CHARACTERIZING ALTERNATIVE SPLICING AND LONG NON-CODING RNA WITH HIGH-THROUGHPUT SEQUENCING TECHNOLOGY

Ao Zhou

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

October 2018

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

Doctoral Committee

_____

Huanmei Wu, Ph.D., Chair

_____

Yunlong Liu, Ph.D.

October 2018

_____

Sarath C. Janga, Ph.D.

_____

Xiaowen Liu, Ph.D.

DEDICATION

To my parents, who grew the love of science in me

and constantly supported me from the beginning of my life.

ACKNOWLEDGEMENTS

The 10 years' odyssey of Ph.D. study was full of ups and downs. During the past decade, I worked with many people, I learned from many people, I received generous help from many people, and I shared happiness with many people.

First of all, I would like to thank Prof. Yunlong Liu, who led me in the gate of next generation sequencing technology, and continuously helped me to improve my scientific knowledge and skills during the past 8 years. Besides his selfless support in academics, his optimism and "can do" spirit inspired me during the down times, and will continue inspiring me in my future career.

I would also like to thank Prof. Jake Chen, who taught me a lot on database construction and management skills. He tremendously influenced me with his scientific way of thinking and logical presentations skills.

During the preparation and writing of this dissertation, I have received invaluable advice from Prof. Huanmei Wu, Prof. Sarath Janga and Prof. Xiaowen Liu. My appreciation also extend to Prof. Howard Edenberg, Prof. Todd Skaar, Prof. Yue Wang, Prof. Lang Li and Prof. Keith Dunker, who shared their precious data and algorithm which made this dissertation possible.

I am also enormously grateful to all those who collaborated with me and helped me during my Ph.D. study. They are Dr. Meng Li, Dr. Hai Lin, Dr. Yangyang Hao, Mr. Ed Simpson, Dr. Weilun Hsu, Dr. Fei Huang, Ms. Bo He, Dr. Fan Zhang and Dr. Xiaogang Wu.

I especially appreciate the help from Ms. Elizabeth Cassell on student affairs and the organization of my dissertation defense.

My special thanks goes to Dr. Christian Haudenschild, my supervisor and mentor at Personalis Inc., who taught me a lot about the genomic industry and generously supported me on finishing this dissertation.

The work in Chapter 2 was supported in part by the Medical and Molecular Genetics, Indiana University School of Medicine Startup Funds, Showalter Trust Award and by the Indiana Clinical and Translational Sciences Institute, funded in part by grant # TR 000006 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award.

The work in Chapter 4 was supported by the grants from the National Institutes of Health AA017941, CA113001, GM085121, and GM088076.

The work in Chapter 5 was supported by a grant from the National Cancer Institute (U24CA126480-01), which is a part of NCI's Clinical Proteomic Technologies Initiative (http://proteomics.cancer.gov).

Ao Zhou

CHARACTERIZING ALTERNATIVE SPLICING

AND LONG NON-CODING RNA

WITH HIGH-THROUGHPUT SEQUENCING TECHNOLOGY

Several experimental methods has been developed for the study of the central dogma since late 20th century. Protein mass spectrometry and next generation sequencing (including DNA-Seq and RNA-Seq) forms a triangle of experimental methods, corresponding to the three vertices of the central dogma, i.e., DNA, RNA and protein. Numerous RNA sequencing and protein mass spectrometry experiments has been carried out in attempt to understand how the expression change of known genes affect biological functions in various of organisms, however, it has been once overlooked that the result data of these experiments are in fact holograms which also reveals other delicate biological mechanisms, such as RNA splicing and the expression of long non-coding RNAs. In this dissertation, we carried out five studies based on high-throughput sequencing data, in an attempt to understand how RNA splicing and differential expression of long non-coding RNAs is associated biological functions.

In the first two studies, we identified and characterized 197 stimulant induced and 477 developmentally regulated alternative splicing events from RNA sequencing data. In the third study, we introduced a method for identifying novel alternative splicing events that were never documented. In the fourth study, we introduced a method for identifying known and novel RNA splicing junctions from protein mass spectrometry data. In the fifth study, we introduced a method for identifying long non-coding RNAs from poly-A selected RNA sequencing data. Taking advantage of these methods, we turned RNA sequencing and protein mass spectrometry data into an information gold mine of splicing and long non-coding RNA activities.

Huanmei Wu, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

dATP	Deoxyadenosine Triphosphate
dGTP	Deoxyguanosine Triphosphate
dCTP	Deoxycytidine Triphosphate
dTTP	Deoxythymidine Triphosphate
ddATP	Dideoxyadenosine Triphosphate
ddGTP	Dideoxyguanosine Triphosphate
ddCTP	Dideoxycytidine Triphosphate
ddTTP	Dideoxythymidine Triphosphate
PPi	Pyrophosphate
RT-base	Reversible Terminator Base
RNA-seq	RNA Sequencing
RPKM	Reads per Kilo-base per Million Mapped Reads
ANOVA	Analysis of Variance
MS	Mass Spectrometry
PDB	Protein Data Bank
MS/MS	Tandem Mass Spectrometry
MALDI	Matrix-Assisted Laser Desorption/Ionization
ESI	Electrospray Ionization
IPI	International Protein Index
AS	Alternative Splicing
SE	Skipped Exon
A5SS	Alternative 5’ Splice Site
A3SS	Alternative 3’ Splice Site
RI	Retained Intron
PSI or Ψ	Percentage Spliced-in

BF	Bayes Factor
MLE	Maximum-Likelihood Estimation
RISC	RNA-Induced Silencing Complexes
lncRNA	Long Non-Coding RNA
RNA pol II	RNA Polymerase II
mRNA	Messenger RNA
H3K27me3	Histone H3 Lysine 27 Trimethylation
H3K9me3	Histone H3 Lysine 9 Trimethylation
H3K4me3	Histone H3 Lysine 4 Trimethylation
H3K36me3	Histone H3 Lysine 36 Trimethylation
EMT	Epithelial to Mesenchymal Transition
IRES	Internal Ribosome Entry Site
UTR	Un-Translated Region
BMSC	Bone Marrow-Derived Mesenchymal Stem Cells
LPS	Lipopolysaccharide
TLR4	Toll-like Receptor 4
QC	Quality Control
MISO	Mixture of Isoform
NMD	Nonsense-Mediated Decay
MoRF	Molecular Recognition Feature
PTM	Post-Translational Modification
RBP	RNA Binding Protein
PPI	Protein-Protein Interaction
Ped	Pediatric
Fet	Fetal
Adu	Adult

ISE	Intronic Splicing Enhancer
ESE	Exonic Splicing Enhancer
ISS	Intronic Splicing Silencers
ESS	Exonic Splicing Silencer
BAM	Binary Sequence Alignment/Map
FDR	False Discovery Rates
EST	Expressed Sequence Tag
SNP	Single-Nucleotide Polymorphism
PEPPI	Peptideomics Database of Protein Isoforms
EXON_KB	Peptide Knowledgebase for Exonic Region
E_E_KB	Exon-Exon Junctions Knowledgebase
E_E_TH	Hypothetical Exon-Exon Junctions
E_I_TH	Hypothetical Exon-Intron Junctions
I_E_TH	Hypothetical Intron-Exon Junctions
ORF	Open Reading Frame
ENCODE	Encyclopedia of DNA Elements Consortium
ncRNA	non-protein-coding RNA
indel	Insertion/Deletion Variation

Chapter 1. Introduction to Sequencing and Data Analysis Methods

The central dogma consists of three major levels, genome, transcriptome and proteome. The advance of high-throughput sequencing technologies for nucleic acid and proteins has provided affordable measures for investigating biological processes, mechanisms and functions at all three major levels of the central dogma. In Chapter 1, we review the current nucleic acid and protein sequencing technologies, as well as the biological background and current data analysis methods of the two major topics we investigated in this dissertation, alternative splicing and long non-coding RNA.

1.1 Next Generation Nucleic Acid Sequencing Technology

1.1.1 Brief History of Nucleic Acid Discovery

Chromosome was the first observed before human realize DNA was its main component. The word chromosome originates from the Greek roots chroma, meaning color, and soma, meaning body. In the 19th century, Schleiden [1], Virchow [2] and Bütschli [3] recognized the structures now known as chromosome. Walther Flemming named this structure “chromatin” [4]. In 1878, he published his discovery on how “chromatin” separate during cell division, also known as mitosis. In 1888, von Waldeyer-Hartz coined the name “chromosome”, describing their strong staining by particular dyes [5]. Aided by Mendel’s earlier work rediscovered in the early 1900s, Boveri pointed out the association between heredity and behavior of chromosomes [6]. The number of human chromosomes, 46, was determined by Joe Hin Tjio in 1956 [7].

In 1869, almost the same time when chromosome was recognized, Friedrich Miescher discovered a microscopic substance he named “nuclein”, which is now known as DNA [8]. In 1878, Albrecht Kossel isolated the pure form of “nuclein”, nucleic acid, and later isolated its five primary nucleobases [9]. In 1919, Phobus Levene identified the base, sugar and phosphate nucleotide units [10]. In 1927, Nikolai Koltssov proposed that the inheritance of traits could be carried via a “giant hereditary molecule” made up with “two

mirror strands that replicate in a semi-conservative fashion using each strand as a template" [11]. In 1953, Watson and Crick suggested the now accepted double-helix model of DNA structure based on X-ray diffraction images [12].

When first studied in the early 1900s, the biological differences between RNA and DNA were not well understood. The role of RNA in protein synthesis was suspected as late as 1939 [13]. The concept of messenger RNA emerged during the late 1950s, and was associated with Crick's description of his "Central Dogma of Molecular Biology", which asserted that genetic information flows from DNA to RNA, and thus led to the synthesis of proteins [14]. During the 1970s, retroviruses and reverse transcriptase were discovered, which later enables RNAs to be sequenced by DNA sequencing technology [15, 16]. In 1977, introns and RNA splicing were discovered by Philip Sharp and Richard Roberts.[17, 18]

1.1.1 Basic DNA Sequencing Methods

Two DNA sequencing technologies were developed around 1977, one is Maxam-Gilbert sequencing, developed by Allan Maxam and Walter Gilbert [19], the other is Sanger sequencing, developed by Frederick Sanger [20, 21].

Maxam-Gilbert sequencing requires radioactive ^{32}P labelling of the 5' end of the purified DNA sequence. Four different chemical treatments are applied to DNA to create breakages on DNA molecules at one or two specific bases (G, A+G, C, C+T). The concentration of chemicals are controlled to induce on average one breakage per DNA molecule. The DNA fragments are then electrophoresed and the resulting gel is exposed to X-ray film for autoradiography. Then the DNA sequence is inferred from presence and absence of certain DNA fragments.

Maxam-Gilbert sequencing was once widely used right after its invention because of certain advantages: 1) PCR amplification on DNA is not required, 2) high accuracy in inferring homopolimetric DNA sequences, and 3) can be used to analyze DNA-protein

interactions and epigenetic modifications. However, it is no longer commonly used today because of certain disadvantages: 1) extensive usage of hazardous chemicals, 2) the complexity in the experiment set-up, and 3) not being capable to analyze sequences with more than 500 base pairs.

Sanger sequencing, or the chain-termination method, involves DNA polymerization with for standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and four radioactively or fluorescently labelled dideoxynucleotides (ddATP, ddGTP, ddCTP and ddTTP). The existence of dideoxynucleotides stops the polymerization of the DNA templates at a random loci, thus creating a library of DNA sequences of different lengths. If the experiment is designed properly, the library should cover each base on the DNA template to be sequenced. The concentration of dideoxynucleotides is low enough to make sure some polymerized DNA fragments covers the whole length of the DNA template to be sequenced. Then the DNA library is electrophoresed and the nucleotide on each loci can be inferred from the bands of the electrophoresis.

Sanger sequencing and Maxam-Gilbert sequencing share similar ideas in their design. Both methods process the DNA molecule and derive shortened fragments of all possible lengths at chosen nucleic acids, then infer the original DNA sequence with electrophoresis. Sanger sequencing was first automated and commercialized by Applied Biosystems. It was widely accepted and prevailed from the 1980s to mid-2000s for its advantages, including its relative ease and reliability, and longer read length (800bp). The Sanger sequencing technology led to the first human genome in 2001. From the late 2000s, high-throughput DNA sequencing methods reached the market and brought the cost per genome from $100 million in 2001 to $100 in 2017.

1.1.2 High-throughput DNA Sequencing Methods

Several high-throughput DNA sequencing methods has been developed in the 21st century. Among these methods, 454 pyrosequencing, Illumina (Solexa) sequencing, ion

torrent semiconductor sequencing and SOLiD sequencing have been widely used in academic and industrial scenarios. The sequencing data analyzed in this dissertation is derived with either the Illumina or the SOLiD technology, which are the most popular sequencing technologies in the late 2000's and early 2010's.

Owned by 454 Life Sciences and later by Roche, the 454 technology is the first widely used high-throughput sequencing technology and was once popular in the 2000s. It was known for its long read length (400-500 bases), but its sequencing depth is lower (400-600M reads per run) than SOLiD and Illumina technologies. The production of 454 sequencers has been shut down by Roche in 2013. The 454 technology separates DNA templates to be sequenced in detached droplets, and adds solutions of A, T, C and G sequentially for synthesis. When the nucleic acid matches the template and the synthesis happens, a pyrophosphate (PPi) will be released. Then the PPi molecules go through a series reaction and finally reacts with the luciferin in the solution and generates visible light in amounts that are proportional to the amount of PPi. After the light signal is captured, the nucleotides and other byproducts are washed away and the reaction restarts with another nucleotide. The DNA sequences can be interpreted from a stack of images captured in each reaction. Similar to Sanger sequencing, the 454 technology also involved DNA synthesis, however, the 454 design allows it to sequence many sequences at once, and the avoidance of electrophoresis potentiates a much faster sequencing speed.

By the year of 2017, the Illumina sequencing technology is the most popular large scale sequencing method among the academia and industry. Currently the Illumina sequencing technology can achieve up to 8-10B reads per run (NovaSeq 6000) with 2x150 bases paired-end reads. Illumina also provides a lightweight benchtop sequencing solution MiSeq, which is capable of achieving 25M reads per run with 2x300 bases each read. In this method, the DNA molecules are first attached to a glass slide (flowcell) and amplified to form clusters. To identify the DNA sequence, four types of reversible terminator bases

(RT-bases) are added once a time. An RT-base is an engineered nucleic acid molecule that can be synthesized to a DNA molecule, but inhibits subsequent synthesis. The RT-base also carries a fluorescent unit that allows detection of synthesis activity by camera. After the camera has recorded the location of clusters with the synthesis activity, the fluorescent unit is chemically removed and a new base is added to the flowcell. After all four bases are added to the flowcell, the synthesis inhibition unit is removed to allow a new cycle of synthesis. The design of RT-base granted two advantages to the Illumina sequencing technology comparing to the 454 technology, 1) the DNA chain is extended one base at a time, which achieves higher accuracy on repetitive DNA sequences; 2) the light signal unit is attached to the DNA instead of released in the solution, which eliminated the need of droplets and simplified the experiment design.

Owned by the Thermo Fisher Scientific, Ion Torrent is another popular lightweight sequencing solution that is comparable to MiSeq in throughput in 2017. It is capable to sequence 9-12M reads with 600 bases per run. Ion Torrent is also a sequencing by synthesis method. The DNA sequence and DNA polymerase are flooded by A, T, C and G dNTP sequentially. When the dNTP matches the DNA sequence, DNA synthesis takes place and a hydrogen ion is released. The hydrogen ion is then detected by a semiconductor sensor and the DNA sequence can be interpreted from the synthesis signals. The Ion Torrent technology suffers from a limitation. If the DNA sequence contains a long repetition of a certain base, then multiple hydrogen ions would be released in the corresponding cycle. Theoretically this would result in a proportionally higher electronic signal, however, the strength difference between two repetitions with different lengths are not easy to determine in practice. Similar to the 454 pyrosequencing technology, this characteristic has limited the accuracy of Ion Torrent technology on repetitive sequences.

Owned by Life Technologies, SOLiD was a popular sequencing technology in the late 2000's. It used to be the only sequencing method on the market that was comparable

against the Illumina technology. However, due to concerns in its cost and accuracy, SOLiD was discontinued from January 2013. Life Technologies was then acquired by Thermo Fisher in January 2014. Different from 454, Illumina and Ion Torrent technologies, SOLiD sequencing is based on DNA ligation instead of synthesis. SOLiD sequencing first ligates adapters with known sequences to the DNA molecules to be sequenced. Then it attaches these DNA molecules to magnetic beads and a PCR process makes clones of only one DNA sequence occupy the surface of each bead. The beads are then covalently bound to a glass slide. The SOLiD method involves 16 di-base probes, which consists of two bases for mapping the DNA molecule to be sequenced, three universal bases for enhanced affinity, and three universal bases with fluorescent dye. The 16 di-base probes are divided in four groups, each group includes 4 probes labeled with a distinct color. Four groups of di-base probes are added to the flowcell in sequence for ligation reactions. Due to the design of the three universal bases in the di-base probes, a series of continuous ligation reactions can only cover two in every five bases. To cover all bases in the DNA molecule, five series of ligation reactions with 1 base offset in each series are implemented. The first base of each series of reactions is known because the reaction initiates on the primer. The whole DNA sequence can be interpreted from the first base and the colored light signals captured during the ligation reactions.

Other nucleic acid sequencing technologies also present in the same era, such as the SMRT sequencing by Pacific Biosciences [22], the Nanoball sequencing by Complete Genomics [23], the HeliScope single molecule sequencing by Helicos [24] and the Nanopore DNA sequencing by Oxford Nanopore [25]. However these technologies are not as widely applied as previous four because of disadvantages in cost, accuracy or technology maturity. The RNA sequencing (RNA-seq) data analyzed in this dissertation are derived from either the SOLiD or the Illumina sequencing technology, which are two of the most accurate and cost efficient technologies around 2010.

1.1.3 Sequencing Reads Alignment

The sequences of the cDNA fragments detected in the sequencing experiment are called reads. The sequencer only produces read sequences in its output files, however the genomic locations of these reads need to be find out through a computational process called alignment. In the alignment step, the reads are usually compared against the corresponding UCSC genome [26].

Suppose the length of the genome is n, then we actually face such a problem in short read alignment: to implement a substring search and tolerate mismatches to a certain level in approximately O(1) time and O(n) memory space. Several short read aligners have been developed to address this problem, including BFAST [27], Bowtie [28] and BWA [29]. All these tools employed the Smith-Waterman algorithm for final alignment and scoring. Bowtie and BWA employed a suffix tree based on the Burrows-Wheeler transform (or FM-index) to speed up the substring match. BFAST also employed a suffix array for genome indexing but as genome positions sorted by suffix sequence.

Short read aligners are only able to implement non-spliced alignment, which means the RNA-seq reads crossing splicing junctions between exons will get a low mapping score because of the huge gap in the intron and will not be considered as a match if aligned directly by BFAST, Bowtie or BWA. To take RNA splicing in consideration, RNA aligner tools need to be used for RNA-seq alignment. Such tools includes Tophat [30] and STAR [31].

1.1.4 Gene Expression Analysis Methods

After the RNA-seq reads are mapped to the genome, the expression intensity of each gene can be inferred from the number of reads mapped. There are two primary factors affecting this number, the length of the gene and the overall sequencing depth of this sample. To normalize these factors, the unit reads per kilo-base per million mapped reads (RPKM) is used to evaluate gene expression intensity. The “per kilo-base” factor

normalizes the length of the gene, it could be either the total length of the gene or the exonic length. The “per million mapped reads” factor normalizes the overall sequencing depth of the sample.

To test whether a gene is differentially expressed across two biological conditions, analysis of variance (ANOVA) or Friedman-test can be used to test RPKM. For these tests, one factor is the grouping based on the biological conditions. Another factor, if applicable, can be batch effect. Batch effect is the random effects during a sequencing experiment, such as time, light, or temperature that may affect the RPKM values. If the same sample is processed twice in two experiments, its RPKM may differ because of batch effects. Batch effects can be removed by adding a batch factor when implementing ANOVA or Friedman-test. ANOVA uses the actual RPKM values for testing, while Friedman-test uses the rank of the RPKM values for testing.

Another model for gene expression analysis is edgeR [32]. Different from ANOVA and Friedman-test, which assumes the gene expression level follows normal distribution or chi-square distribution, edgeR models it with negative binomial distribution. In ideal condition, the natural distribution of the counts of reads falling on a gene should follow a Poisson distribution. However, more variation is observed in practice, which is known as overdispersion, which is induced by measurement errors. Therefore edgeR utilized the negative binomial distribution to model overdispersion. When the overdispersion is 0, the negative binomial distribution will become Poisson distribution. Instead of calculating the RPKM first, edgeR implements gene expression tests with counts of reads directly.

1.2 Protein Mass Spectrometry Technology

1.2.1 Brief History of Protein Science

In 1838, the Dutch chemist Gerardus Johannes Mulder carried out elemental analysis on common proteins and erroneously concluded that proteins might be composed by a single type of large molecule [33]. Mulder’s collaborator, the Swedish chemist Jöns

Jacob Berzelius coined the term “protein”, which is derived from the Greek proteios, which means “primary”, and the suffix -in [34]. Mulder identified the first known amino acid leucine. For a long time, the exact composition of protein is not well known, until the 20th amino acid, threonine, was discovered in 1936. In 1949, the first protein sequencing was successfully implemented on insulin by Frederik Sanger, thus demonstrating that proteins are linear polymers of amino acids [35]. Linus Pauling suggested two main types of protein secondary structure, the α-helix and the β-strand (or β-sheet) in 1951 [36]. In the 1980s, mass spectrometry (MS) was widely applied in high-throughput sequencing and identification of proteins. In November 2017, the UniProt database has collected 556,196 protein entries. At the same time, the protein data bank (PDB) [37] has collected 135,359 protein tertiary structures, including 121,176 derived from X-ray crystallography, 12,032 derived from NMR, and 1,817 derived from electron microscopy.

1.2.2 Protein Mass Spectrometry

A typical mass spectrometry experiment for protein identification involves the following stages: protein sorting, digestion, ionization, and tandem mass spectrometry (MS/MS). Protein sorting can be done by 2 dimensional electrophoresis, in which the first dimension separates proteins by their isoelectric points, and the second dimension separates proteins by their molecular weights. The protein molecules in each band have similar isoelectric points and molecular weights. Then these protein molecules are extracted and digested by a protease such as trypsin. The digested peptides are then ionized by either matrix-assisted laser desorption/ionization (MALDI) or electrospray ionization (ESI). The first stage of MS/MS sorts the peptide ions by their mass-to-charge ratio, and these peptide ions are further fragmented by collision-induced dissociation, photo-dissociation or other processes. The fragment ions then enter the second stage and the final mass spectrum was derived. The mass spectrum was then compared against known protein databases to identify the protein molecules in the sample. Before 2011, the

international protein index (IPI) database was widely used for such purpose, and later the UniProtKB database took its place after IPI's retirement. The mass spectrometry analyses in this dissertation were finished in 2009 and were based on MS/MS and IPI.

1.3 Alternative Splicing

1.3.1 Discovery of Alternative Splicing

During messenger RNA transcription, the RNA polymerase II travels through the gene region of the antisense strand of DNA and produces pre-mRNA, which is an identical copy of the sense strand of DNA. Then the spliceosome cuts out the intron regions on the pre-RNA and splices exons together, this process is called RNA splicing. In 1977, Phillip Sharp, Richard Roberts and Louise T. Chow demonstrated the existence of RNA splicing by visualizing the loops of introns under electron microscope when hybriding a cDNA sequence to its mRNA. The GT-AG sequence pattern on intron boundaries was discovered in 1978 by Breathnach [38]. When investigating the biological mechanism that enables RNA splicing, researchers soon realized that splicing can take place with "alternative splicing pathways" [39].

There are four basic modes of alternative splicing (AS). 1) Skipped exon (SE), or exon skipping, or cassette exon, where an exon may be left out during splicing. This is the most common type of AS in mammals. 2) Alternative 5' splice site (A5SS), where an alternative 3' end of the upstream exon (also the 5' donor site) was used. 3) Alternative 3' splice site (A3SS), where an alternative 5' end of the downstream exon (also the 3' acceptor site) was used. 4) Retained intron (RI) or intron retention, where the whole intron is retained in the final mRNA transcript. (Figure 1.1).

1.3.2 Alternative Splicing Identification Methods

Once the FASTQ or color spaced data are derived from sequencers, sequencing alignment will be implemented with RNA aligners, including Tophat and STAR. The intron regions appear as long deletions on the RNA-seq reads crossing two exons (or splice

junction reads), and such reads will get discarded by DNA aligners because of low scoring. RNA aligners are able to process these splice junction reads and achieve correct alignment. The alignments and scores of the RNA-seq reads are saved in BAM files.

Based on BAM files, two approaches may be used to investigate AS. The first approach is whole mRNA transcript reconstruction. Software tools following this approach include Cufflinks [40] and Scripture [41]. Both of these tools construct transcripts with graph-based methods. Cufflinks builds a connectivity graph with each paired-end read as a node, and then attempts to search a minimal set of paths that covers all reads. Then these paths are scored and prioritized based on the abundance of reads they cover. On the other hand, Scripture first ignores the paired-end information and constructs the exons just based on overlapping sequence, then builds the connectivity graph based on exons and junction reads. These graphs are scored and filtered by read coverage, and then joined with paired-end data. In practice, Cufflinks has better specificity and Scripture has better sensitivity.

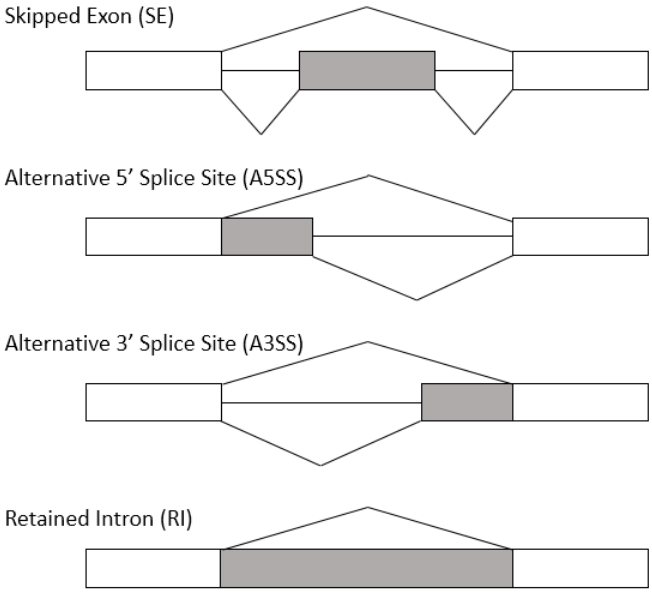

Figure 1.1 Four types of alternative splicing events

Both Cufflinks and Scripture refines their connectivity paths based on mathematical assumptions, which may not be real in biology. Therefore the mRNA transcripts reconstructed from the RNA-seq data may not reveal the real composition of the transcriptome. This is a limitation of the short read RNA-seq and the best solution would be shifting to a long read RNA-seq method, such as the SMRT technology by Pacific Biosciences [22]. However, those sequencing methods were either expensive or not available in the first decade of 21th century. As a workaround, the approach for AS investigation was developed. Instead of attempting to reconstruct the transcriptome, this approach focuses on quantifying the ratio of mRNA transcripts supporting the inclusive and exclusive isoform of a single AS event. Such tools include MISO [42] and rMATS [43]. In this approach, an AS event is modeled as a pair of isoform transcripts, the one with flanking exons and the alternative region is defined as the inclusive isoform, and the one with flanking exons only is defined as the exclusive isoform. Figure 1.1 shows four types of AS events, including SE, A5SS, A3SS and RI. The boxes denotes regions that may be transcribed in the mRNA. White boxes denote constitutive regions that will present in all transcripts. Gray boxes denote alternative regions that only present in some transcripts. The straight line in the middle denotes introns, the curve on the top denotes exon junctions in the exclusive isoform, and the curves at the bottom denote junctions in the inclusive isoform. Both MISO and MATS calculates a percentage spliced-in (PSI or Ψ) value to quantify AS events. The Ψ denotes the percentage of the transcripts including the alternative region (cassette exon, retained intron, etc). Both MISO and MATS are capable to implement sample-wise or group-wise comparison for AS changes.

1.3.3 Algorithms of MISO

MISO [42] is one of the most frequently used AS quantification software tools in this dissertation. Here we introduce MISO's algorithm in Ψ calculation and sample-wise

comparison. The following equation describes the most straightforward way of estimating Ψ:

$$\widehat{\Psi} = \frac{D_I}{D_I + D_E} \quad [1.1]$$

D_I is the density of inclusive reads, D_E is the density of exclusive reads. D_I and D_E can be derived by the following equations:

$$D_I = \frac{N_{Inc}}{a - r + 1 + 2(r + 1 - 2o)} \quad [1.2]$$

$$D_E = \frac{N_{Exc}}{r + 1 - 2o} \quad [1.3]$$

N_{Inc} and N_{Exc} are the number of reads supporting inclusion and exclusion isoforms, respectively. Let *a* be the length of the alternative region, *r* be the length of the RNA-seq reads, and *o* the overhang constraint placed on splice junctions. These two equations normalizes the number of reads mapped to RNA isoforms (the numerator) with the number of the possible positions on each isoform (the denominator).

However this is not the real approach that MISO used to deduce the Ψ value. Beside the inclusive and exclusive reads (or informative reads), MISO also utilized the constitutive reads to improve and stabilize Ψ estimates. The idea is to formulate the probability distribution of Ψ (*a posteriori*) values given the reads detected from mRNA-seq (*a priori*). It can be described by the following equation:

$$P(R_{1:N}|\Psi) = \prod_{n=1}^{N}[P(R_n|Inc)\Psi + \mathrm{P}(R_n|Exc)(1 - \Psi)] \quad [1.4]$$

In this equation, $P(R_n|Inc)$ and $P(R_n|Exc)$ denote the probability of read *n* supporting the inclusive and exclusive transcript, respectively. Their values are determined by two factors, the possible positions of a read can be mapped to a transcript, which is denoted by *p* here, and whether a read is mapped to a transcript, which is denoted by *R*. Therefore we have the following equations:

$$P(R_n|Inc) = p_{Inc}R_n^{Inc} = \frac{R_n^{Inc}}{a - r + 1 + 2(r + 1 - 2o)} \quad [1.5]$$

$$P(R_n|Exc) = p_{Exc}R_n^{Exc} = \frac{R_n^{Exc}}{r + 1 - 2o} \quad [1.6]$$

For one AS event, both p_{Inc} and p_{Exc} are constants, we may simply use this notation for readability. R_n^{Inc} and R_n^{Exc} denote whether the read *n* is mapped to the corresponding transcript isoform. If the read is uniquely mapped then the value is 1, if not then 0. If the read is mapped to the constitutive region then both R_n^{Inc} and R_n^{Exc} equal to 0.5. Thus we have the updated expression of the probability of the given set of reads and Ψ:

$$P(R_{1:N}|\Psi) = \prod_{n=1}^{N}[p_{Inc}R_n^{Inc}\Psi + p_{Exc}R_n^{Exc}(1 - \Psi)] \quad [1.7]$$

This equation describes the probability distribution of Ψ with a given set of reads. The mapping probability (*p*) and the mapping result (*R*) are both fixed, the Ψ value that achieves the maximum $P(R_{1:N}|\Psi)$ will become the most likely Ψ. Here is the expression after taking the log:

$$\hat{\Psi} = \arg\max_{\Psi} \sum_{n=1}^{N} \log[p_{Inc}R_n^{Inc}\Psi + p_{Exc}R_n^{Exc}(1 - \Psi)] \quad [1.8]$$

After taking the derivative, we have the final equation for getting most likely Ψ:

$$\frac{d}{d\Psi}\sum_{n=1}^{N} \log[p_{Inc}R_n^{Inc}\Psi + p_{Exc}R_n^{Exc}(1 - \Psi)] = 0$$
$$\sum_{n=1}^{N} \frac{p_{Inc}R_n^{Inc} - p_{Exc}R_n^{Exc}}{p_{Inc}R_n^{Inc}\Psi + p_{Exc}R_n^{Exc}(1 - \Psi)} = 0 \quad [1.9]$$

Assuming the estimated Ψ value has been evaluated for biological condition A and B, now it is important to compare whether Ψ_A is significantly different from Ψ_B. MISO calculates a Bayes factor (BF) to evaluate the likelihood of differential splicing.

Let δ be the difference between Ψ_A and Ψ_B, then the null hypothesis (H_0) is δ=0, and the alternative hypothesis is δ≠0. The BF is defined as the weight of the evidence in the data D in support of H_1 over H_0:

$$BF = \frac{P(D, H_1)}{P(D, H_0)} = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)} \quad [1.10]$$

The authors of MISO uses Savage-Dickey density ratio [44] to approximately estimate BF:

$$BF \approx \frac{P(\delta = 0|H_1)}{P(\delta = 0|D, H_1)} \quad [1.11]$$

When H1 stands, the probability of δ=0 is 1, so the equation can be simplified as:

$$BF \approx \frac{1}{P(\delta = 0|D, H_1)} \quad [1.12]$$

Associating to the equation of Ψ before, the denominator of the equation above can be stated as:

$$P(\delta = 0|D, H_1) = \int_0^1 P_A(R_{1:M}|\Psi)P_B(R_{1:N}|\Psi) \quad [1.13]$$

MISO's BF approach for differential AS testing requires merging the BAM files of each group. This procedure has ignored the within-group variation and the result may be skewed by outliers.

1.3.4 Algorithms of rMATS

To put the within-group variation in consideration, the rMATS [43] package has been developed in 2014. Its algorithm for Ψ value estimation uses only reads that can be uniquely mapped to inclusive or exclusive isoforms. The equation for calculating Ψ in rMATS is the same as equation [1.1].

In the rMATS model, the observed Ψ of a given AS event is affected by two types of variations, including 1) the biological variation of the replicate Ψ within the sample group, which is modeled by a normal distribution, and 2) the variation of observed reads given a Ψ value, which is modeled by a binomial distribution. The first variation is represented in the following model:

$$\log \Psi_{gk} \sim \text{Normal}(\mu = \log \Psi_g, \sigma^2 = \sigma_g^2) \quad [1.14]$$

In model [1.14], g represents group number, and k represents the sample number within the corresponding group. The second variation is represented in model [1.15]:

$$I_{gk}|\Psi_{gk} \sim \text{Binomial}(n_{gk} = N_{Inc,k} + N_{Exc,k}, p_{gk} = \frac{l_{Inc}\Psi_g}{l_{Inc}\Psi_g + l_{Exc}(1-\Psi_g)}) \quad [1.15]$$

In this model, $N_{Inc,k}$ and $N_{Exc,k}$ represent the numbers of reads mapped to the inclusive and exclusive isoform, respectively. The effective lengths of the inclusive and exclusive isoform are represented with l_{Inc} and l_{Exc}.

Based on the variation models above, it is possible to estimate the likelihood of getting the observed read counts from sample group A and B bases on a given Ψ_A and Ψ_B:

$$\begin{aligned} L &= L_1 L_2 \\ L_1 &= \prod_{k=1}^{M_A} P(I_{Ak}|\Psi_{Ak}, n_{Ak}) \prod_{k=1}^{M_B} P(I_{Bk}|\Psi_{Bk}, n_{Bk}) \\ L_2 &= \prod_{k=1}^{M_A} P(\Psi_{Ak}|\Psi_A, \sigma_A) \prod_{k=1}^{M_B} P(\Psi_{Bk}|\Psi_B, \sigma_B) \end{aligned} \quad [1.16]$$

$P(I_{gk}|\Psi_{gk}, n_{gk})$ can be derived from the binomial distribution:

$$\begin{aligned} \prod_{k=1}^{M_g} P(I_{gk}|\Psi_{gk}, n_{gk}) = \prod_{k=1}^{M_g} \binom{N_{Inc} + N_{Exc}}{N_{Inc}} \\ \times \exp\left(\sum_{k=1}^{M_g} N_{gk,Inc} \log \frac{l_{Inc}\Psi_{gk}}{l_{Inc}\Psi_{gk} + l_{Exc}(1-\Psi_{gk})}\right. \\ \left. + N_{gk,Exc} \log \frac{l_{Exc}(1-\Psi_{gk})}{l_{Inc}\Psi_{gk} + l_{Exc}(1-\Psi_{gk})}\right) \end{aligned} \quad [1.17]$$

$P(\Psi_{gk}|\Psi_g, \sigma_g)$ can be derived from the normal distribution:

$$\prod_{k=1}^{M_g} P(\Psi_{gk}|\Psi_g, \sigma_g) = \quad [1.18]$$

$$\exp\left(\sum_{k=1}^{M_g}\frac{-\log 2\pi}{2}-\log\sigma_g-\frac{\left(\log\Psi_{gk}-\log\Psi_g\right)^2}{2\sigma_g^2}+\log(\Psi_{gk}(1-\Psi_{gk}))\right)$$

Since the goal is to test the difference of mean inclusion level between two sample groups (i.e., Ψ_1-Ψ_2), the Ψ value of each sample Ψ_{1k} and Ψ_{2k} can be treated as latent variables, and the marginal distribution can be transformed with integral of Ψ_{1k} and Ψ_{2k}:

$$f(\Psi_1,\sigma_1,\Psi_2,\sigma_2) = c\left(\prod_{k=1}^{M_1}\int f(\Psi_1,\sigma_1,\Psi_{1k})d\Psi_{1k}\prod_{k=1}^{M_2}\int f(\Psi_2,\sigma_2,\Psi_{2k})d\Psi_{2k}\right) \quad [1.19]$$

In equation [1.19], c is a constant, and f(Ψ_1, σ_1, Ψ_{1k}) can be defined by the following equation:

$$\begin{aligned} f(\Psi_g,\sigma_g,\Psi_{gk}) = \exp\Bigg(&-\log\sigma_g-\frac{\left(\log\Psi_{gk}-\log\Psi_g\right)^2}{2\sigma_g^2}\\ &+\log\left(\Psi_{gk}(1-\Psi_{gk})\right)\\ &+N_{gk,Inc}\log\frac{l_{Inc}\Psi_{gk}}{l_{Inc}\Psi_{gk}+l_{Exc}(1-\Psi_{gk})}\\ &+N_{gk,Exc}\log\frac{l_{Exc}(1-\Psi_{gk})}{l_{Inc}\Psi_{gk}+l_{Exc}(1-\Psi_{gk})}\Bigg) \end{aligned} \quad [1.20]$$

After Laplace approximation, we can derive the following equation:

$$\begin{aligned} \int f(\Psi_g,\sigma_g,\Psi_{gk})d\Psi_{gk} \approx{}& \sqrt{2\pi}\left(\left|\frac{\partial^2 f_1(\Psi_g,\sigma_g,\widehat{\Psi}_{gk})}{\partial\Psi_{gk}^2}\right|\right)^{-0.5}\\ &\times\exp(f_1(\Psi_g,\sigma_g,\widehat{\Psi}_{gk})\\ \frac{\partial^2 f_1(\Psi_g,\sigma_g,\widehat{\Psi}_{gk})}{\partial\Psi_{gk}^2} &\\ =\frac{2\widehat{\Psi}_{gk}-1}{\widehat{\Psi}_{gk}^2(1-\widehat{\Psi}_{gk})^2}&\left(\frac{\log\Psi_g-\log\widehat{\Psi}_{gk}-(2\widehat{\Psi}_{gk}-1)^{-1}}{\sigma_g^2}+1\right) \end{aligned} \quad [1.21]$$

$$-N_{gk,Inc}l_{Exc}\frac{(2l_{Inc}+l_{Exc})\widehat{\Psi}_{gk}+l_{Exc}(1-\widehat{\Psi}_{gk})}{\widehat{\Psi}_{gk}^2\left(l_{Inc}\widehat{\Psi}_{gk}+l_{Exc}(1-\widehat{\Psi}_{gk})\right)^2}$$

$$-N_{gk,Exc}l_{Inc}\frac{(l_{Inc}+2l_{Exc})(1-\widehat{\Psi}_{gk})+l_{Inc}\widehat{\Psi}_{gk}}{(1-\widehat{\Psi}_{gk})^2(l_{Inc}\widehat{\Psi}_{gk}+l_{Exc}(1-\widehat{\Psi}_{gk}))^2}$$

The Laplace's method approximates the distribution of Ψ_{gk} by normal distribution. The estimated values of Ψ_{gk} can be derived in optimization procedure for maximum-likelihood estimation (MLE):

$$\widehat{\Psi}_{gk}=\arg\max_{\Psi_{gk}}\left(\frac{-0.5\left(\log\Psi_{gk}-\log\widehat{\Psi}_g\right)^2}{\sigma_g^2}+\log\Psi_{gk}\right.$$

$$+\log(1-\Psi_{gk})+N_{gk,Inc}\log\left(\frac{l_{Inc}\Psi_{gk}}{l_{Inc}\Psi_{gk}+l_{Exc}(1-\Psi_{gk})}\right) \qquad [1.22]$$

$$\left.+N_{gk,Exc}\log\left(\frac{l_{Exc}(1-\Psi_{gk})}{l_{Inc}\Psi_{gk}+l_{Exc}(1-\Psi_{gk})}\right)\right)$$

Based on the marginal distribution described above, P value of the splicing difference across two groups of samples can be calculated. For each AS event, the null hypothesis is the difference of Ψ across two groups is smaller than or equal to a user defined cutoff, and the alternative hypothesis is the difference of Ψ is larger than the cutoff.

1.4 Long Non-coding RNA

A major portion of eukaryotic genome is covered by DNA sequences that do not code for proteins, however, it is observed that many of these regions are transcribed and produces a huge species of non-coding RNAs (ncRNAs) [45, 46]. Some of ncRNAs are constructively expressed in all cells, such as ribosomal RNA, transfer RNA, and small nuclear and nucleolar RNA, and are known as housekeeping ncRNAs. Besides housekeeping ncRNAs, other ncRNAs can be categorized into small non-coding RNAs, which are 100 nucleotides and shorter, and long non-coding RNAs, which are longer than 200 nucleotides.

Small non-coding RNAs can be subdivided into subgroups (miRNA, siRNA, piRNA, etc.) based on their size, biogenesis and mode of action. It is known that small ncRNAs regulate gene expression by guiding repressive chromatin complexes and RNA-dependent RNA polymerase complexes, which regulates transcription silencing and starting respectively at the transcriptional level, and by guiding RNA-induced silencing complexes (RISCs) to cleave the target mRNA at the post-transcriptional level [47-51].

Comparing with small ncRNAs, long non-coding RNAs (lncRNAs) are less characterized, and their biological functions are poorly investigated. Most of lncRNAs are transcribed by RNA Polymerase II (RNA pol II), and possess a 5' methyl cap and polyA tail, which is the same as messenger RNAs (mRNA). LncRNAs may be transcribed from four possible locations: 1) introns of another gene, 2) intergenic regions, 3) sense strand of a gene or 4) antisense strand of a gene. LncRNAs are less conservative in sequence than protein-coding mRNAs, but still shows positive selection over neutral sequences [52]. LncRNAs may act as transcription silencers via transcriptional interference. A very well-known example is the X inactivation specific transcript (Xist), which deactivates whole X chromosome in female mammals by spreading across the X chromosome and recruiting polycomb repressive complex 2 (PRC2), which hence transfers repressive histone markers (H3K27me3 and H3K9me3) to the chromosome [53]. Besides whole chromosome repression, lncRNA also participates in individual gene repressions. In yeast *S.cerevisiae*, the transcription of a non-coding gene SRG1, which is located upstream of the promoter of another gene, SER3, suppresses the transcription of SER3 [54, 55]. There are two models that may explain lncRNA's repressive effect on gene expression. The first model is the transcription activity causes the occlusion of the transcription machinery. The second model is the lncRNA itself may bind to the transcriptional complex and act as an inhibitor. In some cases, lncRNA may act as transcription activators, if its transcription occludes transcription inhibitors of another gene. For example, the transcription of lncRNA

LINoCR occludes the chromatin insulator protein CTCF, a gene expression repressor, and activates the chicken Lysozyme gene [56]. LncRNA may also regulate mRNA splicing. During epithelial to mesenchymal transition (EMT), the translation of Zeb2 mRNA is prevented in epithelial cells by a splicing event which removes the internal ribosome entry site (IRES) containing the 5' UTR. On the other hand, in mesenchymal cells, an antisense lncRNA binds to the mRNA and prevents this splicing event, and hence allowing the translation of Zeb2 mRNA [57].

