2019

# Machine learning tools for mRNA isoform function prediction

Gaurav Kandoi
*Iowa State University*

**Machine learning tools for mRNA isoform function prediction**

by

**Gaurav Kandoi**

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Julie A. Dickerson, Co-major Professor
Carolyn Lawrence-Dill, Co-major Professor
Iddo Friedberg
Justin Walley
Kris De Brabanter

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2019

## DEDICATION

To my parents, my younger sister and brother for their unconditional support, commitment and encouragement throughout my life.

# TABLE OF CONTENTS

Page

# LIST OF FIGURES

Page

# LIST OF TABLES

Page

## ACKNOWLEDGMENTS

At this point in my five-year long journey through this dissertation, I would like to thank a long list of people, for none of this would be possible without their constant support, encouragement and guidance.

I would like to start by thanking my PhD advisor, Prof. Julie A. Dickerson for her guidance, patience and support throughout the completion of this dissertation. Her valuable insights and critiques have both equally helped grow this dissertation into a worthwhile contribution to the scientific community. The opportunities and freedom that she has provided me during my dissertation has been immensely helpful for evolving me, both professionally and personally.

I thank members of my dissertation committee, Dr. Carolyn Lawrence-Dill, Dr. Iddo Friedberg, Dr. Justin Walley and Dr. Kris De Brabanter for their efforts and contributions throughout this dissertation. The suggestions and feedback that I've received from you has helped improve the quality of research presented in this dissertation. Thank you for also keeping a check on me to make sure I understand the pitfalls and shortcomings of this dissertations.

Akshay Yadav, Anilisa and Manjil, Bekah Starks, Bhakti and Viraj Muthye, Gokul Wimalanathan, Naihui Zhou, Pulkit Kanodia, Sambit Mishra, Saranya and Surya, and Talon Brown, after tolerating all my quirks and still being there, you deserve an honest thank you. Thank you for keeping me sane and making my stay in Ames more eventful, exciting and fun. Thank you, Bekah, for being my cookie buddy. Thank you Naihui, for never reading those drafts.

A heartfelt thanks to my friends' half way across the globe, Abhimanyu Mann, Karnika Taneja, Mehak Rastogi, Preeti Sharma, Pulkit Sharma, Rishabh Agarwal and Vikram Juneja, who I have constantly annoyed for over a decade. Thank you for still being with me and encouraging me to succeed further.

I am also thankful to all my lab members, Divya Mistry, Erin Boggess, Jesse Walsh, and Ruolin Liu for welcoming me into the lab and for the advice and suggestions for a successful dissertation. I also thank, Trish Stauble, for always being there. Your support has been remarkable. Thank you for cooking amazing food and organizing BCB dinners and social events. Thank you for also keeping a tab on my progress throughout the years. I also thank fellow BCB students for their support, notably, Carla Mann, Jennifer Chang and John Hsieh.

A sincere thanks to numerous authors at Stack Overflow and Biostars for their efforts and contributions that have helped me navigate through multiple technical difficulties.

I am thankful to my parents, Mr. Mangturam Kandoi and Mrs. Sweta Kandoi for their unconditional love, support and encouragement to follow my dreams. Your faith in my abilities have been a source of inspiration and allowed me to become a better student and human being. Thank you, Neha and Abhishek, my siblings, for putting up with me during the good and bad times of my journey.

I can thank many more people, but, time, space, and modesty oblige me to stop. If this dissertation fails to provide what you came hoping for, feel free to contact me and I promise to get back to you within reasonable time!

## ABSTRACT

This dissertation is focused on improving mRNA isoform characterization in terms of functional networks, function prediction and tissue-specificity. There are three major challenges in solving these problems. The first is the unavailability of mRNA isoform level functional data which is required to develop machine learning tools. However, the available data, even at the gene level doesn't include all genes, further complicating the matter. The second challenge is the lack of information about tissue-specificity in functional databases such as Gene Ontology, Kyoto Encyclopedia of Genes and Genomes and UniProt. The third challenge is the lack of mRNA isoform level "ground truth" functional annotation data. The scope of this dissertation includes using mRNA isoform and protein sequences, high-throughput RNA-sequencing data and functional annotations at the gene level to develop computational methods for predicting functions for alternative spliced mRNA isoforms in mouse.

To address these challenges, this dissertation develops and describes two computational tools. The first is a supervised learning-based machine learning framework for predicting tissue-specific mRNA isoform functional networks. Tissue-spEcific mrNa iSoform functIonal Networks (TENSION) makes use of single mRNA producing gene annotations and gene annotations tagged with "NOT" to create a high-quality mRNA isoform level functional data. We use these mRNA isoform level functional data to train random forest algorithms to develop mRNA isoform functional network prediction models. By using a leave-one-tissue-out approach and incorporating tissue-specific mRNA isoform level predictors along with those obtained from mRNA isoform and protein sequences, we have developed mRNA isoform level functional networks for 17 mouse tissues. We

identify about 10.6 million tissue-specific functional mRNA isoform interactions and demonstrate the ability of our networks to reveal tissue-specific functional differences of the isoforms of the same genes. We validate our models and predictions by using a series of tests such as 10-fold stratified cross validation, comparison with published method and validating against literature datasets. As a result, we have also generated a high-quality mRNA isoform level functional dataset that can be used for benchmarking future methods.

Next, we describe mRNA Function Recommendation System (mFRecSys), a recommendation system for making tissue-specific function recommendations for mRNA isoforms. In mFRecSys, we consider mRNA isoforms as "users" and Gene Ontology biological process terms as "items". By using explicit contexts for mRNA isoforms, Gene Ontology biological process terms and tissue-specific mRNA isoform expression, mFRecSys is able to make tissue-specific mRNA isoform function recommendations.

This work emphasizes the significance of incorporating diverse biological context to develop better machine learning tools for biology. It also highlights the use of simplified supervised learning methods for biological network prediction. The machine learning models and recommendation systems developed as part of this work also draw attention to the power of simple mRNA isoform sequence-based predictors to improve mRNA isoform function prediction. The methods developed have potential practical applications, for instance as predictive models for distinguishing the functions of different mRNA isoforms of the same gene or identifying tissue-specific functions of mRNA isoforms.

