou ZH, Zhao R. A unified mechanism for intron and exon definition and back-splicing. Nature. 2019;573(7774):375-80. Epub 2019/09/06. doi: 10.1038/s41586-019-1523-6. PubMed PMID: 31485080.
315. Pineda JMB, Bradley RK. Most human introns are recognized via multiple and tissue-specific branchpoints. Genes Dev. 2018;32(7-8):577-91. doi: 10.1101/gad.312058.118. PubMed PMID: 29666160; PMCID: PMC5959240.
316. Sun L, Fazal FM, Li P, Broughton JP, Lee B, Tang L, Huang W, Kool ET, Chang HY, Zhang QC. RNA structure maps across mammalian cellular compartments. Nature structural & molecular biology. 2019;26(4):322-30. doi: 10.1038/s41594-019-0200-7. PubMed PMID: 30886404.
317. Buratti E, Baralle FE. Influence of RNA secondary structure on the pre-mRNA splicing process. Molecular and cellular biology. 2004;24(24):10505-14. doi: 10.1128/MCB.24.24.10505-10514.2004. PubMed PMID: 15572659; PMCID: PMC533984.
318. De La Garza A, Cameron RC, Gupta V, Fraint E, Nik S, Bowman TV. The splicing factor Sf3b1 regulates erythroid maturation and proliferation via TGFbeta

signaling in zebrafish. Blood Adv. 2019;3(14):2093-104. Epub 2019/07/14. doi: 10.1182/bloodadvances.2018027714. PubMed PMID: 31300417; PMCID: PMC6650725.

319. Wang L, Brooks AN, Fan J, Wan Y, Gambe R, Li S, Hergert S, Yin S, Freeman SS, Levin JZ, Fan L, Seiler M, Buonamici S, Smith PG, Chau KF, Cibulskis CL, Zhang W, Rassenti LZ, Ghia EM, Kipps TJ, Fernandes S, Bloch DB, Kotliar D, Landau DA, Shukla SA, Aster JC, Reed R, DeLuca DS, Brown JR, Neuberg D, Getz G, Livak KJ, Meyerson MM, Kharchenko PV, Wu CJ. Transcriptomic Characterization of SF3B1 Mutation Reveals Its Pleiotropic Effects in Chronic Lymphocytic Leukemia. Cancer Cell. 2016;30(5):750-63. Epub 2016/11/08. doi: 10.1016/j.ccell.2016.10.005. PubMed PMID: 27818134; PMCID: PMC5127278.

320. Cretu C, Schmitzova J, Ponce-Salvatierra A, Dybkov O, De Laurentiis EI, Sharma K, Will CL, Urlaub H, Luhrmann R, Pena V. Molecular Architecture of SF3b and Structural Consequences of Its Cancer-Related Mutations. Mol Cell. 2016;64(2):307-19. Epub 2016/10/22. doi: 10.1016/j.molcel.2016.08.036. PubMed PMID: 27720643.

321. Shuai S, Suzuki H, Diaz-Navarro A, Nadeu F, Kumar SA, Gutierrez-Fernandez A, Delgado J, Pinyol M, Lopez-Otin C, Puente XS, Taylor MD, Campo E, Stein LD. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature. 2019. Epub 2019/10/10. doi: 10.1038/s41586-019-1651-z. PubMed PMID: 31597163.

322. Seiler M, Peng S, Agrawal AA, Palacino J, Teng T, Zhu P, Smith PG, Cancer Genome Atlas Research N, Buonamici S, Yu L. Somatic Mutational Landscape of Splicing Factor Genes and Their Functional Consequences across 33 Cancer Types. Cell Rep. 2018;23(1):282-96 e4. Epub 2018/04/05. doi: 10.1016/j.celrep.2018.01.088. PubMed PMID: 29617667; PMCID: PMC5933844.

323. Thompson MW, Beasley KA, Schmidt MD, Seipelt RL. Arginyl aminopeptidase-like 1 (RNPEPL1) is an alternatively processed aminopeptidase with specificity for methionine, glutamine, and citrulline residues. Protein Pept Lett. 2009;16(10):1256-66. Epub 2009/06/11. doi: 10.2174/092986609789071199. PubMed PMID: 19508204.

324. Yu Y, Reed R. FUS functions in coupling transcription to splicing by mediating an interaction between RNAP II and U1 snRNP. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(28):8608-13. Epub 2015/07/01. doi: 10.1073/pnas.1506282112. PubMed PMID: 26124092; PMCID: PMC4507187.