Currently lncRNAs can be identified at a genome-wide scale by active chromatin signatures associated with RNA pol II transcription. Such chromatin signatures include histone H3 lysine 4 trimethylation and histone H3 lysine 36 trimethylation domains (K4-K36 domain) [58, 59].

1.5 Organization of the Dissertation

This dissertation identified and quantified AS events and novel lncRNA in several organisms and cell lines with help of high-throughput sequencing technologies. In Chapter 2, we investigated stimulant induced AS with lipopolysaccharide stimulated bone marrow-derived mesenchymal stem cells (BMSC) [60]. In Chapter 3, we investigated developmentally regulated AS within human liver across three developmental stages (fetal, pediatric and adult). In Chapter 4, we discussed a method for identifying novel AS events from any mRNA-seq experiment [61]. In Chapter 5, we built a peptidomic database for identifying novel AS events from MS/MS data [62]. In Chapter 6, we identified and characterized lncRNAs associated with alcohol dependence from mRNA-seq data [63].

Chapter 2. Stimulant Induced Alternative Splicing

2.1 Background

Alternative splicing (AS) is important for gene regulation and is a major source of proteome diversity in mammals [64] through altering the composition of mRNA transcripts by including or excluding specific exons [65]. AS can further modulate organism complexity not only by effectively increasing regulatory and signaling network complexity, but also by doing so in a temporal- and spatial-specific manner, supporting cell differentiation, developmental pathways, and other processes associated with multicellular organisms. Indeed, AS shows a strong relationship with organism complexity, as estimated by the organism's number of different cell types [66]. The recent ENCODE Project concluded that at least 90% of human genes express multiple mRNAs through alternative splicing of exons or exon segments [67]. As might be expected, deregulation of this process is associated with numerous diseases [68-73].

Bone marrow-derived mesenchymal stem cells (BMSCs) are adult stem cells capable of self-renewal and differentiation into numerous cell lineages, including osteocytes, adipocytes, and chondrocytes [74]. One promising use of BMSCs is repair of ischemia-damaged cardiac tissue. BMSCs are easy to expand in vitro, can be genetically modified and exhibit significant immunotolerance properties [75-77], making BMSCs an attractive candidate for tissue repair/regeneration therapy. Intramyocardial injection of BMSCs reduces inflammation, fibrosis, infarct size, ventricular remodeling, and therefore, improves cardiac function following tissue insult [78-81].

Because the majority of BMSCs are soon lost during after injection, the observed therapeutic effects likely derive from paracrine effects of bioactive molecules released from these cells [78, 79]. Indeed, BMSC-mediated release of cytoprotective protein factors or transfer of intracellular components (e.g.,mRNAs, microRNAs, and proteins) via cell membrane exosomes, represents a novel mechanism of cell-to-cell communication [82].

To date, however, clinical trials have demonstrated that while effective, delivery of BMSCs to ischemic myocardium results in only modest and short-lived benefits [83, 84]. Therefore, there is a critical need to elucidate the mechanisms by which BMSCs mediate their therapeutic benefits, including identification of their specific paracrine factor(s), and conditions under which their functions can be optimized.

Upon injection into damaged heart tissue, BMSCs face a hypoxic, ischemic environment that severely limits their therapeutic efficacy. Thus, preconditioning BMSCs with various growth factors and endogenous or exogenous molecules has been used to improve BMSC therapeutic efficacy [85-87]. Indeed, it has been reported previously that bacterial endotoxin (lipopolysaccharide, LPS) could stimulate BMSCs to release paracrine factors, including angiogenic growth factors, cytokines, and chemokines that facilitate tissue repair [76, 77]. In addition, our previous study suggested that BMSC expression of the LPS receptor, toll-like receptor 4 (TLR4), regulates BMSC paracrine properties and intracellular STAT3 signaling cascades [88]. Moreover, preconditioning of BMSCs with LPS improves their therapeutic efficacy in rodent models of ischemia/reperfusion injury [86]. However, BMSC transcriptomic changes (in particular, alterations in mRNA transcript processing and splicing) that occur following LPS stimulation have been little studied.

Besides use as an attractive therapeutic tool for repairing ischemic heart, BMSCs have been used for numerous other diseases, including graft-versus-host disease, Crohn's disease, stroke, cartilage defects, diabetes and many others [89-94]. With the growing incidence of bacterial endotoxin LPS detected in older or immunocompromised patients with multiple-drug resistant bacteria, diabetes, cancer, indwelling IV catheters, and on complex chemotherapy regimens [95, 96], it is of great importance to study whether the stimulation of these implanted BMSCs by endogenous LPS would alter their therapeutic efficacy. Moreover, because MSCs are present in bone marrow and many other tissues, it

merits extensive investigation whether LPS stimulation of these endogenous MSCs would influence the clinical outcomes of complex therapeutic regimens.

Despite BMSC's strong clinical potential, the role(s) of alternative splicing in LPS response has not been fully explored. The recent development of high-throughput sequencing technology has now made transcriptome-wide profiling of splicing isoforms possible. In this study, we used RNA-seq analysis of BMSCs to identify and characterize gene transcripts whose splicing patterns were altered by LPS treatment.

2.2 Results

To investigate LPS-induced transcriptomic changes in BMSCs due to alternative splicing, RNA-seq analysis was conducted on BMSCs before and after LPS treatment, in triplicate. A strand-directed single-end RNA-seq protocol (75 bp reads) was used with the SOLiD 5500xl instrument.

The total analysis resulted in 326 million reads, with each of the six samples ranging from 43 to 59 million reads. After removing the reads with low sequencing quality (see Methods) and filtering reads mapped to ribosomal RNAs and repeats, the remaining reads were mapped to the standard mouse reference genome (mm9). The total number of mappable reads in each sample ranged from 29 to 36 million, with an average mapping percentage of 59%. Among the mappable reads in each sample, 3.8 to 5.0 million are mapped to protein coding exons, and 2.8 to 4.0 million are mapped to splice junctions. Detailed mapping statistics for the six samples are listed in Table 2.1 Statistics of the RNA sequencing experiment

.

2.2.1 LPS-Induced Alternative Splicing

We applied a MISO (Mixture of Isoform) algorithm [97] to identify alternative splicing events elicited by LPS treatment. Based on a Bayesian inference framework, MISO is a probabilistic framework that quantitates the expression levels of alternatively

spliced genes from RNA-Seq data, and identifies differentially regulated exons across samples. MISO computes Percent Spliced In (PSI, or Ψ) values for each alternative splicing event, representing the fraction of a gene's mRNA that includes the exon. For each event, MISO also calculates a Bayesian Factor (BF) that quantifies the likelihood of the changes. For instance, [BF]=5 indicates it is five times more likely that a specific alternative splicing event occurred than did not occur.

Sample	**Group**	**Sum of Raw FASTQ**	**Sum of Passed Quality Control (QC) Filter**	**% Pass QC**	**Sum of Passed Seq Filter (rRNA/tRNA)**	**% Pass Seq Filter**	**Sum of Mapped**	**% Overall Mapped**
Library1_1	CTR	56576037	50499686	89.26	38033602	75.31	31752992	56.12
Library2_2	CTR	59674412	54819003	91.86	37784090	68.93	31603786	52.96
Library3_3	CTR	43434865	37622163	86.62	33262389	88.41	29070658	66.93
Library4_4	LPS	47452526	42001138	88.51	35363458	84.20	30198874	63.64
Library5_5	LPS	59253111	53152924	89.70	42448510	79.86	36138497	60.99
Library6_6	LPS	59897311	52868581	88.27	40708184	77.00	34247425	57.18

Table 2.1 Statistics of the RNA sequencing experiment

Overall, we identified 197 exons whose splicing patterns differed between control and LPS-treated BMSCs (Bayesian factor [BF]>5 and $|\Delta\Psi|>0.05$). This number represents 2.32% of all 8,475 events whose inclusion percentages could be reliably measured from the RNA-seq data; these genes generally had higher expression levels to generate enough read depth for splicing analysis. For genes with lower expression levels, our RNA-seq experiment did not have enough read depth for such analysis. The 197 LPS-induced alternatively spliced events included 82 cassette exons, 28 alternative donor site events (5'-ss), 45 alternative acceptor site events (3'-ss), and 42 intron retention events. Figure 2.1 demonstrates the magnitude (X-axis) and significance (Y-axis) of LPS-induced splicing pattern changes on all the alternatively exons that could be reliably identified by MISO under both untreated and LPS-treated conditions (Figure 2.1). Among these 197 events (red dots in Figure 2.1), 117 showed positive $\Delta\Psi$ values, indicating that the percentage of transcripts containing the specific exon increased in the LPS-treated samples compared to control samples. Similarly, 80 events showed negative $\Delta\Psi$ values, indicating a decrease in the percentage of transcripts containing specific exons. For each of the four types of splicing events (cassette exons, alternative 5'-donor sites, alternative 3'-acceptor sites, and intron retention), we show one Sashimi plot for the exons with the largest LPS-induced changes (either increases or decreases) in percentage of inclusion in the gene product (Figure 2.2). The Sashimi plot demonstrates the RNA-seq read densities along exons and junctions, in the context of the structure of the gene's isoforms. In addition, the distribution and the confidence intervals of the estimated Ψ under both conditions (LPS vs. untreated) are also included.

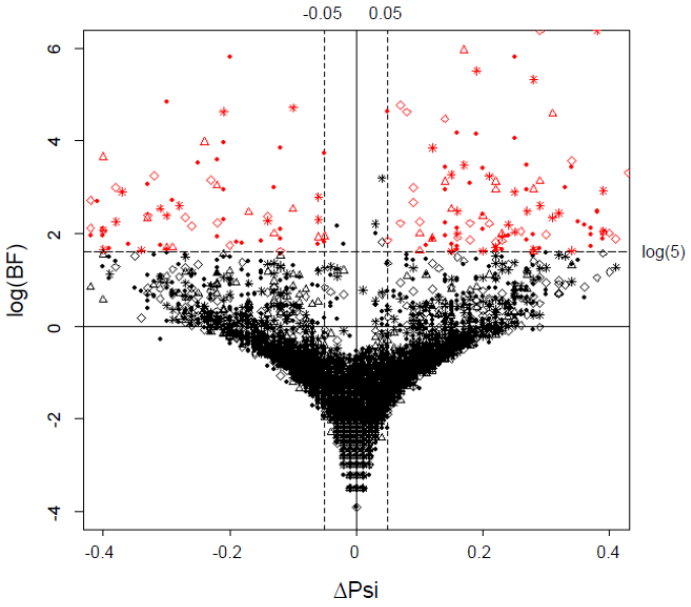

Figure 2.1 LPS-induced alternative splicing events.

Scatter plot of all the AS events identified in MISO. The X-axis represents ΔΨ, and the Y-axis represents log(BF). The shape of the dots indicates the type of the events. Specifically, circle indicates cassette exon events; star indicates intron retention events; triangle indicates alternative 5' splice site events; and diamond indicates alternative 3' splice site events. Alternatively spliced events with BF≥5 are colored in red.

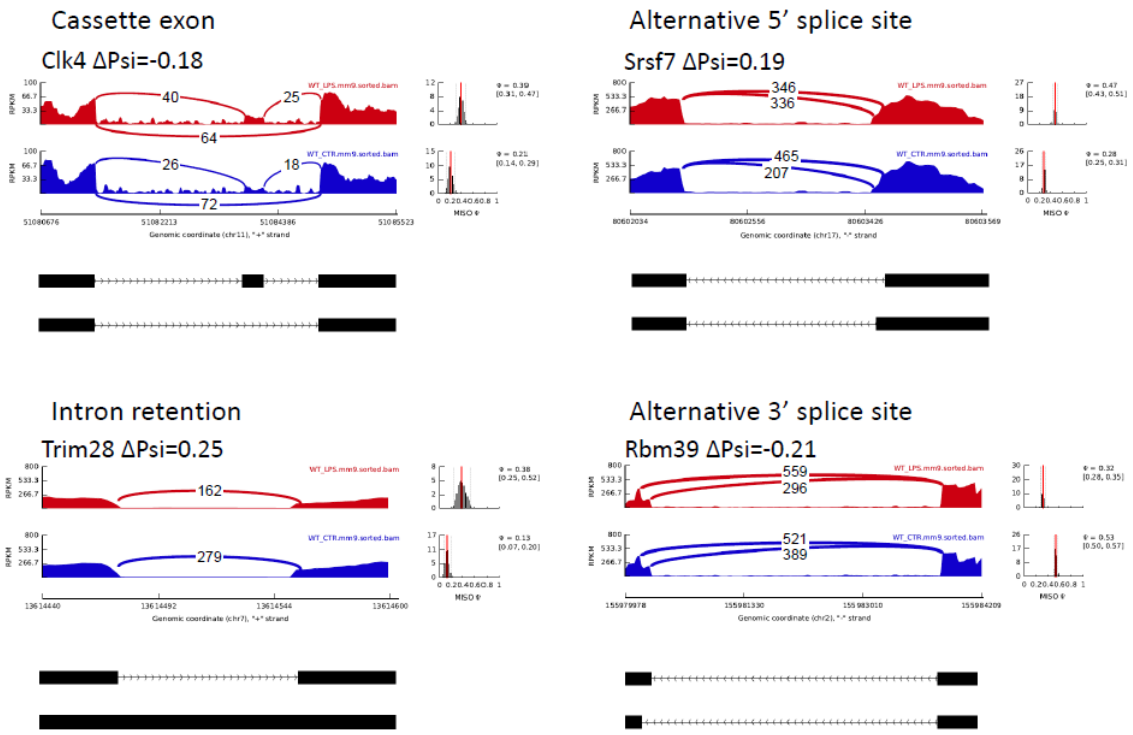

Figure 2.2 Sashimi plots of four types of AS events.

Sashimi plots of four types of AS events were shown, including cassette exon, intron retention, alternative 5' and 3' splice site. The red plots represent the LPS treated condition, and the blue ones represent controls. The X-axes indicate genomic locations, and the Y-axes indicate transcription intensity. In each plot, a "sashimi-like" region indicates a heavily transcribed region, in this case, exonic region. The blank regions between exonic regions indicate intronic regions. The "bridges" crossing exons indicate junction reads. The numbers of junction reads are shown on the "bridges". The exonic structure of each AS event is shown below each Sashimi plot. On the right it displays the estimated Ψ (red line) value and the full posterior distribution (black bars).

To validate whether the alternative splicing events were induced by LPS treatment, we performed RNA-seq on BMSCs derived from MyD88-/- animals before and after LPS treatment. MyD88 is a key signaling molecule responsible for LPS response [98]. Among the 197 LPS-induced alternative splicing events in wild-type BMSCs, 189 did not occur following LPS treatment of MyD88-/- BMSCs (Figure 2.3). This observation indicates that a large majority of BMSC splicing changes were a direct consequence of LPS induction, and such effects were negated in cells whose LPS response is compromised. It should be noted that in addition to MyD88 pathways, LPS also functions through TRIF pathways [99]; the functions of TRIF pathway is intact in the MyD88-/- cells. This partially explains why some LPS-induced splicing effects remained in MyD88-deficient animals.

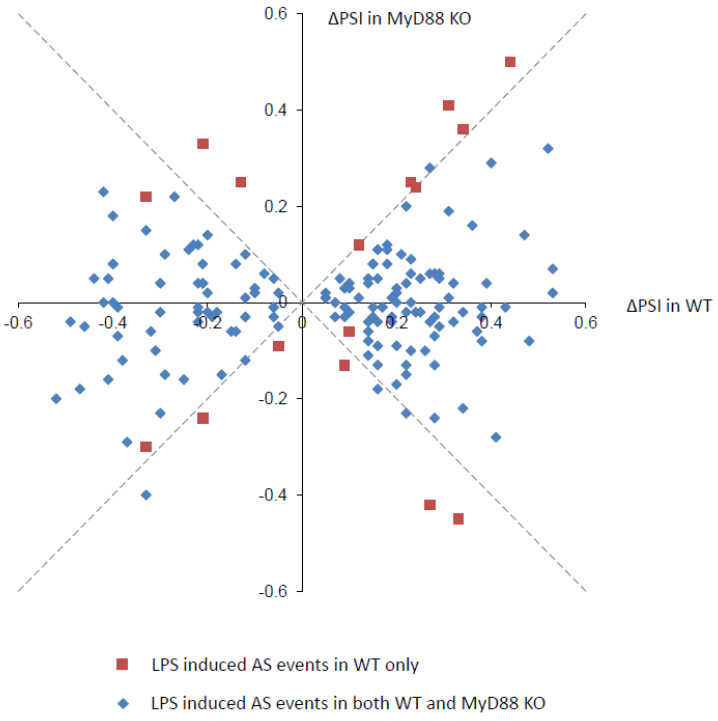

Figure 2.3 LPS-induced splicing changes in wild-type BMSC were repressed in MyD88-/- cells.

The X-axis and Y-axis represents ΔPSI in wild type and MyD88 knock out animals respectively. Blue diamond represents LPS induced AS events in wild type only, and red square represents LPS induced AS events in both wild type and MyD88 knock-out cells.

Among the 197 LPS-induced alternative splicing events, 103 were located in the coding regions of transcripts, and 94 were either in the 5'- or 3'- untranslated regions (UTRs). Among the 103 alternatively spliced coding events, 65 were composed of multiples of three nucleotides, leading to the inclusion or exclusion of specific amino-acid residues in the final protein products. These events could potentially generate multiple viable protein products having the same translation frame. Thirty-eight of the 103 coding exons contained either a premature stop codon, and/or a shift in their translation frames. Such events trigger either nonsense-mediated decay (NMD) mechanisms [100], or a translated protein having a complete different amino acid sequence downstream of the alternatively spliced exon.

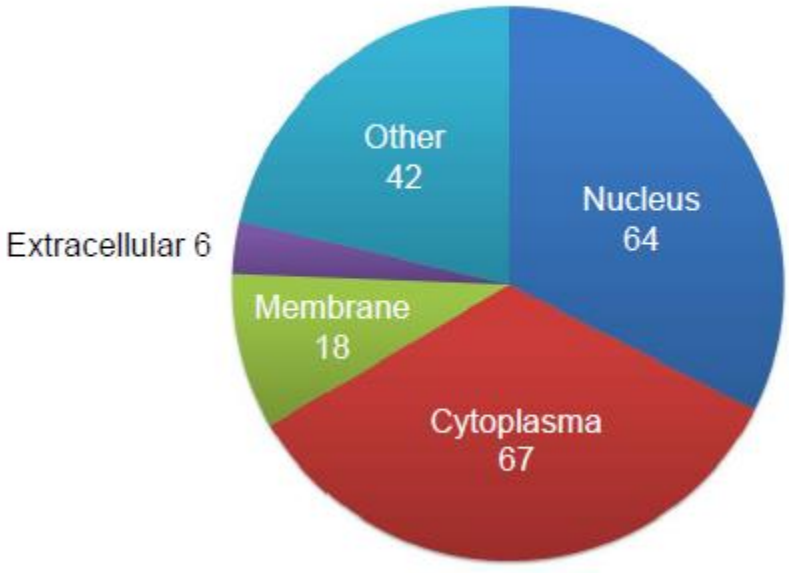

Figure 2.4 Distribution of AS genes in different cellular locations.

	Nucleus	Cytoplasma	Plasma Membrane	Extracellular Space	Other	Total
Enzyme	13	12			4	29
G-protein Coupled Receptor			1			1
Transmembrane Receptor			3			3
Kinase	2	9				11
Ligand-dependent Nuclear Receptor	1					1
Peptidase	1	3			2	6
Phosphatase			1		1	2
Transcription Regulator	17	2			1	20
Transporter		6	1			7
Other	30	35	12	6	34	117
Total	64	67	18	6	42	197

Table 2.2 Functions and cellular locations of AS genes

We then systematically examined the localization and functions of the gene products possessing alternatively spliced exons (Figure 2.4 and Table 2.2). Among them, 64 were nuclear proteins, including 17 transcription regulators, 13 enzymes, 2 kinases, 1 peptidase, and 1 ligand-dependent nuclear receptor. The 67 cytoplasmic alternatively spliced gene products included 12 enzymes, 9 kinases, 6 transporters, 3 peptidases, and 2

translation regulators. In addition, we also observed six potentially secreted proteins and 18 plasma membrane-spanning proteins. A detailed list of the genes in each category is provided in Table S-1. These results strongly suggest that LPS induces splicing changes in highly diverse proteins having a variety of cellular functions.

To understand the biological functions of genes whose splicing patterns were altered by LPS treatment, we conducted functional annotation analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 [101]. Three functional terms in the SP_PIR (Swiss-Prot and Protein Information Resources) category showed significant enrichment in our gene list. Among the 161 genes that could be mapped to DAVID gene annotations, 97 categorized as phosphoproteins (p-value=7.2x10-12, FDR=6.6x10-10). In addition, 26 genes contained zinc finger domain proteins (p-value= 3.6x10-5, FDR=2.2x10-3) whose functions range from DNA or RNA binding to protein-protein interactions and membrane association [102]. Furthermore, 35 genes were involved in protein acetylation (p-value=1.3 x10-3, FDR=3.2x10-2). Together, these results suggest that LPS treatment has major effects on the splicing patterns of signaling proteins.

Both gene expression levels and splicing patterns may be altered by BMSC responses to LPS treatment. While differential gene expression may lead to changes in the abundance of the entire gene product, alternative spicing modifies the structural composition of a specific protein. To evaluate to what extent the two mechanisms interact, we examined the number of genes present in both differentially expressed and alternatively spliced gene sets. We utilized edgeR [32] to identify genes differentially expressed between LPS-treated and control samples. In total, 416 differentially expressed genes were identified using a false discovery rate ≤ 0.05. Surprisingly, only one gene, Plscr2 (Phospholipid Scramblase 2) was both differentially expressed and alternatively spliced. The expression level of Plscr2 increased 1.77-fold in LPS-induced samples with

FDR=0.01, while the percentage of inclusion of one cassette exon in the 3'-untranslated region (3'-UTR) increased by 0.16.

2.2.2 Protein Domains are Differentially Spliced

Alternatively spliced exons residing in known protein domains are more likely to disrupt protein function. Therefore, we systematically searched the overlap between LPS-induced AS events for known protein family domains documented in the pfam database [103]. Among 65 alternatively spliced exons that did not disrupt codon frame, seven overlapped with known protein domains (Table 2.3). In addition, seven other known domains that overlapped flanking exons had functions ranging from RNA and protein binding, enzymatic activities, methyltransferase activity, phosphopantetheinyl transferase activity, RNA editing, and microRNA processing.

Gene Symbol	**AS Type**	**Pfam Domain**	**Domain Description**
Ttc13	cassete exon	TPR_11	TPR repeat
Rabep1	cassete exon	Rabaptin	Rabaptin
Camk1d	cassete exon	Pkinase	Protein kinase domain
Nr1h2	alternative 5' splice site	Hormone_recep	Ligand-binding domain of nuclear hormone receptor
Adarb1	alternative 5' splice site	A_deamin	Adenosine-deaminase (editase) domain
Scoc	alternative 3' splice site	DUF2205	Predicted coiled-coil protein
Ppip5k2	alternative 3' splice site	His_Phos_2	Histidine phosphatase superfamily (branch 2)

Table 2.3 Alternatively spliced genes containing known protein domains

2.2.3 AS in Protein Domains may Affect Protein Interactions

To examine whether alternatively spliced protein domains modulate protein-protein interactions, we searched for their binding partners based on two criteria: (1) at least one experimental study supporting direct interaction between the partner protein and the alternatively spliced protein in a known protein-protein interaction network [104, 105]; and (2) at least one structural study in the Protein Data Bank (PDB) supporting direct interaction between a domain in the binding partner and the domain modified by alternative splicing. For the first criterion, we merged two datasets of experimentally validated direct interactions [104, 105] and compiled a library of 9,795 protein-coding genes with 80,518 experimentally validated interactions. For the second criterion, we derived the domain interactions in PDB from iPfam [103] and then searched for proteins containing these domains in Pfam [103]. In total, 3,573 interactions with structural evidence were found between 13 alternatively spliced coding transcripts and 3103 binding partners. By joining two interaction tables, we identified eight interactions having both experimental and structural evidence. As shown in Figure 2.5, these eight interactions involved three genes with altered splicing domains, Rabep1 (Rab GTPase-binding effector protein 1), Camk1d (Calcium/Calmodulin-Dependent Protein Kinase 1D), and Nr1h2 (nuclear receptor subfamily 1, group H, member 2). The alternatively spliced exons in these genes overlapped with known protein domains, including rabaptin, pkinase, and ligand-binding domain of nuclear hormone receptor.

The differences in the percentage of inclusion for these three events ranged from 14% to 31%. The potential protein partners included Rabep1, Gga1 (Golgi-associated, gamma adaptin ear containing, ARF-binding protein 1), Gga2 (Golgi-associated, gamma adaptin ear containing, ARF binding protein 2), Gga3 (Golgi-associated, gamma adaptin ear

containing, ARF binding protein 3), Camkk1 (calcium/calmodulin-dependent protein kinase kinase 1, alpha), Nr0b2 (nuclear receptor subfamily 0, group B, member 2), Rxra (retinoid X receptor, alpha), and Rxrb (retinoid X receptor, beta). LPS-induced splicing changes could significantly impact these proteins' interactions with their partners. Among these putative protein interaction partners, only one protein, Nr0b2 (nuclear receptor subfamily 0, group B, member 2), was not expressed.

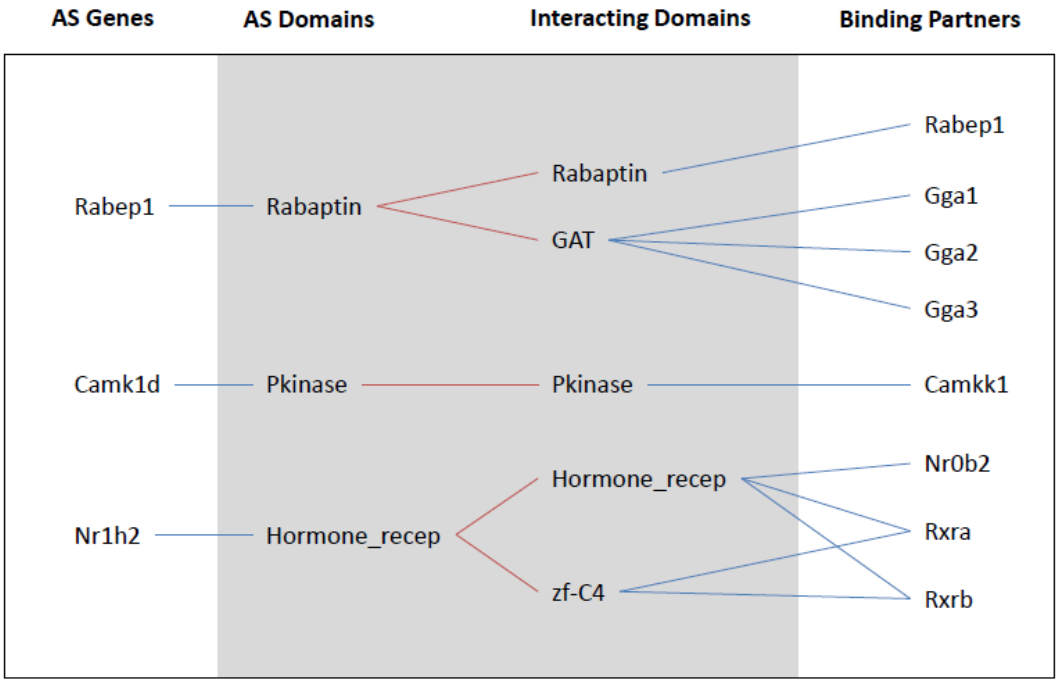

Figure 2.5 PPI with both structural and experimental evidences.

Ten AS gene products involved in protein-protein interactions. Gene symbols are displayed in white regions, and corresponding protein domains are displayed with gray background. Blue line indicates a gene/protein contains a domain, and a red line indicates an interaction between protein domains.

2.2.4 Intrinsic Disorder and MoRF in AS Regions

It was previously reported that alternatively spliced regions are enriched with unfolded protein regions (intrinsic disorder) [106]. To examine these features within LPS-induced alternatively spliced regions (cassette exons, alternative 5'/3' exons and retained introns), we performed disorder prediction on the protein sequences of these regions using VSL2B [107], a bioinformatics algorithm for predicting intrinsically disordered regions based on the biophysical properties of amino acids. Among the alternative regions of 65 protein sequences translated from LPS-induced alternative splicing events, 34 (52.3%) were predicted to be totally disordered, 21 (32.3%) partially disordered, and only 10 (15.3%) totally structured (Figure 2.6). These percentages are consistent with previous reports that alternatively spliced exons tend to locate in intrinsically disordered regions [108].

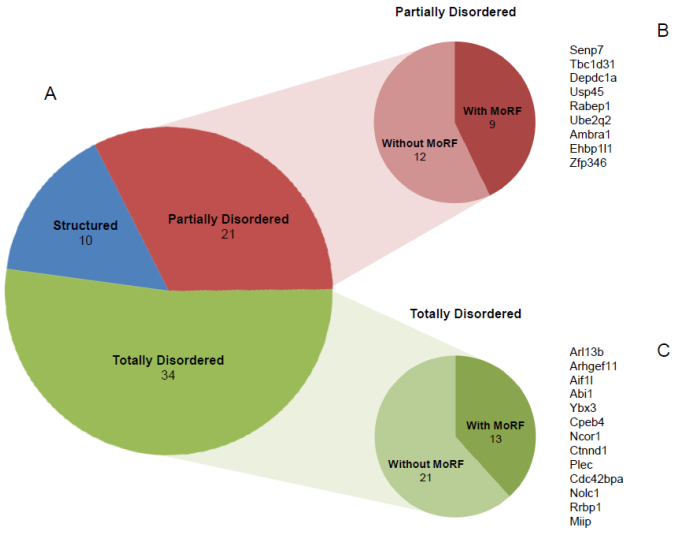

Figure 2.6 Predicted disorder of AS gene products.

(A) The distribution of 65 non-frameshifting protein coding AS genes in three categories, including totally disordered, partially disordered and structured. (B) The MoRF containing and non-containing events among partially disordered AS genes. The list of MoRF containing genes is shown on the right. C) The same distribution and gene list as panel B but it is for totally disordered AS genes.

gene_symbol	gene_description
Arl13b	ADP-ribosylation factor-like 13B
Senp7	SUMO1/sentrin specific peptidase 7
Tbc1d31	TBC1 domain family, member 31
Depdc1a	DEP domain containing 1
Arhgef11	Rho guanine nucleotide exchange factor (GEF) 11
Aif1l	allograft inflammatory factor 1-like
Abi1	abl-interactor 1
Usp45	ubiquitin specific peptidase 45
Ybx3	Y box binding protein 3
Cpeb4	cytoplasmic polyadenylation element binding protein 4
Rabep1	rabaptin, RAB GTPase binding effector protein 1
Ncor1	nuclear receptor corepressor 1
Ube2q2	ubiquitin-conjugating enzyme E2Q family member 2
Ctnnd1	catenin (cadherin-associated protein), delta 1
Plec	plectin
Cdc42bpa	CDC42 binding protein kinase alpha (DMPK-like)
Ambra1	autophagy/beclin-1 regulator 1
Ehbp1l1	EH domain binding protein 1-like 1
Zfp346	zinc finger protein 346
Nolc1	nucleolar and coiled-body phosphoprotein 1
Rrbp1	ribosome binding protein 1
Miip	migration and invasion inhibitory protein

Table 2.4 Alternatively spliced genes containing Molecular Recognition Features (MoRF)

A molecular recognition feature (MoRF) is a region in an RNA that undergoes a disorder-order transformation while bound by another protein. We predicted MoRF regions within the alternative regions using the software tool MoRF2 [106]. As a result, among the 55 alternatively spliced exons in the partial or totally disordered regions, 22 contained regions predicted to be MoRFs (Figure 2.6, Table 2.4); these regions could thus be regarded as potential protein-protein interaction sites.

2.2.5 PTM within Differentially Spliced Regions

We next annotated post-translational modification (PTM) sites in regions affected by LPS-induced alternative splicing. We searched known PTM sites deposited in UniProt, and we also predicted novel ones using ModPred [109]. Three alternatively spliced exons containing known PTM (phorphorylation) sites localized to three genes, Abi1 (abl-interactor 1), Depdc1a (DEP Domain-Containing 1), and Ybx3 (Y box-binding protein 3). In addition, 13 PTMs were predicted to occur in 29 alternatively spliced regions, including proteolytic cleavage, phosphorylation, amidation, hydroxylation, carboxylation, ADP-ribosylation, O-linked glycosylation, acetylation, GPI anchor amidation, palmitoylation, pyrrolidone carboxylic acid, methylation and ubiquitination (Figure 2.7). Proteolytic cleavage sites were the most common PTM sites, appearing in 14 alternative regions. It is possible that LPS affects the signaling activities of these proteins by inclusion or exclusion of the PTM sites in the final protein product (i.e., whether or not it is cleaved).

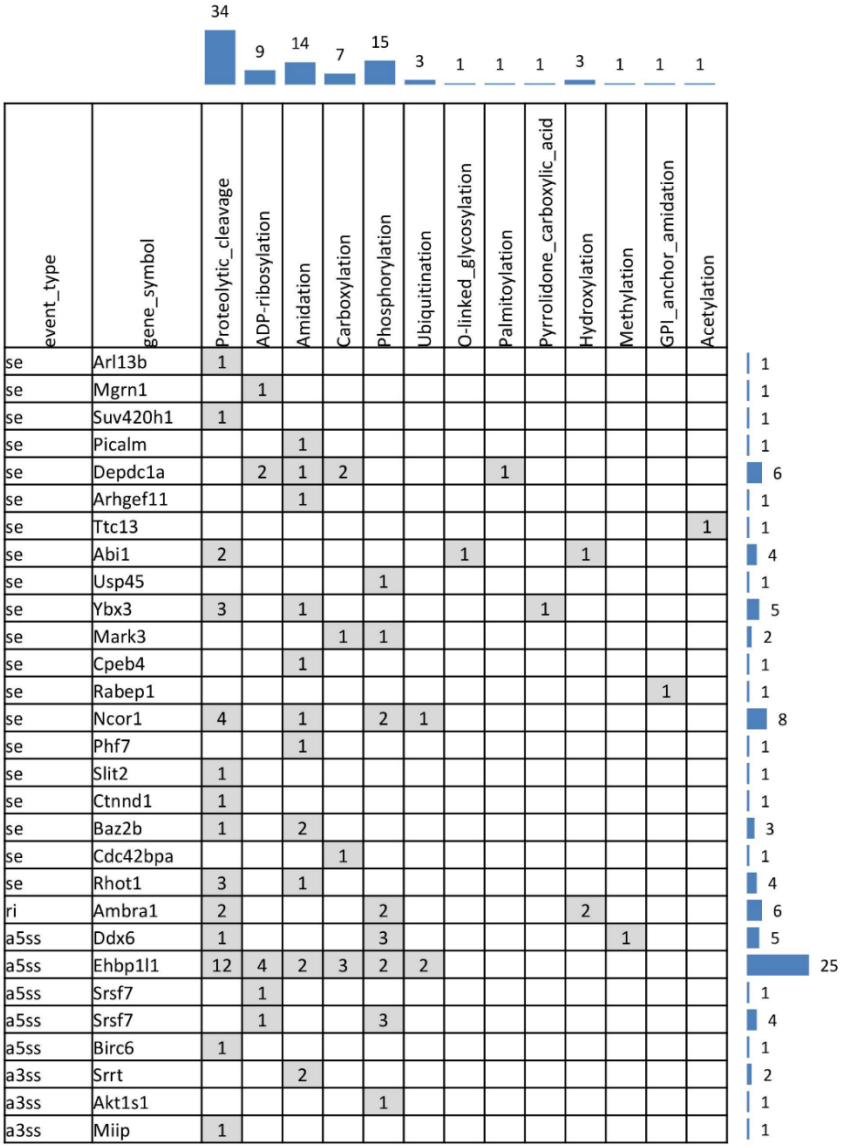

Figure 2.7 Predicted PTM sites in AS regions.

Column displays different types of PTM sites, and row displays the event types and LPS-induced AS genes. The numbers in the shadowed grids on the crossing of gene A and PTM type B shows how many type B PTM sites fall in the AS region of gene A. The total number of PTM sites in each gene is displayed on the right, and the total number of PTM sites in each type is displayed on the top.

2.2.6 Characterization of Potential Splicing Regulators

We defined 7 regulatory regions for each cassette exon event (Figure 2.8), among which Region 1 and 7 are 150nt constitute exon segments, Region 2, 3, 5 and 4 are 300nt intronic segments, and Region 4 is the whole cassette exon. We used FIMO [110] to search for CISBP-RNA [111] motifs within the regulatory regions of both up-regulated and down-regulated cassette exon events. With p-value cutoff of 1E-4 and FDR cutoff of 0.1, we identified 29 RBP motifs in the up-regulated events, and 23 in the down-regulated events. BRUNOL5, BRUNOL4 and RBM38 are the most frequently observed RBPs. Their motifs concentrate in Region 2 and 3 for up-regulated events, and in Region 5 for down-regulated events. These three proteins are all known as RNA-splicing related. Motifs of several other RNA-splicing related proteins, including SRSF2, HNRNPL, HNRNPLL, HNRNPH2 and PCBP2, are observed in both up-regulated and down-regulated cassette exon regulatory regions. Some RBPs (SRSF9, RBM5, PCBP3, PCBP1, ZCRB1, NCL, FUSIP1, PABPN1, TARDBP and NOVA2) are found exclusively in up-regulated cassette exon events, and some (KHDRBS3, BRUNOL6, G3BP2, FXR1, SRSF4, SNRNPA, SNRPB2) are found exclusively in down-regulated events.

2.3 Discussion

Lipopolysaccharide (LPS, endotoxin) is a complex associated with the outer membrane of Gram-negative bacteria, capable of triggering a series of cellular responses in many cell types. One promising advance is to use LPS as a pre-conditioning agent to improve BMSC therapeutic efficacy for repairing ischemic, injured tissues [86, 112]. For such application, because LPS is a potent stimulant for the host immune system, BMSCs should be washed using PBS to completely remove any residual endotoxin before administration. We reported previously that BMSCs treated with LPS produced more angiogenic factors VEGF, IGF-1 and HGF [113, 114] which can spur the formation of new blood vessels in ischemic tissue and survival and differentiation of implanted BMSCs. By

contrast, with the growing incidence of sepsis, in which free LPS can bind to and activate Toll-like receptor 4 on many cell types, the roles of LPS on endogenous BMSCs and other cell types are worth detailed investigation.

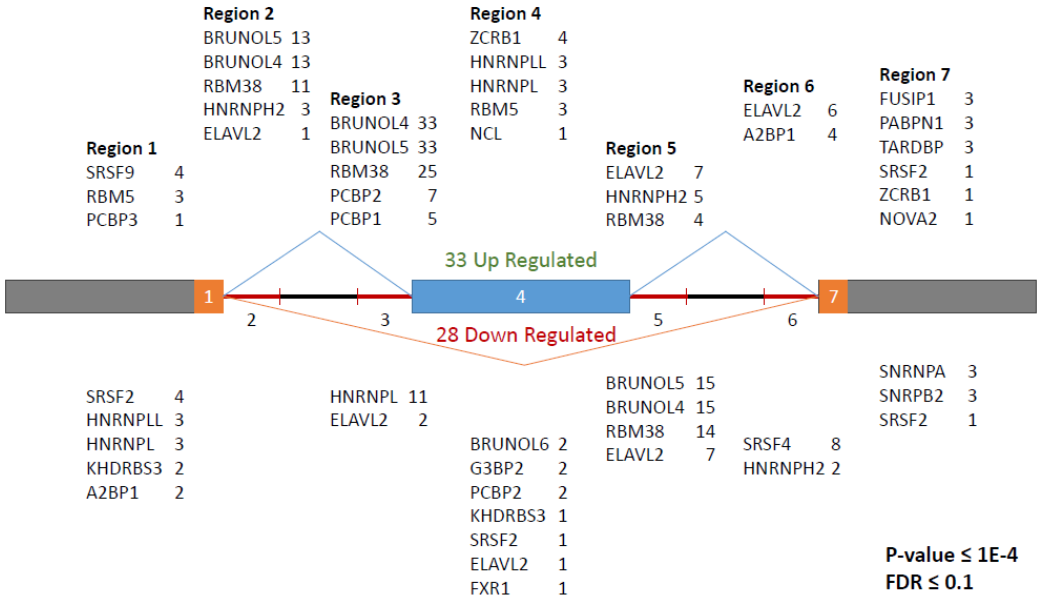

Figure 2.8 RNA binding protein (RBP) motifs in regulatory regions of differentially spliced events.

RBP names and their occurrences are listed adjacent to corresponding regulatory regions.

Microarray studies have reported that expression levels of hundreds of genes can be altered after LPS treatment in different tissues. In recent years, high-throughput RNA sequencing technology has provided a more accurate and comprehensive measurement of RNA transcript levels and their isoforms than historic array-based methods. This technological advance has enabled measuring not only gene expression level alterations amongst different conditions, but also complicated splicing pattern changes in response to specific cellular perturbations. In this study, we systematically identified alternative splicing changes in mouse bone marrow-derived mesenchymal stem cells (BMSCs) in

response to LPS treatment, using RNA-seq technology. We further implemented a series of bioinformatics tools to evaluate the biological functions of alternatively spliced exons and their host genes.

We observed strong enrichment in three functional categories amongst the gene products whose splicing patterns were altered by LPS treatment, phosphoproteins, zinc finger proteins, and proteins subject to acetylation. Most of these proteins were signaling proteins, and the subtle differences in their splicing isoforms could affect their function.

Among 161 gene products containing AS exons, 97 belonged to phosphoprotein families, five of which contained documented phosphorylation sites in their AS regions found in the UniProt database. These proteins included Kansl2 (KAT8 regulatory NSL complex subunit 2), Depdc1a (DEP domain-containing 1), Abi1 (abl-interactor 1), Ybx3 (Y box-binding protein 3), and UBl4a (Slc10a3-Ubl4 readthrough). The functions of these proteins strongly associate with the functions of BMSCs. For instance, Abi1 contains one cassette of exons whose percentage of inclusion increased by 14% after LPS induction ($\Delta\Psi$=0.14), with one phosphorylation site in the AS region documented in the UniProt database. Widely expressed with highest levels in bone marrow, spleen, brain, testes, and embryonic brain, Abi1 may negatively regulate cell growth and transformation by interacting with the nonreceptor tyrosine kinases ABL1 and/or ABL2, thus regulating EGF-induced Erk pathway activation and EGFR signaling. In addition to these five proteins, eight other AS regions were predicted to have phosphorylation sites, based on their amino acid contents. These proteins included Usp45 (ubiquitin-specific peptidase 45), Mark3 (MAP/microtubule affinity-regulating kinase 3), Ncor1 (nuclear receptor corepressor 1), Ctnnd1 (cadherin-associated protein, beta 1), Ambra1 (autophagy/beclin-1 regulator 1), Ddx6 (DEAD (Asp-Glu-Ala-Asp) box helicase 6), Ehbp1l1 (EH domain binding protein 1-like 1), and Akt1s1 (AKT1 substrate 1). Overall, LPS may affect the

functions of these proteins by including/excluding specific domains amenable to phosphorylation.