# CHAPTER 1. GENERAL INTRODUCTION

## 1.1. Alternative Splicing

A gene is a functional unit of heredity, a sequence of DNA within the genome that functions by producing a discrete RNA or a polypeptide product (Krebs, Jocelyn E., 2017). The specific location where a gene resides on a chromosome is formally referred to as a genetic locus (Krebs, Jocelyn E., 2017). A gene can exist in multiple forms, each with small difference (or no difference) in their DNA sequence (US National Library of Medicine, 2018). The alleles are the different forms of the same gene found at its genetic locus (Krebs, Jocelyn E., 2017). A gene is transcribed into a precursor messenger RNA (pre-mRNA), which undergo splicing to generate mature mRNAs that is colinear with the polypeptide product (Krebs, Jocelyn E., 2017). These mature mRNAs are then translated into polypeptide products (Krebs, Jocelyn E., 2017). Gene expression is the process used to synthesize an RNA or polypeptide product using the information from a gene (Krebs, Jocelyn E., 2017).

Alternative Splicing (AS) is a post-transcriptional regulatory mechanism that allows a cell to generate multiple unique mRNA isoforms from a single gene. The generated mRNA isoforms can differ in their coding sequence or untranslated regions (UTRs). These differences in the sequence of different mRNA isoforms of the same gene can result from one of many AS mechanisms. The most common AS mechanisms include intron retention (where an intron is transcribed and present in the mature mRNA), exon skipping (specific exons are not transcribed in the mature mRNA) and the use of alternative 5'/3' donor/acceptor sites. AS occurs as a normal phenomenon in eukaryotes and is more prevalent in higher eukaryotes such as plants and mammals (Keren, Lev-Maor, & Ast, 2010). Recent studies highlight that ~90% multi-exon human genes, ~60% of multi-exon *Drosophila* genes and ~61% intron-containing

*Arabidopsis thaliana* genes undergo AS (Graveley et al., 2011; Syed, Kalyna, Marquez, Barta, & Brown, 2012). The most common types of AS events in animals and plants are not always the same. In plants, intron retention is the most prevalent AS event (~ 40%), but, the least frequent in humans (Marquez, Brown, Simpson, Barta, & Kalyna, 2012). In humans, exon skipping it the most prevalent AS event (> 40%), however, the least frequent in plants (~ 5%) (Keren et al., 2010).

The functional importance of AS should be apparent based on the ubiquitousness of the phenomenon. The consequences of AS are vast and can result in mRNA isoforms which are non-functional to those that perform completely opposite functions. These mRNA isoforms have different biological properties such as subcellular localization, protein-protein interactions and catalytic abilities (Rafalska et al., 2004). Some other functions of AS include generating protein diversity, gene expression regulation, stress response, mRNA stability, developmental and physiological processes. An interesting example of AS is the generation of a stress-initiated exon skipping of SMG1 exon 63 in peripheral leukocytes of male medical students during examination stress (Kurokawa et al., 2010). Not all mRNA isoforms generated as a consequence of AS are functional and can be quickly degraded. This provides the cell with another method of regulating gene expression before translation. Despite the various function of AS, like most other biological processes, the complete roles and mechanisms of AS are still unknown.

### 1.2. Gene Ontology

A controlled vocabulary representing our current knowledge of gene (or gene products) functions is computationally represented in the form of Gene Ontology (GO). The structure of GO can be described as a graph, where GO terms serve as nodes while the edges represent the

relationship between the terms. The GO is semi-hierarchical, where a GO term can have multiple parents. The GO describes three distinct aspects of a gene (or gene products): 1) Cellular Component, 2) Molecular Function, and 3) Biological Process.

The cellular component aspect of GO refers to a cellular anatomy unlike other GO aspects that refer to processes. It describes the cellular locations, either compartments such as mitochondrion or stable macromolecular complexes such as ribosome, where the gene (or gene product) performs a function.

The activities performed by genes (or gene products) at the molecular level is captured by the molecular function aspect of GO. Activities such as “catalysis” or “transport” that occur at the molecular level are some examples of molecular function terms. The GO molecular function terms do not specify the context in which the action takes place. Neither do these terms specify the location and time of the actions. Rather, these terms represent the activities (such as catalytic activity), but not the entities (molecules or complexes).

The larger processes that are made up of multiple molecular activities are defined by the biological process aspect of GO. The complex dependencies or dynamics required to completely define a pathway are not represented in GO. Therefore, it should be noted that a biological process is not an equivalent of a pathway. Some examples of GO biological process terms are glucose transmembrane transport, signal transduction or DNA repair.

The functional annotations of a gene refer to the assignment of one or more GO terms from one or more GO aspects. An evidence code describing how the annotation is supported is included with every annotation. These evidence codes fall under six general categories: 1) Experimental, 2) Phylogenetic, 3) Computational, 4) Author statements, 5) Curatorial statements, and 6) Automatically generated. The experimental evidence code indicates an

experiment-level evidence directly supporting the annotation. The annotations obtained from an explicit gain and loss of gene functions from a specific branch of a phylogenetic tree are supported by a phylogenetic evidence code. Annotations obtained from an *in silico* analysis are indicated by the computational evidence code. If authors make a statement about a functional annotation of a gene (or gene product) in a cited reference, such annotations are supported by the author statement evidence codes. Similarly, if a curator makes a statement about a functional annotation of a gene (or gene product), and such annotations do not fit into other evidence codes, such annotations are supported by the curator statement evidence codes. The automatically generated evidence code is the least supported evidence code since no reviewed analysis of the functional annotation is performed.

### 1.3. Machine Learning

Machine learning is a data-driven approach that has been utilized for developing predictive models in biology for a long time. In its most basic form, the goal of a machine learning system is to find a function (or a mapping) that is able to distinguish one class of entities from another. In doing so, a machine learning system exploits the information characteristic of the entities under investigation. A typical lifecycle of a machine learning task consists of the following: 1) Feature calculation, 2) Label generation, 3) Model training, 4) Model parameter and feature optimization, and 5) Final predictions.

The first two steps, Feature calculation and Label generation are the most crucial part of any machine learning task. These steps involve calculating features or predictors that are most predictive of separating one class of entities from another and defining the class (labels) for a subset of elements in the data. A machine learning model is as accurate as its predictors are capable of distinguishing one class from another and how closely the labels represent the truth.