325. Yamaguchi A, Takanashi K. FUS interacts with nuclear matrix-associated protein SAFB1 as well as Matrin3 to regulate splicing and ligand-mediated transcription. Sci Rep. 2016;6:35195. Epub 2016/10/13. doi: 10.1038/srep35195. PubMed PMID: 27731383; PMCID: PMC5059712.

326. Baechtold H, Kuroda M, Sok J, Ron D, Lopez BS, Akhmedov AT. Human 75-kDa DNA-pairing protein is identical to the pro-oncoprotein TLS/FUS and is able to promote D-loop formation. The Journal of biological chemistry. 1999;274(48):34337-42. Epub 1999/11/24. doi: 10.1074/jbc.274.48.34337. PubMed PMID: 10567410.

327. Bao L, Yuan L, Li P, Bu Q, Guo A, Zhang H, Cui N, Liu B. A FUS-LATS1/2 Axis Inhibits Hepatocellular Carcinoma Progression via Activating Hippo Pathway. Cell Physiol Biochem. 2018;50(2):437-51. Epub 2018/10/12. doi: 10.1159/000494155. PubMed PMID: 30308519.

328. Tatematsu K, Yoshimoto N, Okajima T, Tanizawa K, Kuroda S. Identification of ubiquitin ligase activity of RBCK1 and its inhibition by splice variant RBCK2 and

protein kinase Cbeta. The Journal of biological chemistry. 2008;283(17):11575-85. Epub 2008/02/28. doi: 10.1074/jbc.M706961200. PubMed PMID: 18303026.

329. Tian Y, Zhang Y, Zhong B, Wang YY, Diao FC, Wang RP, Zhang M, Chen DY, Zhai ZH, Shu HB. RBCK1 negatively regulates tumor necrosis factor- and interleukin-1-triggered NF-kappaB activation by targeting TAB2/3 for degradation. The Journal of biological chemistry. 2007;282(23):16776-82. Epub 2007/04/24. doi: 10.1074/jbc.M701913200. PubMed PMID: 17449468.

330. Lu X, Ye K, Zou K, Chen J. Identification of copy number variation-driven genes for liver cancer via bioinformatics analysis. Oncol Rep. 2014;32(5):1845-52. Epub 2014/09/02. doi: 10.3892/or.2014.3425. PubMed PMID: 25174835.

CURRICULUM VITAE

Rajneesh Srivastava

EDUCATION	YEAR
PhD, Bioinformatics (Major) and Genomics & Molecular biology (Minor) Indiana University, USA	2020
M.Sc., Applied Microbiology, Banaras Hindu University, India	2010
B.Sc. Botany, Chemistry, Industrial Microbiology, Banaras Hindu University, India	2008

WORK EXPERIENCE

Graduate Research Scholar, School of Informatics and Computing, Indiana University	2014 – 2020
J1 Research Scholar, School of Informatics and Computing, Indiana University	2012 – 2014
Project Assistant, Proteomics lab, IIT, Bombay	2010 – 2012
Research Intern, AIIMS, New Delhi, India	2009 – 2009
Summer Research Intern, IAS Bangalore, India	2009 – 2009

HONORS & ACHIEVEMENT

John R. Gibbs Scholarship, IUPUI	2015
Graduate Aptitude Test in Engineering (GATE), 98.36 percentile	2012
Summer Research Fellowship from IAS Bangalore, India	2009

PUBLICATIONS

1. Vemuri S, **Srivastava R**, Mir Q, Hashemikhabir S, Dong XC, Janga SC. SliceIt: A genome-wide resource and visualization tool to design CRISPR/Cas9 screens for editing protein-RNA interaction sites in the human genome. Methods. 2019 Sep 5;. doi: 10.1016/j.ymeth.2019.09.004. [Epub ahead of print] PubMed PMID: 31494246.
2. Koh B, Abdul Qayum A, **Srivastava R**, Fu Y, Ulrich BJ, Janga SC, Kaplan MH. A conserved enhancer regulates Il9 expression in multiple lineages. Nat Commun. 2018 Nov 15;9(1):4803. doi: 10.1038/s41467-018-07202-0. PubMed PMID: 30442929; PubMed Central PMCID: PMC6237898.
3. Budak G, Dash S, **Srivastava R**, Lachke SA, Janga SC. Express: A database of transcriptome profiles encompassing known and novel transcripts across multiple development stages in eye tissues. Exp Eye Res. 2018 Mar;168:57-68. doi: 10.1016/j.exer.2018.01.009. Epub 2018 Jan 11. PubMed PMID: 29337142; PubMed Central PMCID: PMC5826895.