Among the proteins containing LPS-induced alternative splicing events, 25 contained multiple types of zinc finger domains, including PHD (Plant Homeo Domain), RING (Really Interesting New Gene), and C2H2-type zinc-finger domains. Four proteins, Phf7 (PHD finger protein 7), Phf20 (PHD finger protein 20), Phf20l1 (PHD finger protein 20-like 1), and Phrf1 (PHD and ring finger domains 1), contained PHD-type zinc finger domains known to recognize trimethylated histone lysines (thus possibly influencing chromatin structure). Four other proteins, Rnf14 (ring finger protein 14), Rad18 (RAD18 homolog), Trim28 (tripartite motif-containing 28), and Trim2 (tripartite motif-containing 2), all contain RING-type zinc fingers, known ligases for ubiquitination enzymes and their substrates. It is well documented that both PHD and RING-type domains are usually involved in protein-protein binding [115, 116], and such binding could possibly be disrupted by splicing variations.

Overall, the LPS-induced AS genes could be classified into several categories (Figure 2.9), including kinases, zinc-finger proteins, transcription, RNA-binding, cytoskeleton, and protein acetylation. Many of these proteins were also phosphoproteins, which play significant roles in cell signaling. Analysis of the relationship between splicing and protein structure has suggested that AS exons play major roles in controlling protein-protein interactions (PPIs) through disrupting either known protein interaction domains or molecular recognition sites, which typically locate in intrinsically disordered regions. Our analysis suggests that LPS-induced alternative splicing could affect PPIs through both mechanisms. In particular, protein interaction domains of three proteins with known PPI partners were disrupted by LPS-induced splicing alterations (Figure 2.5). Interestingly, all three interactive domains could self-interact (forming domain-domain interactions with themselves), and one of these domains facilitates homodimerization of

Rabep1 (RAB GTPase binding effector protein 1). Expressed in embryonic tissues and most types of stem cells, Rabep1 showed abundant expression in BMSCs (about 30 RPKM). Homo-dimerization of this protein is involved in early endosome fusion [117], an event directly related to the paracrine effects of BMSCs, where small vesicles are released when multivesicular endosomes fuse with the plasma membrane [118, 119]. In addition, Rabep1 also moderates intracellular transportation between lysosomes and the Golgi apparatus [120], and between the Golgi apparatus and endoplasmic reticulum [121]. LPS treatment also increased the inclusion of the interaction domain by 14%, which could increase either homodimerization or heterodimerization with other interaction partners.

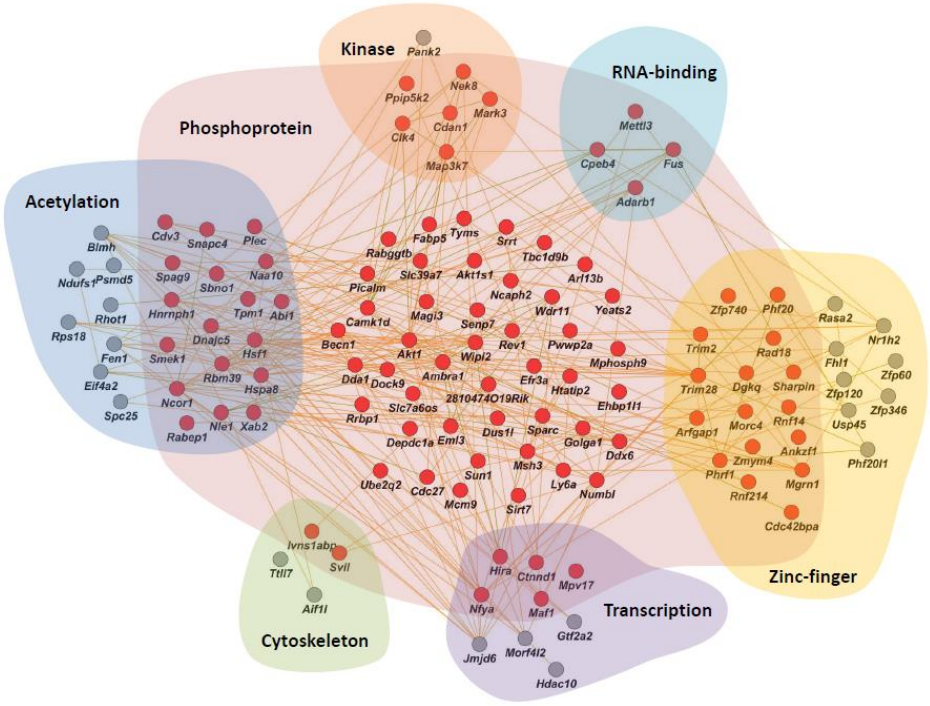

Figure 2.9 Predicted interaction network among LPS-induced AS genes.

Red nodes indicate genes producing phosphoproteins, and gray nodes indicate genes not involved in protein phosphorylation. Genes associated with terms other than phosphoproteins are clustered in corresponding shadowed areas. These terms include acetylation, cytoskeleton, transcription, zinc-finger, RNA-binding and kinase.

We further evaluated how differences in splicing patterns in transcriptional regulators affected their regulatory activity by assessing gene expression changes of their downstream target genes. NFYA (nuclear transcriptional factor Y) contains an alternative acceptor site whose splicing pattern in BMSCs is altered by LPS treatment; the overall percentage of inclusion of the alternative acceptor site decreased by 31% (Sashimi plot for NFYA shown in Figure 2.10). Moreover, the expression of five downstream target genes of NFYA were enriched for genes found differentially expressed (p-value≤0.01) by LPS treatment (FDR≤0.05), including COL11A1 (collagen, type XI, alpha 1), COL5A3 (collagen, type V, alpha 3), FGFR2 (Fibroblast Growth Factor Receptor 2), PGK1 (phosphoglycerate kinase 1) and RGS4 (regulator of G-protein signaling 4). It was previously reported that NFYA activates transcription levels of COL11A1 and FGFR2 [122]; these two genes were both downregulated by LPS, suggesting inhibition of NFYA function by the removal of 18nt (or 6 amino acids) after LPS treatment, thus impacting NFYA downstream effectors.

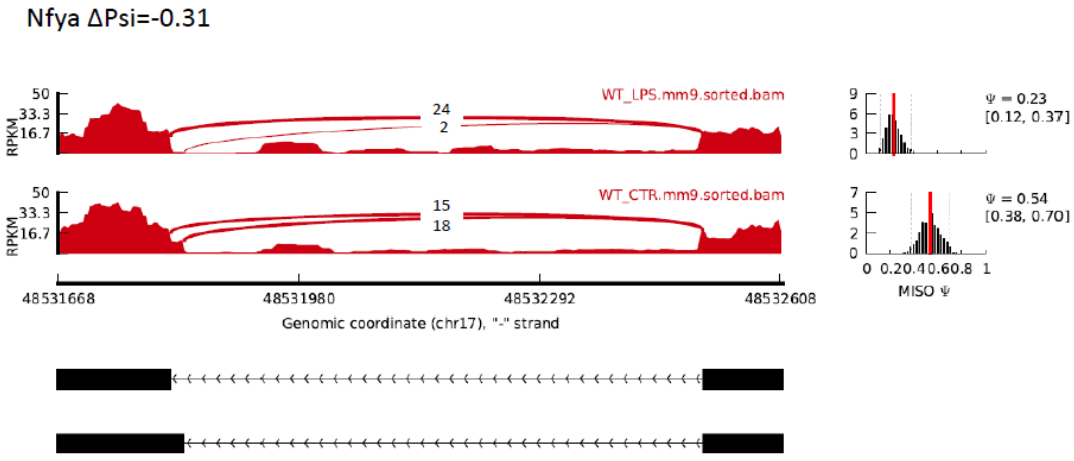

Figure 2.10 Sashimi plot of NFYA.

2.4 Methods

2.4.1 Preparation of Mouse BMSCs

A single-step stem cell purification method was employed as previously described [123]. Briefly, BMSCs were collected from the bilateral femurs and tibias of sacrificed mice by removing the epiphyses and flushing the shaft with complete media, Iscove's Modified Dulbecco's Medium (IMDM; Life Technologies) and 10% fetal bovine serum (Life Technologies), using a syringe with a 26G needle. Cells were disaggregated by vigorous pipetting and passed through a 30-μm nylon mesh to remove any remaining clumps of tissue. Cells were then centrifuged for 5 min at 500 g at 24°C. The cell pellet was then resuspended and cultured in 75 cm2 culture flasks in complete media at 37°C with 5% CO2. Since BMSCs preferentially attach to polystyrene [124], after 48 h, floating non-adherent cells were discarded. Fresh complete media was added and replaced every three or four days thereafter. When the cells reached 90% confluence, MSC cultures were recovered by the addition of a solution of 0.25% trypsin-EDTA (Invitrogen) and passaged. Cell passage was restricted to passages 6–10 for the experiments. To purify BMSCs, the cells were subject to fluorescence-activated cell sorting (FACS) analysis, with collection of cells positive for Sca-1 and CD44 [124], but negative for the hematopoietic stem cell and macrophage marker CD45 [88].

2.4.2 RNA Sample Preparation and RNA-seq Assay

BMSCs were plated at 1×105 cells/well/ml for 24 h and further treated with LPS (200ng/ml) for another 24 h, and total RNA was extracted before and after LPS treatment, following a standard protocol [88]. Experiments were conducted in triplicate.

Standard methods were used for RNA-seq library construction, EZBead preparation, and Next-Gen sequencing, based on the Life Technologies SOLiD 5500xl system. Briefly, 2 μg of total RNA per sample was used for library preparation. The rRNA was first depleted using the standard protocol of RiboMinus Eukaryote Kit for RNA-Seq

(Ambion), and rRNA-depleted RNA was concentrated using a PureLink RNA Micro Kit (Invitrogen) with 1 volume of lysis buffer and 2.5 volumes of 100% ethanol. After rRNA depletion, a whole transcriptome library was prepared and barcoded per sample using the standard protocol of SOLiD Total RNA-seq Kit (Life Technologies). Each barcoded library was quantified by quantitative polymerase chain reaction (qPCR) using SOLiD Library Taqman qPCR Module (Life Technologies) and pooled in equal molarity. EZBead preparation, bead library amplification, and bead enrichment were then conducted using the Life Technologies EZ Bead E80 System. Finally sequencing by ligation was performed using a standard single-read, 5′-3′ strand-specific sequencing procedure (75nt-read) on SOLiD 5500xl.

2.4.3 Bioinformatics Analysis for RNA-seq Data

RNA-seq data analysis included the following steps: quality assessment, sequence alignment, and alternative splicing analysis. The RNA-seq data can be accessed through the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE64568).

2.4.4 Data Processing and Quality Assessment

We used SOLiD Instrument Control Software and SOLiD Experiment Tracking System software for read quality recalibration. Each sequence read was scanned for low-quality regions, and if a 5-base sliding window had an average quality score less than 20, the read was truncated at that position. Any read < 35 bases was discarded. Our experience suggests that this strategy effectively eliminates low-quality reads, while retaining high-quality regions [125-127].

2.4.5 Sequence Alignment

We used BFAST (http://bfast.sourceforge.net) [27] as our primary alignment algorithm due to its high sensitivity for aligning reads on loci containing small insertions and deletions, as compared to the reference genome (mm9). We then used a TopHat-like

strategy [30] to align the sequencing reads containing cross-splicing junctions using NGSUtils (http://ngsutils.org/) [125]. After aligning the reads to a filtering index including repeats, ribosome RNA, and other sequences that were not of interest, we conducted a sequence alignment at three levels: genome, known junctions (University of California Santa Cruz Genome Browser), and novel junctions (based on the enriched regions identified in the genomic alignment). We restricted our analysis to uniquely aligned sequences with no more than two mismatches.

2.4.6 Alternative Splicing Analysis

We used MISO (mixture of isoforms) [97] to identify alternatively spliced exons whose splicing patterns were altered after LPS treatment. We first used Samtools (v0.1.19) to merge six RNA-seq samples into two BAM files according to their biological conditions, i.e., control vs. LPS-treated samples. We then estimated Percent Spliced In (PSI or Ψ), which indicates the proportion of RNA isoforms containing the alternatively spliced exon (inclusive isoforms) among all isoforms (inclusive plus exclusive isoforms). We also computed a Bayes factor (BF) to describe the likelihood of an AS event between the LPS-treated and control conditions. A BF of 5 means that an AS event is 5 times more likely to be differentially spliced than not. Both Ψ and BF values were computed by the software package MISO [97]. The difference between Ψ s across the two conditions was defined as ΔΨ. We required each AS event to have a BF>5 and $|\Delta\Psi|>0.05$ to be considered differentially spliced.

2.4.7 Ontological Annotations

The functions and cellular locations of AS genes were annotated by the pathway analysis tool Ingenuity Pathway Analysis (IPA), and the functional and biochemical properties of these genes were further annotated based on SwissProt and PIR keywords with DAVID v6.7 [101].

2.4.8 Protein Domains Overlapping AS regions

Protein domain information was predicted based on the RNA nucleotide sequences of the alternatively spliced exon, and 30-base flanking sequencings of both upstream and downstream exons. These RNA sequences were then translated into peptides, based on open reading frames (ORFs) documented by Ensembl and Refseq, which were then input into Pfam [103] for identification of protein domains overlapping AS regions.

2.4.9 Identification of Protein Interactions

We also examined whether alternatively spliced exons overlapped with potential protein-protein interaction domains. Based on the protein domains identified in or overlapping AS regions, we retrieved their binding partner domains with iPfam [103], which documents domain-domain interactions in the Protein Data Bank (PDB). We further used Pfam to search for genes encoding partner domains (i.e., potential protein interaction partners). The identified protein interaction partners were verified by two protein-protein interaction databases derived from high-throughput experiments.

2.4.10 Other Characterizations

Protein disorder was predicted with VSL2B [107], a highly regarded protein disorder prediction tool, especially for long regions of disorder [128]. We required the peptides flanking the AS regions to be at least 9 amino acids long for accurate prediction. Potential binding sites were predicted with MoRF2, a software tool that predicts protein-binding sites that undergo a disorder-order transformation while binding another protein molecule [129]. Known post-translational modification (PTM) sites were derived from UniProt, and novel PTM sites were predicted by ModPred [109]. The upstream gene regulator NFYA (Nuclear transcription Factor Y subunit Alpha) [130] was predicted by Ingenuity Pathway Analysis (IPA), based on gene expression data and known regulatory gene interactions.

Chapter 3. Developmentally Regulated Alternative Splicing

3.1 Background

Liver is a vital organ only found in vertebrates [131], and it is involved in many crucial physiological functions, including protein synthesis, detoxification and production of many chemicals necessary for digestion. Liver constitutes 5% of the bodyweight at birth but 2% in adult [132]. During organogenesis, liver originates from both the foregut endoderm and septum transversum mesenchyme, and many morphological and physiological changes take place in liver from embryo to fetus. Liver is the organ where most of drug metabolism processes take place. Drugs are transformed into more water-soluble forms for the ease of excretion after exerting their desired functions. Such processes include phase I and phase II metabolic reactions. Phase I involves structural alteration on drug molecules, including oxidation, reduction and hydrolysis. Phase II involves addition of functional groups, including acetylation, methylation, glucuronidation, sulfation, conjugation with amino acids and glutathione. It is known that many of these drug metabolism processes change over the lifespan of human because of the expression change of certain genes [133-135]. These genes include cytochrome P450s (CYPs), flavin-containing monooxygenases (FMO), monoamine oxidases (MAO), alcohol dehydrogenases (ADH), molybdenum hydroxylases, NADPH-cytochrome P450 reductases (FAD), aldo-ketoreductases, esterases, N-acetyltransferases (NAT), methyltransferases, UDP glucuronosyltransferases (UGT), sulfotransferases (SULT) and cytosolic glutathione S-transferases (GST) [135]. Although a lot of studies profiled the developmental expression change of the genes that are associated with drug metabolism in liver, the role of alternative splicing during the development and drug metabolism in liver is still poorly understood.

Alternative splicing (AS) is a regulated process during RNA transcription that alters the composition of exons within the mRNAs, and thus enables the production of a

variety of protein isoforms from one gene. AS is a common phenomenon in eukaryotes, and virtually 95% of multi-exonic genes go through AS [136]. Besides contributing to the diversity of RNA and proteins, AS also participate in the modulation of cell differentiation [40], developmental pathways [137], drug response [138] and other physiological processes [139-142] by activating in a temporal- and spatial-specific manner. Studies have shown that abnormalities of AS is associated with various diseases [143-145]. As the next generation RNA sequencing (RNA-seq) technology gets more economic and accurate, it has become a primary tool for studying AS. Several software tools are developed for studying AS. These software tools falls in two major categories, 1) whole transcript builders, including Cufflinks [40] and Scripture [41], which rebuild and quantifies whole mRNA transcripts, and 2) splicing event quantifiers, including MISO [97] and MATS [146], which quantifies only the alternative region and adjacent exons. In splicing event quantifiers, four major types of AS events are defined, i.e. cassette exon or skipped exon (SE), intron retention or retained intron (RI), alternative 5' splice site (A5SS) and alternative 3' splice site (A3SS). The count of RNA-seq reads within alternative region and spanning across splicing junctions are used to estimate a percentage-spliced in (PSI or Ψ) value. What's more, MISO utilizes the count of reads within adjacent exons (constitutive reads) to improve the accuracy of estimated Ψ. The Ψ value of an AS event represents the percentage of mRNAs containing the alternative regions among all mRNAs. Both MISO and MATS provided solutions for identifying significant Ψ changes of an AS event across two time points during liver development, MATS uses a multivariate uniform distribution and put within-group variance into consideration.

In this study, we introduced a Bootstrap based method that utilizes the information of both constitutive reads and within-group variance. We applied this method on RNA-seq data of human liver in three developmental stages to identify developmentally regulated AS events.

3.2 Results

To investigate the AS changes during liver development, RNA-seq analysis was conducted on fresh frozen liver samples across fetal, pediatric and adult stages, with 10 samples in each stage. A strand-directed single-end RNA-seq protocol (75 nt reads) was implemented with SOLiD instrument. The sequencing reads were mapped to the standard human genome (hg19), and the analysis resulted in 300 million mappable reads, with each of the 30 samples ranging from 3.3M to 17.9M reads. Among the mappable reads in each sample, about 0.5 to 2.8 million are mapped to protein coding exons, and about 0.2 to 1.6 million are mapped to splice junctions.

3.2.1 Developmentally Regulated Alternative Splicing Events in Liver

We utilized MISO (Mixture of Isoform) algorithm [97] to estimate the probability distribution of Percentage Spliced In (PSI or Ψ) value for each AS event in each liver sample. Ψ is defined as the fraction of the mRNA isoforms that include the alternative region. In total 72,922 AS events were analyzed, including 42,485 SE events, 7,197 RI events, 9,035 A5SS events and 14,205 A3SS events. A Bootstrap approach (see Methods) was then applied to compare the mean of Ψ values across two groups, i.e., Pediatric group (Ped) vs Fetal group (Fet), or Adult group (Adu) vs Pediatric group. In each comparison, the Bootstrap approach calculates a P-value and a False Discovery Rate (FDR) for each event.

Overall, we identified 477 exons that are differentially spliced during the development from fetuses to childhood, which represents 3.52% of all 13,550 measureable AS events, which are covered by no less than 20 reads by MISO's default. On the other hand, we identified only 49 differentially spliced exons during the childhood to adult development, which represents 0.99% of 4,915 measurable AS events by MISO's default. The obvious difference between the percentages of differentially spliced exons reasonably

reveals the drastic changes in transcriptome during early years of a human's life (Table 3.1).

Figure 3.1 demonstrates the magnitude (X-axis) and significance (Y-axis) of liver developmental splicing pattern changes on all AS events that could be reliably identified by MISO during two developmental periods, i.e. fetus to childhood and childhood to adult. Among the 477 AS events differentially spliced during the fetus to childhood development, 194 showed positive ΔΨ values, indicating that the percentage of transcripts containing the specific exon increased in this process. Similarly, 283 showed negative ΔΨ values, indicating decreases in the percentage of transcripts containing specific exons. On the other hand, 49 AS events were differentially spliced during the development from childhood to adult, 23 of which showed positive ΔΨ, and 26 showed negative ΔΨ. Twenty-one AS events changed their splicing patterns from fetus to childhood, and continued changing from childhood to adult. The contrast between the numbers of splicing changes in earlier and later developmental periods indicates the physiology of liver changes prominently during the development from fetus to childhood, while it generally reaches its mature state after the beginning of childhood.

	SE	**RI**	**A5SS**	**A3SS**
Pediatric vs Fetal	303	76	36	62
Adult vs Pediatric	14	20	11	4

Table 3.1 Number of differentially spliced events

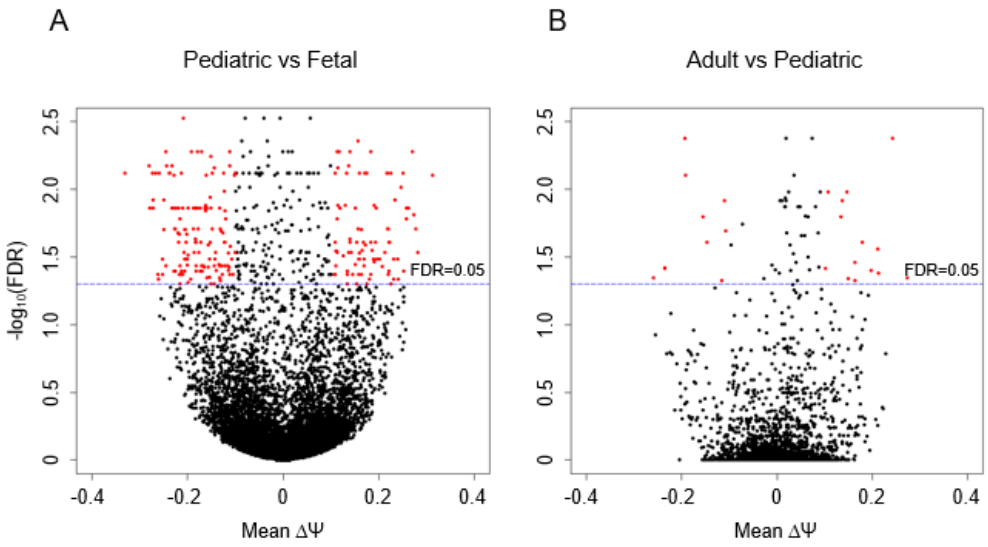

Figure 3.1 Volcano plot of AS events

(A) Comparison between pediatric stage (Ped) and fetal (Fet) stage. (B) Comparison between adult (Adu) stage and pediatric stage. The x-axes denote mean ΔΨ, the y-axes denote –log10(FDR). Each dot represents an AS event, and the red dots represent AS events with $|\Delta\Psi| \geq 0.1$and FDR≤0.05.

Figure 3.2 displays a classic SE event that is differentially spliced during one development period but is not during the other. The alternative region in this event covers the total length of the fn3 domain in fibronectin 1 (FN1), a gene involved in cell adhesion and migration processes in embryogenesis [147-151]. The splicing of this gene is known as developmentally regulated [152, 153]. Its Ψ values consistently dropped (ΔΨ=-0.12, FDR≈0) during the fetus to childhood development, and the Ψ values remain low in the adult stage (Figure 3.2A). To create a straightforward visualization of the RNA-seq data, we merged all samples in each developmental stage and generated the Sashimi plots (Figure 3.2B) [97] for each stage. In the fetal stage, 131 and 94 reads are mapped to the junctions supporting the inclusive transcript, and 183 reads are mapped to the junction

supporting the exclusive transcript. But in the pediatric stage, the number of inclusive junction reads drops to 29 and 22, while the number of exclusive junction reads raised to 298. This leads to a conclusion that the relative dosage of the fn3 domain declined during the fetus to childhood development, which is consistent with the observation of reduced Ψ in Figure 3.1A. In the adult stage, both inclusive and exclusive junction reads reduced due to lower sequencing depth, but the change is not as dramatic as the previous developmental period and is insignificant statistically.

To understand the biological functions of genes whose splicing patterns are developmentally regulated in liver, we implemented functional annotation analysis using the IPA [154] (Figure 3.3A). In the fetus to childhood development, 477 AS events were mapped to 474 genes. Among these genes, 36 are transcription regulators, 22 are kinases, 19 are transporters, 18 are peptidases, 10 are translation regulators, 7 are transmembrane receptors and 121 are other enzymes. Among the 49 AS genes that are differentially spliced during the childhood to adult development, 4 are transcription regulators, 4 are transporters, 1 is peptidase, 1 is translation regulator and 18 are other enzymes. We also implemented an analysis on the cellular locations where these genes exert their biophysical functions (Figure 3.3B). Comparing with the earlier development period (fetus to childhood), the percentages of differentially expressed genes in the later period (childhood to adult) are higher in cytoplasm (55% vs 44%) and nucleus (31% vs 25%), but lower in extracellular space (8% vs 10%) and plasma membrane (2% vs 10%).

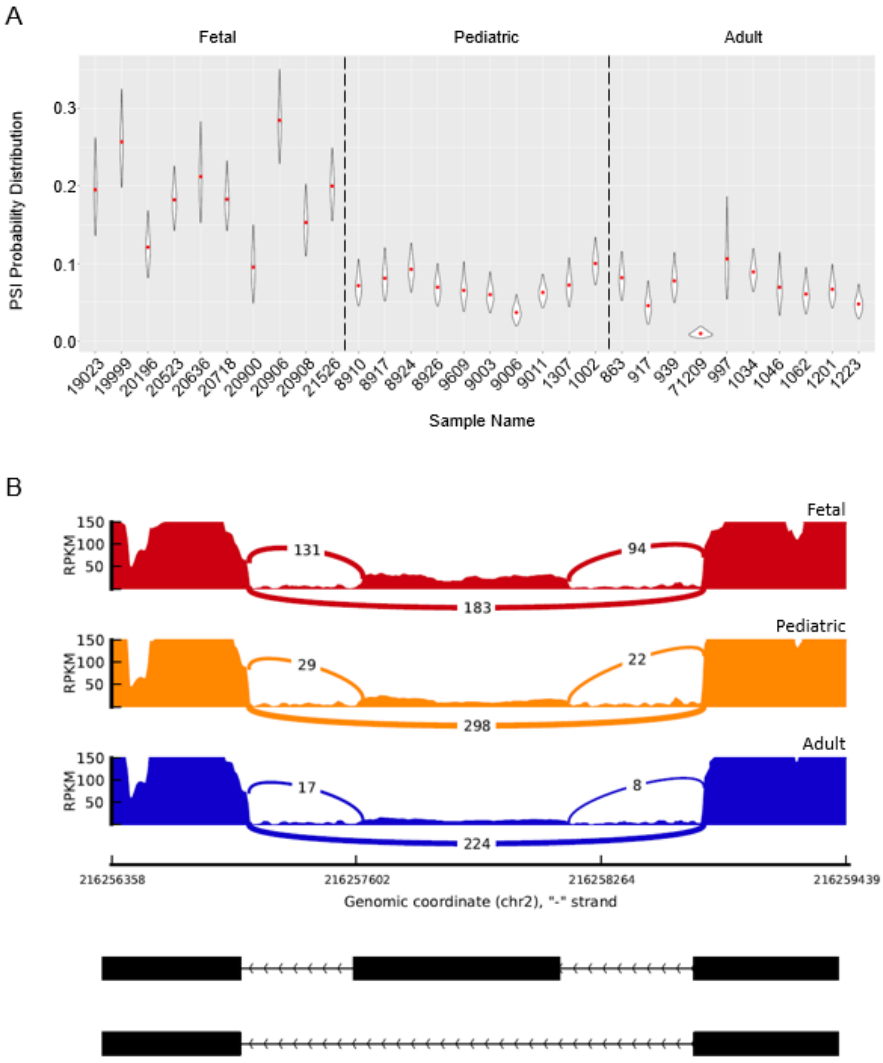

Figure 3.2 An alternative splicing event in fibronectin 1

(A) Ψ distributions of an AS event. The x-axis displays the names of samples. The y-axis denotes the probability distribution of Ψ. The mean of each probability distribution is marked as a red dot. (B) Sashimi plot of the same AS event. The sashimi-like color blocks denote expression intensity distributions of RNA-seq reads on genomic coordinates. The curves connecting color blocks show the numbers of exon-exon junction reads. The genomic structure of the AS event is displayed at the bottom, with black boxes as exons and straight lines as introns. The upper structure shows the transcript including the alternative region, and the lower one shows the transcript without the alternative region.

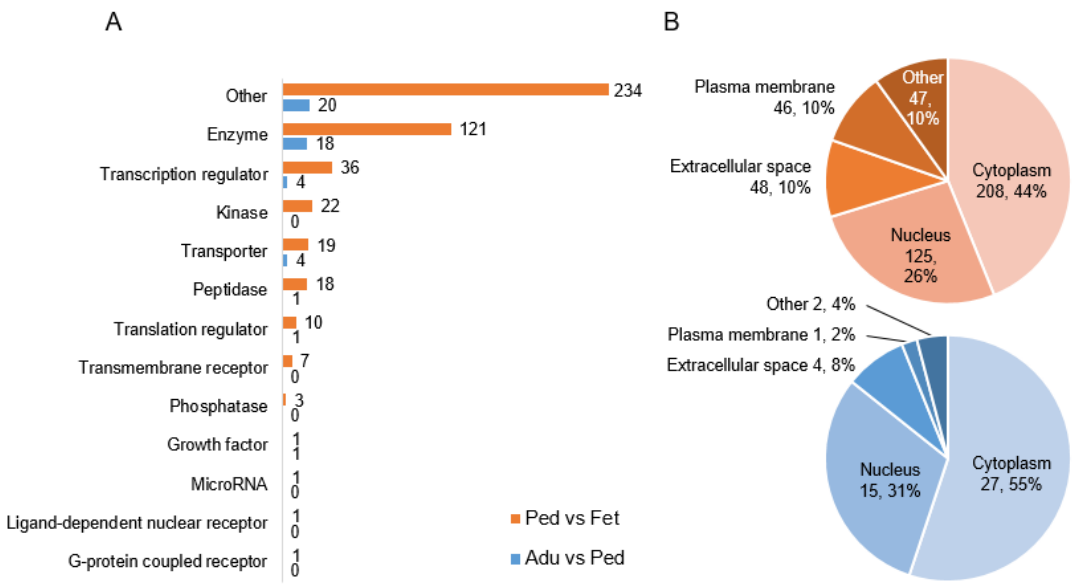

Figure 3.3 Functions and locations of AS events

(A) Cellular locations of differentially spliced events. (B) Function of differentially spliced events.

3.2.2 Transporters Associated with Diseases and Drug Metabolism

Transporter genes play an important role in liver physiology and drug metabolism [155-157]. During the fetus to childhood development, 19 splicing pattern changes took place in 16 transporter genes, among which 8 locate in cytoplasm, 5 locate on plasma membrane, 1 in extracellular space and 1 in nucleus (Table 3.2). ABCB6 is a porphyrin transporter, and it belongs to a protein family (ATP-binding cassette) known to be associated with drug resistances [158]. AP2M1 is a transporting vesicle coat component, which is involved in cargo selection and vesicle formation, but can also be hijacked by HPV for viral assembly [159]. Similarly, ARFGAP1, a protein required in vesicle and target compartment fusion, can also be hijacked by HPV to create a phosphatidylinositol 4-phosphate enriched microenvironment [160]. SEC14L2 is a lipid transporter and is reported to enhance vitamin E mediated inhibition of lipid peroxidation and thus promote HCV replication [161]. SLC37A4 transports glucose-6-phosphate into the lumen of

endoplasmic reticulum and inorganic phosphate to the opposite direction. It is reported to be associated with glycogen storage disease type Ib [162, 163]. SLC13A5 is a sodium-dependent citrate transporter that locates on plasma membrane. The intake of citrate facilitates synthesis of fatty acids and cholesterol. Its transcription can be stimulated by the drug rifampicin and induce lipid accumulation in hepatocytes [164]. SLC38A4 is a sodium-dependent amino acid transporter, and a recent genome-wide association study identifies it as risk loci for alcoholic hepatitis [165].

One gene AP1G1 is differentially spliced during the development from childhood to adult. This gene is an important component of clathrin-coated vesicles that transport ligand-receptor complexes from plasma membrane or trans-Golgi network to lysosomes. Mutations in this gene may induce multiple abnormalities during early development [166].

Comparison	Cellular Location	Gene Symbol	Gene Name
Ped vs Fet	Cytoplasm	ABCB6	ATP binding cassette subfamily B member 6 (Langereis blood group)
		AP2M1	Adaptor related protein complex 2 mu 1 subunit
		ARFGAP1	ADP ribosylation factor GTPase activating protein 1
		SEC14L2	SEC14-like lipid binding 2
		SLC25A16	Solute carrier family 25 (mitochondrial carrier), member 16
		SLC37A4	Solute carrier family 37 (glucose-6-phosphate transporter), member 4
		STAU1	Staufen double-stranded RNA binding protein 1
		USO1	USO1 vesicle transport factor
	Plasma Membrane	ATP11C	ATPase, class VI, type 11C
		SLC13A5	Solute carrier family 13 (sodium-dependent citrate transporter), member 5
		SLC38A4	Solute carrier family 38 member 4
		SLC39A7	Solute carrier family 39 (zinc transporter), member 7
		TFRC	Transferrin receptor
	Extracellular Space	APOC4	Apolipoprotein C-IV
	Nucleus	XPO7	Exportin 7
Adu vs Ped	Cytoplasm	AP1G1	Adaptor related protein complex 1 gamma 1 subunit

Table 3.2 Cellular location of transporter proteins

3.2.3 Cytochrome P450s

Cytochromes P450 proteins are a group of monooxygenases that catalyze a variety of reactions and play important roles in drug metabolism [167-169], and can be potential targets for personal drug treatments [170, 171]. We have found CYP2E1 and CYP3A5 differentially spliced during the fetus to childhood development, and CYP3A5 differentially spliced during the childhood to adult development. CYP3A5 is involved in the metabolism of several drugs, including irinotecan (CPT-11) [172], paclitaxel [173], cyclosporine [174, 175], tacrolimus [174, 176] and statins [177]. Two AS events containing premature stop codons change their splicing patterns in CYP3A5, and the dosage of the alternative region increases during the fetus to childhood development ($\Delta\Psi_1$=0.32, $\Delta\Psi_2$=0.25) and decreases during the childhood to adult development ($\Delta\Psi_1$=-0.17, $\Delta\Psi_2$=-0.19). The premature stop codons in these transcripts will result in truncated P450 domain in translated proteins, and thus change the percentage of functional P450 domains, which may affect drug response. CYP2E1 is known for its association with oxidative stress and drug toxicity [178, 179]. It significantly contributes to the formation of a toxin N-acetyl-p-benzoquinone imine (NAPQI), and is the major cause of liver necrosis during acetaminophen overdose [180]. Two SE events changed their splicing pattern during the fetus to childhood development ($\Delta\Psi_1$=0.11, $\Delta\Psi_2$=0.17). Both of the alternative exons are components of the P450 domain, but the length of neither exon can be evenly divided by 3, which means exclusion of these exons will induce frame shift and result in non-functional protein isoforms.

3.2.4 Potential Disease-causing Genes

Insertion and deletion of highly conservative and structured regions in proteins may change its function and behavior [181-183], and may cause diseases [184]. To further identify AS events with higher biological importance, ExonImpact [185] to calculate functional impact scores (FIS) for each AS region based on its PhyloP scores,

secondary structure, accessible surface area (ASA), disorderness, protein domain coverage and post translational modification (PTM) sites. With FIS cutoff 0.82 [185], an A3SS event and an SE event in PGS1 and MYL6 are identified as events with higher functional impact in fetus to childhood development and childhood to adult development, respectively. Both AS events are highly structured in the alternative regions with average structured probability scores 0.82 and 0.85, and average disorder probability both as 0. PGS1 is a gene that functions in the cardioipin and glycerrophospholipid biosynthesis pathways. The A3SS region locates in the protein coding region, and down regulated 25.0% from fetus to childhood. MYL6 is a gene in the myosin family and functions as molecular motors. The alternative exon in MYL6 encodes a part of the EF-hand 6, a protein domain that interacts with Ca^{2+} ions [186]. The SE region in this gene down regulated 14.8% from childhood to adult.

3.2.5 PPIs are Developmentally Regulated through Alternative Splicing

To examine whether developmentally regulated alternative splicing events modulate protein-protein interactions (PPI), we searched for interactions between the differentially spliced protein domains and their binding partners based on two criteria: (1) at least one yeast two-hybrid study [104, 187, 188] supports the direct interaction between the pair of interacting proteins and (2) at least one structural study in the Protein Data Bank (PDB) supporting direct interaction between the same pair of proteins (See Methods).

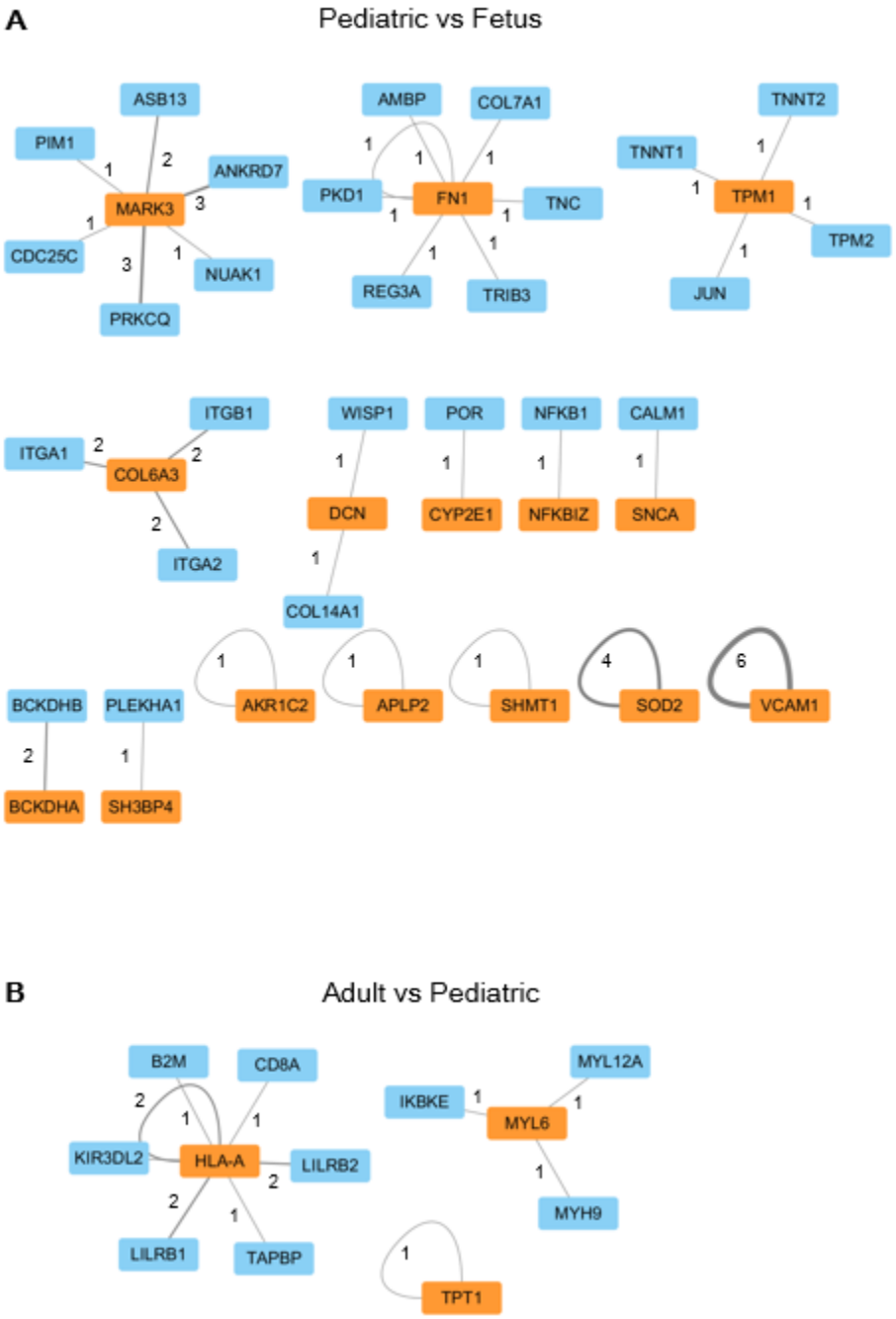

Figure 3.4 Protein-protein interactions that may be disrupted by splicing change

The orange nodes indicate differentially spliced genes during development, and the blue nodes indicate binding partners that interacts with the alternatively spliced protein domains of the genes in the previous category. The numbers and the width of the edges indicate the number of interacting domains in the partner protein.

For the developmental stages from fetus to childhood, we identified 32 PPIs, including 15 genes of which the splicing is developmentally regulated and 26 other binding partner proteins, as well as 6 self-interactions. For the childhood to adult development, we identified 11 PPIs, involving 3 genes that contains developmentally regulated AS events and 9 binding partner proteins, and 2 self-interactions (Figure 3.4). The interaction between HLA-A (major histocompatibility complex, class I, A) and KIR3DL2 (killer cell immunoglobulin like receptor, three Ig domains and long cytoplasmic tail 2) is a part of the natural killer cell mediated cytotoxity pathway (KEGG ID: map04650). This indicates this pathway may be affected by the developmental regulation of HLA-A's splicing.

3.3 Discussion

Liver is an important organ where most drug metabolism take place. Each step in drug metabolism could potentially be affected by differences in gene variation, expression and RNA splicing. While gene variations in somatic cells mostly stay unchanged during the lifespan of human, gene expression and RNA splicing changes over time. In this study we analyzed the RNA-seq result of fetus, pediatric and adult stages, and identified 477 and 49 AS events during the fetus to childhood development and childhood to adult development respectively. The overt discrepancy in the number of differentially spliced AS events in these two developmental periods implies much fewer tramscriptome activity changes in the development after childhood comparing to that before childhood. We found the splicing of drug metabolism genes, including transcriptors, cytochrome P450s, and two potential disease causing genes (PGS1 and MYL6) are developmentally regulated in liver. This implies that AS is one of the mechanisms that cells employ to regulate drug metabolism and other biophysical functions during development. We also found 32 and 11 protein binding domains disrupted by developmentally regulated AS during the fetus to childhood and childhood

to adult development, respectively. One among these PPIs can be mapped to a known natural killer cell mediated cytotoxity pathway in KEGG, this demonstrates that developmentally regulated AS events can exert their biophysical functions through regulating PPIs in cellular pathways.