Machine learning algorithms can be categorized into two groups based on the type of predictions they make. When the target classes are categorical, such as functional vs non-functional mRNA, such machine learning prediction problems are referred to as classification problems. At the same time, if the target class is continuous, such as metabolic flux through a reaction or a pathway, such machine learning problems are referred to as regression problems.

Machine learning algorithms can also be categorized based on how they are trained. If a machine learning algorithm is trained using the known classes for a subset of elements in the data, such machine learning problems are an example of supervised learning. At the same time, if there is no prior knowledge about the classes for a subset of elements in the data, such machine learning problems are considered a part of unsupervised learning. Some commonly used supervised learning algorithms include generalized linear models, logistic regression, random forest, and support vector machines. Algorithms such as those used for clustering (hierarchical clustering, k-means, and mixture models), anomaly detection, and techniques for blind signal separation (principal components analysis, singular vector decomposition, and non-negative matrix factorization) are some popular unsupervised learning algorithms.

The utility of machine learning in bioinformatics and computational biology is enormous. Some common applications include gene function prediction, drug target identification, protein-protein interaction prediction, protein structure prediction (secondary and tertiary structures) and active site prediction (Dale, Popescu, & Karp, 2010; Demerdash, Daily, & Mitchell, 2009; Kandoi, Acencio, & Lemke, 2015; Kandoi, Leelananda, Jernigan, & Sen, 2017; Kandoi & Dickerson, 2019; Mishra, Kandoi, & Jernigan, 2018, 2019; Petrova & Wu, 2006).

### 1.4. Recommendation Systems

While the tools and techniques used in machine learning have been applied to biological problems for a long time, another set of techniques collectively known as recommendation systems are yet to be explored for problems in biology. Recommendation systems are a set of tools and techniques capable of providing suggestions, of some sort, for "items" useful to a "user". In the context of biology, the "users" will typically be a molecular entity such as genes or mRNA, while the "items" represent a biological property such as function or structure. A recommendation system can be formulated as either a classification or regression problem, or as a supervised or unsupervised problem making them a set of very powerful tools and techniques.

Some desirable features of a good recommendation system for use in biology include: 1) user as well as item context, 2) personalized as well as novel recommendations, and 3) the ability to work with limited and sparse datasets. The explosive growth in available biological data generated from omics technologies can overwhelm many traditional machine learning frameworks. However, such an information overload is rather perfect for a recommendation system, which relies mostly on highly efficient, scalable and parallelizable matrix calculations. Recommendation systems also allow us to incorporate information pertaining to the biological property under study, which is often difficult in traditional machine learning frameworks. Nevertheless, it should be noted that like most machine learning frameworks, recommendation systems also suffer from sparsity that is inherent in biological data.

At the core of any recommendation system, there is matrix factorization (MF). Matrix factorization is a class of techniques used to identify a low-dimensional representation (latent space) of an otherwise large data while preserving as much information as possible. A very popular and commonly used form of recommendation systems is matrix factorization for

collaborative filtering. In this approach, the user-item association matrix is projected into two latent spaces, whose dot product is an estimate of the original user-item associations.

While the basic MF techniques have been useful for several other problems (e.g. Movie recommendation) (Koren, Bell, & Volinsky, 2009), it is not ideal for many biological problems for few reasons. First, there is a huge difference in the number of biological molecules (such as mRNA and protein) and biological properties (e.g. GO terms). This makes projecting the users and the items onto same latent feature space difficult. Second, a drawback of the basic MF approach is that it doesn't allow us to incorporate explicit biological context. Third, the amount of known true labels for most biological problems is very limited. This insufficient information leads to the cold-start problem for test entities, where we don't have enough information to make relevant recommendations.

However, the tri-factorization approach proposed previously for predicting multi-relational dyadic data (Nickel, Tresp, & Kriegel, 2011) can be used for many biological problems, including mRNA isoform function prediction. In a tri-factorization approach, the user as well as the items are respectively projected into latent spaces of different sizes. After that, a third mapping will associate them, leading to the final recommendations. The advantage of using tri-factorization approach as opposed to an MF based collaborative filtering is that we can introduce explicit biological context and can use non-linear mappings.

## 1.5. Problem Formulations

**Problem 1. Tissue-specific mRNA isoform level functional network prediction**

With identifying and computing properties of mRNA isoforms characteristic of their tissue-specific function, and a way to propagate sufficient gene level functions at the mRNA

isoform level, the problem is to develop models capable of predicting whether or not two mRNA isoforms will be involved in a common function in a tissue-specific manner.

Few methods have been previously developed to predict mRNA isoform level functional networks (H.-D. D. Li et al., 2016; Tseng et al., 2015) with great success. The aim is to improve upon these methods while also creating a publicly accessible high-quality mRNA isoform level functional dataset. Some limitations of these studies, that have been overcome in the current work include: 1) Predicting novel mRNA isoform interactions with no gene level interaction information in current biological databases (a limitation of (Tseng et al., 2015)), 2) Predicting tissue-specific mRNA isoform level functional networks, 3) Limiting bias in the machine learning model by using a more biologically sound way of defining non-functional (negative pairs) mRNA isoform pairs, and 4) Formulating the task of mRNA isoform level functional network prediction as a simple supervised learning task.

**Problem 2. Tissue-specific mRNA isoform level function recommendation**

Several recently developed methods have greatly advanced our understanding of mRNA isoform functions by tackling the problem of mRNA isoform function prediction (Eksi et al., 2013; W. Li et al., 2014; Luo et al., 2017; Panwar et al., 2016; Shaw, Chen, & Jiang, 2018). Despite their success in mRNA isoform level function prediction, there are several shortcomings in these methods. For instance, IsoPred (Eksi et al., 2013) and IsoFunc (Panwar et al., 2016) maintains all evidence codes, assumes unannotated genes as non-functional (negative), initializes all mRNA isoforms of the same gene as functional (positive), uses only mRNA expression profile as predictors, and do not utilize information other than the obvious hierarchical relationship between the GO terms.