4. **Srivastava R**, Budak G, Dash S, Lachke SA, Janga SC. Transcriptome analysis of developing lens reveals abundance of novel transcripts and extensive splicing alterations. Sci Rep. 2017 Sep 14;7(1):11572. doi: 10.1038/s41598-017-10615-4. PubMed PMID: 28912564; PubMed Central PMCID: PMC5599659.
5. Budak G, **Srivastava R**, Janga SC. Seten: a tool for systematic identification and comparison of processes, phenotypes, and diseases associated with RNA-binding proteins from condition-specific CLIP-seq profiles. RNA. 2017 Jun;23(6):836-846. doi: 10.1261/rna.059089.116. Epub 2017 Mar 23. PubMed PMID: 28336542; PubMed Central PMCID: PMC5435856.
6. Gollapalli K, Ghantasala S, Kumar S, **Srivastava R**, Rapole S, Moiyadi A, Epari S, Srivastava S. Subventricular zone involvement in Glioblastoma - A proteomic evaluation and clinicoradiological correlation. Sci Rep. 2017 May 3;7(1):1449. doi: 10.1038/s41598-017-01202-8. PubMed PMID: 28469129; PubMed Central PMCID: PMC5431125.
7. **Srivastava R**, Zhang Y, Xiong X, Zhang X, Pan X, Dong XC, Liangpunsakul S, Janga SC. Prediction and Validation of Transcription Factors Modulating the Expression of Sestrin3 Gene Using an Integrated Computational and Experimental Approach. PLoS One. 2016;11(7):e0160228. doi: 10.1371/journal.pone.0160228. eCollection 2016. PubMed PMID: 27466818; PubMed Central PMCID: PMC4965051.
8. **Srivastava R**, Micanovic R, El-Achkar TM, Janga SC. An intricate network of conserved DNA upstream motifs and associated transcription factors regulate the expression of uromodulin gene. J Urol. 2014 Sep;192(3):981-9. doi: 10.1016/j.juro.2014.02.095. Epub 2014 Mar 1. PubMed PMID: 24594405.
9. Rao AA, Patkari M, Reddy PJ, **Srivastava R**, Pendharkar N, Rapole S, Mehra S, Srivastava S. Proteomic analysis of Streptomyces coelicolor in response to Ciprofloxacin challenge. J Proteomics. 2014 Jan 31;97:222-34. doi: 10.1016/j.jprot.2013.08.013. Epub 2013 Aug 28. PubMed PMID: 23994098.
10. **Srivastava R**, Ray S, Vaibhav V, Gollapalli K, Jhaveri T, Taur S, Dhali S, Gogtay N, Thatte U, Srikanth R, Srivastava S. Serum profiling of leptospirosis patients to investigate proteomic alterations. J Proteomics. 2012 Dec 5;76 Spec No.:56-68. doi: 10.1016/j.jprot.2012.04.007. Epub 2012 Apr 17. PubMed PMID: 22554907.
11. Ray S, **Srivastava R**, Tripathi K, Vaibhav V, Patankar S, Srivastava S. Serum proteome changes in dengue virus-infected patients from a dengue-endemic area of India: towards new molecular targets?. OMICS. 2012 Oct;16(10):527-36. doi: 10.1089/omi.2012.0037. Epub 2012 Aug 23. PubMed PMID: 22917478; PubMed Central PMCID: PMC3459427.
12. Ray S, Renu D, **Srivastava R**, Gollapalli K, Taur S, Jhaveri T, Dhali S, Chennareddy S, Potla A, Dikshit JB, Srikanth R, Gogtay N, Thatte U, Patankar S, Srivastava S. Proteomic investigation of falciparum and vivax malaria for identification of surrogate protein markers. PLoS One. 2012;7(8):e41751. doi: 10.1371/journal.pone.0041751. Epub 2012 Aug 9. PubMed PMID: 22912677; PubMed Central PMCID: PMC3415403.
13. Gollapalli K, Ray S, **Srivastava R**, Renu D, Singh P, Dhali S, Bajpai Dikshit J, Srikanth R, Moiyadi A, Srivastava S. Investigation of serum proteome alterations

in human glioblastoma multiforme. Proteomics. 2012 Aug;12(14):2378-90. doi: 10.1002/pmic.201200002. PubMed PMID: 22684992.

14. Ray S, Kamath KS, **Srivastava R**, Raghu D, Gollapalli K, Jain R, Gupta SV, Ray S, Taur S, Dhali S, Gogtay N, Thatte U, Srikanth R, Patankar S, Srivastava S. Serum proteome analysis of vivax malaria: An insight into the disease pathogenesis and host immune response. J Proteomics. 2012 Jun 6;75(10):3063-80. doi: 10.1016/j.jprot.2011.10.018. Epub 2011 Nov 7. PubMed PMID: 22086083.