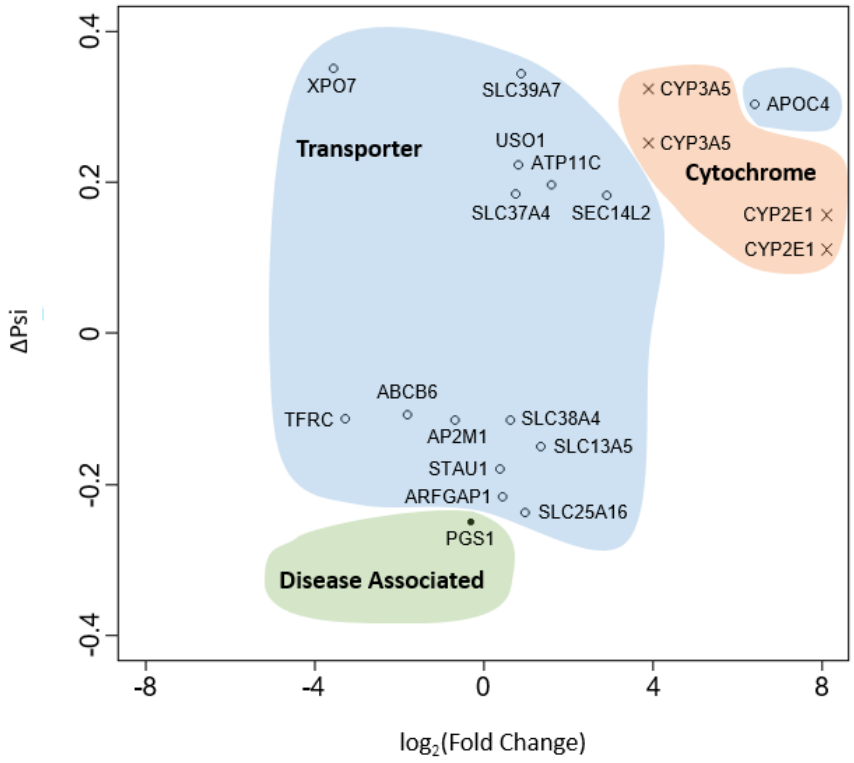

Figure 3.5 Differentially spliced genes that are essential to liver functions during fetus to childhood development

Each data points represents an AS event, and the gene name is labeled next to data points. The x-axis is the logarithm of gene fold change, while the y-axis is ΔΨ, the change of AS.

While AS is an important factor that regulates liver functions, gene expression could also be a regulator from another dimension. To investigate the co-occurrence of developmentally regulated AS and gene expression, we visualized the ΔΨ and fold change scatter plot for functionally important AS genes that are differentially spliced

during the fetus to child development (Figure 3.5). For transporter genes, the ΔΨ range from -0.25 to 0.35, and the log2-fold-change of gene expression range from -3.5 to 6.4, which demonstrate evenly scattered pattern on both AS and gene expression dimensions. The ΔΨ of cytochrome P450s range from 0.11 to 0.25, which indicates an increase of the alternative region in childhood comparing to fetus. The log2-fold-change of gene expression in the same developmental period range from 3.90 to 8.13, showing that CYP3A5 and CYP2E1 are both up regulated in childhood comparing with fetal stage. The disease associated gene PGS1 has a ΔΨ of -0.24, indicating the percentage of the mRNA transcripts containing the alternative region down regulated 24% during the fetus to childhood development. Its log2-fold-change of gene expression is -0.3, which is not as a dramatic change as the cytochrome P450 genes and some of the transporters. Interestingly, the three main categories of genes, transporters, cytochrome P450s and predicted disease associated genes, can be clearly separated according to their ΔΨ and log2-fold-change values, except APOC4, an outlier of the transporter category. This distinct difference among gene categories may be induced by shared regulatory mechanisms within categories, or shared evolutionary pressure among genes with similar functions.

To investigate which RNA binding proteins (RBP) may have contributed to developmentally regulate AS, we implemented an RBP motif analysis on 303 SE events that are differentially spliced during fetus to childhood development. Seven regulatory regions were defined for each SE event (Figure 3.6). Region 1 and 7 were defined as the 150 nucleotides long sequence in the exon starting from the splicing site, region 2, 3, 5 and 6 were defined as 300 nucleotides long sequence in the intron starting from the splicing site, and region 4 was the full length of the alternative exon. We retrieved the genomic sequence of these 7 region from each SE event, and searched for RBP motifs against the RBP motifs from the CISBP-RNA database [189] with FIMO [110]. We found

6 RBPs contributing to the upregulation of SE events. Among these RBPs, HNRNPC, PABPC1 and SART3 are known as AS regulators. The motif “ATTTTTG” is a potential intronic splicing enhancer (ISE), “AGAAAAA” and “AAAAAAA” are potential exonic splicing enhancers (ESE) (Figure 3.6A). On the other hand, all 6 RBPs contributing to the downregulation of SE events are known as AS regulators. Motifs “TTTTTTC”, “GGGAGGC”, “GGGAGGA”, “TGTGTGT”, “GTGTGTG” and “GGGTGTG” are potential intronic splicing silencers (ISS), and “GAAGGAA” is a potential exonic splicing silencer (ESS) (Figure 3.6A).

To identify whether an RBP is associated with the up-regulation or down-regulation of a cassette exon during development, we implemented a principal component analysis on the percentage of regulated cassette exons among four categories of cassette exon events, including the up-regulated or down-regulated during the fetus to childhood development and the fetus to adulthood development (Ped vs Fet Up, Ped vs Fet Down, Adu vs Fet Up and Adu vs Fet Down). Figure 3.6B demonstrates these RBPs clearly separated into two groups. In the figure, each dot represents an RBP. The group on the left (colored orange) represents the RBPs associated with down regulated cassette exons during development, while the group on the right (colored cyan) represents the RBPs associated with up regulated cassette exons. It is important to clarify that an RBP’s association of up or down regulation does not lead to the interpretation of an enhancing or repressing function, because it may also be achieved by the inhibition or activation of a repressing RBP.

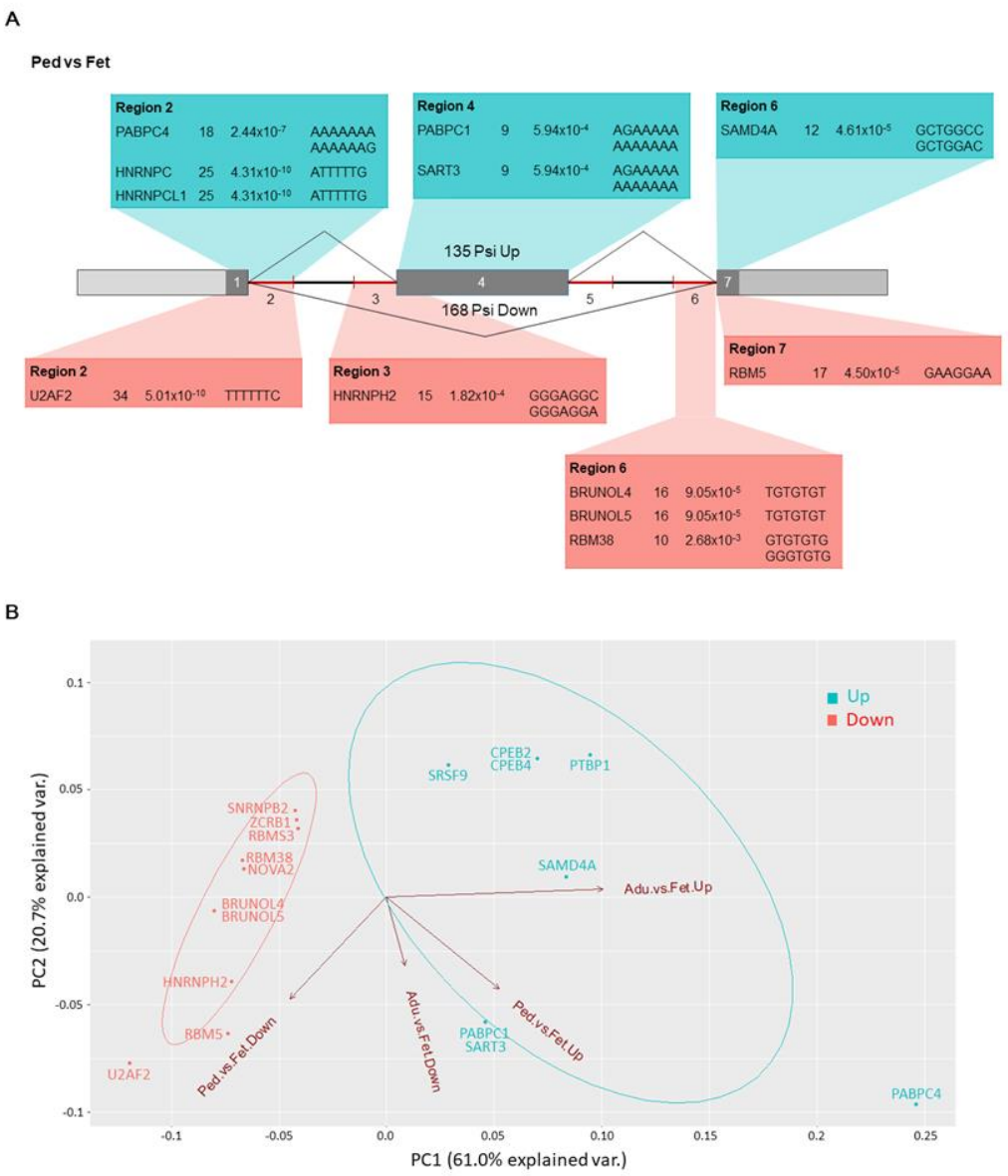

Figure 3.6 RBP and RBP binding motifs

(A) Regulatory motifs of SE events. For each RBP row, the values are RBP name, number of motif matches, p-value and motif sequence. The green tables in the upper half of the figure show potential splicing enhancer motifs, and the orange tables in the lower half show potential splicing silencers motifs. (B) RBPs associated with up-regulated and down-regulated cassette exons are separated in principle component analysis.

3.4 Methods

3.4.1 Bootstrap Approach for Differential Alternative Splicing Detection

The MISO Analysis (MISO Run) component calculates sample-wise Ψ probability distributions for each AS event, and these distributions are distinct from one sample to another. This renders it impractical to apply classical statistical tests (e.g. Student T-test or Friedman test), which assumes a uniformed type of distribution across all samples, to detect differentially spliced events.

The MISO [97] algorithm has a built-in component - MISO Comparison that compares the Ψ distribution of a pair of samples and calculates a Bayes Factor (BF) to describe the likelihood of splicing change. To implement group-wise comparisons on multiple samples with MISO, users have to merge all samples in each group and then apply MISO Comparison on the merged Binary Sequence Alignment/Map (BAM) files. This method overlooks within group variance and the difference of the Ψ probability distribution across samples. On the other hand, MATS [146] is capable to implement group-wise comparison without sample merging, however it uses only junction reads to evaluate Ψ, which misses valuable information in reads mapped to alternative regions and constitutive exons. To combine the advantages of MISO and MATS, we developed a Bootstrap approach to estimate the significance of AS change.

The Bootstrap approach consists of two steps. The first step is to calculate the weighted mean on Ψ values randomly sampled from the probability distributions from each group (Fetal, Pediatric or Adult), then calculate the difference between mean Ψ values (ΔΨ) across two groups. Let the number of samples in two comparing groups be *m* and *n*, and let *i* be the current round of iteration and *e* be the number of the current AS event, then we can derive ΔΨ for event e in iteration *i* with the following equation:

$$\Delta\Psi_{ei} = \sum_{j=1}^{m} W_{ej}\Psi_{eji} - \sum_{k=1}^{n} W_{ek}\Psi_{eki}$$

W denotes the weight of each Ψ value, and it is calculated with the following equation:

$$W_{es} = 1 - C_{es}$$

C_{es} is the width of the 95% confidence interval of the Ψ probability distribution of event *e* in sample *s*.

The second step is to iterate the first step for a certain amount of times (10,000 times in this work) and calculate the p-value based on the alternative hypothesis that the mean of ΔΨ is not 0:

$$p = \frac{N_{Tail}}{N_{All}}$$

Assuming most of the iterations generated ΔΨ larger than 0, then N_{Tail} is the number of iterations that generated ΔΨ less than 0, and vice versa. N_{All} is the total number of iterations. False discovery rates (FDR) are then calculated based on p-values with the Benjamini-Hochberg approach [190]. AS events with $|\Delta\Psi| \geq 0.1$ and $FDR \leq 0.05$ are considered as differentially spliced.

3.4.2 Identification of PPI Affected by Differential Splicing

We examined protein-protein interactions (PPI) potentially affected by differential splicing with both structural and experimental evidences. We firstly identified Pfam [191] domains within or overlapped with alternative regions, then mapped them to binding partner domains with iPfam [103], which documents domain-domain interactions found in Protein Data Bank (PDB) [37]. These PPIs are then verified with a non-redundant PPI dataset (172,911 protein pairs) derived from three yeast two-hybrid screening datasets [104, 187, 188].

Chapter 4. Novel Alternative Splicing Events in Transcriptome

4.1 Background

Alternative splicing is an important level of gene regulation that greatly contributes to proteome diversity [192]. It enables one gene to produce multiple isoforms that can have different biological functions. In humans, more than 90% of genes encode multiple protein isoforms [183], and many diseases are caused by the dysregulation of splicing patterns [193]. Traditionally, EST (Expressed Sequence Tags) databases and microarray technologies have been utilized to study splicing regulation [194-197]. In recent years, high-throughput RNA sequencing (RNA-seq) technology has revolutionized functional genomics by offering the most comprehensive and accurate measurements of RNAs. In addition to previously known splicing events, RNA-seq technology can be used to identify novel splicing events.

Many bioinformatics tools have been developed to derive splicing patterns from RNA-seq data. For instance, dozens of strategies have been designed for aligning RNA-seq reads. Using various strategies, such tools, including TopHat [30], MMES [198], SpliceMap [199], SplitSeek [200], G-Mo-R-Se [201], GSNAP [202] and SAW [203], enable alignment of short sequencing reads over splice junction sites even across large intronic regions. Based on such splicing-sensitive alignments, follow-up algorithms, such as Cufflinks [204] and Scripture [205] have been developed to reconstruct transcript isoforms using a genome-guided approach. Although the idea of reconstructing the whole transcriptome is intriguing, a quantitative estimate of the expression levels of each isoform is difficult, particularly for transcripts expressed at low levels and/or when more than a few isoforms exist. In addition, isoform-based approaches increase the complexity of studying splicing regulation when many isoforms are present in the sample. Event-based approaches, however, only focus on the inclusion and exclusion of individual splicing events, regardless of membership in different isoforms. This greatly

reduces the computational complexity, and offers a direct path for studying splicing regulation. Based on the sequencing reads supporting inclusion and exclusion events, MISO (mixture of isoforms) [42] is designed to estimate the percentage of inclusion for every previously documented alternative-splicing event in a sample. It further offers a probabilistic framework for detecting differentially regulated exons, and provides functional insights into pre-mRNA processing.

One requirement for implementing MISO is to provide a pre-defined alternative event annotation. Such an annotation heavily relies on previous knowledge, and is not complete or even available for many species. For instance, in the official MISO release, alternative splicing annotation library [42] is only available for human, mouse, and Drosophila genomes, and does not allow event-based analysis on datasets from other species. In addition, even for the species whose alternative splicing has been heavily investigated, identifying novel splicing events can be important. Therefore, having a tool for detecting novel splicing events directly from RNA-seq data is desirable.

In this study, we developed a tool, Alt Event Finder, for generating de novo annotation for alternative splicing events from a map of transcripts and isoforms reconstructed from RNA-seq experiments. In conjunction with upstream alignment and isoform reconstruction tools, we demonstrated that Alt Event Finder has the ability to identify novel cassette exon events that are not documented in the established databases. We evaluated the performance of this strategy with different combinations of alignment and transcript reconstruction algorithms, using a human dataset where alternative splicing events have been extensively investigated. We further implemented this tool on an RNA-seq dataset from rat genome, for which alternative-splicing annotation is not available.

4.2 Results

4.2.1 Workflow

As shown in Figure 4.1A, the input to Alt Event Finder is a mixture of RNA isoforms identified from transcriptome reconstruction tools, such as Cufflinks [204] or Scripture [205]. The output is a list of alternative splicing events directly derived from isoform annotation. Alt Event Finder includes two major steps. First, based on a GTF or BED file for isoform annotation, the unions of the exon regions are split into the smallest units that do not overlap with each other in the genome space, or minimum non-overlapping exon units. This design is similar to the PSR (probe selection regions) definition for Affymetrix exon arrays [206], and can reflect the complexity of the exon structures where alternatively spliced exons from the same gene may overlap (i.e. alternative donor or acceptor sites). Second, individual transcript isoforms (identified from transcriptome reconstruction tools such as Cufflinks and Scripture) will be projected to the non-overlapping exon units (Figure 4.1B). The number of isoforms containing each unit is recorded. Special strings of such numeric patterns will be used for deriving different types of splicing events. For instance, for a gene with two isoforms, a string of [2-0-1-0-2] indicates the presence of a cassette exon. Although this report focuses on cassette exons, such simple design allows extension to other types of events easily.

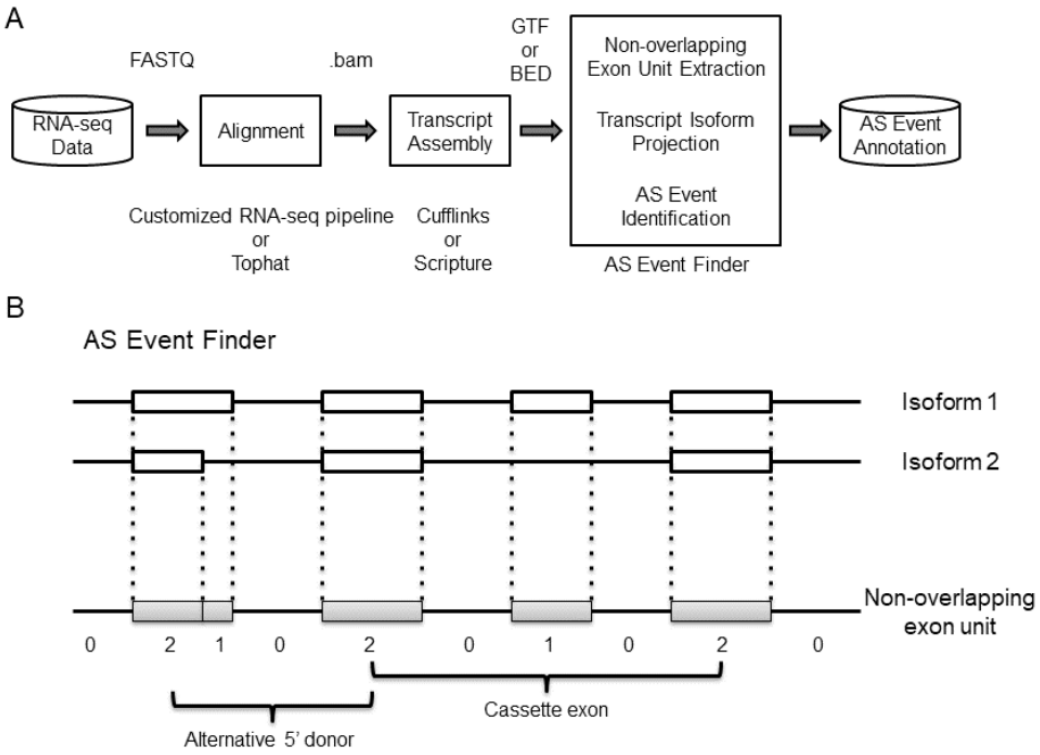

Figure 4.1 Workflow of the alternative splicing event identification pipeline

(A) RNA-seq-derived transcriptome data was aligned using a customized RNA-seq pipeline based on known splicing junctions or Tophat. Cufflinks or Scripture was used for isoform annotation. Based on the data-derived transcript annotation, Alt Event Finder was applied to identify the novel alternative events. (B) Strategies of Alt Event Finder for de novo event detection.

4.2.2 Alternative Splicing Event Annotations from Human Liver Data

To test the performance of our strategy, we implemented Alt Event Finder on a RNA-seq dataset derived from human primary hepatocytes; the RNA-seq experiment was conducted using the SOLiD 5500xl system (Life Technologies). The dataset consists of 7 pairs of samples derived from 7 individuals. Each pair includes a drug exposed sample and a control sample. To test the performance of Alt Event Finder on data with various sequencing depths, in addition to the 14 RNA-seq samples, we created 7 patient-specific

datasets by merging the exposed and control samples from the same individual; and 1 hepatocyte-specific dataset by merging all the 14 samples together.

We used BFAST [207] as the primary aligner of short reads, due to its higher sensitivity on color-space data [208]. The alignment was conducted on both genomic DNA sequences and a junction library including all the combinations of known junction boundaries (within a 100 kb span) annotated in the UCSC Gene database. The total number of mappable reads in each sample ranged from 6.6 million to 19.5 million. We then used Cufflinks [204] to reconstruct transcript isoforms. From that, we applied Alt Event Finder. The number of identified alternative splicing events increased as a function of depth of coverage (Figure 4.2A); events ranged from 433 to 1,049 in individual samples, from 761 to 1,298 for patient-specific datasets (combining control and treated data), and was 1,771 for all the samples combined.

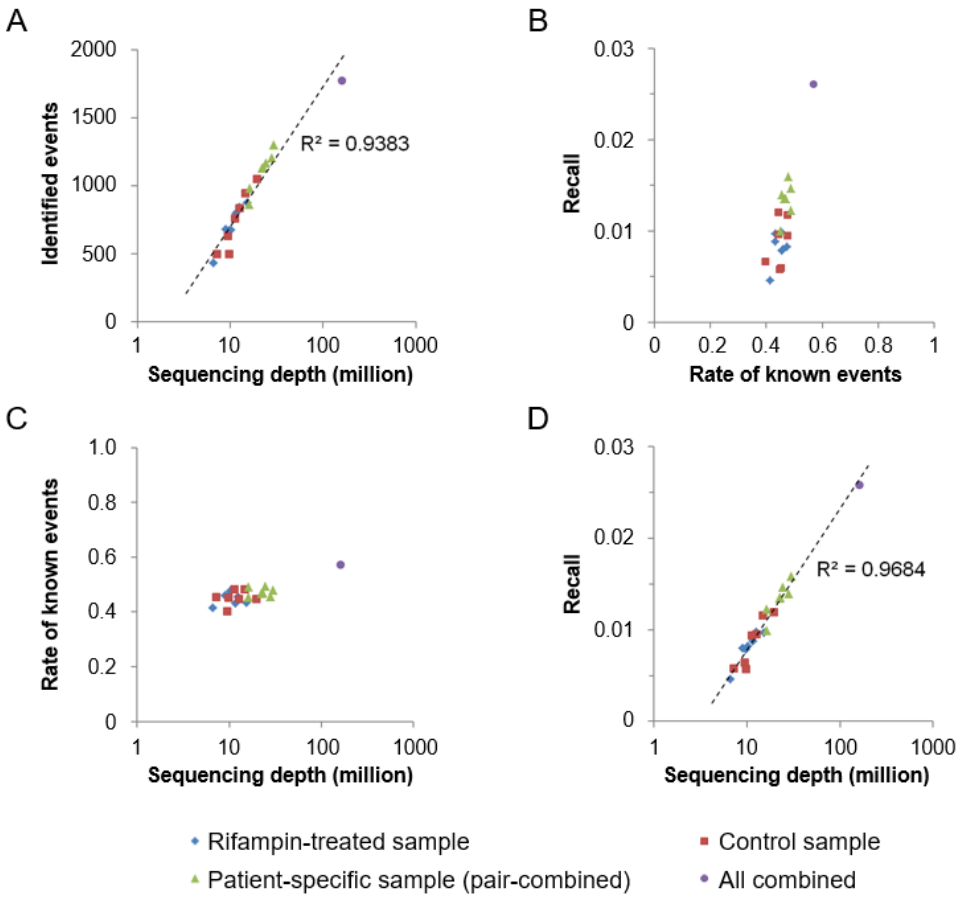

Figure 4.2 Performance assessment for the Alt Event Finder

(A) The total number of identified events increases with sequencing depth. Each dot corresponds to a sample; the color and the shape of the sample denote its biological condition; the regression line is displayed in dashed line with an R-squared value of 0.9383; (B) Performance for the Alt Event Finder pairing with a customized alignment pipeline and Cufflinks. The X-axis (rate of known events) is defined by the number of overlapping events divided by the number of data-driven events, and Y-axis (recall) is defined as the number of overlapping events divided by the number of events in the official MISO annotation library; (C) The rate of known events does not change with sequencing depth; (D) The recall rate of the Alt Event Finder increase linearly with the logarithmic transformation of the total mappable reads. The regression line is displayed in dashed line with an R-squared value of 0.9684.

To evaluate the performance of the proposed strategy, we compared our data-derived events with known alternative splicing events documented in the MISO release (based on UCSC hg19 assembly) [42]. For each sample, we calculated a rate of known events (RKE), which measures the percentage of identified events that were in the known splicing events annotation, and a recall value, which was calculated as the percentage of known splicing events that were recovered by our strategy. As shown in Figure 4.2B, the rate of known events varies from 0.4 to 0.57. This indicates that a significant portion of splicing events we detected was not documented in the current database, although junction reads were found in support of their existence. The recall values, however, are low, ranging from 0.004 to 0.025. This is not surprising since the known event annotation aims at completeness, and therefore documents events from many tissues with a variety of biological conditions; most of these events should not be present in one tissue under one or two biological conditions. We further evaluated the relationship between sequence depth and rate of known events (Figure 4.2C) and recall values (Figure 4.2D). Rate of known events do not show apparent changes, suggesting that the genes expressed at lower levels contain a similar percentage of novel events as the more abundant transcripts, but they require greater sequencing depth to identify. The recall, however, increases almost linearly with logarithmic transformation of the total number of mappable reads. These results (Figure 4.2C and D) indicate that many more events will be identified with deeper sequenced samples, while the percentage of novel events doesn’t change. Therefore, more novel events will be identified from deeper sequenced data.

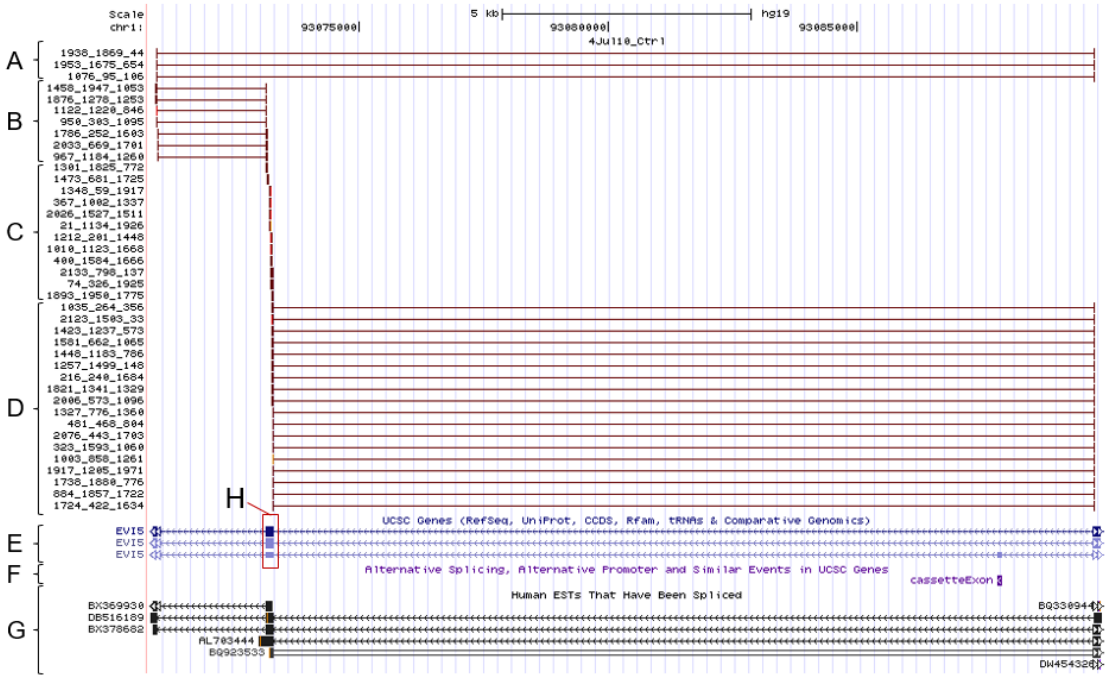

Figure 4.3 Screenshot of one of the novel cassette exon events

The diagram demonstrates reads supporting a cassette exon event that was not previously documented. This gene locates on the reverse strand. The genomic loci of the cassette exon, the 5' and 3' constitutive exon are chr1: 93073138-93073284, chr1: 93089733-93089891, and chr1: 93070864-93070959, respectively. (A) Sequencing reads supporting the junction of the 5' and 3' constitutive exons, which indicates exon exclusion; (B) Sequencing reads supporting the junction of the cassette exon and the 3' constitutive exon; (C) Sequencing reads supporting the transcription of the cassette exon; (D) Sequencing reads supporting the junction of the 5' constitutive exon and the cassette exon; (E) UCSC gene annotation track; (F) Alternative splicing annotation track; (G) Human EST track; H) The cassette exon in UCSC gene annotations. No previous evidence of this cassette exon event was shown in the gene annotations, alternative splicing event annotations or the EST records.

Figure 4.3 is the screen shot of one of the novel cassette exons (chr1: 93073138-93073284) not documented in either the official MISO annotation library (based on UCSC hg19 assembly) [42] or the Alt Event track (Figure 4.3F) in the UCSC Genome Browser (GRCh37/hg19, Feb. 2009) [209]. As shown in the figure, 40 reads are identified around this exon (Figure 4.3ABCD), of which 37 (Figure 4.3BCD) support inclusion events (exonic reads on the alternative exon, and junction reads connecting the upstream or downstream exon with the alternative exon), and 3 (Figure 4.3A) support exclusion events (reads connecting upstream and downstream exons directly), respectively. Importantly, the presence of 28 exclusive junction reads provides a strong evidence for the presence of this novel event.

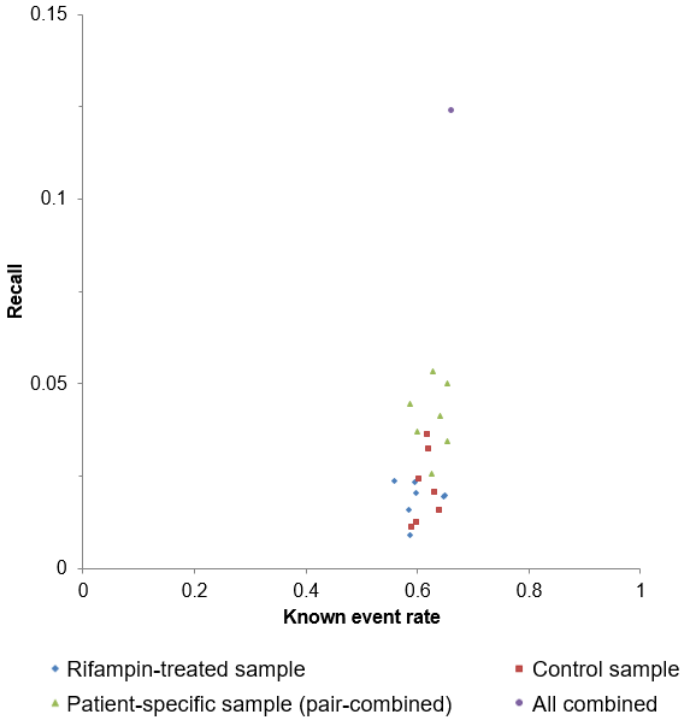

Figure 4.4 Performance with adjusted known event annotation

The rate of known events and recall values were calculated only based on the events that have at least 10 junction reads supporting the inclusive event and 1 junction read supporting the exclusive event.

Due to the tissue specific nature of gene expression and alternative splicing, using all the known events in the human genome cannot fairly evaluate the sensitivity of the proposed approach. This is either due to the absence of certain isoforms in a specific tissue, hepatocytes in this case, or because the overall gene expression levels are too low to be detected given a specific sequencing coverage. We therefore removed the events either with low expression levels, or with extremely unbalanced inclusion/exclusion ratio, from the overall alt event library. For the latter, it is possible that the RNA-seq data can only detect the isoforms with inclusion or exclusion events, but not both. To fairly evaluate the performance, we derived exon inclusion and exclusion ratios using MISO, based on all the annotated splicing events. We further filtered the annotated events containing no less than 10 reads supporting inclusion and 1 read supporting exclusion. After applying this filtering, for the hepatocyte-specific sample (merging reads from all the 14 samples), 83.4% of the 39,232 total annotated cassette exon events were removed. The adjusted recall rate is shown in Figure 4.4. Clearly, for individual samples, the recall remains low, ranging from 0.9% to 3.6%. This number increased for patient-specific samples (merging control and drug treatment for one individual), ranges from 2.6% to 5.0%, and 12.4% for all the 14 samples combined. This low recall rate may be due to the stringent threshold of Cufflinks, which aims at maximizing specificity.

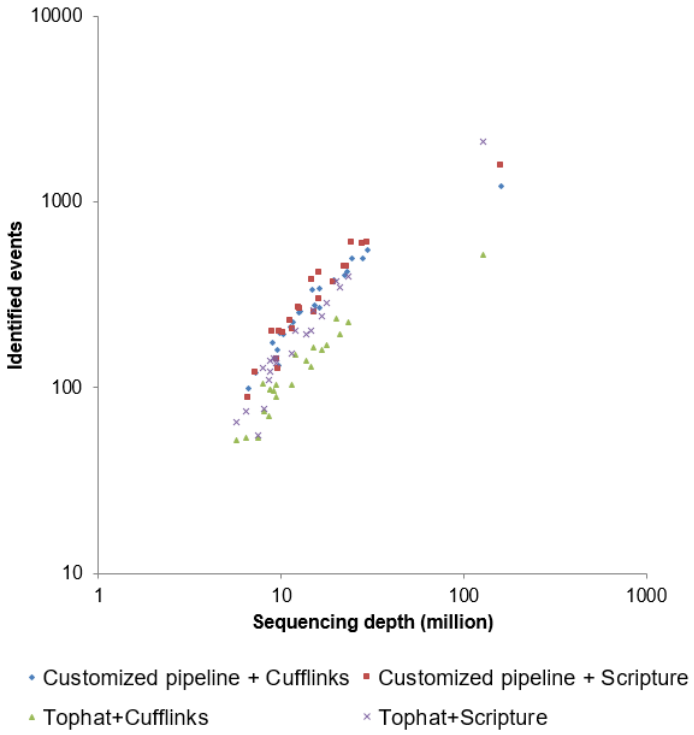

Figure 4.5 The number of identified events differs with different combinations of alignment and transcript reconstruction algorithms

Each dot describes a sample. X and Y axes denotes sequencing depth and the total number of identified events. Samples were color-coded based on their combinations of upstream algorithms.

4.2.3 Selection of Alignment and Transcriptome Reconstruction Tools

We further evaluated how the performance of the Alt Event Finder is influenced by the alignment and transcriptome reconstruction tools. For the alignment tool, in addition to our customized RNA-seq pipeline which focus on known splicing junctions, we also tested TopHat [30], one of the most widely used RNA-seq alignment software. For the transcriptome reconstruction tool, in addition to Cufflinks [204], which aims at maximizing specificity, we have also tested Scripture [205], a computational algorithm aiming at higher sensitivity.

The total number of events identified based on 4 different strategies (Customized RNA-seq pipeline and Cufflinks, Customized RNA-seq pipeline and Scripture, Tophat and Cufflinks, and Tophat and Scripture) varies significantly (Figure 4.5). At low sequencing coverage, the customized RNA-seq pipeline (using BFAST and annotated exon boundaries) consistently identified more events. When the sequencing depth is higher than 100 million reads, however, our AS identification pipeline offers significantly more events when Tophat is partnering with Scripture (Figure 4.5). When comparing two transcriptome reconstruction tools, Scripture offers higher number of events regardless of the sequencing depth and sequencing alignment algorithm (Figure 4.5). Among all the four strategies, the combination of Tophat and Scripture at high sequencing coverage identified highest number of events.

4.2.4 Identify Alternative Splicing Events in the Rat Genome

We applied Alt Event Finder to study the alcohol-induced alternative splicing changes in liver tissue, using alcohol-preferring rats as a model system. Seven female rats were heavily exposed to alcohol for 10 weeks followed by 2 weeks without alcohol, and another 7 were not subjected to alcohol exposure (controls). An RNA-seq experiment was conducted on the liver tissues. After sequence alignment using TopHat, 123,017,701 and 92,389,972 total reads were mapped in the 7 control and 7 alcohol-exposed animals, respectively. Scripture was used for transcript reconstruction. Alt Event Finder identified 505 candidate events with a mixture of multiple isoforms in the combined sample of all 14 rats. With a MISO isoform differential expression test, we found 75 were alternatively spliced at Bayesian Factor (BF) [42] larger than 2; this number implies that it is twice as likely for the events to be alternatively spliced than not. A more stringent cutoff derived 55 events with BF>5.

Figure 4.6 shows the Sashimi plots [42] for three events with apparent alcohol-induced splicing changes in genes highly expressed in liver tissues, LOC691397

(similar to PI-3-kinase-related kinase SMG-1), Glycerol kinase, and CD47 (Figure 4.6A). For LOC691397, 40 junction-reads support exon inclusion in the control samples, and 17 support exclusion. In the alcohol exposed samples, however, these numbers changed to 8 and 15, respectively. This pattern indicates that chronic alcohol exposure induces higher relative expression levels of the isoforms without the cassette exon, with a BF value 15.37. Similarly, CD47 showed lower inclusion ratio after alcohol exposure (Figure 4.6C), while alcohol drinking induces exon inclusion for the glycerol kinase (Figure 4.6B).

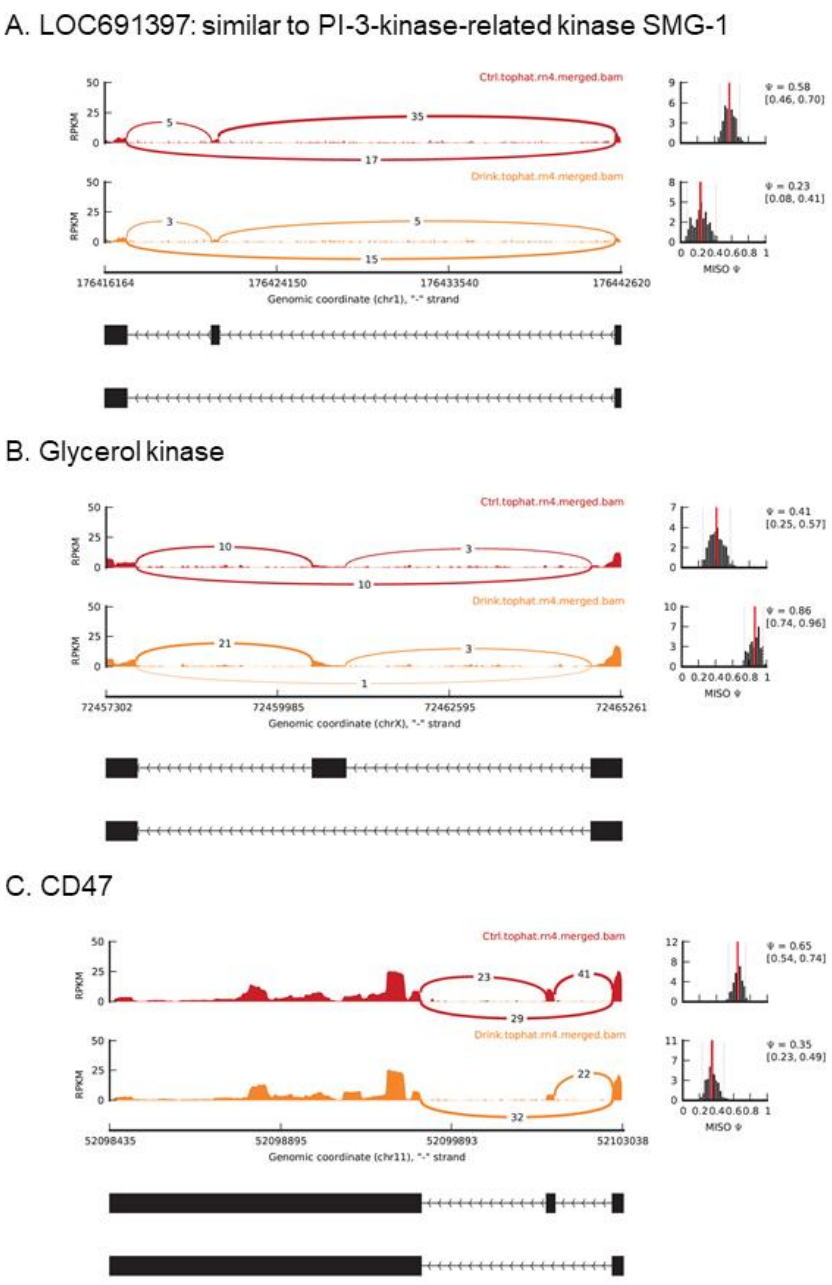

Figure 4.6 Sashimi plot for three novel events that are alternatively spliced in rat liver with chronic alcohol exposure.

The RNA-seq read densities supporting inclusion and exclusion events are shown in the figure. The estimated percentage of inclusion for the alternative events and their estimated confidence intervals are also demonstrated. The Sashimi plot is produced by the MISO package.

4.3 Discussion

In this study, we developed a tool, Alt Event Finder, which generates splicing event annotations from RNA-seq data. Most event-based analysis, such as MISO [42], cannot work without a library of known event annotations. Therefore they cannot be implemented on a genome for which annotation is unavailable, such as the rat genome. Even for a genome for which alternative splicing has been extensively studied, such as human or mouse, lack of a de novo event finding tool limits the power of studying events that are not previously documented. Alt Event Finder bridges the gap between event-based analysis and isoform-based transcriptome reconstruction algorithms, such as Cufflinks and Scripture. It's an important addition to the current AS analysis toolset.

Our algorithm extracts "minimum non-overlapping exon units" (Figure 4.1B) from RNA-seq-derived transcript isoform annotation based on Cufflinks or Scripture, and further identifies potential alternative events. This strategy greatly increases the flexibility of our methods. Although the current study focuses on cassette exon, it can be easily extended for other types splicing events, such as intron retention, alternative 5' donor, alternative 3' acceptor, and so on. This is important because certain types of events can be more prevalent in specific tissue types. For instance, cassette exons are dominant in brain tissues, while alternative 5' donor and 3' acceptor events are more abundant in liver tissues [210].

Alt Event Finder relies on upstream alignment and isoform reconstruction tools. We have evaluated how different tool combinations affect the ability to discover novel splicing events. We found that a customized alignment pipeline based on known exon boundaries perform better in low sequencing coverage (<100 Million reads), while TopHat did better for high sequencing coverage. This is because TopHat derives exon structures mainly based on the accumulation of RNA sequencing reads. Since it does not rely on existing exon annotations, at lower coverage, the data may not have adequate

power to properly identify low expressed exons. For higher coverage, however, TopHat will not only have enough power to precisely map the boundaries of known exons, but also be more suitable for identifying novel exons. We have also found that we can generally identify more AS events using Scripture as isoform reconstruction tool, compared to using Cufflinks, because Scripture aims at maximizing sensitivity, while Cufflinks aims at specificity. Overall, we recommend using the mapping algorithm based on known exon annotation and Scripture combination at a low sequencing depth, and the TopHat and Scripture strategy with high sequencing coverage.

To find out the cause of the low recall rate, we investigated the AS events that were identified with the official MISO library but not found in our annotations. One of the major causes of such events is lack of junction reads between the cassette exon and constitutive exons, which makes the inclusive isoform not detectable by Cufflinks and Scripture, but still quantifiable by MISO since reads are covering the cassette exon. Another cause is additional alternative spliced 3' and 5' sites on a cassette exon event, which make an event in our annotation different from the official MISO annotation.

Since Alt Event Finder is a data-driven approach, its power highly depends on the sequencing depth. When the sequencing depth is low, a lot of junction read will be missed, and a lot of low expressed exons could be "disconnected"; this will significantly decrease the power of the transcriptome reconstruction algorithm for rebuilding the isoforms from RNA-seq data, therefore affect the performance of Alt Event Finder. Therefore, when possible, increasing the sequencing depth can significantly elevate the power of novel event identification.