The aim is to overcome some of these shortcomings and develop recommendation systems for tissue-specific mRNA isoform level GO biological process recommendation. Some limitations of previous studies, that have been overcome in the current work include: 1) Exploiting characteristics of mRNA isoforms apart from their expression profile, 2) Recommending tissue-specific mRNA isoform level functions, 3) Limiting bias in the training and testing process by using a more biologically sound way of defining non-functional (negative pairs) mRNA isoform pairs, 4) Formulating the task of mRNA isoform function prediction as a recommendation system, and 5) Incorporating the relations between the GO terms apart from the obvious hierarchical relations.

### 1.6. Dissertation Organization

This dissertation is divided into 5 chapters. Figure 1 shows the analysis structure of this work. A brief description of other chapters is provided below.

**Chapter 1** provides an introduction and details the background for this dissertation. It also describes the specific problems being addressed in this work.

**Chapter 2** includes a published manuscript motivating the need to study alternatively spliced mRNA isoforms. By analyzing RNA-Seq datasets obtained from *Arabidopsis thaliana* subjected to heat and cold stress, we show that more knowledge can be gained regarding biological regulation by using differentially alternatively spliced genes (DASGs) in addition to differentially expressed genes (DEGs). The manuscript has been published under the title, "*Differential alternative splicing patterns with differential expression to computationally extract plant molecular pathways*" by *Gaurav Kandoi* and *Julie A. Dickerson* as part of the *Integrative Data Analysis in Systems Biology* workshop during *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Kandoi & Dickerson, 2017).

**Chapter 3** includes a manuscript currently under review at the peer reviewed journal, *Scientific Reports*. By utilizing information from mRNA isoform sequence, protein sequence and mRNA isoform expression profile and exploiting the GO annotations involving single mRNA producing genes and those tagged with "NOT", we have developed 17 tissue-specific mRNA isoform level functional networks in addition to an organism level network for mouse. The manuscript is titled, "*Tissue-specific mouse mRNA isoform networks*". All code and data associated with the manuscript is also freely available through DataShare: Iowa State University's Open Research Data Repository through doi: https://doi.org/10.25380/iastate.c.4275191 (Dickerson & Kandoi, 2019).

**Chapter 4** includes a manuscript currently under preparation for submission to a peer reviewed journal. We have developed recommendation systems for tissue-specific mRNA isoform level function recommendations for mouse. The recommendation system predicts the association of mRNA isoforms with GO biological process terms by utilizing input information from mRNA isoform sequences, protein sequences, mRNA isoform expression profile and the semantic similarity between GO biological process terms. The system also exploits the GO annotations involving single mRNA producing genes and those tagged with "NOT" for generating the training and testing labels. The manuscript is titled, "*mFRecSys: mRNA Function Recommendation System*" by *Gaurav Kandoi* and *Julie A. Dickerson*. All code and data associated with this manuscript will be freely available through DataShare: Iowa State University's Open Research Data Repository.

**Chapter 5** outlines the general conclusions of this dissertation and also suggests future directions with improvements that can be made to the dissertation.

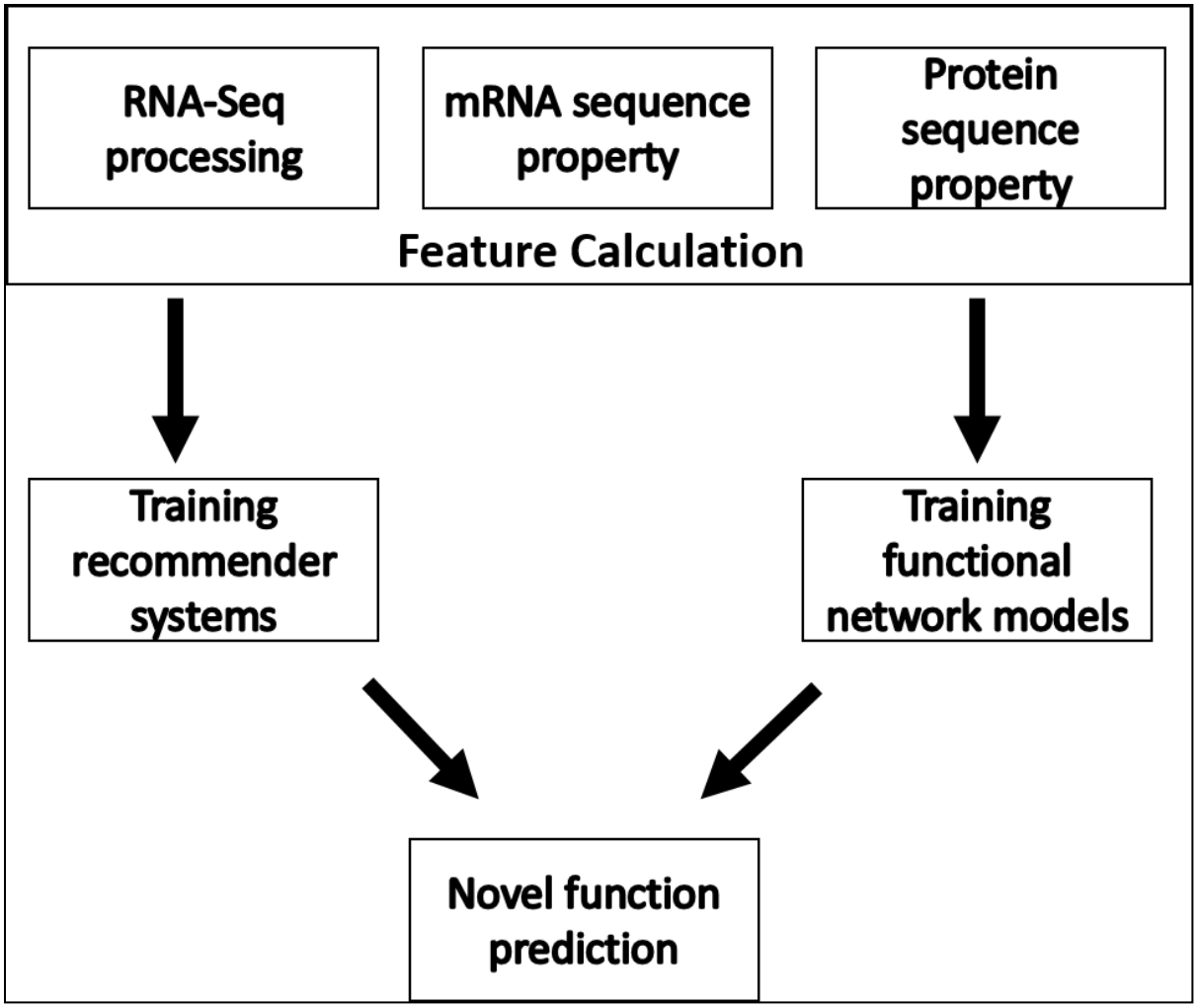


Figure 1.1 **An overview of this dissertation.** Both problems being addressed in this dissertation, 1) developing tissue-specific mRNA isoform level functional networks, and 2) developing tissue-specific mRNA isoform function recommendation systems lead to the characterization of mRNA isoforms of the same gene.