When deep sequencing data is not available, at the de novo event identification step, we recommend pooling sequencing reads from all the samples. This will enable identification of the events that lowly expressed in individual samples. It will also enable us to identify the events that have complete inclusion in one condition, but exclusion in

another. These events cannot be identified within individual samples, but the inclusion/exclusion switches are enormously interesting.

4.4 Methods

4.4.1 Dataset

RNA-seq dataset. We used two RNA-seq datasets for de novo alternative splicing event identification, human hepatocytes and rat liver cells. In the human study, primary hepatocytes were isolated from seven individual subjects, and treated with Rifampin. Total RNA from both control and treated samples were extracted. RNA-seq experiments were conducted using the SOLiD 5500xl system with the standard protocol. In the rat study, RNA-seq experiments were conducted on liver tissues from 7 non-drinking alcohol-preferring rats, and 7 alcohol-preferring rats that were heavily exposed to alcohol for 10 weeks followed by 2 weeks without alcohol. The experiment was conducted on the SOLiD 4 system with the standard protocol.

Known splicing event annotation. The known alternative splicing event annotation for human genome was retrieved from the official MISO library (based on UCSC hg19 assembly). The annotation file was generated based on transcript annotation using an EST database; a splicing event was considered alternative if it was supported by multiple ESTs.

4.4.2 RNA-seq Alignment

We used two RNA-seq alignment pipelines, TopHat [30] and a customized strategy using BFAST [207] as primary aligner and known splicing sites documented in UCSC Known Gene database [211]. TopHat v1.4.0 was used with standard parameter settings on color space data. The customized pipeline uses BFAST [207] as a primary aligner due to its computability with small insertions/deletions, and reported higher sensitivity on color space data [208]. The overall alignment of our customized RNA-seq pipeline includes two levels, alignment on genomic DNA sequences, and alignment on a

junction library based on all possible exon combinations within a 100,000-bp span, based on documented exon boundaries. This is different from TopHat strategy, which uses sequencing reads enrichment and splicing sequence features (GU…AG) for exon boundary detection.

4.4.3 Other Algorithms for Splicing Analysis

Based on the alignment output from TopHat or the customized pipeline, Cufflinks v1.2.1 [204] and Scripture [205] were used for isoform reconstruction. fastMISO (Mixture of Isoforms) [42] was used to calculate the percentage of inclusion for annotated and novel alternative splicing events. Standard parameter settings were used for all the three programs.

4.4.4 *De novo* Alternative Splicing Event Identification

As shown in Figure 4.1A, Alt Event Finder uses transcript isoform annotation from Cufflinks (GTF format) or Scripture (BED format) as input. The output is the data-derived alternative event annotation in GFF3 format, which can be used as MISO input. From the isoform annotation, the Alt Event Finder extracts "minimum non-overlapping exon regions" as expression units (Figure 4.1B), counts the number of isoforms that include each expression unit, and further derives appropriate AS events based on the string of counts (Figure 4.1B). In this study, we focus on cassette exons.

4.4.5 Performance Assessment

The ability of Alt Event Finder was evaluated by comparing with the splicing event annotation in the MISO library. Events from two annotations are considered consistent only if the genomic loci of the alternative exon (cassette exon) and their 5' upstream and 3' downstream exons are identical. This ensures the most conservative evaluation. The performance of Alt Event Finder is assessed by using three measurements, the total number of identified events, and the rate of known events and the recall of the overall finding. The rate of known events is defined by the percentage of known events

within data-driven ones, and recall is defined as the percentage of data-driven events within known ones.

Chapter 5. Novel Alternative Splicing Events in Proteome

5.1 Background

Human cells benefit from elaborate mechanisms to modify proteins, creating many protein variants (isoforms), both to increase the diversity of functions and to regulate the activities of proteins. A protein isoform is any of several different forms of the same protein. Different forms of a protein may be produced from related genes such as single-nucleotide polymorphisms (SNPs) or may arise from the same gene by alternative splicing or post-translational modifications (PTM). Alternative splicing and SNPs expands the number of messenger RNAs to about 88,000 mRNA variants during transcription of these genes. About 8% of these protein isoforms are generated from mRNA transcripts affected by alternative splicing or SNPs, whereas over 90% of protein isoforms are created through post-translational modifications (PTMs) after the mRNA is translated into a protein [212]. Recent studies have shown that the identification, analysis and characterization of these individual protein isoforms (Alternative Splicing, SNPs and PTMs) could improve understanding of diseases improve disease diagnosis or interventions [213-222].

Recent advances in clinical proteomics technology, particularly liquid chromatography-coupled tandem mass spectrometry (LC-MS/MS), have enabled biomedical researchers to characterize thousands of proteins in parallel in biological samples[223]. Identifying disease-related protein isoforms using tandem mass spectrometry, therefore, can provide hope for improving both the sensitivity and the specificity of candidate disease biomarkers, because proteomics identification, instead of quantification, of the same set of protein isoforms is often sufficient to distinguish between disease samples and controls.

However, identifying protein isoforms using current MS proteomics search databases and software tools has been challenging, primarily because of the smaller size

of known or common alternatively spliced protein isoforms relative to several orders of magnitude larger size of MS search databases, which makes exhaustive novel peptide identification computationally inefficient for routine proteomics studies. Up to 80% of all MS spectra peaks in a typical proteomics experiment may remain uncharacterized when searched against a standard MS database with little protein isoform information. Such standard MS search databases include: the IPI database [224], the NCBI-nr database, and the UniProt knowledge base [225]. These databases integrate more than a dozen public protein and DNA sequence databases into a non-redundant list of both known and predicted protein sequences, with only publicly known splice variant transcripts represented. MS search software such as SEQUEST [226], Mascot [227], X!Tandem [228], and OMSSA [229]. may further allow customized identification of limited types of PTM-derived peptides and proteins. However, these protein sequence databases do not contain information about alternatively spliced transcripts or theoretically possible "mis-spliced" protein isoforms; nor do they contain peptide variants arising from SNPs that result in amino acid changes. Therefore, they are ill-suited for comprehensive protein isoform identification purposes.

Although there are several publicly available alternative splicing mRNA transcript databases and SNP databases including ASTD [230], EID [231, 232], ASPicDB [233], ECgene [234], MutDB [235, 236], and dbSNP [237], none of these databases can be readily used for identification of novel peptides derived from uncharacterized protein isoforms. Since predictions of gene splicing patterns in all the methods are based on alignments of transcript data (mostly expressed sequence tags, ESTs) to a genomic sequence, some limitations exist in all these methods mostly due to the sequence errors frequently occurring in ESTs and to the repetitive structure of the genome sequence. Moreover, all the databases mentioned contain a rather small set of alternatively spliced peptides because they are either manually curated or literature-based data sets, as well as

poor annotation of splice events, which are inadequate for the identification of alternatively spliced protein isoforms. To explore the huge solution space of all possible alternatively spliced combination of exons and potentially coding introns, one must generate virtual peptides exhaustively so that uncharacterized MS spectra can be searched against them. In addition, the database of virtual peptides should be expanded to accommodate the amino acid alterations introduced by each SNP.

In this paper, we describe the development of a Peptideomics Database of Protein Isoforms (PEPPI), which consists of systematically generated virtual peptides that cover alternative splicing events and known SNP variations, for identifying protein isoforms in large-scale proteomics results. In the PEPPI database, we introduce a peptidomics approach to integrating genome, transcriptome, proteome and SNP information for human proteomics studies. The database contains a comprehensive set of peptides derived from all known annotated human genes in the Genome Reference Consortium Human genome build 37 by generating alternative splicing events and incorporating non-synonymous SNPs. It is the first comprehensive database that can be used to characterize novel protein isoforms derived from alternative splicing and SNP variations in MS spectra. The database has a web user interface that allows its users to query a gene/protein and compare all its above-mentioned types of protein isoforms and associated virtual peptides online.

5.2 Results

5.2.1 Database Content

Drawn from Ensembl's genomic data [238], the PEPPI database contains a comprehensive set of peptides derived from all known human protein-coding genes and was constructed by generating both annotated and hypothetical alternative splicing events and incorporating non-synonymous SNPs. In addition to representing an in-frame peptide for each exonic region (EXON_KB) of human proteins, four types of PEPPI splice

junctions are also captured for all possible combinations of each coding sequence of gene: annotated exon-exon junctions (E_E_KB type), hypothetical exon-exon junctions (E_E_TH type), hypothetical exon-intron junctions (E_I_TH type), and hypothetical intron-exon junctions (I_E_TH type). An exonic region or a splice junction is defined as a peptide region. For each peptide region, we also include hypothetical peptides translated with each known non-synonymous SNP. By cataloguing each peptide configurations in the PEPPI database, users can study alternative splicing events such as exon skipping, alternative donor site, alternative accepter site, and intron retention at the proteome level. They can also batch-download the peptide annotation and sequences in FASTA format for MS data searching. The current PEPPI database includes human data only. As of April 2010, it is comprised of 7,848,236 PEPPI peptide entries derived from 23,491 protein-coding genes and 66,384 proteins, incorporating 150,054 non-synonymous SNPs (Table 1).

A peptide-protein mapping is also captured for comparing the MS search results derived with the PEPPI and conventional protein sequence databases. In total 613,591 peptides are mapped to 66,384 IPI [224] proteins (Table 1).

5.2.2 General Online Features

In Figure 5.1, we show the user interfaces of the web-based online version of the PEPPI database. It allows searching by Ensembl Gene ID, gene symbol, UniProt ID, IPI AC, peptide sequence, PEPPI Peptide Region ID and PEPPI Peptide ID. With the cross-links users can easily link to Ensembl [238], IPI [224], UniProt [239], HAPPI [240] and HPD [241] and get access to much more detailed information about genes, proteins, protein-protein interactions and human pathways. The peptide annotations and sequences are freely available for batch-download in FASTA format on the download page.

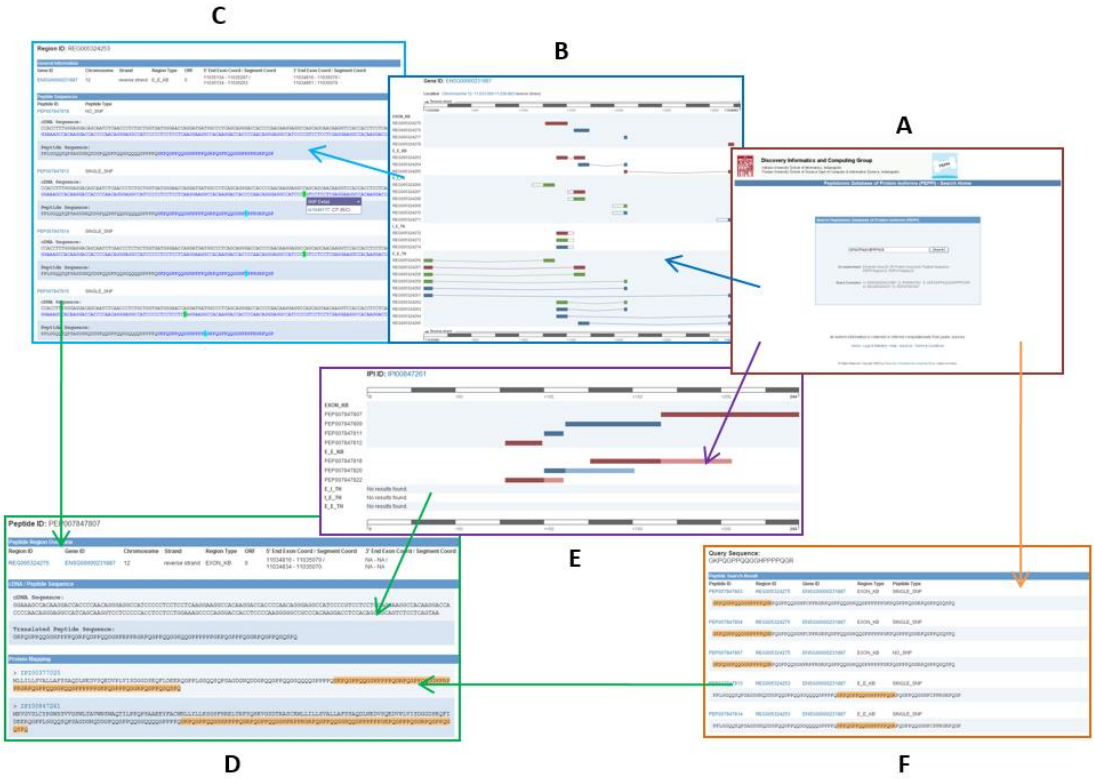

Figure 5.1 Web Interface Structure

(A) Search Home: main search page allowing five types of query string: Ensembl gene ID, IPI protein accession number, peptide sequence, PEPPI region ID and peptide ID. (B) Gene View: search result page visualizing peptide regions within a gene. (C) Region View: search result page displaying peptides within a peptide region. (D) Peptide View: PEPPI peptide information page. (E) Protein View: search result page of PEPPI peptides mapped to an IPI protein. (F) Sequence Search: search result page of PEPPI peptides mapped to a query peptide sequence.

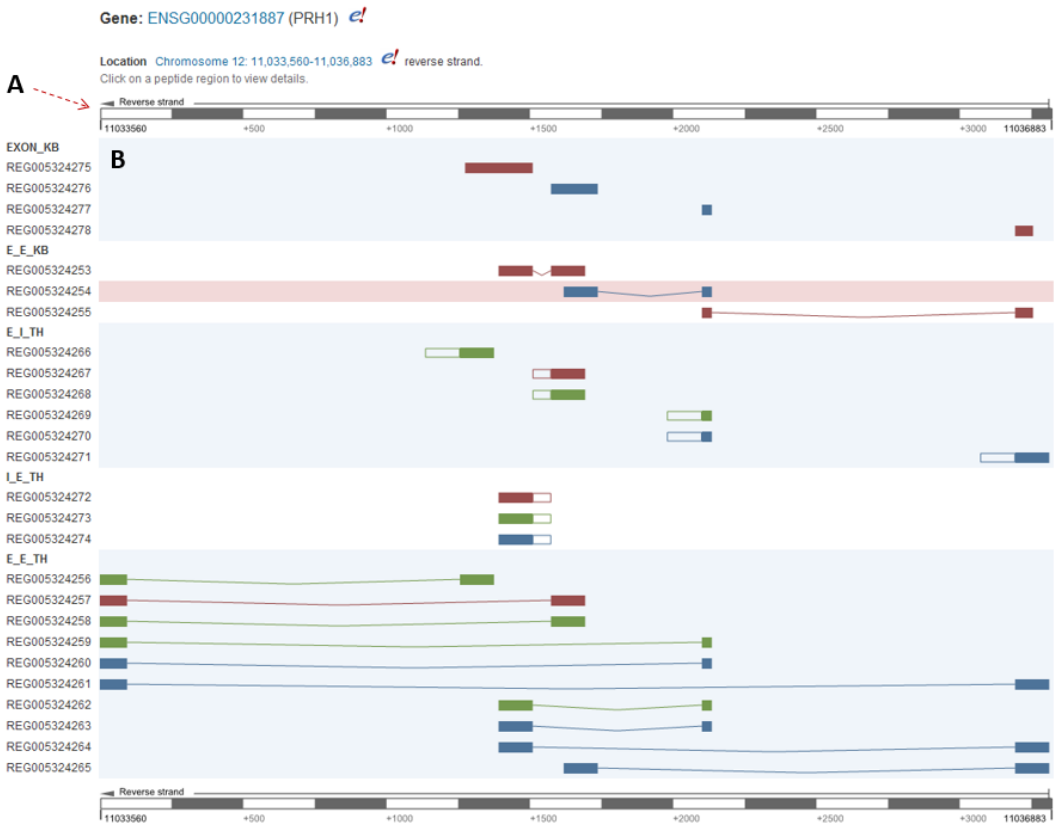

Figure 5.2 Gene View

(A) Gene scale. It shows the user the chromosome coordinates and strand of the gene. With the gene scale, users can read the approximate position of the peptide regions. (B) Peptide regions. It shows the user which five types of peptide regions within current given gene that the peptide belong to. The coloring of the peptide regions indicates the ORF (Red: 0, Green: 1, Blue: 2). The solid bars indicate exons, and the blank bars indicate introns. The curve between two exons means the exons are spliced with each other.

Figure 5.3 Region View

(A) The different colors of the cDNA and peptide sequences indicate two different exons, or an exon and an intron. The amino acid letter colored in red overlaps with the splice site. (B) Green and light cyan backgrounds are used to indicate SNP in cDNA and peptide sequences. (C) By clicking on a SNP in sequence, users can see the details of the SNP. A link to the dbSNP database is provided.

Figure 5.4 Peptide View

(A) The cDNA and peptide sequences that the current PEPPI peptide is mapped to. Same color theme is used in the region view. By clicking on SNP in sequence, users can access detailed information of the SNP. (B) In the protein mapping list, the view displays all the proteins mapped to the current PEPPI peptide. The peptide sequence is highlighted from within the protein sequence.

5.2.3 Case Study 1: Browsing PEPPI Peptides and Relating Information

For users who would like to overview all the peptide regions and peptides within a gene of interest, we provide the standard gene search procedure. In this case study we show how to browse the PEPPI peptide regions, PEPPI peptides and related information within gene PRH1. Users will start from the Search Home (Figure 5.1A), go through the Gene View (Figure 5.1B), Region View (Figure 5.1C), and finally navigate to the Peptide View (Figure 5.1D).

By searching with gene PRH1 in the standard query box provided at the PEPPI database home page, users can retrieve all peptide regions corresponding to this gene (Figure 5.2). In the Gene View, the PEPPI database visualizes all the peptide regions that

can be mapped to this gene. The “Location” section shows this gene is located on chromosome 12, from 11,033,560 bp to 11,036,883 bp. Links to Ensembl are provided on the gene ID and location. A scale of chromosome coordinate is provided on the top and bottom of the visualization. The arrow on the chromosome coordinate scale shows this gene is located on the reverse strand, so the 5’ end of the gene should be the right end. Peptide regions are displayed in five categories, including EXON_KB, E_E_KB, E_I_TH, I_E_TH and E_E_TH. The color of the region indicates the protein translation open reading frame (ORF) of the corresponding cDNA. By clicking on the peptide region REG005324254, the browser will be re-directed to the Region View.

The Region View (Figure 5.3) displays detailed information of the peptide region and the peptide sequences within this region. In the “Peptide Region Overview” section, the exon coordinate is the chromosome coordinate of the source exon, and the segment coordinate is the coordinate of the flanking sequence beside the splice site. The peptide without SNP is displayed on the top of the “cDNA/Peptide Sequence” section, and the peptides with SNPs are displayed below. In the sequences, black and blue are used to color different exons/introns. An amino acid residue overlapping a splice site is colored in red. SNPs are highlighted by green and light cyan. By clicking on the highlighted SNPs, the SNP ID will be shown along with the nucleotide change and amino acid change. A link to the corresponding page in dbSNP is also provided. By clicking on a PEPPI peptide ID, e.g., “PEP007847820”, the browser will be navigated to the Peptide View.

In the Peptide View (Figure 5.4), detailed information of the peptide region and a peptide-protein mapping is shown in the “Peptide Region Overview” section. The cDNA and peptide sequence is displayed in the same pattern as the Region View. The “Protein Mapping” section lists the proteins mapped to the current peptide. The result shows that IPI00847261 is the only protein mapped to the peptide PEP007847820. The annotation on IPI states that IPI00847261 is one of the protein products of PRH1. Since the peptide

PEP007847820 contains a mutant non-synonymous SNP allele, we can infer that the mapped protein IPI00847261 is not the wild-type.

5.2.4 Case Study 2: Identifying Genomic Origins of AS Events

For users, especially MS proteomics scientists, who want to start the query from a peptide sequence or a protein, we provided a peptide sequence search function (Figure 5.1F) and the Protein View (Figure 5.1E). In this case study we demonstrate that the PEPPI database can help identify the genomic origins of peptides detected from MS data, and can help characterize the alternative splicing events related to these peptides.

The MS peptides can be derived from Healthy Human Individual's Integrated Plasma Proteome Database (HIP-2) [242] by inputting its protein ID. For this example, by entering “IPI00023636” as the query, a mapping table with several MS peptides identified by the MS data analysis program will be returned (Figure 5.5A). To identify the genomic region that encodes a specific peptide sequence, we can search the peptide sequence on the PEPPI database’s search home.

As shown in Figure 5.5B and C, peptide 1 is mapped to “PEP000841715”, which is an E_E_KB peptide, and peptide 2 is mapped to “PEP000841692” which is an EXON_KB peptide. This indicates peptide 1 is coded by an exon-exon junction, and peptide 2 is coded by a single exon.

To study the related alternative splicing events, we then compared the number of proteins which can be mapped to these peptides. By clicking on the peptide ID, we can get access to the proteins mapped to each peptide. We found 4 proteins mapped to peptide 1, and 5 proteins mapped to peptide 2. Interestingly, only one protein (IPI00745806) was differentially mapped. By looking up the protein information in IPI, we found that the proteins involved are five different alternatively spliced isoforms of MP2K7_HUMAN, and IPI00745806 is the third isoform. Therefore, it is likely that only a specific

alternative splicing event that takes place is annotated and can be mapped onto the protein sequence IPI00745806.

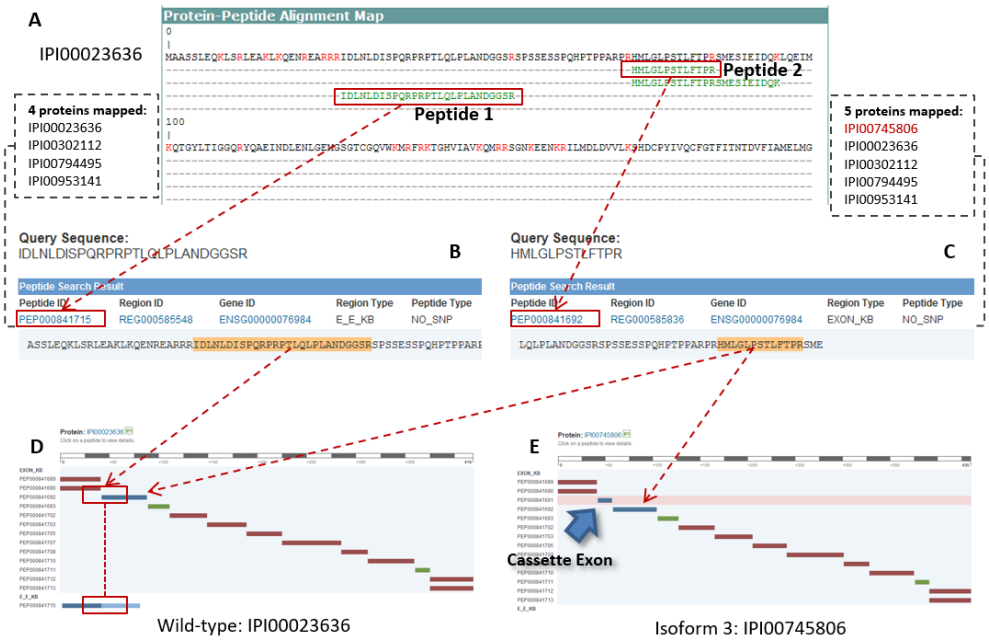

Figure 5.5 Identifying The Genomic Origin of MS Detected Peptides and The Relating Alternative Splicing Event

(A) The HIP-2 search result page of protein IPI00023636, displaying the evidence peptides detected in MS experiments. (B) The PEPPI sequence search result page of peptide 1, indicating the query peptide is produced from an exon-exon combination region. The corresponding PEPPI peptide can be mapped to 4 proteins. (C) The sequence search result page of peptide 2, indicating the peptide comes from an exon, and can be mapped to 5 proteins. (D) The search result of the wild-type MP2K7_HUMAN, showing the regions mapped by the peptides. Peptide 1 crosses the splice site of two exons (PEP000841690 and PEP000841692). Peptide 2 is produced from a single exon, PEP000841692. (E) The search result of the 3rd isoform of MP2K7_HUMAN, the protein that is mapped to the peptide 2 but not mapped to peptide 1. That is because the insertion of a cassette exon (PEP000841691) changed the sequence of the protein.

To verify our suspicion on the existence of the alternative splicing event, we compared the protein-peptide mapping of the wild-type and the third isoform of MP2K7_HUMAN. In the wild-type MP2K7_HUMAN (Figure 5.5D), "PEP000841715" contains the splice junction of two exons (PEP000841690 and PEP000841692), and peptide 1 just crosses the splice site. Nevertheless, in the MP2K7_HUMAN isoform 3 (Figure 5.5E), we found a unique cassette exon (PEP000841691) spliced between the two exons, which hampered the coding of sub-sequence mappable to peptide 1. Meanwhile, peptide 2 is only mapped to a single exon (PEP000841692), which exists in all five proteins and unaffected by any splice events. Thus we have confirmed the suspicion that a cassette exon event caused the protein mapping difference between two MS peptides, and have shown the PEPPI database's ability to help infer alternative splicing events from peptides detected from MS experiments.

5.2.5 Case Study 3: Identifying New Peptide Isoforms for Human

Human fetal liver can evolve into a major site of embryonic hematopoiesis; therefore, protein profiling may help researchers understand how the interaction between hepatic and hematopoietic systems and the migration of the hematopoietic system during mammalian development take place. We collected four human fetal liver cytoplasm proteome data sets from the human fetal liver project (http://hlpic.hupo.org.cn /dblep). SDS-PAGE with different cross-linking percentages 15%, 10%, and 7.5% was used for protein separation to obtain a full representation of proteins ranging from 5 kDa to more than 300 kDa. After these gels were stained with Colloidal Coomassie Blue R250 and the gel lanes were manually excised from loading position to the bottom of the gel, the extracted peptide mixtures were loaded onto nanoscale LC-ESI-Q-TOF MS or micro-LC-ion trap MS systems for protein identification [243].

In order to show that the PEPPI database can be used to identify additional novel peptide isoforms than the traditional protein database, we downloaded the protein

database IPI and created three datasets using the PEPPI database: 1) annotated exonic peptides and exon-exon combinations without SNP (PEPPI_KB), 2) all PEPPI peptides without SNP (PEPPI_without_SNP), and 3) all PEPPI peptides including peptides with SNP (PEPPI_with_SNP). PEPPI_KB consists of peptides of both the EXON_KB and E_E_KB region types without additional SNP permutations; PEPPI_without_SNP consists of peptides of the EXON_KB, E_E_KB, E_I_TH, I_E_TH, and E_E_TH region types without additional SNP permutations; PEPPI_with_SNP consists of peptides of all types, with or without SNPs, in the PEPPI database. We also created four corresponding inverse sequence datasets to evaluate the false discovery rate with a target-decoy search strategy [244]. The four peak list files of human fetal liver from LC-ESI-Q-TOF MS or micro-LC-ion trap MS raw files were searched by OMSSA[229] against the four databases and their four inverse databases in order to compare the results among them.

OMSSA reports hits ranked by E-value. An E-value for a hit is a score that is the expected number of random hits from a search library to a given spectrum, such that the random hits have an equal or better score than the hit. For example, a hit with an E-value of 1.0 implies that one hit with a score equal to or better than the hit being scored would be expected at random from a sequence library search [229]. The search results with OMSSA can vary substantially with differing search parameters, sequence libraries, and samples [244]. Therefore, we adopted the MS/MS false discovery rate (FDR) instead of E-value as scoring criterion for evaluating the four databases, and this method is based on commonly used scoring methodologies and the target-decoy search strategy [244]. All other OMSSA search parameters [229] for the four databases are the same. To increase identification accuracy, only peptides/proteins with at least two hits of different samples was recognized as true peptides/proteins.

A comparison of search results against four MS databases, i.e., IPI, PEPPI_KB, PEPPI_without_SNP, and PEPPI_with_SNP, is shown in Table 2. Results are shown only

at a commonly used 1% MS/MS FDR for each database. Compared to the traditional IPI database, the elapsed time for PEPPI_KB decreased although the dataset size increased by two and a half times. And with the increase of sizes, the elapsed time increases significantly linearly from PEPPI_KB to PEPPI_without_SNP to PEPPI_with_SNP (Intercept=8.17021, slope=0.04738 , and adjusted R2=0.9975).

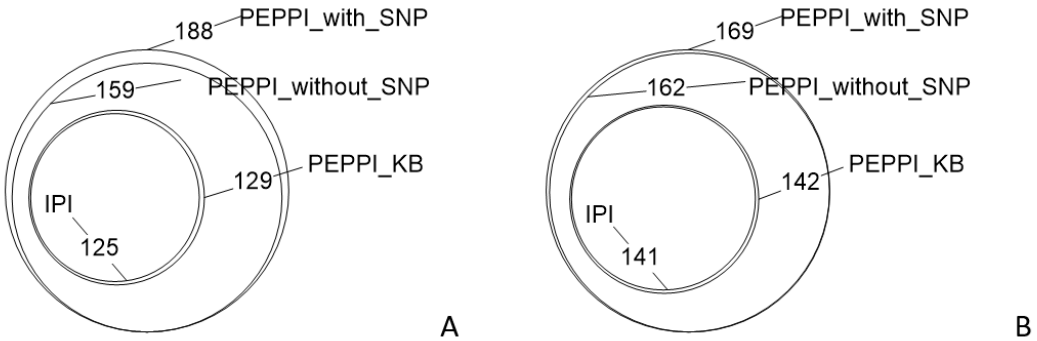

Figure 5.6 Overlap of Peptides/Genes Identified by Four Search Databases.
(A) Peptides identified from two or more samples. (B) Proteins identified from two or more samples.

Under the criteria of MS/MS FDR 0.01, the target MS/MS hits markedly increases with the increase of database size, and target peptide hits, target protein hits, and target PEPPI hits all increase while the corresponding FDRs remained approximate. The overlap of genes identified by each database is shown graphically by Venn diagram in Figure 5.6. The results show that the PEPPI database can be used to identify more peptides/proteins under the same false positive rate than the traditional IPI database.

From the four human fetal liver MS data sets, we identified 63 peptides which mapped to 74 PEPPI peptides and 9 SNP events using PEPPI_with_SNP (See Table S-2). Among the 74 PEPPI peptides, 55 EXON_KB type peptides were also annotated previously in IPI, and 19 peptides were novel peptides uniquely identified with the PEPPI

database (13 E_E_KB type peptides, 2 E_E_TH type peptides, 1 I_E_TH type peptide, and 3 E_I_TH type peptides)

The peptide hit matrix shows the number of PEPPI peptides mapped to the peptides detected from the samples, and the number of samples (N) in which the peptide is detected (See Table S-3).

5.3 Discussion

We created a comprehensive PEPPI database of both annotated and hypothetical peptides representing human protein isoforms for MS analysis. The PEPPI database made it possible for high-throughput identification of gene variations, exon expression, and alternative splicing events at the proteome level. We also constructed a web-server for searching and visualizing the peptides. With the user-friendly interface and powerful search functions, users can easily study the alternative splicing events and gene variations related to any gene, protein, or peptide sequence of interest.

A comparison between the PEPPI database and conventional MS methods is shown in Table 3. An MS approach with the PEPPI database uses the same samples, equipments and analysis software as a conventional MS approach. To use the PEPPI database, users just need to set the PEPPI database or a subset of the PEPPI database as the user defined sequence database in the MS search software. With the PEPPI database, users can gain information on the expression of exons, alternative splicing events, SNPs, and protein existence from the proteome, while the conventional MS approach can only derive the protein existence information. Users can opt to use different subsets of the PEPPI database for different study purposes. The computational cost of adopting the PEPPI database approach over a conventional approach is kept low, due to the use of computing cycles without human intervention.

Since the database incorporated a large number of hypothetical peptides, it is possible that the search result contains false positives due to the noise. To solve this

problem, we will design an optimized routine for MS data analysis. For example, users may analyze several samples at one time, and only keep the peptides detected in more than one sample. On the other hand, a recently published article [245] reported tissue dependent splicing patterns, which make it possible to generate tissue specific PEPPI peptides and reduce the chance of incorporating false positives.

5.4 Methods

5.4.1 Genome Data Source

The PEPPI peptides were generated from the human genome. The source genome data was downloaded from Ensembl Version 55 [238] with BioMart. As shown in Figure 5.7A, four tables, including Un-Translated Region (UTR), Gene Sequence, Gene-Protein Mapping and Gene Structure, were pulled from the Ensembl Homo sapiens Genes Dataset. The UTR table describes the coordinates of all the transcript UTRs. The Gene Sequence table contains chromosome coordinates and sequences in FASTA format. The Gene Structure table contains the exon annotation information, including Exon ID, Gene ID, Transcript ID, as well as the genome coordinate and translation phase of the exons. The SNP table was derived from the Ensembl Homo sapiens Variation Dataset, and contains the SNP's chromosome coordinate and nucleotide shift. The PEPPI database incorporated 44,285 genes and 16,489,577 SNPs.

5.4.2 Data Pre-Processing

A data pre-processing procedure (Figure 5.7B) was implemented to remove non-coding genes and SNPs which are not in exonic regions.

Firstly we imported all the source tables into a SQLite3 database with the SQLite3 command-line interface, and then we used SQL statements to remove non-coding genes. The Gene-Protein Mapping was utilized as a filter, and genes not mapped to proteins were considered as non-coding genes. The Coding Gene Sequence

and Coding Gene Structure table was derived after filtering, and 21,351 protein-coding genes were captured.

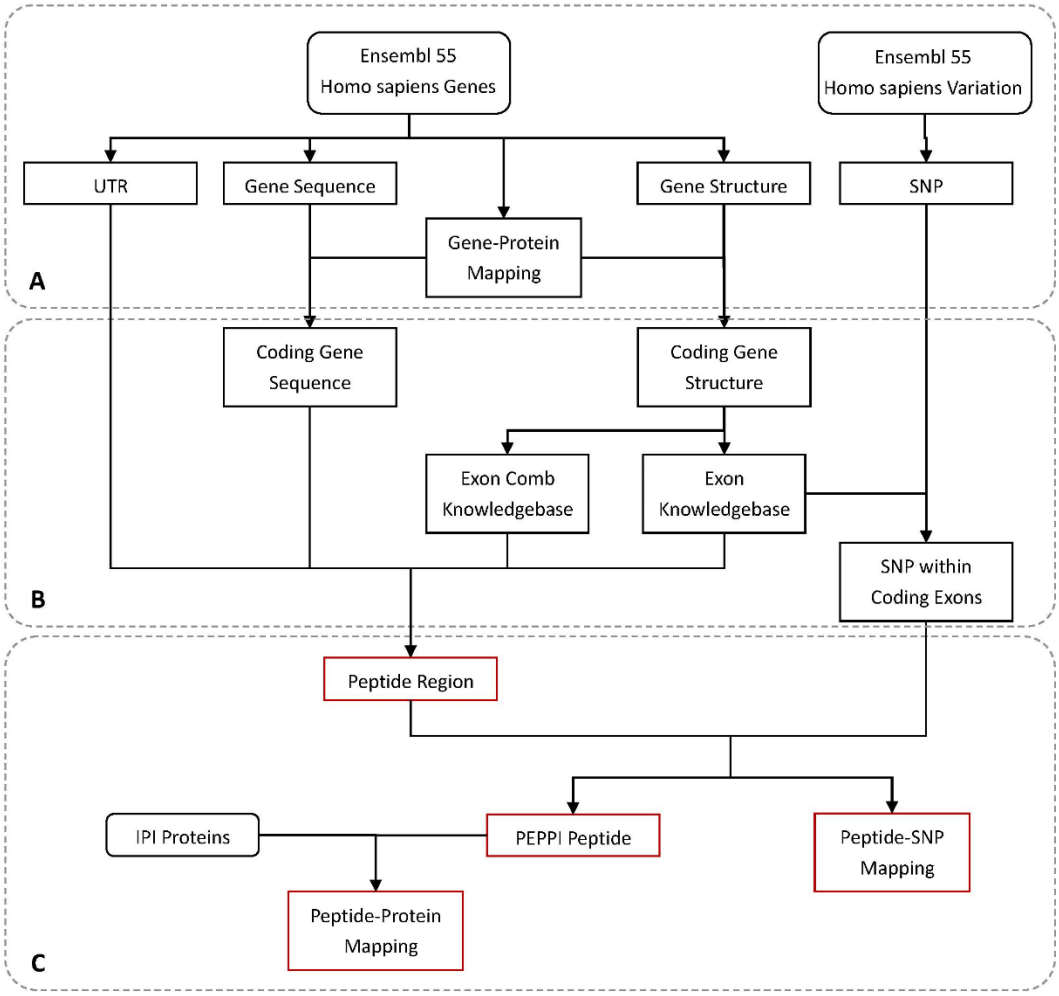

Figure 5.7 Data Generation Process.

The whole data generation process was divided into three steps: (A) deriving the genome data from Ensembl; (B) pre-processing of the data to select protein-coding genes and SNPs within coding exons; and (C) generation of peptide regions and PEPPI peptides. The result datasets are colored in red.

Then we compiled a C program with the SQLite3 library to extract the annotated transcription information from the Coding Gene Structure table, and produced two tables. The Exon Knowledgebase table describes all the protein-coding exons, and the Exon Comb Knowledgebase table describes all the exon-exon combinations found in the annotated transcripts. Then the SNP table was searched against the Exon Knowledgebase table, and 390,539 SNPs within the annotated coding exons were retrieved for peptide generation.

5.4.3 Peptide Region Generation

We compiled a pipeline program with C and the SQLite3 library to generate peptide regions (Figure 5.7C). The program first generated the wild-type cDNA sequences of the peptide region, and then translated the cDNA sequences into peptides. The derived peptides were estimated by the program according to a set of protocols, and un-qualified peptides and the corresponding region were discarded. Different cDNA generation procedures and peptide estimation protocols were implemented on different types of peptide regions.

For the EXON_KB type, the chromosome coordinate of the exons were derived directly from the Exon Knowledgebase table, and the whole length of the exon cDNA sequence was captured from the gene sequence. Then the exon's cDNA sequences were translated into peptides according to the annotated ORF. In the peptide estimation process, if a stop codon existed anywhere except the end of the peptide, the corresponding region was considered invalid and was discarded.

Similar to the EXON_KB type, the chromosome coordinates of the two exons in the E_E_KB type were derived directly from the Exon Comb Knowledgebase table. Then the derived cDNA sequences were translated into peptides according to the annotated ORF of the exon on the 5' end. The same peptide estimation protocol used with the EXON_KB type was applied to the E_E_KB type.

For the E_I_TH and I_E_TH type, the program derived the chromosome coordinates of exons from the Exon Knowledgebase table and spliced them with the adjacent introns. The cDNA flanking sequence on both side of the splice site was limited to 120 nucleotides. If the exon/intron is shorter than 120 nucleotides, the program will pull out the actual sequence. This limit was set according to the longest peptide in the HIP-2 database [242], which had a length of 80 amino acids and corresponded to 240 nucleotides in the cDNA sequence. This represents the length of the longest peptide that can be identified from an MS experiment. In this way we made sure that any MS identified peptide that crosses the splice site can be captured by the PEPPI database. In the peptide estimation process, a stop codon is tolerated in the intron of E_I_TH, but not tolerated anywhere expect the 3' end in I_E_TH.

When producing the E_E_TH type of peptide regions, all possible exon-exon combinations were generated and searched against the Exon Comb table. Any combinations that cannot be found in the Exon Comb table were captured as an E_E_TH type candidate. Then each E_E_TH type cDNA was translated in all 3 ORFs, and if a stop codon was found in the 5' end exon, the peptide was discarded. Note if more than one E_E_TH peptide derived from the 3 ORFs were considered valid, then a peptide region was created for each ORF.

5.4.4 PEPPI Peptide Generation

After the generation of peptide region, PEPPI peptides were produced by inserting non-synonymous SNPs into the wild-type peptide of the corresponding region (Figure 5.7C). All the non-synonymous SNPs within a peptide region were first captured in a list, and then inserted into the wild-type cDNA according to their chromosome coordinates. Each cDNA sequence with SNP was then translated into peptides. The peptides were then estimated according to the peptide estimation protocol of its own region type, and invalid

ones were discarded. During peptide generation, a table of Peptide-SNP Mappings was also generated. The wild type peptides were also deposited in the PEPPI Peptide table.

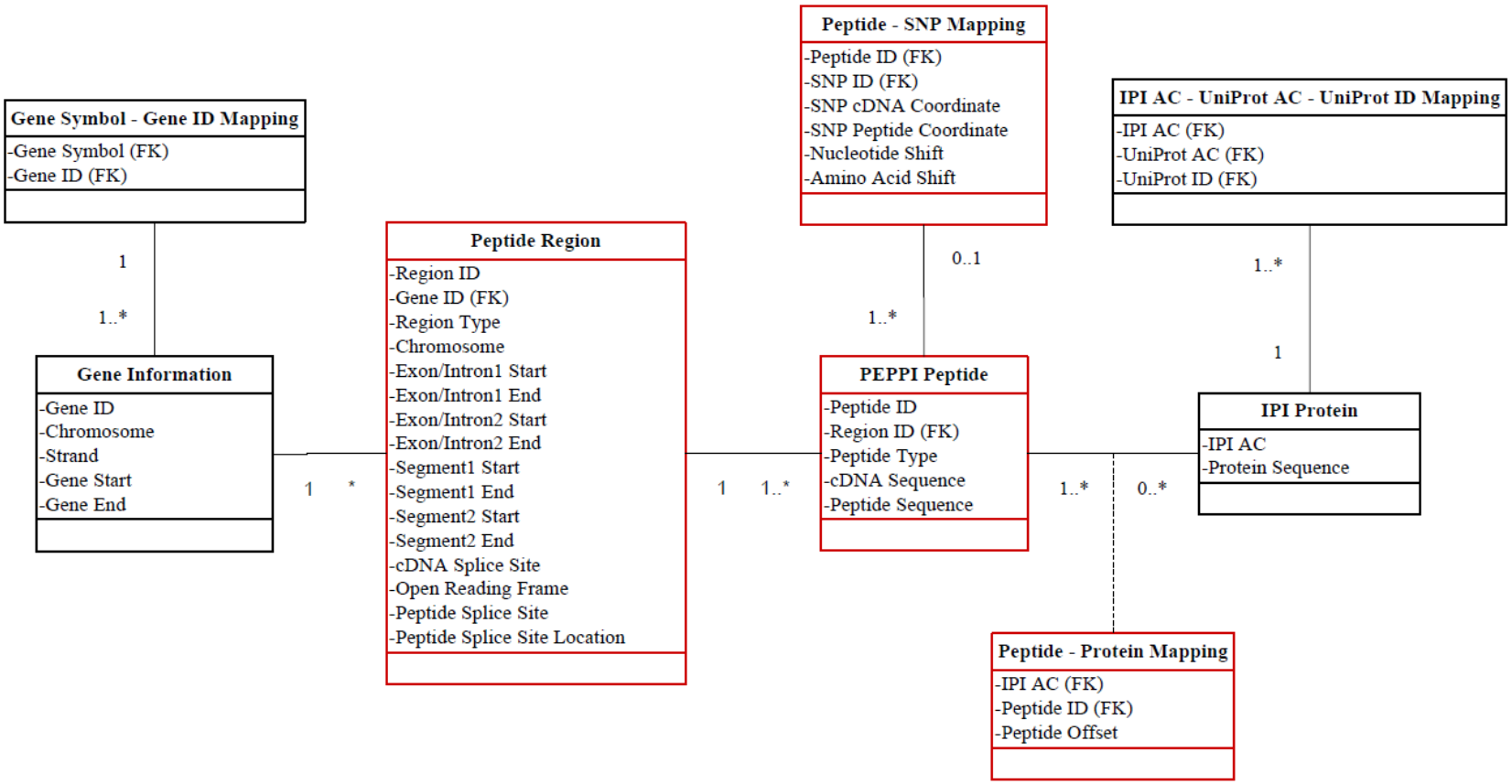

Figure 5.8 The UML of Database Backend

The datasets derived by the data generation pipeline are colored in red, and the datasets derived from other databases are colored in black.