## References

Dale, J. M., Popescu, L., & Karp, P. D. (2010). Machine learning methods for metabolic pathway prediction. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-11-15

Demerdash, O. N. A., Daily, M. D., & Mitchell, J. C. (2009). Structure-based predictive models for allosteric hot spots. PLoS Computational Biology, 5(10). https://doi.org/10.1371/journal.pcbi.1000531

Dickerson, J. A., & Kandoi, G. (2019). Tissue-spEcific mrNa iSoform functIOnal Networks (TENSION) Collection. https://doi.org/10.25380/iastate.c.4275191

Eksi, R., Li, H. D., Menon, R., Wen, Y., Omenn, G. S., Kretzler, M., & Guan, Y. (2013). Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. PLoS Computational Biology, 9(11). https://doi.org/10.1371/journal.pcbi.1003314

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., … Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. Nature, 471(7339), 473–479. https://doi.org/10.1038/nature09715

Kandoi, G., Acencio, M. L., & Lemke, N. (2015). Prediction of druggable proteins using machine learning and systems biology: A mini-review. Frontiers in Physiology, Vol. 6. https://doi.org/10.3389/fphys.2015.00366

Kandoi, G., & Dickerson, J. A. (2017). Differential alternative splicing patterns with differential expression to computationally extract plant molecular pathways. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017-Janua, 2144–2151. https://doi.org/10.1109/BIBM.2017.8217993

Kandoi, G., & Dickerson, J. A. (2019). Tissue-specific mouse mRNA isoform networks. BioRxiv Bioinformatics, 558361. https://doi.org/10.1101/558361

Kandoi, G., Leelananda, S. P., Jernigan, R. L., & Sen, T. Z. (2017). Predicting protein secondary structure using consensus data mining (CDM) based on empirical statistics and evolutionary information. In Methods in Molecular Biology (Vol. 1484). https://doi.org/10.1007/978-1-4939-6406-2_4

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews. Genetics, 11(5), 345–355. https://doi.org/10.1038/nrg2776

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30–37. https://doi.org/10.1109/MC.2009.263

Krebs, Jocelyn E., E. al. (2017). Lewin's GENES XII. Retrieved from https://books.google.com/books?id=U0B4DgAAQBAJ

Kurokawa, K., Kuwano, Y., Tominaga, K., Kawai, T., Katsuura, S., Yamagishi, N., … Rokutan, K. (2010). Brief naturalistic stress induces an alternative splice variant of SMG-1 lacking exon 63 in peripheral leukocytes. Neuroscience Letters, 484(2), 128–132. https://doi.org/10.1016/j.neulet.2010.08.031

Li, H.-D. D., Menon, R., Eksi, R., Guerler, A., Zhang, Y., Omenn, G. S., & Guan, Y. (2016). A Network of Splice Isoforms for the Mouse. Scientific Reports, 6(April), 1–11. https://doi.org/10.1038/srep24507

Li, W., Kang, S., Liu, C. C., Zhang, S., Shi, Y., Liu, Y., & Zhou, X. J. (2014). High-resolution functional annotation of human transcriptome: Predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Research, 42(6), e39–e39. https://doi.org/10.1093/nar/gkt1362

Luo, T., Zhang, W., Qiu, S., Yang, Y., Yi, D., Wang, G., … Wang, J. (2017). Functional Annotation of Human Protein Coding Isoforms via Non-convex Multi-Instance Learning. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17, 345–354. https://doi.org/10.1145/3097983.3097984

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Research, 22(6), 1184–1195. https://doi.org/10.1101/gr.134106.111

Mishra, S. K., Kandoi, G., & Jernigan, R. L. (2018). Dynamic Communities in Proteins: Allosteric Hotspots and Functional Modules. Biophysical Journal, 114(3), 421a. https://doi.org/10.1016/j.bpj.2017.11.2334

Mishra, S. K., Kandoi, G., & Jernigan, R. L. (2019). Coupling Dynamics and Evolutionary Information with Structure to Identify Protein Regulatory and Functional Binding Sites. Proteins: Structure, Function, and Bioinformatics, prot.25749. https://doi.org/10.1002/prot.25749

Nickel, M., Tresp, V., & Kriegel, H.-P. (2011). A Three-Way Model for Collective Learning on Multi-Relational Data. ICML, 809--816.

Panwar, B., Menon, R., Eksi, R., Li, H.-D., Omenn, G. S., & Guan, Y. (2016). Genome-Wide Functional Annotation of Human Protein-Coding Splice Variants Using Multiple Instance Learning. Journal of Proteome Research, 15(6), 1747–1753. https://doi.org/10.1021/acs.jproteome.5b00883

Petrova, N. V., & Wu, C. H. (2006). Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties. BMC Bioinformatics, 7. https://doi.org/10.1186/1471-2105-7-312

Rafalska, I., Zhang, Z., Ben-Ari, S., Stamm, S., Thanaraj, T. A., Toiber, D., … Soreq, H. (2004). Function of alternative splicing. Gene, 344, 1–20. https://doi.org/10.1016/j.gene.2004.10.022

Shaw, D., Chen, H., & Jiang, T. (2018). DeepIsoFun: a deep domain adaptation approach to predict isoform functions. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty1017

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., & Brown, J. W. S. (2012, October). Alternative splicing in plants - coming of age. Trends in Plant Science, Vol. 17, pp. 616–623. https://doi.org/10.1016/j.tplants.2012.06.001