5.4.5 Online PEPPI Server Design

The online version of PEPPI database is a typical 3-tier web application, with a MySQL database at the backend database service layer, Apache/PHP server scripts to the middleware application web server layer, and CSS driven web pages presented on the browser. The Javascript library uuCanvas is used to render the real-time data visualizations in the gene view and the protein view.

The result tables derived from the data generation step were imported into the MySQL database (Figure 5.8). The chromosome coordinate information was deposited in the Peptide Region table, and the sequence information was deposited in the PEPPI Peptide table. The ID mapping tables for genes and proteins enable users to query the database with different IDs.

Chapter 6. Long Non-coding RNAs in Transcriptome

6.1 Background

Alcohol is the third leading cause of preventable death in the United States [246]. Alcohol misuse negatively affects the quality of life for millions of Americans, and has profound sociological and economic impacts. The neurobiological basis underlying alcohol dependence is not fully understood, but extensive evidence indicates that genetic factors play key roles in influencing the risk of alcohol dependence [247-254]. Over the past decades, several specific genes have been implicated in the risk of alcoholism [249, 255-265]. In addition, recent studies suggest that epigenetic processes play a critical role in affecting the risks of alcohol dependence [266-268].

Deep sequencing data from the Encyclopedia of DNA Elements Consortium (ENCODE) suggests that over 90% of the human genome can be transcribed, and non-protein-coding RNAs (ncRNA) exceed the number of protein-coding genes [67]. The recent discovery of over 200 ncRNAs significantly enriches the portfolio of potential genetic factors [269]. Rather than being transcriptional noise, many ncRNAs serve as master regulators that affect expression levels of dozens or even hundreds of target genes [270, 271]. These regulatory RNAs integrate signals from both genetic and environmental factors, and therefore can play major roles in controlling alcohol preference. Most notably, a strong association of epigenetic marks with long non-coding RNAs (lncRNAs, >200 nucleotides) in humans and mice was recently described [272]. Many lncRNAs contain conserved elements and show spatiotemporally restricted expression patterns, implying that they are functional and regulated [273]. These lncRNAs are reported to regulate dosage compensation, imprinting, and development by establishing chromatin domains in an allele- and cell-type specific manner [274-276]. It is also reported that lncRNAs are involved in post-transcriptional regulations [277].

It is now possible to identify novel lncRNAs from the high-throughput sequencing data. Guttman et al [34] found that genes being transcribed by RNA polymerase II (Pol II) are marked by trimethylation of lysine 4 of histone 3 (H3K4me3) in the promoters and trimethylation of lysine 36 of histone 3 (H3K36me3) along the transcribed regions. They defined such structure as "K4-K36 domains" and identified more than 1600 previously unknown K4-K36 domains from mouse by CHIP-sequencing; these transcription active regions represent either protein coding genes or lncRNAs.

In the current study, we designed an RNA-sequencing experiment and a computer approach to identify and characterize novel lncRNAs that are actively transcribed and correlated with alcohol preference in rat. We conducted a scan on the transcriptional intensities within the rat orthologous regions of the mouse K4-K36 domains published by Guttman et al [58], and focused on "intergenic" lncRNAs, i.e., lncRNAs residing outside all known protein-coding genes. We identified 420 novel lncRNAs, among which 37 were differentially expressed between P (alcohol preferring) and NP (alcohol non-preferring) rats. Our pathway analysis on the differentially expressed lncRNAs demonstrated that many of them had shown significant association with neural functions. Our method is also applicable to other diseases and species.

6.2 Results

In order to understand the role of lncRNA in alcohol preference, we conducted RNA sequencing and bioinformatics analysis on P (alcohol preferring) and NP (alcohol non-preferring) rat strains. Our analysis includes four major steps: (i) deriving the rat orthologous regions of the K4-K36 domains in mouse; (ii) acquiring the transcriptome from the hippocampus of P and NP rats by means of next-generation sequencing; (iii) identifying potential regulatory lncRNAs associated with alcohol consumption, based on the RNA sequencing and epigenetic marker information; and (iv) inferring the functions of lncRNAs differentially expressed in P and NP rat strains (Figure 6.1).

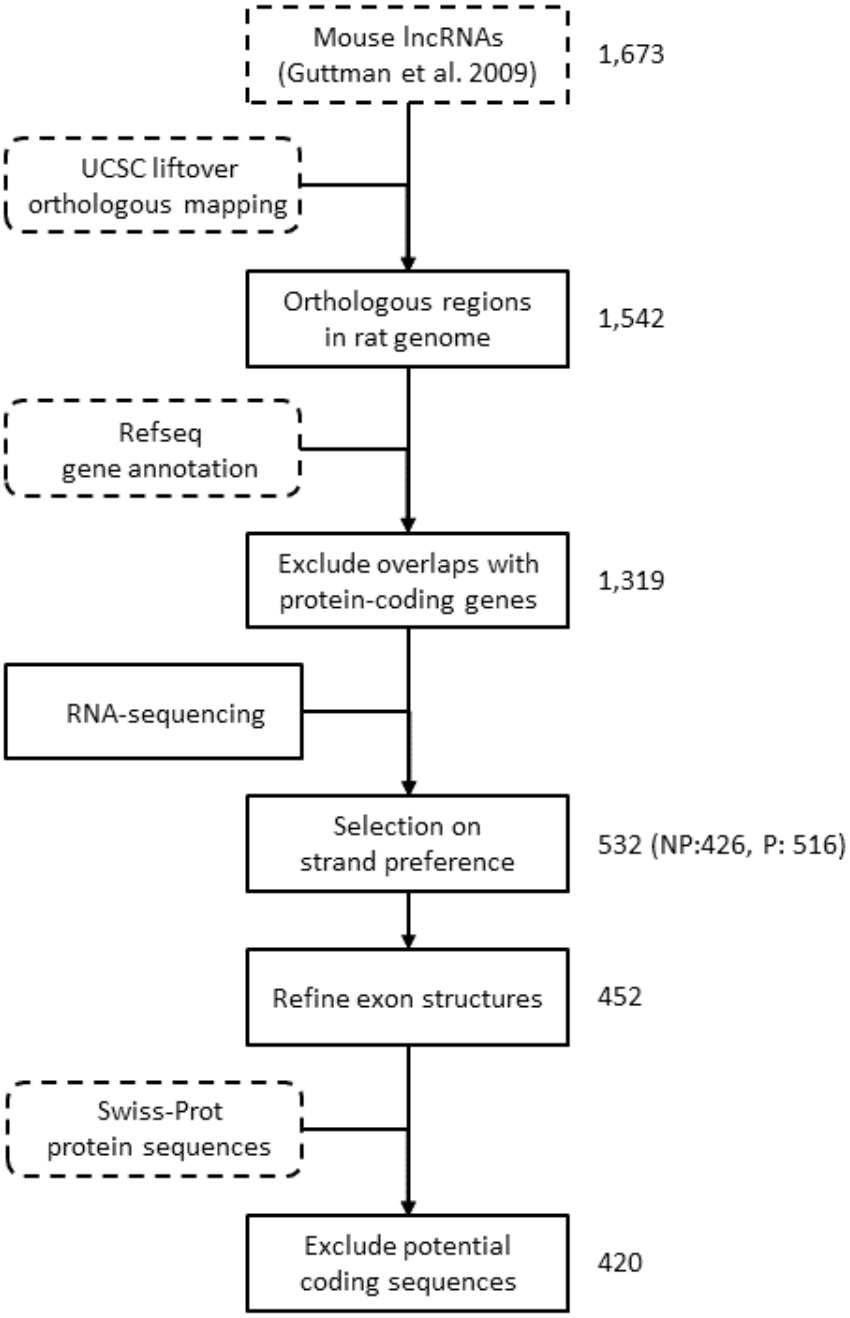

Figure 6.1 The Workflow of LncRNA Annotation.

The dashed boxes indicate extern data source, and solid boxes indicate results generated in our analysis. The numbers at the right of the boxes are the number of putative lncRNAs after each step of filtering. The filtering begins from 1673 K4-K36 domain in mouse and ends up with 420 putative lncRNA regions in rat.

6.2.1 Rat Genomic Regions Orthologous to Mouse K4-K36 Domains

Guttman et al [58] reported 1673 K4-K36 domains in the mouse genome that may include lncRNAs. To identify rat lncRNAs, we mapped these K4-K36 domains to the rat genome with UCSC LiftOver [278]; 1542 putative lncRNA regions were identified.

We discarded or truncated the rat orthologous domains to eliminate overlaps with (i) known protein-coding genes in rat, or (ii) orthologous regions of known protein-coding genes in mouse and human. We focused on the remaining 1319 putative lncRNA regions, in which all the known protein coding sequences were excluded.

6.2.2 Hippocampus Rranscriptomes of P and NP Rats

To examine the transcription activity of these regions in alcohol-preference, we implemented an RNA sequencing experiment on P and NP rats. P and NP rats [279] are a pair of model animals developed for alcohol dependence research, traits other than alcohol preference were strictly controlled. Total RNA was extracted from the hippocampus of 8 non-inbred P and 8 NP rats, poly-adenylated RNA was selected and reverse transcribed. The resulting cDNA was sequenced using the Illumina Genome Analyzer IIe, with the strand of the RNA transcripts restrained. RNA from each individual rat was sequenced in one Illumina lane that produced 2.8 to 12.8 million mappable reads.

6.2.3 Potential Regulatory lncRNA Regions in P and NP Rats

Among the regions that were transcribed, we assumed that the strand preference for each transcript should be consistent across all the samples, and discarded those that were not. With this filtering, 516 and 426 transcripts were derived from the putative rat lncRNA regions from P and NP rats, respectively (Figure 6.2A). By uniting these two sets of transcripts and removing duplicates, we derived 532 putative lncRNA transcripts with strand specificity.

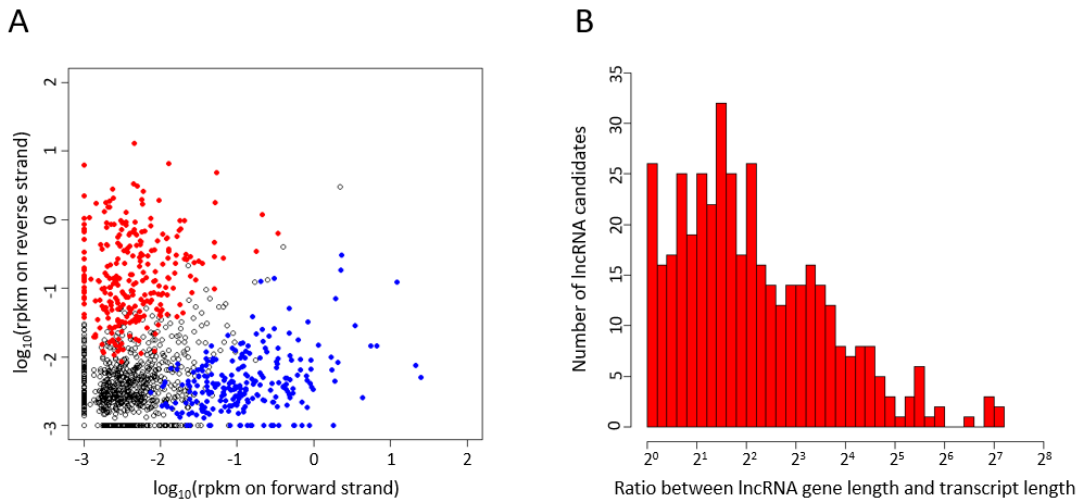

Figure 6.2 Features of identified lncRNAs.

(A) Strand preference. The horizontal and vertical axes denote transcription activity on forward and reverse strand, respectively. The circles denotes the lncRNA candidates not showing significant strand preference, while the blue and red dots denotes the lncRNA candidates that are transcribed on the forward and reverse strand, respectively. (B) Distribution of the ratio between lncRNA gene length and exonic region length.

We used a computational algorithm to annotate exons in the putative lncRNA regions based on the transcriptional intensity. Within each exon, we required at least 8 reads, with a maximum distance between two reads of 25 nucleotides. By discarding the putative lncRNA regions of which the total exonic lengths were less than 200 nucleotides, the candidate pool was reduced to 452 putative lncRNA regions.

We aligned the exonic sequences of putative lncRNAs and known proteins with BlastX [280, 281], and then eliminated a small portion (≈7%) of putative lncRNAs that included exons showing protein-coding capacity (Methods). Eventually, we derived 420 novel lncRNAs with significant transcriptional activity and no significant potential to code for proteins.

	Novel lncRNA	**Known lncRNA**	**Protein-coding genes**
Number of regions	420	99	13892
Length of longest transcript (nt)	72075	83437	17599
Length of shortest transcript (nt)	200	374	105
Mean of all transcript lengths (nt)	4053	4947	2131
Median of all transcript lengths (nt)	1939	3240	1767
Maximum exon number	244	48	106
Minimum exon number	1	1	1
Mean of exon number	14	5	9
Median of exon number	8	4	7
Length of longest exon (bp)	10454	83437	11972
Length of shortest exon (bp)	32	20	3
Mean exon length (bp)	283	913	244
Median of exon length (bp)	208	150	132
Maximum expression intensity (rpkm)	1104.17	N/A	2905.00
Minimum expression intensity (rpkm)	0	N/A	0
Mean expression intensity (rpkm)	3.73	N/A	15.44
Median of expression intensity (rpkm)	0.94	N/A	2.26

Table 6.1 Statistics of Predicted LncRNA, Known LncRNA and Protein-coding Genes

Novel lncRNA indicates the lncRNAs identified by our pipeline; known lncRNAs include known lncRNAs in both mouse and rat; protein-coding genes refers to rat protein-coding genes only.

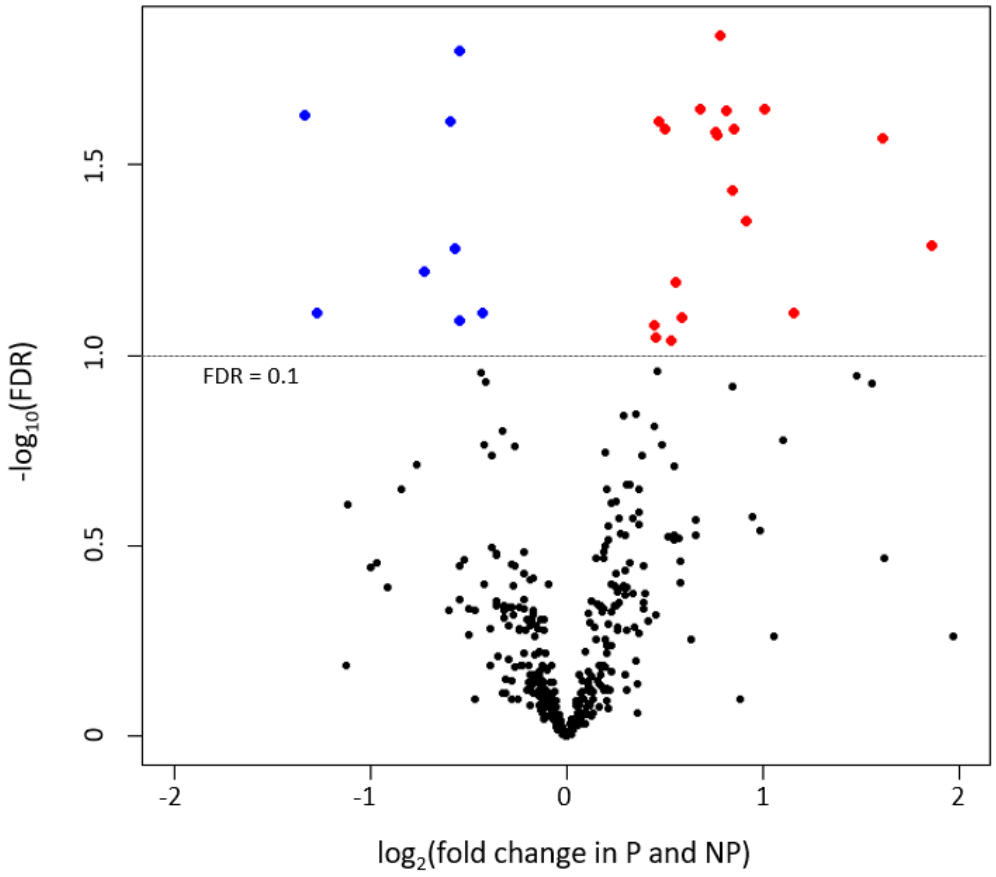

Figure 6.3 Volcano plot of differential expression in P and NP samples.

The black dots denote the lncRNAs that are not differentially expressed. Red and blue dots denote lncRNAs that are significantly higher expressed in P and NP rats, respectively.

These novel lncRNAs are equally distributed along different chromosomes, with 43 and 9 on chr1 and chr12, respectively, which are the longest and shortest chromosomes of rat. The transcript lengths of novel lncRNAs fell between 200 and 72,075 nucleotides, and the ratio of lncRNA transcripts and lncRNA genes ranges from 1 to 165 (Figure 6.2B), which are similar to known lncRNAs (Table 6.1). The mean

transcriptional intensity of novel lncRNAs is 3.7 RPKM, while the average RPKM of protein-coding genes is 15.44.

6.2.4 lncRNA Functions and Associations with Alcohol Preference

Among 420 lncRNAs identified in P or NP rat hippocampus, 37 were differentially expressed at a false discovery rate of 0.1 in our Friedman test (Figure 6.3). Among the differentially expressed lncRNAs, expression levels of 26 are higher in P rats, while 11 are higher in NP rats. This trend is significantly different from protein-coding genes ($p \leq 0.001$), where expression levels of 1401 and 2009 genes were high in P and NP rats respectively (Table 6.2). This is consistent with the observations that most known lncRNAs exert their functions by repressing the expression of protein-coding genes.

	Up-regulated	**Down-regulated**
lncRNA	26	11
Protein coding gene	1401	2009

P-value=0.0006 (Chi-square)

Table 6.2 Chi-square Test of LncRNA Negative Regulation on Protein-coding Genes
Up-regulated indicates the lncRNA or gene is up-regulated in P rats vs NP rats; conversely, down-regulated indicates the lncRNA or gene is down-regulated in P rats vs NP rats.

We used a generalized linear model to characterize the correlations between the transcription levels of lncRNAs and genes. Among the 2873 genes significantly correlating (FDR<0.2, p-value<0.005) with the differentially expressed lncRNAs, 120 are correlated with more than 3 lncRNAs. Several of lncRNA correlating genes are known as associated with alcohol dependence, including ALDH1A1, ALDH9A1, GABRA2,

CHRM2, PDYN and CNR1. We conducted a pathway analysis on these 120 genes with Ingenuity Pathway Analysis (IPA) and found that physiological function most frequently associated to lncRNAs is nervous system development and functions.

Among all the differentially expressed lncRNAs, 22 correlate with more than 30 genes each. 15 of these lncRNAs correlate with genes enriched in nervous system development and function, 8 with neurological diseases, and 6 with behaviors. Moreover, 10 of the 22 lncRNAs are associated to genetic disorder, which may reveal the hereditary nature of alcoholism.

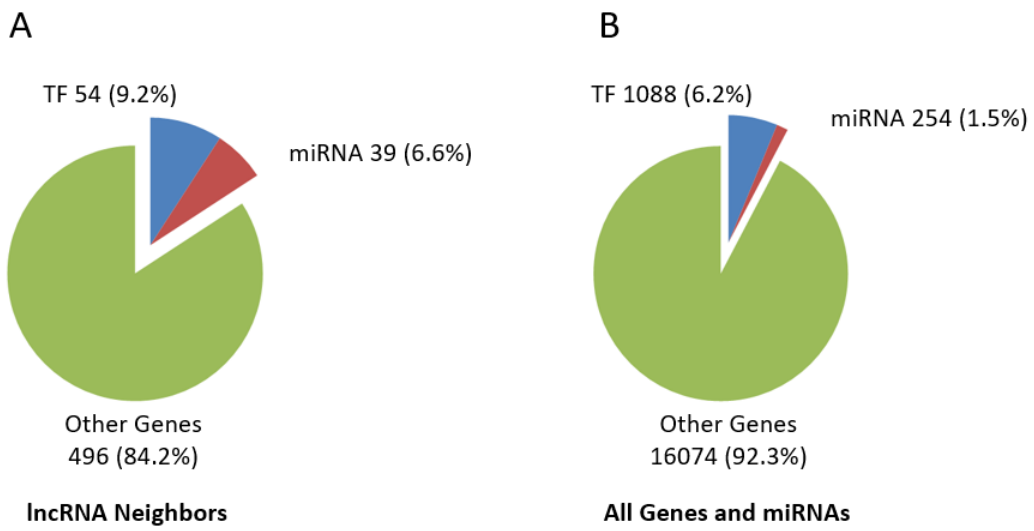

Figure 6.4 Potential cis-regulation of lncRNA.

The pie-chart demonstrates the percentage of transcription factors, miRNA and other genes in lncRNA neighbors and in all genes. The proportion of transcription factors and miRNAs in lncRNA neighbors is significantly higher ($p=2.2\times10\text{-}16$) than that in all genes.

Among genes proximal to lncRNAs differentially expressed between P and NP, 9.2% are annotated as transcription factors, and 6.6% as miRNA; for genes not proximal

to differentially expressed lncRNAs, however, this proportion dropped to 6.2% and 1.5% for transcription factors and miRNAs. This difference is significant with $\chi 2$ p-value 2.2×10-16, suggesting that many of the novel lncRNAs may be associated with neighboring transcription factors and miRNAs, and work in a cis-acting manner.

6.3 Discussion

We report an RNA-seq experiment on the hippocampus of P and NP rats, and a bioinformatics strategy to identify lncRNAs from the RNA-seq information and characterize their roles in alcohol preference. Our strategy includes four components, orthologous lncRNA region mapping from mouse to rat, RNA-seq on P and NP rats, lncRNA annotation and pathway characterization. We identified 420 lncRNAs, 37 of which are differentially expressed across P and NP rats. By applying a generalized linear model to differentially expressed lncRNAs and protein-coding genes, we derived 3699 significantly correlated lncRNA-gene pairs involving 2873 genes. We created a set of significantly correlated genes for each lncRNA, and inferred their functional roles by pathway analysis. The result revealed that 15 are significantly correlated with nervous system development and function. Our statistical analysis also revealed that the proportion of TF and miRNA are significantly higher among the lncRNA neighboring genes than other genes, implying a cis-acting mechanism (Figure 6.4).

Evidence was found supporting the existence and potential regulation functions of the differentially expressed lncRNAs. Region1384_rev is significantly correlated with 824 genes, of which 61 are significantly associated with nervous system development and function; it is located in the promoter (179 nt upstream of transcription start site) of a protein-coding gene CHD2, whose product alters gene expression by modification of chromatin structures [282]. Given the observations that many lncRNAs locating in the promoter of protein-coding genes possess regulation functions on the corresponding genes, the location of region1384_rev suggests a tremendous possibility of a regulatory

role upon CHD2, and thus regulating a large group of genes by chromatin modification. Moreover, we observed several rat ESTs and orthologous non-coding genes of mouse and human within this region, verifying the existence of this lncRNA. (Figure 6.5)

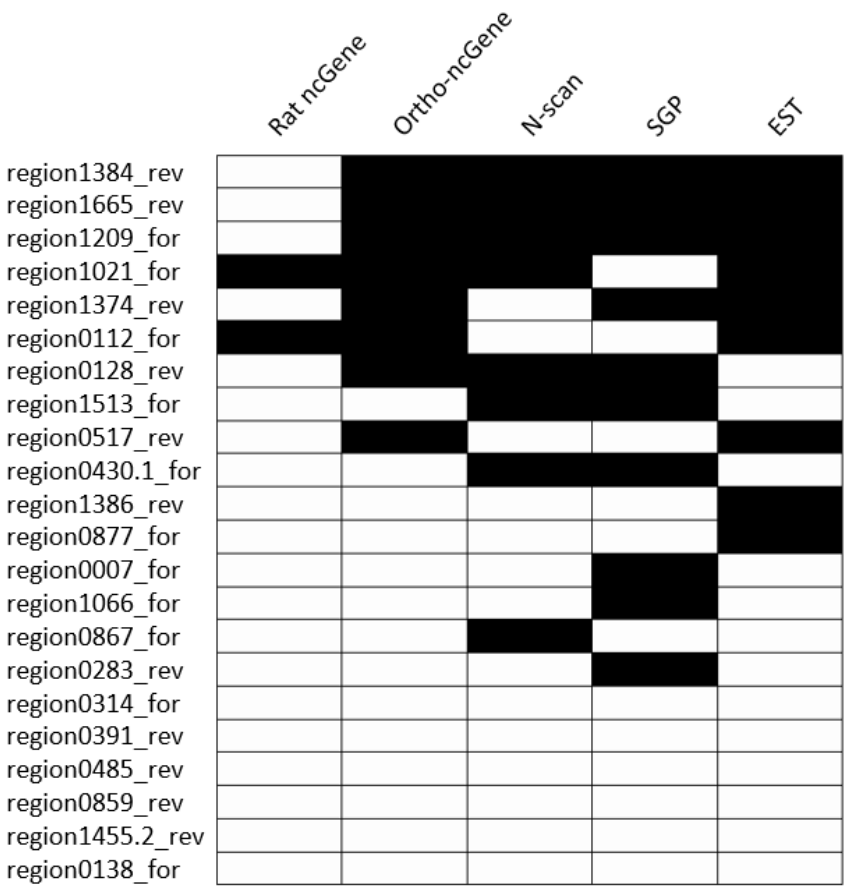

Figure 6.5 Observations supporting the existence of lncRNA.

A dark grid indicates that evidence was found for the corresponding lncRNA, while a white grid indicates no such evidence was found. Rat ncGene stands for rat non-coding genes; Ortho-ncGene stands for orthologous non-coding genes; N-scan and SGP stands for N-scan and SGP gene prediction, respectively; EST stands for expressed sequence tags. The lncRNAs are sorted by the number of evidences found.

Evidence was also found in support of the existence of other differentially expressed lncRNAs and their potential functions on gene regulation and signal transduction. Four lncRNAs (region0877_for, region1066_for, region0112_for, and region0138_for) were identified sequentially adjacent to zinc-finger proteins, which were generally found as a component of transcription factors; besides, region0867_for was identified neighboring a gene coding transcription factor, and region0283_rev was found near a gene coding a DNA binding protein; these observations indicates that lncRNA may regulate the expression of transcription factor genes. Moreover, two lncRNAs (region0314_for and region0007_for) were identified adjacent to genes coding G-protein regulation proteins (Rho GTPase activating protein 5 and Rho guanine nucleotide exchange factor), which implies that lncRNAs may also play roles in cell signaling. Region1374_rev was found next to a small-nuclear RNA that is involved in snRNA modifications, and region0430.1_for was next to an RNA motif binding protein; this indicates that lncRNAs may also be involved in RNA regulations.

The “K4-K36” domain only represents a transcription active region that may be a gene, it is unable to differentiate an intronic lncRNA from a novel exon. Therefore our strategy focused on intergenic ncRNAs only, we may have missed the lncRNAs located in intronic regions and untranslated regions. In addition, we required the expression of all candidate lncRNAs be higher on one of the two strands to eliminate noises and error in sequencing data. Because of this criterion, we may also have missed the lncRNAs that are transcribed on an antisense strand of a non-coding gene. Nevertheless, once we have new transcriptome data with deeper sequencing and longer reads, we will be able to identify more lncRNAs from intragenic regions.

In the future, we plan to conduct RT-PCR experiments to validate the transcriptional activity of the lncRNAs which are likely to be associated to gene regulations in alcohol preference. We also plan to sequence more brain regions of rats at

different drinking levels, thus to characterize the tissue related functions of lncRNA and their association with different drinking scores.

6.4 Methods

6.4.1 Eliminating Protein-coding Regions

The annotation library of rat genes and orthologous regions in human and mouse were downloaded from UCSC table browser. To remove non-coding genes from the library, the genes that cannot be mapped to UniProt accessions were eliminated. Then the putative lncRNA regions were superimposed to the protein-coding genes and classified into three categories according to the overlap length, 1) non-overlapping, 2) complete overlapping and 3) partial overlapping. The putative lncRNA regions not overlapping with protein-coding regions were retained for next step, while the complete overlapping regions were discarded. For partially overlapped regions, we truncated the overlap and shortened the lncRNA regions, if transcriptional activity were detected outside the overlap.

6.4.2 Determining Transcriptional Strand Preference

To determine the transcriptional strand preference within the putative lncRNA regions, we firstly calculated two transcription intensities in RPKM for each lncRNA region in each sample, one is for forward strand and the other is for reverse strand. Then we conducted a Friedman test to compare the transcriptional intensities on different strands. A putative lncRNA region with p-value<0.01 was defined as significantly strand biased and the higher expressed strand were defined as the sense strand, while the insignificant regions were eliminated. Both the RPKM calculation and Friedman test were implemented with Partek® Genomics Suite® software, version 6.6 Copyright ©; 2016 Partek Inc., St. Louis, MO, USA.

6.4.3 Refining Exon Structures

To reduce computational time, we eliminated all the RNA-seq reads outside the putative lncRNA regions and on the antisense strand. Then SAMtools was used to pileup all the reads. A list of expressed chunks was generated based on the pileup file. If the distance between two chunks was less than 25 basepairs, they were merged into one chunk. If the number of reads covered by a chunk was less than 8, this chunk was discarded. The derived chunks were defined as putative exons and mapped to the lncRNA regions. A refflat file was generated for each lncRNA regions, annotating their coordinates and putative exon structures.

6.4.4 Detecting Protein-coding Potential

The sequences of the putative exons were extracted and aligned against SwissProt with BlastX. In the cases that several alignments were generated for one exon sequence, we only retained the one with the highest alignment score. Then we calculated the average log10(E-value) of all exons for each lncRNA to evaluate their protein-coding potential. LncRNA regions with an average log10(E-value) < -3 were discarded; the rest regions were defined as non-coding potential regions.

6.4.5 Deriving Significantly Correlated LncRNA-gene Pairs

It is reasonable to assume that number of sequence tags identified in each lncRNA region follows a Poisson distribution; we therefore used a generalized linear model to infer the relationship between the expression intensity of lncRNAs and protein-coding genes.

$$Y_{ik} \sim Poisson(\mu_{ijk})$$

$$\log \mu_{ijk} = N_{ik} + \alpha + \beta_1 s_k + \beta_2 g_{jk} + \beta_3 b_k$$

In this model, i, j, and k are the indices of lncRNA, gene and animal. Y_{ik} is the number of RNA-seq read counts in the region of lncRNA i in animal k; μ_{ijk} is the expected value of Y_{ik}; Nik is a constant value that serves as a normalization factor to

balance sample and lncRNA specific variation. Here, Nij=log(K_i)+log(M_k), where K_i is the length of exon model of lncRNA i and Mk is the total number of mappable reads for sample k. s_k is the strain of animal k (P or NP); g_{ik} is the transcription intensity (RPKM) of gene i in animal k; b_k is the batch effect of the experiment (Run1 or Run2). The significance of β_2 was used to evaluate the correlation between lncRNA i and gene j.

To simplify the model and derive the most trustworthy results, we assumed that there is no interaction between strain and gene transcription intensity. To rule out the cases that the interaction may exist, we used another generalized linear model to identify these cases.

$$Y_{ik} \sim Poisson(\mu_{ijk})$$

$$\log \mu_{ijk} = N_{ik} + \alpha + \beta_1 s_k + \beta_2 g_{jk} + \beta_3 b_k + \beta_4 s_k g_{jk}$$

If the coefficient of strain-gene interaction term β4 was significant ($p \leq 0.05$), then the corresponding lncRNA-gene pair was discarded.

To increase the reproducibility of the results, we required that all lncRNAs and genes should have transcriptional intensities more than 0.5 RPKM in at least 8 samples. All lncRNA-gene pairs involving ineligible lncRNA or genes were discarded.

Chapter 7. Conclusions and Discussions

7.1 Research Summary and Contributions

About 150 years after the first observation of nucleic acids, the advance of high-throughput sequencing technologies has unleashed numerous possibilities in genetic and genomic research. This dissertation demonstrated 5 use cases in transcriptome research facilitated by mRNA sequencing and protein mass spectrometry.

Chapter 2, 3 and 4 covered identification methods for both known and novel alternative splicing (AS) events, and made functional insights on these AS events that can be utilized in hypothesis generation for wet-lab research. In Chapter 2, we identified and characterized stimulant induced AS events. In Chapter 3, we improved the AS event identification technique and applied it for identifying AS events that are regulated across three human development stages (fetal, pediatric and adult). In the two chapters above, we implemented functional analysis on AS events from several aspects, including gene function, disruption on protein structure and interaction, as well as RNA-binding proteins (RBP) and genomic regions that may be involved in AS regulation. In Chapter 4, we developed a method for identifying novel AS events from RNA-seq results. We validated the novel AS events by visualizing their Sashimi plots [97].

In Chapter 5, we extended our investigation on AS to the proteome level. We established a peptide sequence database PEPPI [62] for exons and exon junctions. This database enables researchers to identify splicing junctions from MS/MS results.

In Chapter 6, we developed a set of methodologies to identify long non-coding RNAs (lncRNAs) from RNA-seq data. Most lncRNAs share the same transcription mechanism with mRNAs and has poly-A tails, which enables them to present in mRNA sequencing data derived with poly-A selection. However, many analyses only focused on mRNAs and overlooked lncRNAs, rendering this precious information into a forgotten

treasure. Therefore we developed these methods and turned mRNA sequencing data into a gold mine of lncRNA information.

7.2 Future Research Directions

7.2.1 Biochemical Validation on Discovered Splicing Events and lncRNA

In this dissertation, we discovered and characterized differentially regulated alternative splicing events and lncRNAs that are associated to biological functions, and validated them with prior knowledge published in publications and databases. However, we haven't got the time and resource to implement biochemical validations on these discoveries. The presence of these alternative splicing events and lncRNA can be validated by PCR and Sanger sequencing experiments. The protein-protein interactions may be validated with yeast two-hybrid approach or chip-based analysis. The RBP-RNA interactions may be validated with CLIP sequencing. Novel lncRNAs may be knocked off or knocked down to examine their biological functions.

7.2.2 Better Sequencing Technology Enables Better Results

All RNA-seq data utilized in this dissertation are single-ended, with the longest read length of 75nt. By the year of 2017, Illumina provides paired-end sequencing solutions with 100~150nt on each end at affordable prices. Most of exons are less than 200bp in length [283]. Therefore most of the paired-end reads will be covering at least one splice junction. Employing the latest sequencing technology will enormously improve the number and accuracy of identifiable AS events. It will allow researchers to accurately deduct the genomic structure of lncRNAs.

On the other hand, emerging long read length sequencing technologies may become rather helpful for alternative splicing and lncRNA research. The Oxford Nanopore sequencing technology is capable of achieving the median read length around 1200bp [284, 285]. Pacific Biosciences sequencing technology can achieve an even longer median read length of 30,000bp [22, 286]. These sequencing technologies will

provide informative long length reads covering multiple splicing junctions and even the whole length transcript. With such information, alternative splicing research will not be limited to individual splicing events. Instead, the sequence and the expression intensity of different RNA isoforms can both be derived accurately.

7.2.3 DNA Variations Affect AS

The regulation of AS depends on the facilitation of AS related RNA-binding proteins (RBP). Change of the sequence in binding regions (ESE, ESS, ISE, ISS) on RNA may affect the binding affinity between the RNA molecule and RBP. While the exon sequence is generally available from RNA-seq data, the intron sequences are mostly unavailable. DNA sequencing allows users to identify variations in introns, and thus systematically associate variations in AS regulatory regions and change of AS.

7.2.4 Prioritization of AS Events

In this dissertation we characterized AS from several aspects, including gene function, PTM sites, effect on protein structure and interactions. Such information can also be collected for disease causing insertion/deletion variations. The effect of insertion/deletion variations (indels) is similar to AS at the protein level. Therefore we may use the genomic and protein structural information of disease causing indels as a training set for prioritizing AS events that may be associated with disease.

7.2.5 Algorithm for AS Event Validation

MISO [42] and rMATS [43] provided solutions for quantifying the percentage of splicing isoforms (PSI or Ψ) as well as the significance of AS change across two biological conditions. However, we observed many “significantly changed AS events” showing unrealistic base coverages in Sashimi-plots [42]. Such unrealistic base coverages are generally induced by the variability and limitations of the RNA-seq technology. In this dissertation, we manually reviewed the Sashimi-plot for each AS event to make sure it is biologically valid. This will become a bottleneck if we need to automate AS identification

and characterization. In the future, we may utilize machine learning methods in image analysis to automate this scanning process on the base coverages of AS events.

Supplementary Materials

Table S-1 The function and localization of alternatively spliced genes

AS Type	Gene Symbol	Gene Description	Gene Location	Gene Type	ΔΨ
cassete exon	Mgrn1	mahogunin ring finger 1, E3 ubiquitin protein ligase	Cytoplasm	enzyme	-0.25
cassete exon	Smox	spermine oxidase	Cytoplasm	enzyme	0.37
cassete exon	Rhot1	ras homolog family member T1	Cytoplasm	enzyme	-0.15
cassete exon	Magi3	membrane associated guanylate kinase, WW and PDZ domain containing 3	Cytoplasm	kinase	-0.42
cassete exon	Pank2	pantothenate kinase 2	Cytoplasm	kinase	-0.21
cassete exon	Mark3	MAP/microtubule affinity-regulating kinase 3	Cytoplasm	kinase	-0.22
cassete exon	Camk1d	calcium/calmodulin-dependent protein kinase ID	Cytoplasm	kinase	-0.31
cassete exon	Cdc42bpa	CDC42 binding protein kinase alpha (DMPK-like)	Cytoplasm	kinase	-0.12
cassete exon	Plscr2	phospholipid scramblase 2	Cytoplasm	other	0.16
cassete exon	Fopnl	FGFR1OP N-terminal like	Cytoplasm	other	0.11
cassete	Cyb561a	cytochrome b561 family,	Cytoplasm	other	0.37

exon	3	member A3			
cassete exon	Mpv17	MpV17 mitochondrial inner membrane protein	Cytoplasm	other	0.23
cassete exon	Tbc1d31	TBC1 domain family, member 31	Cytoplasm	other	0.22
cassete exon	Picalm	phosphatidylinositol binding clathrin assembly protein	Cytoplasm	other	-0.06
cassete exon	Numbl	numb homolog (Drosophila)-like	Cytoplasm	other	-0.39
cassete exon	Arhgef11	Rho guanine nucleotide exchange factor (GEF) 11	Cytoplasm	other	0.29
cassete exon	Abi1	abl-interactor 1	Cytoplasm	other	0.14
cassete exon	Spc25	SPC25, NDC80 kinetochore complex component	Cytoplasm	other	0.15
cassete exon	Fhl1	four and a half LIM domains 1	Cytoplasm	other	-0.05
cassete exon	Mob4	MOB family member 4, phocein	Cytoplasm	other	0.25
cassete exon	Spag9	sperm associated antigen 9	Cytoplasm	other	0.2
cassete exon	Mphosph 9	M-phase phosphoprotein 9	Cytoplasm	other	-0.36
cassete exon	Plec	plectin	Cytoplasm	other	0.05
cassete	Xpnpep3	X-prolyl aminopeptidase	Cytoplasm	peptid	0.27

exon		(aminopeptidase P) 3, putative		ase	
cassete exon	Blmh	bleomycin hydrolase	Cytoplasm	peptidase	0.25
cassete exon	Fabp5	fatty acid binding protein 5 (psoriasis-associated)	Cytoplasm	transporter	0.1
cassete exon	Copg2	coatomer protein complex, subunit gamma 2	Cytoplasm	transporter	0.2
cassete exon	Rabep1	rabaptin, RAB GTPase binding effector protein 1	Cytoplasm	transporter	0.14
cassete exon	Pctp	phosphatidylcholine transfer protein	Cytoplasm	transporter	0.39
cassete exon	Arl13b	ADP-ribosylation factor-like 13B	Extracellular Space	other	-0.29
cassete exon	Slit2	slit homolog 2 (Drosophila)	Extracellular Space	other	0.53
cassete exon	Baz2b	bromodomain adjacent to zinc finger domain, 2B	Extracellular Space	other	-0.39
cassete exon	Suv420h1	suppressor of variegation 4-20 homolog 1 (Drosophila)	Nucleus	enzyme	0.12
cassete exon	Clk4	CDC-like kinase 4	Nucleus	kinase	0.18
cassete exon	Uty	ubiquitously transcribed tetratricopeptide repeat gene, Y chromosome	Nucleus	other	0.25
cassete	Sun1	Sad1 and UNC84 domain	Nucleus	other	-0.21

exon		containing 1			
cassete exon	Maf1	MAF1 homolog (S. cerevisiae)	Nucleus	other	-0.2
cassete exon	Morf4l2	mortality factor 4 like 2	Nucleus	other	0.17
cassete exon	Hmgxb4	HMG box domain containing 4	Nucleus	other	-0.22
cassete exon	Phf20	PHD finger protein 20	Nucleus	other	-0.13
cassete exon	Zfp120	zinc finger protein 932	Nucleus	other	-0.33
cassete exon	Ift122	intraflagellar transport 122 homolog (Chlamydomonas)	Nucleus	other	0.38
cassete exon	Phf7	PHD finger protein 7	Nucleus	other	-0.46
cassete exon	Ctnnd1	catenin (cadherin-associated protein), delta 1	Nucleus	other	-0.22
cassete exon	Rad18	RAD18 homolog (S. cerevisiae)	Nucleus	other	0.23
cassete exon	Senp7	SUMO1/sentrin specific peptidase 7	Nucleus	peptidase	-0.52
cassete exon	Depdc1a	DEP domain containing 1	Nucleus	TR	0.16
cassete exon	Ybx3	Y box binding protein 3	Nucleus	TR	-0.12
cassete	Ncor1	nuclear receptor corepressor	Nucleus	TR	-0.33

exon		1			
cassete exon	Kansl2	KAT8 regulatory NSL complex subunit 2	Other	enzyme	0.22
cassete exon	Ube2q2	ubiquitin-conjugating enzyme E2Q family member 2	Other	enzyme	0.31
cassete exon	Kansl2	KAT8 regulatory NSL complex subunit 2	Other	enzyme	0.22
cassete exon	Rnf214	ring finger protein 214	Other	other	0.37
cassete exon	Tmem161b	transmembrane protein 161B	Other	other	0.19
cassete exon	Zfp740	zinc finger protein 740	Other	other	-0.3
cassete exon	Ubl4a	Slc10a3-Ubl4 readthrough	Other	other	-0.12
cassete exon	Asb7	ankyrin repeat and SOCS box containing 7	Other	other	0.36
cassete exon	Lins	lines homolog (Drosophila)	Other	other	0.24
cassete exon	Ttc13	tetratricopeptide repeat domain 13	Other	other	-0.21
cassete exon	Slx4ip	SLX4 interacting protein	Other	other	-0.3
cassete exon	Ppp4r1l-ps	protein phosphatase 4, regulatory subunit 1-like,	Other	other	-0.41

		pseudogene			
cassete exon	Smim8	small integral membrane protein 8	Other	other	0.27
cassete exon	Zmym4	zinc finger, MYM-type 4	Other	other	0.47
cassete exon	Smim8	small integral membrane protein 8	Other	other	0.27
cassete exon	Svil	supervillin	Other	other	0.22
cassete exon	Adprm	ADP-ribose/CDP-alcohol diphosphatase, manganese-dependent	Other	other	0.34
cassete exon	Tbc1d9b	TBC1 domain family, member 9B (with GRAM domain)	Other	other	-0.4
cassete exon	Pwwp2a	PWWP domain containing 2A	Other	other	0.53
cassete exon	Adprm	ADP-ribose/CDP-alcohol diphosphatase, manganese-dependent	Other	other	0.34
cassete exon	Smim8	small integral membrane protein 8	Other	other	0.27
cassete exon	2810474 O19Rik	KIAA1551	Other	other	0.28
cassete exon	Usp45	ubiquitin specific peptidase 45	Other	peptid ase	-0.41

cassete exon	Prepl	prolyl endopeptidase-like	Other	peptidase	0.14
cassete exon	Rnf14	ring finger protein 14	Other	TR	-0.19
cassete exon	Tmem11	transmembrane protein 11	PM	GPCR	0.16
cassete exon	Ttll7	tubulin tyrosine ligase-like family, member 7	PM	other	0.38
cassete exon	Aif1l	allograft inflammatory factor 1-like	PM	other	0.32
cassete exon	Aif1l	allograft inflammatory factor 1-like	PM	other	0.32
cassete exon	Dnajc5	DnaJ (Hsp40) homolog, subfamily C, member 5	PM	other	-0.18
cassete exon	Cpeb4	cytoplasmic polyadenylation element binding protein 4	PM	other	0.52
cassete exon	Tpm1	tropomyosin 1, alpha	PM	other	-0.05
cassete exon	Jmjd6	jumonji domain containing 6	PM	TMR	0.14
A3SS	Eci2	enoyl-CoA delta isomerase 2	Cytoplasm	enzyme	0.16
A3SS	Cyp4f16	cytochrome P450, family 4, subfamily f, polypeptide 16	Cytoplasm	enzyme	-0.49
A3SS	Rabggtb	Rab geranylgeranyltransferase,	Cytoplasm	enzyme	0.09

		beta subunit			
A3SS	Akt1	v-akt murine thymoma viral oncogene homolog 1	Cytoplasm	kinase	0.29
A3SS	Ppip5k2	diphosphoinositol pentakisphosphate kinase 2	Cytoplasm	kinase	0.26
A3SS	Cdv3	CDV3 homolog (mouse)	Cytoplasm	other	0.05
A3SS	Dock9	dedicator of cytokinesis 9	Cytoplasm	other	0.48
A3SS	Akt1s1	AKT1 substrate 1 (proline-rich)	Cytoplasm	other	0.08
A3SS	Scoc	short coiled-coil protein	Cytoplasm	other	-0.23
A3SS	Becn1	beclin 1, autophagy related	Cytoplasm	other	-0.1
A3SS	Blmh	bleomycin hydrolase	Cytoplasm	peptidase	0.25
A3SS	Rrbp1	ribosome binding protein 1	Cytoplasm	transporter	-0.26
A3SS	Arfgap1	ADP-ribosylation factor GTPase activating protein 1	Cytoplasm	transporter	-0.4
A3SS	Sparc	secreted protein, acidic, cysteine-rich (osteonectin)	Extracellular Space	other	0.1
A3SS	Bod1l	biorientation of chromosomes in cell division 1-like 1	Extracellular Space	other	0.34
A3SS	Mettl3	methyltransferase like 3	Nucleus	enzyme	0.45
A3SS	Tyms	thymidylate synthetase	Nucleus	enzyme	0.09

A3SS	Mcm9	minichromosome maintenance complex component 9	Nucleus	enzyme	-0.46
A3SS	Rev1	REV1, polymerase (DNA directed)	Nucleus	enzyme	-0.22
A3SS	Zfp346	zinc finger protein 346	Nucleus	other	0.39
A3SS	Chchd1	coiled-coil-helix-coiled-coil-helix domain containing 1	Nucleus	other	-0.12
A3SS	Fra10ac1	fragile site, folic acid type, rare, fra(10)(q23.3) or fra(10)(q24.2) candidate 1	Nucleus	other	0.43
A3SS	Srrt	serrate RNA effector molecule homolog (Arabidopsis)	Nucleus	other	0.22
A3SS	Ivns1abp	influenza virus NS1A binding protein	Nucleus	other	0.07
A3SS	Hnrnpr	heterogeneous nuclear ribonucleoprotein R	Nucleus	other	0.18
A3SS	Cdc27	cell division cycle 27	Nucleus	other	-0.2
A3SS	Phrf1	PHD and ring finger domains 1	Nucleus	other	-0.33
A3SS	Nolc1	nucleolar and coiled-body phosphoprotein 1	Nucleus	TR	0.14
A3SS	Htatip2	HIV-1 Tat interactive protein 2, 30kDa	Nucleus	TR	0.4
A3SS	Rbm39	RNA binding motif protein	Nucleus	TR	-0.21

		39			
A3SS	Hdac10	histone deacetylase 10	Nucleus	TR	-0.44
A3SS	Nfya	nuclear transcription factor Y, alpha	Nucleus	TR	-0.32
A3SS	Ankzf1	ankyrin repeat and zinc finger domain containing 1	Nucleus	TR	-0.42
A3SS	Fus	fused in sarcoma	Nucleus	TR	0.21
A3SS	Sbno1	strawberry notch homolog 1 (Drosophila)	Other	enzyme	0.3
A3SS	4833420G17Rik	chromosome 5 open reading frame 34	Other	other	-0.38
A3SS	Eml3	echinoderm microtubule associated protein like 3	Other	other	-0.42
A3SS	Phf20l1	PHD finger protein 20-like 1	Other	other	-0.27
A3SS	Ubl4a	Slc10a3-Ubl4 readthrough	Other	other	-0.12
A3SS	Dda1	DET1 and DDB1 associated 1	Other	other	0.07
A3SS	Ttpal	tocopherol (alpha) transfer protein-like	Other	other	0.41
A3SS	Miip	migration and invasion inhibitory protein	Other	other	0.18
A3SS	Ppp4r1	protein phosphatase 4, regulatory subunit 1	Other	phosphatase	0.23
A3SS	Ly6a	lymphocyte antigen 6 complex, locus A	PM	other	0.44
A3SS	Prnd	prion protein 2 (dublet)	PM	other	-0.45

A5SS	Ndufs1	NADH dehydrogenase (ubiquinone) Fe-S protein 1, 75kDa (NADH-coenzyme Q reductase)	Cytoplasm	enzyme	-0.24
A5SS	Birc6	baculoviral IAP repeat containing 6	Cytoplasm	enzyme	0.28
A5SS	Nit1	nitrilase 1	Cytoplasm	enzyme	-0.33
A5SS	Map3k7	mitogen-activated protein kinase kinase kinase 7	Cytoplasm	kinase	-0.4
A5SS	Ehbp1l1	EH domain binding protein 1-like 1	Cytoplasm	other	-0.05
A5SS	Cmc2	COX assembly mitochondrial protein 2 homolog (S. cerevisiae)	Cytoplasm	other	0.1
A5SS	Spc25	SPC25, NDC80 kinetochore complex component	Cytoplasm	other	0.15
A5SS	Dnajb12	DnaJ (Hsp40) homolog, subfamily B, member 12	Cytoplasm	other	-0.52
A5SS	Mrps33	mitochondrial ribosomal protein S33	Cytoplasm	other	-0.22
A5SS	Mkks	McKusick-Kaufman syndrome	Cytoplasm	other	-0.17
A5SS	Ddx6	DEAD (Asp-Glu-Ala-Asp) box helicase 6	Nucleus	enzyme	0.14
A5SS	Fen1	flap structure-specific	Nucleus	enzym	0.1

		endonuclease 1		e	
A5SS	Ddx6	DEAD (Asp-Glu-Ala-Asp) box helicase 6	Nucleus	enzyme	0.14
A5SS	Naa10	N(alpha)-acetyltransferase 10, NatA catalytic subunit	Nucleus	enzyme	0.28
A5SS	Adarb1	adenosine deaminase, RNA-specific, B1	Nucleus	enzyme	0.23
A5SS	Nr1h2	nuclear receptor subfamily 1, group H, member 2	Nucleus	ligand-dependent nuclear receptor	0.29
A5SS	Srsf7	serine/arginine-rich splicing factor 7	Nucleus	other	0.19
A5SS	Srsf7	serine/arginine-rich splicing factor 7	Nucleus	other	0.19
A5SS	Zfp60	zinc finger protein 60	Nucleus	other	-0.29
A5SS	Hira	histone cell cycle regulator	Nucleus	TR	0.15
A5SS	Yeats2	YEATS domain containing 2	Nucleus	TR	-0.47
A5SS	Tmem234	transmembrane protein 234	Other	other	0.31
A5SS	Arl16	ADP-ribosylation factor-like 16	Other	other	-0.3
A5SS	Zfp740	zinc finger protein 740	Other	other	-0.3

A5SS	Efr3a	EFR3 homolog A (S. cerevisiae)	PM	other	-0.1
A5SS	Smek1	SMEK homolog 1, suppressor of mek1 (Dictyostelium)	PM	other	0.22
A5SS	Inpp5a	inositol polyphosphate-5-phosphatase, 40kDa	PM	phosphatase	0.2
A5SS	Jmjd6	jumonji domain containing 6	PM	TMR	0.14
retained intron	Hspa8	heat shock 70kDa protein 8	Cytoplasm	enzyme	0.12
retained intron	Trim2	tripartite motif containing 2	Cytoplasm	enzyme	0.17
retained intron	Coq6	coenzyme Q6 monooxygenase	Cytoplasm	enzyme	0.38
retained intron	Dgkq	diacylglycerol kinase, theta 110kDa	Cytoplasm	kinase	0.29
retained intron	Yipf2	Yip1 domain family, member 2	Cytoplasm	other	0.32
retained intron	Rasa2	RAS p21 protein activator 2	Cytoplasm	other	0.24
retained intron	Wipi2	WD repeat domain, phosphoinositide interacting 2	Cytoplasm	other	0.19
retained intron	Rps18	ribosomal protein S18	Cytoplasm	other	0.09

retained intron	Wdr11	WD repeat domain 11	Cytoplasm	other	-0.21
retained intron	Golga1	golgin A1	Cytoplasm	other	0.15
retained intron	Becn1	beclin 1, autophagy related	Cytoplasm	other	-0.1
retained intron	Lrrc45	leucine rich repeat containing 45	Cytoplasm	other	-0.28
retained intron	Ambra1	autophagy/beclin-1 regulator 1	Cytoplasm	other	-0.37
retained intron	Eif4a2	eukaryotic translation initiation factor 4A2	Cytoplasm	TLR	0.09
retained intron	Eif4a2	eukaryotic translation initiation factor 4A2	Cytoplasm	TLR	0.09
retained intron	Tmem214	transmembrane protein 214	Extracellular Space	other	-0.4
retained intron	Msh3	mutS homolog 3	Nucleus	enzyme	0.31
retained intron	Nle1	notchless homolog 1 (Drosophila)	Nucleus	enzyme	0.39
retained intron	Sirt7	sirtuin 7	Nucleus	enzyme	0.27
retained intron	Nek8	NIMA-related kinase 8	Nucleus	kinase	-0.38
retained intron	Mis18a	MIS18 kinetochore protein homolog A (S. pombe)	Nucleus	other	0.16

retained intron	9930012 K11Rik	chromosome 8 open reading frame 58	Nucleus	other	0.25
retained intron	Ncaph2	non-SMC condensin II complex, subunit H2	Nucleus	other	-0.08
retained intron	Lrif1	ligand dependent nuclear receptor interacting factor 1	Nucleus	other	0.28
retained intron	Xab2	XPA binding protein 2	Nucleus	other	0.23
retained intron	Sapcd2	suppressor APC domain containing 2	Nucleus	other	0.39
retained intron	Cdan1	codanin 1	Nucleus	other	0.34
retained intron	Hnrnph1	heterogeneous nuclear ribonucleoprotein H1 (H)	Nucleus	other	-0.06
retained intron	Gtf2a2	general transcription factor IIA, 2, 12kDa	Nucleus	TR	0.28
retained intron	Hsf1	heat shock transcription factor 1	Nucleus	TR	-0.34
retained intron	Tfe3	transcription factor binding to IGHM enhancer 3	Nucleus	TR	0.16
retained intron	Trim28	tripartite motif containing 28	Nucleus	TR	0.25
retained intron	Snapc4	small nuclear RNA activating complex, polypeptide 4, 190kDa	Nucleus	TR	-0.31
retained	Mroh1	maestro heat-like repeat	Other	other	0.2

intron		family member 1			
retained intron	Morc4	MORC family CW-type zinc finger 4	Other	other	-0.3
retained intron	Dus1l	dihydrouridine synthase 1-like (S. cerevisiae)	Other	other	0.15
retained intron	Slc7a6os	solute carrier family 7, member 6 opposite strand	Other	other	-0.14
retained intron	Psmd5	proteasome (prosome, macropain) 26S subunit, non-ATPase, 5	Other	other	-0.06
retained intron	Sharpin	SHANK-associated RH domain interactor	PM	other	0.22
retained intron	Lmbr1l	limb development membrane protein 1-like	PM	other	-0.4
retained intron	Jmjd6	jumonji domain containing 6	PM	TMR	0.14
retained intron	Slc39a7	solute carrier family 39 (zinc transporter), member 7	PM	transporter	-0.1

TR: transcription regulator

TMR: transmembrane receptor

TLR: translation regulator

GPCR: G-protein coupled receptor

PM: plasma membrane

Table S-2 Peptides identified by PEPPI database

Peptide	peppi	SNP	Gene	Splicing	FDR	Mass	Theo Mass
AGEYGAEALER	PEP006194756	rs33943087(A/G)	ENSG00000188536	E_E_KB	0.0032	1165.19	1164.54
APVPTGEVYFADSFDR	PEP003042584		ENSG00000127022	EXON_KB	0.0073	1769.79	1769.83
ASSVVVSGTPIR	PEP001086304		ENSG00000103876	EXON_KB	0.0100	1173.11	1171.66
ATSFLLALEPELEAR	PEP004603386		ENSG00000163902	EXON_KB	0.0031	1660.01	1658.89
AVFPSIVGR	PEP006209453	rs11549244(C/T)	ENSG00000184009	EXON_KB	0.0031	945.881	944.543
AVFVDLEPTVLDEVR	PEP006626718	rs36215077(G/C)	ENSG00000198033	E_E_KB	0.0067	1701.74	1700.9
AVFVNLEPTVIDEVR	PEP002507811	rs11546624(G/A)	ENSG00000123416	E_E_KB	0.0066	1700.66	1699.91
AVNTLNEALEF	PEP000166109		ENSG00000021821	E_E_KB	0.0088	1418.72	1418.74

AK			26				
DLLDDLKSELTGK	PEP004712877		ENSG00000164111	EXON_KB	0.0062	1445.69	1445.76
DPQLVPILIEAAR	PEP003888536		ENSG00000147383	EXON_KB	0.0036	1435.05	1433.83
EGDVLTLLESER	PEP007510716		ENSG00000227097	EXON_KB	0.0090	1359.8	1359.69
EGDVLTLLESER	PEP007749514		ENSG00000233927	EXON_KB	0.0090	1359.8	1359.69
ELFSNLQEFAGPSGK	PEP006264479		ENSG00000185432	EXON_KB	0.0035	1622.75	1622.79
ETTIQGLDGLSER	PEP003467192		ENSG00000136872	E_E_KB	0.0066	1417.67	1417.71
FGGLLLTEKPIVLK	PEP004835823		ENSG00000160870	EXON_KB	0.0032	1526.87	1526.94
FLEQQDQVLQTK	PEP005639485	rs61726455(A/G)	ENSG00000172867	EXON_KB	0.0089	1477.5	1475.76
FPGQLN	PEP001		ENSG00	EXO	0.0085	1131.11	1129.59

ADLR	260206		0001011 62	N_K B			
FPGQLN ADLR	PEP001 635184		ENSG00 0001048 33	EXO N_K B	0.0085	1131.11	1129.59
FPGQLN ADLR	PEP003 177389		ENSG00 0001372 67	EXO N_K B	0.0085	1131.11	1129.59
FPGQLN ADLR	PEP003 198554	rs1054 331(G/ C)	ENSG00 0001372 85	EXO N_K B	0.0085	1131.11	1129.59
FVTVQT ISGTGA LR	PEP002 709222		ENSG00 0001251 66	EXO N_K B	0.0100	1448.73	1448.8
GALQNI IPASTG AAK	PEP001 993065	rs1154 9348(G /C)	ENSG00 0001116 40	E_E_ TH	0.0066	1411.99	1410.78
GLGTDE DTIIDIIT HR	PEP006 091881		ENSG00 0001970 43	E_E_ KB	0.0033	1767.84	1767.9
GLSEDT TEETLK	PEP002 153670		ENSG00 0001150 53	EXO N_K B	0.0025	1322.28	1321.63
GTEDFI VESLDA SFR	PEP002 703216		ENSG00 0001247 83	EXO N_K B	0.0071	1684.73	1684.79

GTLYIIKLSADIR	PEP002266542		ENSG00000115593	I_E_TH	0.0046	1463.68	1461.86
GTVTDFPGFDER	PEP004712880		ENSG00000164111	EXON_KB	0.0031	1341.46	1339.6
IFTSIGEDYDER	PEP005147259		ENSG00000167085	EXON_KB	0.0000	1444.87	1443.65
IITLTGPTNAIFK	PEP005100585		ENSG00000169564	EXON_KB	0.0039	1389.78	1387.81
ILGVGPDDPDLVR	PEP004526503		ENSG00000160285	EXON_KB	0.0014	1366.67	1364.73
IPNPDFFEDLEPFR	PEP003042599		ENSG00000127022	EXON_KB	0.0066	1734.8	1734.83
IWHHTFYNELR	PEP000663301		ENSG00000075624	EXON_KB	0.0062	1514.73	1514.74
KIPNPDFFEDLEPFR	PEP003042599		ENSG00000127022	EXON_KB	0.0098	1862.9	1862.92
KPEEVDDEVFYS	PEP000367628		ENSG00000003640	EXON_K	0.0072	1708.73	1708.8

PR			73	B			
LASDLLEWIR	PEP000700443		ENSG00000072110	E_E_KB	0.0062	1214.65	1214.67
LASDLLEWIR	PEP002700953		ENSG00000130402	E_E_KB	0.0062	1214.65	1214.67
LATQSNEITIPVTFESR	PEP001796472		ENSG00000106211	EXON_KB	0.0055	1904.91	1904.99
LAVDEEENADNNTK	PEP000643567		ENSG00000072274	EXON_KB	0.0050	1560.65	1560.69
LDETDDPDDYGDR	PEP000150404		ENSG00000014216	EXON_KB	0.0081	1524.78	1524.59
LGANSLLDLVVFGR	PEP007587517	rs6898538(A/G)	ENSG00000232961	E_I_TH	0.0081	1473.02	1472.83
LGGPEAGLGEYLFER	PEP000526191		ENSG00000087086	EXON_KB	0.0072	1606.99	1606.8
LLDNWDSVTSTFSK	PEP002177464		ENSG00000118137	EXON_KB	0.0100	1611.71	1611.78
LLSGED	PEP000		ENSG00	EXO	0.0094	1545.73	1545.73

VGQDEGATR	839802		000070182	N_KB			
LPLQDVYK	PEP001557134		ENSG00000101210	E_I_TH	0.0028	974.937	974.544
LPLQDVYK	PEP001557050		ENSG00000101210	EXON_KB	0.0028	974.937	974.544
LPLQDVYK	PEP004163143		ENSG00000156508	EXON_KB	0.0028	974.937	974.544
LQEAAELEAVELPVPIR	PEP000044823		ENSG00000004939	EXON_KB	0.0081	1876.24	1876.03
LRVDPVNFK	PEP006194445	rs41515552(G/A)	ENSG00000188536	EXON_KB	0.0098	1086.58	1086.62
LSESHPDATEDLQR	PEP004546571		ENSG00000163554	EXON_KB	0.0081	1597.72	1596.74
LVNVVLGAHNVR	PEP006276907		ENSG00000196415	EXON_KB	0.0032	1291.44	1289.76
NIEDVIAQGIGK	PEP006028640		ENSG00000177600	E_E_KB	0.0057	1255.63	1255.68

NILGGTVFR	PEP003130712		ENSG00000138413	EXON_KB	0.0062	977.345	975.549
QEYDESGPSIVHR	PEP000663267		ENSG00000075624	EXON_KB	0.0072	1515.61	1515.7
QEYDESGPSIVHR	PEP006209424		ENSG00000184009	EXON_KB	0.0072	1515.61	1515.7
QEYDESGPSIVHR	PEP006500247		ENSG00000188219	EXON_KB	0.0072	1515.61	1515.7
QITLNDLPVGR	PEP002213772		ENSG00000123131	EXON_KB	0.0046	1225.31	1224.68
QNQIAVDEIR	PEP000031946		ENSG00000005381	EXON_KB	0.0088	1186.55	1184.62
RLFEGNALLR	PEP005143967		ENSG00000170889	E_E_KB	0.0090	1187.7	1187.68
RLSEDYGVLK	PEP005474171		ENSG00000167815	EXON_KB	0.0072	1178.61	1178.63
SFAAVIQALDG	PEP000671849		ENSG00000079	E_E_KB	0.0025	1506.71	1506.75

EMR			59				
SLLEGEGSSGGGGR	PEP006172793		ENSG00000186395	E_E_KB	0.0031	1262.46	1261.59
SNPEDQILYQTER	PEP001630051		ENSG00000110917	EXON_KB	0.0062	1591.97	1591.75
SNPEDQILYQTER	PEP001630055		ENSG00000110917	E_E_KB	0.0062	1591.97	1591.75
SQIHDIVLVGGSTR	PEP001592926		ENSG00000109971	EXON_KB	0.0067	1480.77	1480.8
SYELPDGQVITIGNER	PEP000664177		ENSG00000075624	E_I_TH	0.0049	1789.73	1789.89
SYELPDGQVITIGNER	PEP001946883		ENSG00000107796	EXON_KB	0.0049	1789.73	1789.89
TGAIVDVPVGEELLGR	PEP004285208		ENSG00000152234	EXON_KB	0.0072	1624.07	1623.88
THLAPYSDELR	PEP002177456		ENSG00000118137	EXON_KB	0.0097	1300.62	1300.64
TPAQY	PEP005		ENSG00	EXO	0.0025	1221.54	1221.59

DASELK	624473		0001827 18	N_K B			
VEYHFL SPYVSP K	PEP000 643542		ENSG00 0000722 74	EXO N_K B	0.0082	1564.75	1564.79
VGGVQ SLGGTG ALR	PEP002 477459		ENSG00 0001200 53	EXO N_K B	0.0023	1271.8	1270.7
VLSGDL GQLPTG IR	PEP001 210793		ENSG00 0001008 89	EXO N_K B	0.0090	1424.95	1424.8
VTQWA EER	PEP003 255754		ENSG00 0001371 77	E_E_ TH	0.0067	1019.25	1017.49
YLSYTL NPDLIR	PEP005 271414		ENSG00 0001668 25	EXO N_K B	0.0000	1467.65	1466.78

Table S-3 Peptide hit matrix

	PRIDE1780	PRIDE1781	PRIDE1783	PRIDE1784	PRIDE1785	PRIDE1786	PRIDE1848	PRIDE1849	N
AGEYGAEALER	1	0	0	0	0	0	0	7	2
APVPTGEVYFADSFDR	0	0	5	2	0	2	0	0	3
ASSVVVSGTPIR	0	2	0	0	0	0	2	0	2
ATSFLLALEPELEAR	1	0	0	0	1	0	0	0	2
AVFPSIVGR	0	1	0	0	0	0	2	2	3
AVFVDLEPTVLDEVR	1	1	1	0	1	0	1	0	5
AVFVNLEPTVIDEVR	3	4	0	0	5	0	3	0	4
AVNTLNEALEFAK	0	0	0	4	0	0	25	0	2
DLLDDLKSELTGK	0	0	0	2	0	0	0	4	2
DPQLVPILIEAAR	0	1	1	0	1	0	0	0	3
EGDVLTLLESER	0	0	0	0	1	0	0	1	2
ELFSNLQEFAGPSGK	0	0	1	1	0	0	0	0	2
ETTIQGLDGLSER	0	0	0	2	0	0	3	0	2
FGGLLLTEKPIVLK	0	0	5	0	0	1	0	0	2
FLEQQDQVLQTK	7	4	0	0	0	2	1	0	4
FPGQLNADLR	0	1	2	1	2	0	0	0	4
FVTVQTISGTGALR	0	0	0	0	0	1	4	0	2
GALQNIIPASTGAAK	0	0	4	3	0	1	3	0	4
GLGTDEDTIIDIITHR	0	0	3	3	0	0	0	0	2
GLSEDTTEETLK	0	0	0	0	0	0	6	1	2
GTEDFIVESLDASFR	0	0	1	1	0	0	0	0	2

GTLYIIKLSADIR	0	1	0	0	0	0	0	1	2
GTVTDFPGFDER	1	3	0	0	0	0	0	0	2
IFTSIGEDYDER	1	1	0	0	1	4	0	0	4
IITLTGPTNAIFK	0	0	0	0	0	0	1	1	2
ILGVGPDDPDLVR	0	3	0	0	1	3	0	0	3
IPNPDFFEDLEPFR	0	0	5	2	0	0	0	0	2
IWHHTFYNELR	0	0	0	2	0	0	3	0	2
KIPNPDFFEDLEPFR	0	0	5	0	0	2	0	0	2
KPEEVDDEVFYSPR	0	0	1	0	0	0	2	0	2
LASDLLEWIR	0	0	0	1	0	0	17	0	2
LATQSNEITIPVTFESR	0	0	0	2	0	1	0	0	2
LAVDEEENADNNTK	0	0	4	5	0	1	0	0	3
LDETDDPDDYGDR	0	1	0	0	0	0	1	0	2
LGANSLLDLVVFGR	0	0	0	1	1	0	0	0	2
LGGPEAGLGEYLFER	0	0	4	2	2	0	0	0	3
LLDNWDSVTSTFSK	0	0	0	2	0	0	0	5	2
LLSGEDVGQDEGATR	0	0	0	8	0	0	1	0	2
LPLQDVYK	0	0	0	0	0	0	4	3	2
LQEAAELEAVELPVPIR	0	0	0	4	2	0	0	0	2
LRVDPVNFK	0	0	3	0	0	0	0	7	2
LSESHPDATEDLQR	0	2	0	0	0	0	3	0	2
LVNVVLGAHNVR	1	1	0	0	0	0	0	0	2
NIEDVIAQGIGK	0	0	2	0	0	0	0	1	2
NILGGTVFR	0	1	0	0	0	0	6	0	2
QEYDESGPSIVHR	0	0	0	0	0	2	4	3	3

QITLNDLPVGR	0	1	0	0	1	0	0	0	2
QNQIAVDEIR	2	2	0	0	0	0	0	0	2
RLFEGNALLR	0	2	1	0	1	0	0	0	3
RLSEDYGVLK	0	0	0	2	0	0	0	3	2
SFAAVIQALDGEMR	0	0	1	1	0	0	0	0	2
SLLEGEGSSGGGGR	9	0	0	0	0	3	0	0	2
SNPEDQILYQTER	0	0	0	0	1	1	0	0	2
SQIHDIVLVGGSTR	0	0	1	0	0	0	11	0	2
SYELPDGQVITIGNER	0	0	0	2	0	2	0	0	2
TGAIVDVPVGEELLGR	0	0	0	3	4	0	0	0	2
THLAPYSDELR	0	0	0	2	0	0	0	5	2
TPAQYDASELK	0	0	2	4	0	0	0	0	2
VEYHFLSPYVSPK	0	0	4	5	0	0	0	0	2
VGGVQSLGGTGALR	0	1	0	0	0	0	7	0	2
VLSGDLGQLPTGIR	0	0	0	0	1	0	3	0	2
VTQWAEER	1	1	0	0	0	0	0	0	2
YLSYTLNPDLIR	0	4	1	0	0	2	0	0	3

References

1. Schleyden MJ: **Microscopical researches into the accordance in the structure and growth of animals and plants**; 1847.
2. R. V: **Die Cellularpathologie in ihrer Begründung auf physiologische und pathologische Gewebelehre**: Berlin : Verlag von August Hirschwald; 1859.
3. Fokin SI: **Otto Bütschli (1848–1920): Where we will genuflect?** *Protistology* 2013, **8**(1):22-35.
4. Paweletz N: **Walther Flemming: pioneer of mitosis research**. *Nature reviews Molecular cell biology* 2001, **2**(1):72-75.
5. Waldeyer-Hartz: **Über Karyokinese und ihre Beziehungen zu den Befruchtungsvorgängen**. *Archiv für mikroskopische Anatomie und Entwicklungsmechanik* 1888, **32**(27).
6. Baltzer F: **Theodor Boveri**. *Science* 1964, **144**(3620):809-815.
7. Tjio JH: **The chromosome number of man**. *Hereditas* 1956, **42**(1-2):1-6.
8. Dahm R: **Discovering DNA: Friedrich Miescher and the early years of nucleic acid research**. *Human genetics* 2008, **122**(6):565-581.
9. Jones ME: **Albrecht Kossel, a biographical sketch**. *The Yale journal of biology and medicine* 1953, **26**(1):80-97.
10. Levene PA: **The Structure of Yeast Nucleic Acid: IV. Ammonia Hydrolysis**. *J Biol Chem*, **40**(2):415-424.
11. Koltsov NK: **Physical-chemical fundamentals of morphology**. *Prog Exp Biol* 1927, **B**(7):3-31.
12. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**(4356):737-738.
13. J CTS: **Pentose nucleotides in the cytoplasm of growing tissues**. *Nature* 1939, **143**(3623):602-603.

14. Crick FH: **On protein synthesis**. *Symposia of the Society for Experimental Biology* 1958, **12**:138-163.

15. **Central dogma reversed**. *Nature* 1970, **226**(5252):1198-1199.

16. Baltimore D: **RNA-dependent DNA polymerase in virions of RNA tumour viruses**. *Nature* 1970, **226**(5252):1209-1211.

17. Chow LT, Gelinas RE, Broker TR, Roberts RJ: **An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA**. *Cell* 1977, **12**(1):1-8.

18. Berget SM, Moore C, Sharp PA: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA**. *P Natl Acad Sci USA* 1977, **74**(8):3171-3175.

19. Maxam AM, Gilbert W: **A new method for sequencing DNA**. *Proc Natl Acad Sci U S A* 1977, **74**(2):560-564.

20. Sanger F, Coulson AR: **A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase**. *Journal of molecular biology* 1975, **94**(3):441-448.

21. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proc Natl Acad Sci U S A* 1977, **74**(12):5463-5467.

22. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**(5910):133-138.

23. Porreca GJ: **Genome sequencing on nanoballs**. *Nat Biotechnol* 2010, **28**(1):43-44.

24. Thompson JF, Steinmann KE: **Single molecule sequencing with a HeliScope genetic analysis system**. *Current protocols in molecular biology* 2010, **Chapter 7**:Unit7 10.

25. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H: **Continuous base identification for single-molecule nanopore DNA sequencing**. *Nature nanotechnology* 2009, **4**(4):265-270.

26. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ *et al*: **The UCSC Genome Browser Database**. *Nucleic acids research* 2003, **31**(1):51-54.

27. Homer N, Merriman B, Nelson SF: **BFAST: an alignment tool for large scale genome resequencing**. *PloS one* 2009, **4**(11):e7767.

28. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature methods* 2012, **9**(4):357-359.

29. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

30. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.

31. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2013, **29**(1):15-21.

32. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**(1):139-140.

33. Perrett D: **From 'protein' to the beginnings of clinical proteomics**. *Proteomics Clinical applications* 2007, **1**(8):720-738.

34. Hartley H: **Origin of the word 'protein'**. *Nature* 1951, **168**(4267):244.

35. Stretton AO: **The first sequence. Fred Sanger and insulin**. *Genetics* 2002, **162**(2):527-532.

36. Pauling L, Corey RB, Branson HR: **The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain**. *Proc Natl Acad Sci U S A* 1951, **37**(4):205-211.

37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic acids research* 2000, **28**(1):235-242.

38. Breathnach R, Benoist C, O'Hare K, Gannon F, Chambon P: **Ovalbumin gene: evidence for a leader sequence in mRNA and DNA sequences at the exon-intron boundaries**. *Proc Natl Acad Sci U S A* 1978, **75**(10):4853-4857.

39. Lewin B: **Alternatives for splicing: recognizing the ends of introns**. *Cell* 1980, **22**(2 Pt 2):324-326.

40. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**(5):511-515.

41. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**. *Nat Biotechnol* 2010, **28**(5):503-510.

42. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nat Methods* 2010, **7**(12):1009-U1101.

43. 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**. *Proc Natl Acad Sci U S A* 2014, **111**(51):E5593-5601.

44. Kendall MG, O'Hagan A, Forster J: **Kendall's advanced theory of statistics**, 2nd edn. London: Arnold; 2004.

45. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization**. *Nature reviews Genetics* 2007, **8**(6):413-423.

46. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.

47. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.

48. Bartel DP: **MicroRNAs: target recognition and regulatory functions**. *Cell* 2009, **136**(2):215-233.

49. Grewal SI, Jia S: **Heterochromatin revisited**. *Nature reviews Genetics* 2007, **8**(1):35-46.

50. Malone CD, Hannon GJ: **Small RNAs as guardians of the genome**. *Cell* 2009, **136**(4):656-668.

51. Ghildiyal M, Zamore PD: **Small silencing RNAs: an expanding universe**. *Nature reviews Genetics* 2009, **10**(2):94-108.

52. Ponjavic J, Ponting CP, Lunter G: **Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs**. *Genome research* 2007, **17**(5):556-565.

53. Payer B, Lee JT: **X chromosome dosage compensation: how mammals keep the balance**. *Annual review of genetics* 2008, **42**:733-772.

54. Martens JA, Laprade L, Winston F: **Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene**. *Nature* 2004, **429**(6991):571-574.

55. Martens JA, Wu PY, Winston F: **Regulation of an intergenic transcript controls adjacent gene transcription in Saccharomyces cerevisiae**. *Genes & development* 2005, **19**(22):2695-2704.

56. Lefevre P, Witham J, Lacroix CE, Cockerill PN, Bonifer C: **The LPS-induced transcriptional upregulation of the chicken lysozyme locus involves CTCF eviction and noncoding RNA transcription**. *Molecular cell* 2008, **32**(1):129-139.

57. Beltran M, Puig I, Pena C, Garcia JM, Alvarez AB, Pena R, Bonilla F, de Herreros AG: **A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition**. *Genes & development* 2008, **22**(6):756-769.

58. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals**. *Nature* 2009, **458**(7235):223-227.

59. Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A *et al*: **Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression**. *Proc Natl Acad Sci U S A* 2009, **106**(28):11667-11672.

60. Zhou A, Li M, He B, Feng W, Huang F, Xu B, Dunker AK, Balch C, Li B, Liu Y *et al*: **Lipopolysaccharide treatment induces genome-wide pre-mRNA splicing pattern changes in mouse bone marrow stromal stem cells**. *BMC genomics* 2016, **17 Suppl 7**:509.

61. Zhou A, Breese MR, Hao Y, Edenberg HJ, Li L, Skaar TC, Liu Y: **Alt Event Finder: a tool for extracting alternative splicing events from RNA-seq data**. *BMC genomics* 2012, **13 Suppl 8**:S10.

62. Zhou A, Zhang F, Chen JY: **PEPPI: a peptidomic database of human protein isoforms for proteomics experiments**. *BMC bioinformatics* 2010, **11 Suppl 6**:S7.

63. Ao Z, Yadong W, Yunlong L, Weixing F, Edenberg HJ: **Characterizing the roles of long non-coding RNA in rat alcohol preference**. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 15-18 Dec. 2016 2016*; 2016: 167-173.

64. Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO: **Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity**. *Molecular biology and evolution* 2014, **31**(6):1402-1413.

65. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H: **Function of alternative splicing**. *Gene* 2005, **344**:1-20.

66. Chen L, Bush SJ, Tovar-Corona JM, Castillo-Morales A, Urrutia AO: **Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity**. *Molecular biology and evolution* 2014.

67. Consortium EP: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**(7414):57-74.

68. Diamond RH, Du K, Lee VM, Mohn KL, Haber BA, Tewari DS, Taub R: **Novel delayed-early and highly insulin-induced growth response genes. Identification of HRS, a potential regulator of alternative pre-mRNA splicing**. *J Biol Chem* 1993, **268**(20):15185-15192.

69. Du K, Peng Y, Greenbaum LE, Haber BA, Taub R: **HRS/SRp40-mediated inclusion of the fibronectin EIIIB exon, a possible cause of increased EIIIB expression in proliferating liver**. *Mol Cell Biol* 1997, **17**(7):4096-4104.

70. Huang R, Xing Z, Luan Z, Wu T, Wu X, Hu G: **A specific splicing variant of SVH, a novel human armadillo repeat protein, is up-regulated in hepatocellular carcinomas**. *Cancer Res* 2003, **63**(13):3775-3782.

71. Kanai Y, Saito Y, Ushijima S, Hirohashi S: **Alterations in gene expression associated with the overexpression of a splice variant of DNA methyltransferase 3b, DNMT3b4, during human hepatocarcinogenesis**. *J Cancer Res Clin Oncol* 2004, **130**(11):636-644.

72. Kurahashi H, Takami K, Oue T, Kusafuka T, Okada A, Tawa A, Okada S, Nishisho I: **Biallelic inactivation of the APC gene in hepatoblastoma**. *Cancer Res* 1995, **55**(21):5007-5011.

73. Saito Y, Kanai Y, Sakamoto M, Saito H, Ishii H, Hirohashi S: **Overexpression of a splice variant of DNA methyltransferase 3b, DNMT3b4, associated with DNA hypomethylation on pericentromeric satellite regions during human hepatocarcinogenesis**. *Proc Natl Acad Sci U S A* 2002, **99**(15):10060-10065.

74. Pevsner-Fischer M, Morad V, Cohen-Sfady M, Rousso-Noori L, Zanin-Zhorov A, Cohen S, Cohen IR, Zipori D: **Toll-like receptors and their ligands control mesenchymal stem cell functions**. *Blood* 2007, **109**(4):1422-1432.

75. Gnecchi M, Danieli P, Cervio E: **Mesenchymal stem cell therapy for heart disease**. *Vascular pharmacology* 2012, **57**(1):48-55.

76. van den Akker F, de Jager SC, Sluijter JP: **Mesenchymal stem cell therapy for cardiac inflammation: immunomodulatory properties and the influence of toll-like receptors**. *Mediators of inflammation* 2013, **2013**:181020.

77. Lu X, Liu T, Gu L, Huang C, Zhu H, Meng W, Xi Y, Li S, Liu Y: **Immunomodulatory effects of mesenchymal stem cells involved in favoring type 2 T cell subsets**. *Transplant immunology* 2009, **22**(1-2):55-61.

78. Gnecchi M, He H, Liang OD, Melo LG, Morello F, Mu H, Noiseux N, Zhang L, Pratt RE, Ingwall JS *et al*: **Paracrine action accounts for marked protection of ischemic heart by Akt-modified mesenchymal stem cells**. *Nat Med* 2005, **11**(4):367-368.

79. Gnecchi M, He H, Noiseux N, Liang OD, Zhang L, Morello F, Mu H, Melo LG, Pratt RE, Ingwall JS *et al*: **Evidence supporting paracrine hypothesis for Akt-modified mesenchymal stem cell-mediated cardiac protection and functional improvement**. *Faseb J* 2006, **20**(6):661-669.

80. Kocher AA, Schuster MD, Szabolcs MJ, Takuma S, Burkhoff D, Wang J, Homma S, Edwards NM, Itescu S: **Neovascularization of ischemic myocardium by human bone-marrow-derived angioblasts prevents cardiomyocyte apoptosis, reduces remodeling and improves cardiac function**. *Nat Med* 2001, **7**(4):430-436.

81. Steele A, Steele P: **Stem cells for repair of the heart**. *Curr Opin Pediatr* 2006, **18**(5):518-523.

82. Tomasoni S, Longaretti L, Rota C, Morigi M, Conti S, Gotti E, Capelli C, Introna M, Remuzzi G, Benigni A: **Transfer of growth factor receptor mRNA via exosomes unravels the regenerative effect of mesenchymal stem cells**. *Stem cells and development* 2013, **22**(5):772-780.

83. Abdel-Latif A, Bolli R, Tleyjeh IM, Montori VM, Perin EC, Hornung CA, Zuba-Surma EK, Al-Mallah M, Dawn B: **Adult bone marrow-derived cells for cardiac repair: a systematic review and meta-analysis**. *Arch Intern Med* 2007, **167**(10):989-997.

84. Dai W, Hale SL, Martin BJ, Kuang JQ, Dow JS, Wold LE, Kloner RA: **Allogeneic mesenchymal stem cell transplantation in postinfarcted rat myocardium: short- and long-term effects**. *Circulation* 2005, **112**(2):214-223.

85. Gnecchi M, Zhang Z, Ni A, Dzau VJ: **Paracrine mechanisms in adult stem cell signaling and therapy**. *Circulation research* 2008, **103**(11):1204-1219.

86. Yao Y, Zhang F, Wang L, Zhang G, Wang Z, Chen J, Gao X: **Lipopolysaccharide preconditioning enhances the efficacy of mesenchymal stem cells transplantation in a rat model of acute myocardial infarction**. *Journal of biomedical science* 2009, **16**:74.

87. Herrmann JL, Wang Y, Abarbanell AM, Weil BR, Tan J, Meldrum DR: **Preconditioning mesenchymal stem cells with transforming growth factor-alpha improves mesenchymal stem cell-mediated cardioprotection**. *Shock* 2010, **33**(1):24-30.

88. Wang Y, Abarbanell AM, Herrmann JL, Weil BR, Manukyan MC, Poynter JA, Meldrum DR: **TLR4 inhibits mesenchymal stem cell (MSC) STAT3 activation and thereby exerts deleterious effects on MSC-mediated cardioprotection**. *PloS one* 2010, **5**(12):e14206.

89. Ma A, Jiang L, Song L, Hu Y, Dun H, Daloze P, Yu Y, Jiang J, Zafarullah M, Chen H: **Reconstruction of cartilage with clonal mesenchymal stem cell-acellular dermal matrix in cartilage defect model in nonhuman primates**. *International immunopharmacology* 2013, **16**(3):399-408.

90. El-Badri N, Ghoneim MA: **Mesenchymal stem cell therapy in diabetes mellitus: progress and challenges**. *Journal of nucleic acids* 2013, **2013**:194858.

91. Wang LQ, Lin ZZ, Zhang HX, Shao B, Xiao L, Jiang HG, Zhuge QC, Xie LK, Wang B, Su DM *et al*: **Timing and dose regimens of marrow mesenchymal stem**

cell transplantation affect the outcomes and neuroinflammatory response after ischemic stroke. *CNS neuroscience & therapeutics* 2014, **20**(4):317-326.

92. Uccelli A, Moretta L, Pistoia V: **Mesenchymal stem cells in health and disease**. *Nature reviews Immunology* 2008, **8**(9):726-736.

93. Kurtzberg J, Prockop S, Teira P, Bittencourt H, Lewis V, Chan KW, Horn B, Yu L, Talano JA, Nemecek E *et al*: **Allogeneic human mesenchymal stem cell therapy (remestemcel-L, Prochymal) as a rescue agent for severe refractory acute graft-versus-host disease in pediatric patients**. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation* 2014, **20**(2):229-235.

94. Dalal J, Gandy K, Domen J: **Role of mesenchymal stem cell therapy in Crohn's disease**. *Pediatric research* 2012, **71**(4 Pt 2):445-451.

95. Martin GS: **Sepsis, severe sepsis and septic shock: changes in incidence, pathogens and outcomes**. *Expert review of anti-infective therapy* 2012, **10**(6):701-706.