Tseng, Y. T., Li, W., Chen, C. H., Zhang, S., Chen, J. J. W., Zhou, X. J., & Liu, C. C. (2015). IIIDB: A database for isoform-isoform interactions and isoform network modules. BMC Genomics, 16(Suppl 2), 1–7. https://doi.org/10.1186/1471-2164-16-S2-S10

US National Library of Medicine. (2018). What is a gene? - Genetics Home Reference - NIH. Retrieved July 10, 2019, from US National Library of Medicine website: https://ghr.nlm.nih.gov/primer/basics/gene

# CHAPTER 2. DIFFERENTIAL ALTERNATIVE SPLICING PATTERNS WITH DIFFERENTIAL EXPRESSION TO COMPUTATIONALLY EXTRACT PLANT MOLECULAR PATHWAYS

Gaurav Kandoi, and Julie A. Dickerson

2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)

doi: 10.1109/BIBM.2017.8217993

## Author's contributions

GK leads this study. GK and JD contribute to the design of the study and the interpretation of the results. GK and JD together wrote the manuscript. GK wrote the programs and performed data analysis. All the authors read and approved the final manuscript.

## Abstract

Alternative splicing (AS) produces multiple messenger RNAs by combining different regions of the precursor transcript to produce diversity in gene products. Under stress conditions, many genes produce transcripts that are not otherwise produced during normal conditions. Plant growth and development are extensively affected by environmental stresses. In this study, we combine Differentially Alternatively Spliced Genes (DASGs) with Differentially Expressed Genes (DEGs) to discover important metabolic networks in the presence of environmental stress. Using publicly available RNA-Seq datasets from Arabidopsis thaliana (Col-0) subjected to heat stress conditions, we extracted several molecular pathways associated with temperature stress-response using genes that are either differentially alternatively spliced or differentially expressed. Most DASGs are linked with

biological processes such as splicing, circadian rhythm, and metabolic processes. In contrast, most DEGs are linked with cell cycle and division, and transport. These differences in the biological processes highlight the importance of integrating differential splicing information along with differential expression to extract important metabolic pathways. Our analysis suggests that the exon/intron usage of the transcripts involved in key metabolic pathways significantly changes during heat stress conditions.

## Introduction

Alternative splicing (AS) produces multiple messenger RNAs using different combinations of introns and exons of the precursor transcript to produce diversity in gene products. In some extreme cases, thousands of splice isoforms for one gene can be produced by AS (Celotto & Graveley, 2001). Apart from generating transcript diversity, AS is also crucial for many processes such as regulation of gene expression, protein diversity, developmental changes and response to environmental stresses (Syed, Kalyna, Marquez, Barta, & Brown, 2012). Not all AS isoforms are functional and most may result in loss of function.

The study of AS has greatly benefited from the ability to better view the transcriptome using data obtained from RNA sequencing technologies. By generating archives of short sequence reads and mapping them back to the genome and transcriptome, we can define exon-intron structures while quantifying exon/intron usage and discovering novel gene-products. These studies illuminate the differences between sets of gene regulated by AS and those regulated by differential expression (Pan et al., 2004). This suggests the complementary nature of differential alternative splicing and differential expression in regulating biological processes. Downstream analyses of these genes can lead to the identification of important

biological pathways. This knowledge of gene expression and alternative splicing can also help better understand mechanisms of biological disorders in plants and animals alike.

AS is a natural process in eukaryotes and is more prevalent in higher eukaryotes than in lower eukaryotes (Keren, Lev-Maor, & Ast, 2010). It has been widely studied at the functional and protein level in animals, but not as much in plants (Blencowe, 2006). Observed rates of AS in multi-exon genes is as high as 95% in humans and 60% in plants (Marquez, Brown, Simpson, Barta, & Kalyna, 2012; Pan, Shai, Lee, Frey, & Blencowe, 2008). In plants, different types of AS occur as compared with mammals. For example, intron retention is most frequently observed (~40% of multi-exon genes) and exon skipping is least common (~5%). However, the opposite is true for humans, suggesting different ways of recognizing exons and introns in plants and humans (Reddy, Rogers, Richardson, Hamilton, & Ben-Hur, 2012).

Several studies have shown the influence of environmental conditions on AS including genes responsible for modulating stress (Mach, 2009; Sugliani, Brambilla, Clerkx, Koornneef, & Soppe, 2010). The exact processes by which these changes occur and their functional consequences are still widely unknown. Many alternatively spliced isoforms are functionally distinct (Black, 2003) and it is thus important to discover fundamental alternative splicing networks.

Differential alternative splicing and gene expression are both key components of gene regulation. Using a heat stress RNA-Seq dataset in Arabidopsis thaliana, we show that using both differential expression and differential alternative splicing leads to better understanding of perturbations in biological pathway than using only differential expression. Further, by performing gene enrichment analysis, we demonstrate that DEGs are involved in different sets of biological process and molecular functions than DASGs

## Methods

### A. Heat Stress Dataset

The heat stress RNA-Seq dataset used in this study was taken from the Gene Expression Omnibus database (GEO accession GSE85281) (Pajoro, Severing, Angenent, & Immink, 2017). Total RNA from shoot apical meristem enriched tissues was isolated from ~10 plants for each sample. The dataset has three biological replicates for different temperature conditions. All plants were initially grown in short day conditions (8h light/16h dark) at 16℃ for five weeks. The plants were either grown at 16℃ or moved to 25℃ (heat stress). Total RNA was extracted at Day 1 after temperature change from plants growing at 16℃ and 25℃ and at Day 3 and Day 5 after temperature change from the plants growing at 25℃ (hereafter referred to as 16C, 25.1C, 25.3C and 25.5C respectively).

### B. Processing Of RNA Sequencing Reads

The raw sequence reads were mapped to the Arabidopsis thaliana TAIR10 genome using the ultra-fast aligner STAR v2.5.3a (Dobin et al., 2013) (with --outSAMtype BAM SortedByCoordinate --outSAMstrandField intronMotif --chimOutType WithinBAM --chimSegmentMin 20). Read quality check was performed using FastQC (Simon Andrews, 2010). The aligned RNA-Seq reads were then assembled and quantified using StringTie v1.2.4 (default parameters) (Pertea, Kim, Pertea, Leek, & Salzberg, 2016). To compare the impact of annotations on differential expression and differential alternative splicing, we repeat the analysis using Araport11 (Cheng et al., 2017) and AtRTD2 (Zhang et al., 2017) annotations. For all downstream analyses, we don't consider novel transcript predictions.