96. Balk RA: **Severe sepsis and septic shock. Definitions, epidemiology, and clinical manifestations**. *Critical care clinics* 2000, **16**(2):179-192.

97. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nature methods* 2010, **7**(12):1009-1015.

98. Rhee SH, Hwang D: **Murine TOLL-like receptor 4 confers lipopolysaccharide responsiveness as determined by activation of NF kappa B and expression of the inducible cyclooxygenase**. *J Biol Chem* 2000, **275**(44):34035-34040.

99. Fitzgerald KA, Rowe DC, Barnes BJ, Caffrey DR, Visintin A, Latz E, Monks B, Pitha PM, Golenbock DT: **LPS-TLR4 signaling to IRF-3/7 and NF-kappaB**

involves the toll adapters TRAM and TRIF. *The Journal of experimental medicine* 2003, **198**(7):1043-1055.

100. Chang YF, Imam JS, Wilkinson MF: **The nonsense-mediated decay RNA surveillance pathway**. *Annual review of biochemistry* 2007, **76**:51-74.

101. Huang da W, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources**. *Nature protocols* 2009, **4**(1):44-57.

102. Laity JH, Lee BM, Wright PE: **Zinc finger proteins: new insights into structural and functional diversity**. *Curr Opin Struct Biol* 2001, **11**(1):39-46.

103. Finn RD, Miller BL, Clements J, Bateman A: **iPfam: a database of protein family and domain interactions found in the Protein Data Bank**. *Nucleic acids research* 2014, **42**(Database issue):D364-373.

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

105. Wang X, Wei X, Thijssen B, Das J, Lipkin SM, Yu H: **Three-dimensional reconstruction of protein networks provides insight into human genetic disease**. *Nature biotechnology* 2012, **30**(2):159-164.

106. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, Oldfield CJ, Cortese MS, Sickmeier M, LeGall T, Obradovic Z *et al*: **Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms**. *Proc Natl Acad Sci U S A* 2006, **103**(22):8390-8395.

107. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z: **Length-dependent prediction of protein intrinsic disorder**. *BMC bioinformatics* 2006, **7**:208.

108. Buljan M, Chalancon G, Dunker AK, Bateman A, Balaji S, Fuxreiter M, Babu MM: **Alternative splicing of intrinsically disordered regions and rewiring of protein interactions**. *Curr Opin Struct Biol* 2013, **23**(3):443-450.

109. Pejaver V, Hsu WL, Xin FX, Dunker AK, Uversky VN, Radivojac P: **The structural and functional signatures of proteins that undergo multiple events of post-translational modification**. *Protein Sci* 2014, **23**(8):1077-1093.

110. Grant CE, Bailey TL, Noble WS: **FIMO: scanning for occurrences of a given motif**. *Bioinformatics* 2011, **27**(7):1017-1018.

111. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A *et al*: **A compendium of RNA-binding motifs for decoding gene regulation**. *Nature* 2013, **499**(7457):172-177.

112. Wang ZJ, Zhang FM, Wang LS, Yao YW, Zhao Q, Gao X: **Lipopolysaccharides can protect mesenchymal stem cells (MSCs) from oxidative stress-induced apoptosis and enhance proliferation of MSCs via Toll-like receptor(TLR)-4 and PI3K/Akt**. *Cell biology international* 2009, **33**(6):665-674.

113. Crisostomo PR, Wang Y, Markel TA, Wang M, Lahm T, Meldrum DR: **Human mesenchymal stem cells stimulated by TNF-alpha, LPS, or hypoxia produce growth factors by an NF kappa B- but not JNK-dependent mechanism**. *American journal of physiology Cell physiology* 2008, **294**(3):C675-682.

114. Wang M, Tan J, Wang Y, Meldrum KK, Dinarello CA, Meldrum DR: **IL-18 binding protein-expressing mesenchymal stem cells improve myocardial protection after ischemia or infarction**. *Proc Natl Acad Sci U S A* 2009, **106**(41):17499-17504.

115. Pawson T, Nash P: **Assembly of cell regulatory systems through protein interaction domains**. *Science* 2003, **300**(5618):445-452.

116. Pena PV, Davrazou F, Shi X, Walter KL, Verkhusha VV, Gozani O, Zhao R, Kutateladze TG: **Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2**. *Nature* 2006, **442**(7098):100-103.

117. Stenmark H, Vitale G, Ullrich O, Zerial M: **Rabaptin-5 is a direct effector of the small GTPase Rab5 in endocytic membrane fusion**. *Cell* 1995, **83**(3):423-432.

118. Harding C, Heuser J, Stahl P: **Receptor-mediated endocytosis of transferrin and recycling of the transferrin receptor in rat reticulocytes**. *The Journal of cell biology* 1983, **97**(2):329-339.

119. Johnstone RM, Adam M, Hammond JR, Orr L, Turbide C: **Vesicle formation during reticulocyte maturation. Association of plasma membrane activities with released vesicles (exosomes)**. *J Biol Chem* 1987, **262**(19):9412-9420.

120. Hirst J, Lui WW, Bright NA, Totty N, Seaman MN, Robinson MS: **A family of proteins with gamma-adaptin and VHS domains that facilitate trafficking between the trans-Golgi network and the vacuole/lysosome**. *The Journal of cell biology* 2000, **149**(1):67-80.

121. Valsdottir R, Hashimoto H, Ashman K, Koda T, Storrie B, Nilsson T: **Identification of rabaptin-5, rabex-5, and GM130 as putative effectors of rab33b, a regulator of retrograde traffic between the Golgi apparatus and ER**. *FEBS letters* 2001, **508**(2):201-209.

122. Pallai R, Simpkins H, Chen J, Parekh HK: **The CCAAT box binding transcription factor, nuclear factor-Y (NF-Y) regulates transcription of human aldo-keto reductase 1C1 (AKR1C1) gene**. *Gene* 2010, **459**(1-2):11-23.

123. Peister A, Mellad JA, Larson BL, Hall BM, Gibson LF, Prockop DJ: **Adult stem cells from bone marrow (MSCs) isolated from different strains of inbred mice vary in surface epitopes, rates of proliferation, and differentiation potential**. *Blood* 2004, **103**(5):1662-1668.

124. Secunda R, Vennila R, Mohanashankar AM, Rajasundari M, Jeswanth S, Surendran R: **Isolation, expansion and characterisation of mesenchymal stem**

cells from human bone marrow, adipose tissue, umbilical cord blood and matrix: a comparative study. *Cytotechnology* 2014.

125. Breese MR, Liu Y: **NGSUtils: a software suite for analyzing and manipulating next-generation sequencing datasets**. *Bioinformatics* 2013, **29**(4):494-496.

126. Juan L, Wang G, Radovich M, Schneider BP, Clare SE, Wang Y, Liu Y: **Potential roles of microRNAs in regulating long intergenic noncoding RNAs**. *BMC medical genomics* 2013, **6 Suppl 1**:S7.

127. Todd AG, Lin H, Ebert AD, Liu Y, Androphy EJ: **COPI transport complexes bind to specific RNAs in neuronal cells**. *Human molecular genetics* 2013, **22**(4):729-736.

128. Peng ZL, Kurgan L: **Comprehensive comparative assessment of in-silico predictors of disordered regions**. *Current protein & peptide science* 2012, **13**(1):6-18.

129. Cheng YG, Oldfield CJ, Meng JW, Romero P, Uversky VN, Dunker AK: **Mining alpha-helix-forming molecular recognition features with cross species sequence alignments**. *Biochemistry-Us* 2007, **46**(47):13468-13477.

130. Zhu J, Zhang Y, Joe GJ, Pompetti R, Emerson SG: **NF-Ya activates multiple hematopoietic stem cell (HSC) regulatory genes and promotes HSC self-renewal**. *Proc Natl Acad Sci U S A* 2005, **102**(33):11728-11733.

131. Elias H, Bengelsdorf H: **The structure of the liver of vertebrates**. *Acta anatomica* 1952, **14**(4):297-337.

132. de Zwart LL, Haenen HE, Versantvoort CH, Wolterink G, van Engelen JG, Sips AJ: **Role of biokinetics in risk assessment of drugs and chemicals in children**. *Regulatory toxicology and pharmacology : RTP* 2004, **39**(3):282-309.

133. Blanco JG, Harrison PL, Evans WE, Relling MV: **Human cytochrome P450 maximal activities in pediatric versus adult liver**. *Drug metabolism and disposition: the biological fate of chemicals* 2000, **28**(4):379-382.

134. Socha P, Horvath A, Vajro P, Dziechciarz P, Dhawan A, Szajewska H: **Pharmacological interventions for nonalcoholic fatty liver disease in adults and in children: a systematic review**. *Journal of pediatric gastroenterology and nutrition* 2009, **48**(5):587-596.

135. Benedetti MS, Whomsley R, Canning M: **Drug metabolism in the paediatric population and in the elderly**. *Drug discovery today* 2007, **12**(15-16):599-610.

136. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nature genetics* 2008, **40**(12):1413-1415.

137. Lopez AJ: **Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation**. *Annual review of genetics* 1998, **32**:279-305.

138. Medina MW, Gao F, Ruan WM, Rotter JI, Krauss RM: **Alternative splicing of 3-hydroxy-3-methylglutaryl coenzyme a reductase is associated with plasma low-density lipoprotein cholesterol response to simvastatin**. *Circulation* 2008, **118**(4):355-362.

139. Clower CV, Chatterjee D, Wang ZX, Cantley LC, Heiden MGV, Krainer AR: **The alternative splicing repressors hnRNP A1/A2 and PTB influence pyruvate kinase isoform expression and cell metabolism**. *P Natl Acad Sci USA* 2010, **107**(5):1894-1899.

140. Nakagawa N, Sakurai N: **A mutation in At-nMat1a, which encodes a nuclear gene having high similarity to group II intron maturase, causes impaired splicing of mitochondrial NAD4 transcript and altered carbon metabolism in Arabidopsis thaliana**. *Plant Cell Physiol* 2006, **47**(6):772-783.

141. Kaminska D, Kuulasmaa T, Venesmaa S, Kakela P, Vaittinen M, Pulkkinen L, Paakkonen M, Gylling H, Laakso M, Pihlajamaki J: **Adipose Tissue TCF7L2 Splicing Is Regulated by Weight Loss and Associates With Glucose and Fatty Acid Metabolism**. *Diabetes* 2012, **61**(11):2807-2813.

142. Yu CY, Theusch E, Lo K, Mangravite LM, Naidoo D, Kutilova M, Medina MW: **HNRNPA1 regulates HMGCR alternative splicing and modulates cellular cholesterol metabolism**. *Hum Mol Genet* 2014, **23**(2):319-332.

143. Caceres JF, Kornblihtt AR: **Alternative splicing: multiple control mechanisms and involvement in human disease**. *Trends in genetics : TIG* 2002, **18**(4):186-193.

144. Garcia-Blanco MA, Baraniak AP, Lasda EL: **Alternative splicing in disease and therapy**. *Nat Biotechnol* 2004, **22**(5):535-546.

145. Tazi J, Bakkour N, Stamm S: **Alternative splicing and disease**. *Biochimica et biophysica acta* 2009, **1792**(1):14-26.

146. Shen S, Park JW, Huang J, Dittmar KA, Lu ZX, Zhou Q, Carstens RP, Xing Y: **MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data**. *Nucleic acids research* 2012, **40**(8):e61.

147. Wang J, Armant DR: **Integrin-mediated adhesion and signaling during blastocyst implantation**. *Cells Tissues Organs* 2002, **172**(3):190-201.

148. George EL, Georges-Labouesse EN, Patel-King RS, Rayburn H, Hynes RO: **Defects in mesoderm, neural tube and vascular development in mouse embryos lacking fibronectin**. *Development* 1993, **119**(4):1079-1091.

149. Bronson RA, Fusi FM: **Integrins and human reproduction**. *Molecular human reproduction* 1996, **2**(3):153-168.

150. Shimomura Y, Ando H, Furugori K, Kajiyama H, Suzuki M, Iwase A, Mizutani S, Kikkawa F: **Possible involvement of crosstalk cell-adhesion mechanism by**

endometrial CD26/dipeptidyl peptidase IV and embryonal fibronectin in human blastocyst implantation. *Molecular human reproduction* 2006, **12**(8):491-495.

151. Wang J, Mayernik L, Armant DR: **Trophoblast adhesion of the peri-implantation mouse blastocyst is regulated by integrin signaling that targets phospholipase C**. *Dev Biol* 2007, **302**(1):143-153.

152. Goossens K, Van Soom A, Van Zeveren A, Favoreel H, Peelman LJ: **Quantification of Fibronectin 1 (FN1) splice variants, including two novel ones, and analysis of integrins as candidate FN1 receptors in bovine preimplantation embryos**. *Bmc Dev Biol* 2009, **9**.

153. Kornblihtt AR, Umezawa K, Vibe-Pedersen K, Baralle FE: **Primary structure of human fibronectin: differential splicing may generate at least 10 polypeptides from a single gene**. *Embo J* 1985, **4**(7):1755-1759.

154. Kramer A, Green J, Pollard J, Tugendreich S: **Causal analysis approaches in Ingenuity Pathway Analysis**. *Bioinformatics* 2014, **30**(4):523-530.

155. Faber KN, Muller M, Jansen PLM: **Drug transport proteins in the liver**. *Adv Drug Deliv Rev* 2003, **55**(1):107-124.

156. van Montfoort JE, Hagenbuch B, Groothuis GMM, Koepsell H, Meier PJ, Meijer DKF: **Drug uptake systems in liver and kidney**. *Curr Drug Metab* 2003, **4**(3):185-211.

157. Hagenbuch B: **Drug uptake systems in liver and kidney: a historic perspective**. *Clin Pharmacol Ther* 2010, **87**(1):39-47.

158. Szakacs G, Annereau J-P, Lababidi S, Shankavaram U, Arciello A, Bussey KJ, Reinhold W, Guo Y, Kruh GD, Reimers M *et al*: **Predicting drug sensitivity and resistance: profiling ABC transporter genes in cancer cells**. *Cancer Cell* 2004, **6**(2):129-137.

159. Neveu G, Barouch-Bentov R, Ziv-Av A, Gerber D, Jacob Y, Einav S: **Identification and targeting of an interaction between a tyrosine motif within hepatitis C virus core protein and AP2M1 essential for viral assembly**. *PLoS Pathog* 2012, **8**(8):e1002845.

160. Li H, Yang X, Yang G, Hong Z, Zhou L, Yin P, Xiao Y, Chen L, Chung RT, Zhang L: **Hepatitis C virus NS5A hijacks ARFGAP1 to maintain a phosphatidylinositol 4-phosphate-enriched microenvironment**. *Journal of virology* 2014, **88**(11):5956-5966.

161. Saeed M, Andreo U, Chung HY, Espiritu C, Branch AD, Silva JM, Rice CM: **SEC14L2 enables pan-genotype HCV replication in cell culture**. *Nature* 2015, **524**(7566):471-475.

162. Choi R, Park HD, Ko JM, Lee J, Lee DH, Hong SJ, Ki CS, Lee SY, Kim JW, Song J *et al*: **Novel SLC37A4 Mutations in Korean Patients With Glycogen Storage Disease Ib**. *Annals of laboratory medicine* 2017, **37**(3):261-266.

163. Dissanayake VH, Jayasinghe JD, Thilakaratne V, Jayasekara RW: **A novel mutation in SLC37A4 gene in a Sri Lankan boy with glycogen storage disease type Ib associated with very early onset neutropenia**. *J Mol Genet Med* 2011, **5**:262-263.

164. Li L, Li H, Garzel B, Yang H, Sueyoshi T, Li Q, Shu Y, Zhang J, Hu B, Heyward S *et al*: **SLC13A5 is a novel transcriptional target of the pregnane X receptor and sensitizes drug-induced steatosis in human liver**. *Mol Pharmacol* 2015, **87**(4):674-682.

165. Atkinson S, Way M, McQuillin A, Morgan M, Thursz M: **A Genome-Wide Association Study Identifies Pnpla3 and Slc38a4 as Risk Loci for Alcoholic Hepatitis**. *J Hepatol* 2016, **64**:S134-S134.

166. Johnson KR, Gagnon LH, Chang B: **A hypomorphic mutation of the gamma-1 adaptin gene (Ap1g1) causes inner ear, retina, thyroid, and testes abnormalities in mice**. *Mamm Genome* 2016, **27**(5-6):200-212.

167. Guengrich FP: **Cytochrome P-450 3A4: Regulation and role in drug metabolism**. *Annu Rev Pharmacol* 1999, **39**:1-17.

168. Miners JO, Birkett DJ: **Cytochrome P4502C9: an enzyme of major importance in human drug metabolism**. *Brit J Clin Pharmaco* 1998, **45**(6):525-538.

169. Zanger UM, Schwab M: **Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation**. *Pharmacol Therapeut* 2013, **138**(1):103-141.

170. Ingelman-Sundberg M, Oscarson M, McLellan RA: **Polymorphic human cytochrome P450 enzymes: an opportunity for individualized drug treatment**. *Trends Pharmacol Sci* 1999, **20**(8):342-349.

171. Ingelman-Sundberg M: **Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future**. *Trends Pharmacol Sci* 2004, **25**(4):193-200.

172. Santos A, Zanetta S, Cresteil T, Deroussent A, Pein F, Raymond E, Vernillet L, Risse ML, Boige V, Gouyette A *et al*: **Metabolism of irinotecan (CPT-11) by CYP3A4 and CYP3A5 in humans**. *Clin Cancer Res* 2000, **6**(5):2012-2020.

173. Henningsson A, Marsh S, Loos WJ, Karlsson MO, Garsa A, Mross K, Mielke S, Vigano L, Locatelli A, Verweij J *et al*: **Association of CYP2C8, CYP3A4, CYP3A5, and ABCB1 polymorphisms with the pharmacokinetics of paclitaxel**. *Clin Cancer Res* 2005, **11**(22):8097-8104.

174. Haufroid V, Mourad M, Van Kerckhove V, Wawrzyniak J, De Meyer M, Eddour DC, Malaise J, Lison D, Squifflet JP, Wallemacq P: **The effect of CYP3A5 and MDR1 (ABCB1) polymorphisms on cyclosporine and tacrolimus dose**

requirements and trough blood levels in stable renal transplant patients. *Pharmacogenetics* 2004, **14**(3):147-154.

175. Anglicheau D, Thervet E, Etienne I, De Ligny BH, Le Meur Y, Touchard G, Buchler M, Laurent-Puig P, Tregouet D, Beaune P *et al*: **CYP3A5 and MDR1 genetic polymorphisms and cyclosporine pharmacokinetics after renal transplantation**. *Clinical Pharmacology & Therapeutics* 2004, **75**(5):422-433.

176. Zheng HX, Webber S, Zeevi A, Schuetz E, Zhang J, Bowman P, Boyle G, Law Y, Miller S, Lamba J *et al*: **Tacrolimus dosing in pediatric heart transplant patients is related to CYP3A5 and MDR1 gene polymorphisms**. *Am J Transplant* 2003, **3**(4):477-483.

177. Kivisto KT, Niemi M, Schaeffeler E, Pitkala K, Tilvis RT, Fromm MF, Schwab M, Eichelbaum M, Strandberg T: **Lipid-lowering response to statins is affected by CYP3A5 polymorphism**. *Pharmacogenetics* 2004, **14**(8):523-525.

178. Gonzalez FJ: **Role of cytochromes P450 in chemical toxicity and oxidative stress: studies with CYP2E1**. *Mutat Res-Fund Mol M* 2005, **569**(1-2):101-110.

179. Caro AA, Cederbaum AI: **Oxidative stress, toxicology, and pharmacology of CYP2E1**. *Annu Rev Pharmacol* 2004, **44**:27-42.

180. Manyike PT, Kharasch ED, Kalhorn TF, Slattery JT: **Contribution of CYP2E1 and CYP3A to acetaminophen reactive metabolite formation**. *Clinical Pharmacology & Therapeutics* 2000, **67**(3):275-282.

181. Steinberg SF: **Structural basis of protein kinase C isoform function**. *Physiological reviews* 2008, **88**(4):1341-1378.

182. Bridges D, Moorhead GB: **14-3-3 proteins: a number of functions for a numbered protein**. *Science's STKE : signal transduction knowledge environment* 2004, **2004**(242):re10.

183. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.

184. Faustino NA, Cooper TA: **Pre-mRNA splicing and human disease**. *Genes & development* 2003, **17**(4):419-437.

185. Li M, Feng W, Zhang X, Yang Y, Wang K, Mort M, Cooper DN, Wang Y, Zhou Y, Liu Y: **ExonImpact: Prioritizing Pathogenic Alternative Splicing Events**. *Human mutation* 2017, **38**(1):16-24.

186. Lewit-Bentley A, Rety S: **EF-hand calcium-binding proteins**. *Current opinion in structural biology* 2000, **10**(6):637-643.

187. Wang XJ, Wei XM, Thijssen B, Das J, Lipkin SM, Yu HY: **Three-dimensional reconstruction of protein networks provides insight into human genetic disease**. *Nat Biotechnol* 2012, **30**(2):159-164.

188. Rolland T, Tasan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R *et al*: **A Proteome-Scale Map of the Human Interactome Network**. *Cell* 2014, **159**(5):1212-1226.

189. Weirauch MT, Yang A, Albu M, Cote A, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K *et al*: **Determination and inference of eukaryotic transcription factor sequence specificity**. *Cell* 2014, **158**(6):1431-1443.

190. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing**. *J Roy Stat Soc B Met* 1995, **57**(1):289-300.

191. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J *et al*: **Pfam: the protein families database**. *Nucleic acids research* 2014, **42**(Database issue):D222-230.

192. Graveley BR: **Alternative splicing: increasing diversity in the proteomic world**. *Trends Genet* 2001, **17**(2):100-107.

193. Garcia-Blanco MA, Baraniak AP, Lasda EL: **Alternative splicing in disease and therapy**. *Nature biotechnology* 2004, **22**(5):535-546.

194. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P: **Alternative splicing and genome complexity**. *Nat Genet* 2002, **30**(1):29-30.

195. Modrek B, Lee C: **A genomic view of alternative splicing**. *Nat Genet* 2002, **30**(1):13-19.

196. Blencowe BJ: **Alternative splicing: New insights from global analyses**. *Cell* 2006, **126**(1):37-47.

197. Brett D, Hanke J, Lehmann G, Haase S, Delbruck S, Krueger S, Reich J, Bork P: **EST comparison indicates 38% of human mRNAs contain possible alternative splice forms**. *Febs Lett* 2000, **474**(1):83-86.

198. Wang LG, Xi YX, Yu J, Dong LP, Yen LS, Li W: **A Statistical Method for the Detection of Alternative Splicing Using RNA-Seq**. *Plos One* 2010, **5**(1).

199. Au KF, Jiang H, Lin L, Xing Y, Wong WH: **Detection of splice junctions from paired-end RNA-seq data by SpliceMap**. *Nucleic Acids Res* 2010, **38**(14):4570-4578.

200. Ameur A, Wetterbom A, Feuk L, Gyllensten U: **Global and unbiased detection of splice junctions from RNA-seq data**. *Genome biology* 2010, **11**(3).

201. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C *et al*: **Annotating genomes with massive-scale RNA sequencing**. *Genome biology* 2008, **9**(12).

202. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads**. *Bioinformatics* 2010, **26**(7):873-881.

203. Ning K, Fermin D: **SAW: A Method to Identify Splicing Events from RNA-Seq Data Based on Splicing Fingerprints**. *Plos One* 2010, **5**(8).

204. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**(5):511-U174.

205. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**. *Nat Biotechnol* 2010, **28**(5):503-U166.

206. Affymetrix_Inc.: **Affymetrix Technical Notes: Array Design for the Human Exon 1.0 ST Array.** In.

207. Homer N, Merriman B, Nelson SF: **BFAST: An Alignment Tool for Large Scale Genome Resequencing**. *Plos One* 2009, **4**(11):A95-A106.

208. David M, Dzamba M, Lister D, Ilie L, Brudno M: **SHRiMP2: Sensitive yet Practical Short Read Mapping**. *Bioinformatics* 2011, **27**(7):1011-1012.

209. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC**. *Genome Res* 2002, **12**(6):996-1006.

210. Yeo G, Holste D, Kreiman G, Burge CB: **Variation in alternative splicing across human tissues**. *Genome biology* 2004, **5**(10).

211. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A *et al*: **The UCSC Genome Browser database: update 2011**. *Nucleic Acids Res* 2011, **39**:D876-D882.

212. Kim P, Kim N, Lee Y, Kim B, Shin Y, Lee S: **ECgene: genome annotation for alternative splicing**. *Nucleic Acids Res* 2005, **33**(Database issue):D75-79.

213. Lixia M, Zhijian C, Chao S, Chaojiang G, Congyi Z: **Alternative splicing of breast cancer associated gene BRCA1 from breast cancer cell line**. *J Biochem Mol Biol* 2007, **40**(1):15-21.

214. Zhu Z, Xing S, Cheng P, Zeng F, Lu G: **Modification of alternative splicing of Bcl-x pre-mRNA in bladder cancer cells**. *J Huazhong Univ Sci Technolog Med Sci* 2006, **26**(2):213-216.

215. Ku TH, Hsu FR: **Mining colon cancer specific alternative splicing in EST database**. *AMIA Annu Symp Proc* 2005:1012.

216. Ogawa T, Shiga K, Hashimoto S, Kobayashi T, Horii A, Furukawa T: **APAF-1-ALT, a novel alternative splicing form of APAF-1, potentially causes impeded ability of undergoing DNA damage-induced apoptosis in the LNCaP human prostate cancer cell line**. *Biochem Biophys Res Commun* 2003, **306**(2):537-543.

217. Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh C, Gervais-Bird J, Lapointe E, Froehlich U, Durand M *et al*: **Identification of alternative splicing markers for breast cancer**. *Cancer Res* 2008, **68**(22):9525-9531.

218. Shastry BS: **SNP alleles in human disease and evolution**. *J Hum Genet* 2002, **47**(11):561-566.

219. Baenziger JU: **A major step on the road to understanding a unique posttranslational modification and its role in a genetic disease**. *Cell* 2003, **113**(4):421-422.

220. Bierhaus A, Nawroth PP: **Posttranslational modification of lipoproteins--a fatal attraction in metabolic disease?** *J Alzheimers Dis* 2005, **7**(4):315-317.

221. Gyorgy B, Toth E, Tarcsa E, Falus A, Buzas EI: **Citrullination: a posttranslational modification in health and disease**. *Int J Biochem Cell Biol* 2006, **38**(10):1662-1677.

222. Wang JZ, Grundke-Iqbal I, Iqbal K: **Glycosylation of microtubule-associated protein tau: an abnormal posttranslational modification in Alzheimer's disease**. *Nature medicine* 1996, **2**(8):871-875.

223. Klampfl CW: **Review coupling of capillary electrochromatography to mass spectrometry**. *J Chromatogr A* 2004, **1044**(1-2):131-144.

224. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R: **The International Protein Index: an integrated database for proteomics experiments**. *Proteomics* 2004, **4**(7):1985-1988.

225. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R *et al*: **The Universal Protein Resource (UniProt): an expanding universe of protein information**. *Nucleic Acids Res* 2006, **34**(Database issue):D187-191.

226. Eng JK, McCormack AL, Yates Iii JR: **An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database**. *Journal of the American Society for Mass Spectrometry* 1994, **5**(11):976-989.

227. Koenig T, Menze BH, Kirchner M, Monigatti F, Parker KC, Patterson T, Steen JJ, Hamprecht FA, Steen H: **Robust prediction of the MASCOT score for an improved quality assessment in mass spectrometric proteomics**. *J Proteome Res* 2008, **7**(9):3708-3717.

228. Bjornson RD, Carriero NJ, Colangelo C, Shifman M, Cheung KH, Miller PL, Williams K: **X!!Tandem, an improved method for running X!tandem in parallel on collections of commodity computers**. *J Proteome Res* 2008, **7**(1):293-299.

229. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH: **Open mass spectrometry search algorithm**. *J Proteome Res* 2004, **3**(5):958-964.

230. Koscielny G, Texier VL, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F, Stanley E, Fallsehr C, Hofmann O, Kull M *et al*: **ASTD: The Alternative Splicing and Transcript Diversity database**. *Genomics* 2008.

231. Shepelev V, Fedorov A: **Advances in the Exon-Intron Database (EID)**. *Brief Bioinform* 2006, **7**(2):178-185.

232. Saxonov S, Daizadeh I, Fedorov A, Gilbert W: **EID: the Exon-Intron Database-an exhaustive database of protein-coding intron-containing genes**. *Nucleic Acids Res* 2000, **28**(1):185-190.

233. Castrignano T, D'Antonio M, Anselmo A, Carrabino D, D'Onorio De Meo A, D'Erchia AM, Licciulli F, Mangiulli M, Mignone F, Pavesi G *et al*: **ASPicDB: a database resource for alternative splicing analysis**. *Bioinformatics* 2008, **24**(10):1300-1304.

234. Lee Y, Kim B, Shin Y, Nam S, Kim P, Kim N, Chung WH, Kim J, Lee S: **ECgene: an alternative splicing database update**. *Nucleic Acids Res* 2007, **35**(Database issue):D99-103.

235. Mooney SD, Altman RB: **MutDB: annotating human variation with functionally relevant data**. *Bioinformatics* 2003, **19**(14):1858-1860.

236. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R, Mooney SD: **MutDB: update on development of tools for the biochemical analysis of genetic variation**. *Nucleic Acids Res* 2008, **36**(Database issue):D815-819.

237. 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):308-311.

238. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T *et al*: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**(1):38-41.

239. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M *et al*: **UniProt: the Universal Protein knowledgebase**. *Nucleic Acids Res* 2004, **32**(Database issue):D115-119.

240. Chen JY, Mamidipalli S, Huan T: **HAPPI: an online database of comprehensive human annotated and predicted protein interactions**. *BMC Genomics* 2009, **10 Suppl 1**:S16.

241. Chowbina SR, Wu X, Zhang F, Li PM, Pandey R, Kasamsetty HN, Chen JY: **HPD: an online integrated human pathway database enabling systems biology studies**. *BMC Bioinformatics* 2009, **10 Suppl 11**:S5.

242. Saha S, Harrison SH, Shen C, Tang H, Radivojac P, Arnold RJ, Zhang X, Chen JY: **HIP2: an online database of human plasma proteins from healthy individuals**. *BMC Med Genomics* 2008, **1**:12.

243. Ying W, Jiang Y, Guo L, Hao Y, Zhang Y, Wu S, Zhong F, Wang J, Shi R, Li D *et al*: **A dataset of human fetal liver proteome identified by subcellular fractionation and multiple protein separation and identification technology**. *Mol Cell Proteomics* 2006, **5**(9):1703-1707.

244. Balgley BM, Laudeman T, Yang L, Song T, Lee CS: **Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy**. *Mol Cell Proteomics* 2007, **6**(9):1599-1608.

245. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ: **Deciphering the splicing code**. *Nature*, **465**(7294):53-59.

246. Mokdad AH, Marks JS, Stroup DF, Gerberding JL: **Actual causes of death in the United States, 2000**. *Jama* 2004, **291**(10):1238-1245.

247. Blackman PH: **Actual causes of death in the United States**. *Jama* 1994, **271**(9):659-660; author reply 660-651.

248. Edenberg HJ, Reynolds J, Koller DL, Begleiter H, Bucholz KK, Conneally PM, Crowe R, Goate A, Hesselbrock V, Li TK *et al*: **A family-based analysis of whether the functional promoter alleles of the serotonin transporter gene HTT affect the risk for alcohol dependence**. *Alcoholism, clinical and experimental research* 1998, **22**(5):1080-1085.

249. Dick DM, Edenberg HJ, Xuei X, Goate A, Kuperman S, Schuckit M, Crowe R, Smith TL, Porjesz B, Begleiter H *et al*: **Association of GABRG3 with alcohol dependence**. *Alcoholism, clinical and experimental research* 2004, **28**(1):4-9.

250. Edenberg HJ, Strother WN, McClintick JN, Tian H, Stephens M, Jerome RE, Lumeng L, Li TK, McBride WJ: **Gene expression in the hippocampus of inbred alcohol-preferring and -nonpreferring rats**. *Genes, brain, and behavior* 2005, **4**(1):20-30.

251. Dick DM, Agrawal A, Schuckit MA, Bierut L, Hinrichs A, Fox L, Mullaney J, Cloninger CR, Hesselbrock V, Nurnberger JI, Jr. *et al*: **Marital status, alcohol dependence, and GABRA2: evidence for gene-environment correlation and interaction**. *Journal of studies on alcohol* 2006, **67**(2):185-194.

252. Edenberg HJ, Xuei X, Chen HJ, Tian H, Wetherill LF, Dick DM, Almasy L, Bierut L, Bucholz KK, Goate A *et al*: **Association of alcohol dehydrogenase genes with alcohol dependence: a comprehensive analysis**. *Hum Mol Genet* 2006, **15**(9):1539-1549.

253. Liang T, Kimpel MW, McClintick JN, Skillman AR, McCall K, Edenberg HJ, Carr LG: **Candidate genes for alcohol preference identified by expression profiling in alcohol-preferring and -nonpreferring reciprocal congenic rats**. *Genome biology* 2010, **11**(2):R11.

254. Edenberg HJ: **Common and rare variants in alcohol dependence**. *Biological psychiatry* 2011, **70**(6):498-499.

255. Song J, Koller DL, Foroud T, Carr K, Zhao J, Rice J, Nurnberger JI, Jr., Begleiter H, Porjesz B, Smith TL *et al*: **Association of GABA(A) receptors and alcohol dependence and the effects of genetic imprinting**. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 2003, **117B**(1):39-45.

256. Edenberg HJ, Dick DM, Xuei X, Tian H, Almasy L, Bauer LO, Crowe RR, Goate A, Hesselbrock V, Jones K *et al*: **Variations in GABRA2, encoding the alpha 2 subunit of the GABA(A) receptor, are associated with alcohol dependence and with brain oscillations**. *American journal of human genetics* 2004, **74**(4):705-714.

257. Wang JC, Hinrichs AL, Stock H, Budde J, Allen R, Bertelsen S, Kwon JM, Wu W, Dick DM, Rice J *et al*: **Evidence of common and specific genetic effects: association of the muscarinic acetylcholine receptor M2 (CHRM2) gene with alcohol dependence and major depressive syndrome**. *Hum Mol Genet* 2004, **13**(17):1903-1911.

258. Dick DM, Wang JC, Plunkett J, Aliev F, Hinrichs A, Bertelsen S, Budde JP, Goldstein EL, Kaplan D, Edenberg HJ *et al*: **Family-based association analyses of alcohol dependence phenotypes across DRD2 and neighboring gene ANKK1**. *Alcoholism, clinical and experimental research* 2007, **31**(10):1645-1653.

259. Edenberg HJ, Xuei X, Wetherill LF, Bierut L, Bucholz K, Dick DM, Hesselbrock V, Kuperman S, Porjesz B, Schuckit MA *et al*: **Association of NFKB1, which**

encodes a subunit of the transcription factor NF-kappaB, with alcohol dependence. *Hum Mol Genet* 2008, **17**(7):963-970.

260. !!! INVALID CITATION !!!

261. Edenberg HJ, Wang J, Tian H, Pochareddy S, Xuei X, Wetherill L, Goate A, Hinrichs T, Kuperman S, Nurnberger JI, Jr. *et al*: **A regulatory variation in OPRK1, the gene encoding the kappa-opioid receptor, is associated with alcohol dependence**. *Hum Mol Genet* 2008, **17**(12):1783-1789.

262. Wang JC, Grucza R, Cruchaga C, Hinrichs AL, Bertelsen S, Budde JP, Fox L, Goldstein E, Reyes O, Saccone N *et al*: **Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence**. *Molecular psychiatry* 2009, **14**(5):501-510.

263. Chen AC, Tang Y, Rangaswamy M, Wang JC, Almasy L, Foroud T, Edenberg HJ, Hesselbrock V, Nurnberger J, Jr., Kuperman S *et al*: **Association of single nucleotide polymorphisms in a glutamate receptor gene (GRM8) with theta power of event-related oscillations and alcohol dependence**. *American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics* 2009, **150B**(3):359-368.

264. Chen AC, Manz N, Tang Y, Rangaswamy M, Almasy L, Kuperman S, Nurnberger J, Jr., O'Connor SJ, Edenberg HJ, Schuckit MA *et al*: **Single-nucleotide polymorphisms in corticotropin releasing hormone receptor 1 gene (CRHR1) are associated with quantitative trait of event-related potential and alcohol dependence**. *Alcoholism, clinical and experimental research* 2010, **34**(6):988-996.

265. Derringer J, Krueger RF, Manz N, Porjesz B, Almasy L, Bookman E, Edenberg HJ, Kramer JR, Tischfield JA, Bierut LJ: **Nonreplication of an association of SGIP1 SNPs with alcohol dependence and resting theta EEG power**. *Psychiatric genetics* 2011, **21**(5):265-266.

266. Nieratschker V, Batra A, Fallgatter AJ: **Genetics and epigenetics of alcohol dependence**. *Journal of molecular psychiatry* 2013, **1**(1):11.

267. Zhang R, Miao Q, Wang C, Zhao R, Li W, Haile CN, Hao W, Zhang XY: **Genome-wide DNA methylation analysis in alcohol dependence**. *Addiction biology* 2013, **18**(2):392-403.

268. Farris SP, Pietrzykowski AZ, Miles MF, O'Brien MA, Sanna PP, Zakhari S, Mayfield RD, Harris RA: **Applying the new genomics to alcohol dependence**. *Alcohol* 2015, **49**(8):825-836.

269. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, Gloss BS, Dinger ME: **lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs**. *Nucleic acids research* 2015, **43**(Database issue):D168-173.

270. Goodrich JA, Kugel JF: **Non-coding-RNA regulators of RNA polymerase II transcription**. *Nature reviews Molecular cell biology* 2006, **7**(8):612-616.

271. Martianov I, Ramadass A, Serra Barros A, Chow N, Akoulitchev A: **Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript**. *Nature* 2007, **445**(7128):666-670.

272. Mercer TR, Dinger ME, Mattick JS: **Long non-coding RNAs: insights into functions**. *Nature reviews Genetics* 2009, **10**(3):155-159.

273. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome research* 2005, **15**(8):1034-1050.

274. Zhao J, Sun BK, Erwin JA, Song JJ, Lee JT: **Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome**. *Science* 2008, **322**(5902):750-756.

275. Ogawa Y, Sun BK, Lee JT: **Intersection of the RNA interference and X-inactivation pathways**. *Science* 2008, **320**(5881):1336-1341.

276. Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, Tempst P, Rosenfeld MG, Glass CK, Kurokawa R: **Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription**. *Nature* 2008, **454**(7200):126-130.

277. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW: **The antisense transcriptomes of human cells**. *Science* 2008, **322**(5909):1855-1857.

278. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR *et al*: **The UCSC Genome Browser database: extensions and updates 2011**. *Nucleic acids research* 2012, **40**(Database issue):D918-923.

279. L. Lumeng TDH, and T. K. Li: **NEW STRAINS OF RATS WITH ALCOHOL PREFERENCE AND NONPREFERENCE**. *Alcohol and Aldehyde Metabolizing Systems* 1977:537-544.

280. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *Journal of molecular biology* 1990, **215**(3):403-410.

281. Gish W, States DJ: **Identification of protein coding regions by database similarity search**. *Nature genetics* 1993, **3**(3):266-272.

282. Rodriguez D, Bretones G, Quesada V, Villamor N, Arango JR, Lopez-Guillermo A, Ramsay AJ, Baumann T, Quiros PM, Navarro A *et al*: **Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia**. *Blood* 2015, **126**(2):195-202.

283. Sakharkar MK, Chow VT, Kangueane P: **Distributions of exons and introns in the human genome**. *In silico biology* 2004, **4**(4):387-393.

284. Fuselli S, Baptista RP, Panziera A, Magi A, Guglielmi S, Tonin R, Benazzo A, Bauzer LG, Mazzoni CJ, Bertorelle G: **A new hybrid approach for MHC**

genotyping: high-throughput NGS and long read MinION nanopore sequencing, with application to the non-model vertebrate Alpine chamois (Rupicapra rupicapra). *Heredity* 2018.

285. Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien FA, Swinkels W, Koelewijn A, Palstra AP, Pelster B, Spaink HP *et al*: **Rapid de novo assembly of the European eel genome from nanopore sequencing reads**. *Scientific reports* 2017, **7**(1):7213.

286. Huang DW, Raley C, Jiang MK, Zheng X, Liang D, Rehman MT, Highbarger HC, Jiao X, Sherman B, Ma L *et al*: **Towards Better Precision Medicine: PacBio Single-Molecule Long Reads Resolve the Interpretation of HIV Drug Resistant Mutation Profiles at Explicit Quasispecies (Haplotype) Level**. *J Data Mining Genomics Proteomics* 2016, **7**(1).

Curriculum Vitae

Ao Zhou

EDUCATION

Ph.D. in Bioinformatics, Indiana University, USA, 2008 – 2018

M.S. in Biophysics, Nanjing University, China, 2005 – 2008

B.S. in Physics, Nanjing University, China, 2001 – 2005

WORK EXPERIENCE

Senior Bioinformatics Engineer, Personalis Inc., May 2018 – Present

Bioinformatics Engineer II, Personalis Inc., May 2017 – April 2018

Bioinformatics Engineer, Personalis Inc., January 2015 – April 2017

- LIMS maintenance and development: architecting and automation of LIMS workflows with lab director, integrating LIMS with lab instruments and data analysis pipeline, development of LIMS extension software, LIMS data extraction, visualization and analysis for management and accounting, sample labeling and freezer management automation, technical support for lab staff.
- Clearing bioinformatics roadblocks in OPS: troubleshooting demultiplexing and data analysis pipeline issues, identifying mistakes in lab, LIMS and code in pipeline

Research Assistant, Indiana University, August 2008 – December 2014

- RNA sequencing analysis: RNA-seq reads mapping, filtering, quality control, expression intensity quantification and comparison, alternative splicing identification and characterization
- miRNA analysis: miRNA discovery, adaptor trimming and functional analysis
- Proteomics: Built splicing junction peptide database PEPPI for mass spectrometry analysis

PUBLICATIONS

Ao Zhou, Hai Lin, Todd Skaar, and Yunlong Liu

"Identification and characterization of pre-mRNA splicing change in human liver across three developmental stages", *under revision*.

Ao Zhou, Meng Li, Bo He, Fei Huang, Bing Xu, A. Keith Dunker, Curt Balch, Baiyan Li, Yunlong Liu, and Yue Wang

"Lipopolysaccharide treatment induces genome–wide pre–mRNA splicing pattern changes in mouse mesenchymal stem cells", *BMC Genomics* 2016.

Ao Zhou, Marcus R. Breese, Yangyang Hao, Howard J. Edenberg, Lang Li, Todd C. Skaar and Yunlong Liu

"Alt Event Finder: A tool for extracting alternative splicing events from RNA–seq data", *BMC Genomics* 2012.

Ao Zhou, Weixing Feng, Yadong Wang, Howard J. Edenberg, and Yunlong Liu

"Characterizing the roles of long noncoding RNA in alcohol dependence", *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2016.

Ao Zhou, Fan Zhang and Jake Y. Chen

"PEPPI: a peptidomic database of human protein isoforms for proteomics experiments", *BMC Bioinformatics* 2010.

Yu Su, **Ao Zhou**, Xuefeng Xia, Wen Li, Zhirong Sun

"Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction", *Protein Science* 2010.

Ao Zhou, David Miller, Doug Rusch, Tim Huang, Ken Nephew, and Yunlong Liu

"Identification and characterization of estrogen induced alternative splicing in MCF7 cells*", Cancer Research Day at IU* 2014 (Poster).