## C. Differential Gene Expression Analysis

Genes with significant changes in their expression levels across conditions are referred to as Differentially Expressed Genes (DEGs). Analysis of differential expression was carried out using Ballgown v2.4.3 (default parameters) (Pertea et al., 2016) in R v3.3.1. Genes were considered DEGs if their fold change was greater than 2 at a false-discovery rate of 0.05. We perform pairwise analysis for all four conditions leading to six comparisons (16C vs 25.1C; 16C vs 25.3C; 16C vs 25.5C; 25.1C vs 25.3C; 25.1C vs 25.5C and 25.3C vs 25.5C). We repeat the analysis with both Araport11 and AtRTD2.

## D. Differential Gene Alternative Splicing Analysis

Genes with significantly different exon/intron splicing patterns across conditions are referred to as Differentially Alternatively Spliced Genes (DASGs). Analysis of differential alternative splicing was carried out using rMATS v3.2.5 (default parameters) (S. Shen et al., 2014). For selecting DASGs, genes with at least one type of differential alternative splicing event at a cutoff of >10% for splicing differences and a false-discovery rate of 0.05 was used. We perform pairwise analysis for all four conditions (16C vs 25.1C; 16C vs 25.3C; 16C vs 25.5C; 25.1C vs 25.3C; 25.1C vs 25.5C and 25.3C vs 25.5C). Again, we repeat the analysis with both Araport11 and AtRTD2.

## E. Gene Ontology And Pathway Analysis

For the DEGs and DASGs, gene ontology and pathway enrichment analysis was performed with ThaleMine at Araport (Krishnakumar et al., 2015) using Gene Ontology annotations (dated: 8/01/2016) and GenomeNet KEGG pathways data set v.79.0. We use hypergeometric test with multiple testing correction using Holm-Bonferroni at a significance

level of 0.05. The DEGs and DASGs are also used to extract important molecular networks from KEGG. Pathways from KEGG are identified by mapping the set of DEGs and DASGs onto the KEGG pathways using KEGG Mapper. For comparison across the gene sets, all analyses are performed individually using the DEGs, DASGs and combination of the two sets.

### F. Conserved Domain Analysis

To assess the impact of alternative splicing on gene functions, the transcripts from DASGs mapped to the pathways are annotated with the domains in Araport (Cheng et al., 2017; Krishnakumar et al., 2015) and novel predictions from NCBI's Conserved Domain Database (CDD) (Marchler-Bauer et al., 2015).

## Results

### A. Genome Annotations Impact The Results Of Differential Alternative Splicing Analysis

The number of DEGs are similar using both Araport11 and AtRTD2. More genes are differentially alternatively spliced than those that are differentially expressed in the AtRTD2 results, but the opposite can be seen for Araport11 results. Because the Araport11 annotations don't describe many alternatively spliced transcripts, the number of significant DASGs is less.

AtRTD2 describes about 82000 transcripts (74000 from protein coding genes), but there are only about 48000 (from protein coding genes) in Araport11 annotations. AtRTD2 contains 30538 and 18801 transcript annotations from Araport11 and TAIR10 respectively. IR accounts for ~ 45% of all (41759) AS events followed by A3SSS (25%), SE (16%) and A5SSS (14%) in all AtRTD2 comparisons. Slightly different frequency is observed in Araport11 comparisons (IR: 46%; SE: 21%; A3SSS: 19%; A5SSS: 12% of 25772 total events)

A summary of the numbers of DEGs and DASGs found using Ballgown and rMATS using Araport11 and AtRTD2 is presented in Table 1. The number of DEGs are comparable across both annotation sets, Araport11 and AtRTD2. Most of the significant genes are consistent between the corresponding results of Araport11 and AtRTD2 (Fig. 1). For instance, out of 342 genes in AtRTD2 16C vs 25.5C and 323 genes in the corresponding Araport11 comparison, 272 genes are present in both sets. This correspondence between the DEGs using either of the two annotations suggest that the effects of the choice of annotations for differential gene expression analysis are small.

However, the difference in DASGs is very high. Table 1 shows that there are many more DASGs from AtRTD2 than Araport11. This large difference in the number of significant DASGs can be attributed to the fact that AtRTD2 contains many transcripts which are not annotated in Araport11. Most of the genes reported to be differentially alternatively spliced in Araport11 are also found in AtRTD2 (Fig. 2).

There is little overlap (of genes) between the DEGs and DASGs (Fig. 3). Biological processes are regulated by both differential expression and differential alternative splicing (Sheng et al., 2015). Combining DASGs with DEGs may provide new insight to biological pathways. Alternatively, the minimal overlap between DEGs and DASGs could be because of the difference in annotation alone.

Since there are more annotated transcripts in the AtRTD2 annotations, all further analyses were performed only on the results obtained from AtRTD2.

## B. Differentially Alternatively Spliced Genes Help Discover Important Biological Pathways

Most bioinformatics analyses use the set of DEGs to extract and study molecular networks. To be able to study sets of genes from the perspective of a system (pathways in our case) is a primary goal of systems biology. However, differential expression is only one piece of the complex biological regulation process. There exists another yet complementary method to regulate biological processes by differential alternative splicing of a gene (Pan et al., 2004). With this purview, we use DASGs in addition to DEGs to extract important biological pathways.

We perform KEGG pathway enrichment for the DASGs and DEGs for all 6 comparisons using Araport's ThaleMine (Krishnakumar et al., 2015). The significant pathways (p-value < 0.05) are summarized in Table 2. The pathways that remain significant after testing for multiple correction are marked in bold. Many common pathways are found to be enriched for the DASGs (significant in at least 3 comparisons). Some of these include the Spliceosome, Sulfur relay system, and Folate biosynthesis pathways.

Fewer significant pathways were found using DEGs and most pathways are specific to a single comparison. No common pathway was found to be significant in at least three comparisons. There are only three instances of pathways that are significant (p-value < 0.05) in both the DASGs and DEGs for the same comparison. These include Spliceosome and Peroxisome in the 16C vs 25.1C comparison and Purine Metabolism in the 25.1C vs 25.5C comparison as shown in Fig. 4-6.

There is no unique DEG in the Spliceosome (Fig. 4), but two genes which are both differentially expressed as well as differentially alternatively spliced (brown font on yellow background). In the Peroxisome pathway (Fig. 5), several genes are differentially alternatively

spliced, one gene is differentially expressed and three are both DEGs and DASGs. Purine metabolism has multiple genes which are differentially expressed and differential alternative splicing as well (Fig. 6). And finally, the genes encoding for the enzymes DNA-directed RNA polymerase (EC: 2.7.7.6) are both differentially expressed as well as differentially alternatively spliced (Fig. 6).

It appears that pathways are regulated at various steps by both differential expression and differential alternative splicing (Fig. 4-6). If we only look at one of these two mechanisms independently, we lose the information about the regulatory impact of the other mechanism. However, using both differential expression and differential alternative splicing can help us study the biological regulation at a finer resolution.

### C. Cell Cycle And Division Associated Genes Are Differentially Expressed Under Heat Stress

Gene set functional enrichment analysis serves the dual purposes of verifying the functional relevance of the genes in the experimental condition and to discover unanticipated shared function between these genes. We perform gene ontology enrichment after multiple testing correction using Holm-Bonferroni at p-value < 0.05 for the DASGs and DEGs for all six comparisons using Araport's ThaleMine (Krishnakumar et al., 2015). The first revealing observation was that there are far more significant ontology terms in the DEGs than the DASGs, despite DASGs being far greater in number. One of the many reasons to help explain this large bias could be the much more prevalent analysis of DEGs than DASGs. Alternatively, it is also possible that these DASGs don't mutually regulate biological processes.

The most common enriched biological processes in the DEGs (Table 3) include those related to cell cycle and division, and transport (water, fluid, and polyol etc.). A similar effect

of heat stress has also been reported in apple (Malus domestica) fruitlets where several core cell-cycle and cell-expansion genes are differentially expressed (Flaishman et al., 2015). The biological processes enriched in the DASGs (Table 3) are mostly different than those enriched in DEGs. Some common enriched biological processes in the DASGs include processes related to splicing, circadian rhythm, and metabolic processes. This enrichment of biological processes in the DASGs is similar to the enriched pathways observed in the DASGs.

The DEGs and DASGs do not share common enriched molecular functions (Table 4). The most significant molecular functions in the DEGs are associated with cell cycle and division. Apart from cell cycle and division related terms other enriched functions in the DEGs include water channel activity, histone kinase activity, and glycerol channel activity etc. Again, very few significant molecular function terms are found in the DASGs. Some common terms include small molecule binding, nucleotide binding, and nucleoside phosphate binding (Table 4).

Both DEGs and DASGs are part of the biological regulatory machinery but their target biological processes and molecular functions seem different. These differences in the enriched gene ontology terms suggest differing ways by which DEGs and DASGs modulate stress response. Using both DEGs and DASGs gives additional insights into the complex regulatory processes that might be missed when using only DEGs or DASGs.

### D. Isoforms Of Differentially Alternatively Spliced Genes Have Different Domain Architectures

Alternative splicing can lead to gain or loss of gene function under different conditions (Black, 2003; Staiger & Brown, 2013). By analyzing the alternative isoforms of a gene, we can get a better understanding of the underlying mechanisms of the gene functions. It is

therefore important to consider the contribution of these alternatively spliced isoforms when studying the regulation of biological pathways in addition to DEGs.

We perform domain analysis for all DASGs associated with the spliceosome pathway from the 16C vs 25.1C comparison to gain more insights about the mechanisms by which the DASGs regulate biological pathways. While most isoforms of these DASGs have same domain structure, there are few genes whose isoforms have different domains. Three genes, AT5G52040, AT1G20920 and AT2G35340 produce splice isoforms with at least one domain that is different from the annotated proteins.

Most notable of these is the protein produced by AT2G35340.2 splice isoform that lacks Smc (COG1196: ATPases that help in cell cycle control, cell division and chromosome partitioning); YL1 superfamily (PF05764: DNA-binding and a possible transcription factor (Horikawa, Tanaka, Yuasa, Suzuki, & Oshimura, 1995)); MAP7 (pfam05672: microtubule-stabilizing protein) and Cwf_Cwc_15 (pfam04889: part of the spliceosome and potentially involved in mRNA splicing). Using CDD we predicted that AT2G35340.2 splice isoform has domains (pfam11600: chromatin assembly factor and TIGR01622: splicing factor) that are not found in the abundant AT2G35340.1 isoform.

## Discussion

Alternate usage of exons and introns alters the protein amino acid sequences and functional domains affecting protein function. Additionally, alternative isoforms can have different RNA structures that can further have regulatory implications in its decay and translation. Several pathways such as Spliceosome, Peroxisome and Purine Metabolism are affected by these differential alternative splicing and expression events. The importance of such pathways in response to stress has been reported in the literature (Corpas, Barroso, & Del

Río, 2001; Staiger & Brown, 2013). This suggests that in-silico extraction of such differential networks that include analyzing both DEGs and DASGs, can infer a better understanding of the regulatory machinery. The sets of DEGs and DASGs have different biological process, cellular components and molecular function suggesting a global stress response system.

While we speculate that these DASGs affect the pathways, it is unknown whether these alternate isoforms completely switch off the pathways or lead to an alternate pathway. Another possibility is that the pathway is up- or down-regulated leading to metabolic fluctuations. By integrating this piece of the regulatory machinery with the information from differential expression, we can extract important metabolic pathways which aren't significant using only DEGs. A logical next step is to investigate the effect of changes in AS pattern on the metabolic networks. Biochemical and functional assays investigating the role of the different isoforms of these genes in response to stress are required for a detailed knowledge about their function.

## Conclusions

By using differential splicing in addition to differential expression, we are able to better infer metabolic pathways significantly altered by temperature stress in Arabidopsis thaliana. Differential splicing and differential expression provide differing yet complementary information about regulation of biological processes. Changes in splicing patterns and expression profiles are both essential for modulating stress responses and by using them together we can learn more about the underlying biology of stress response and tolerance.

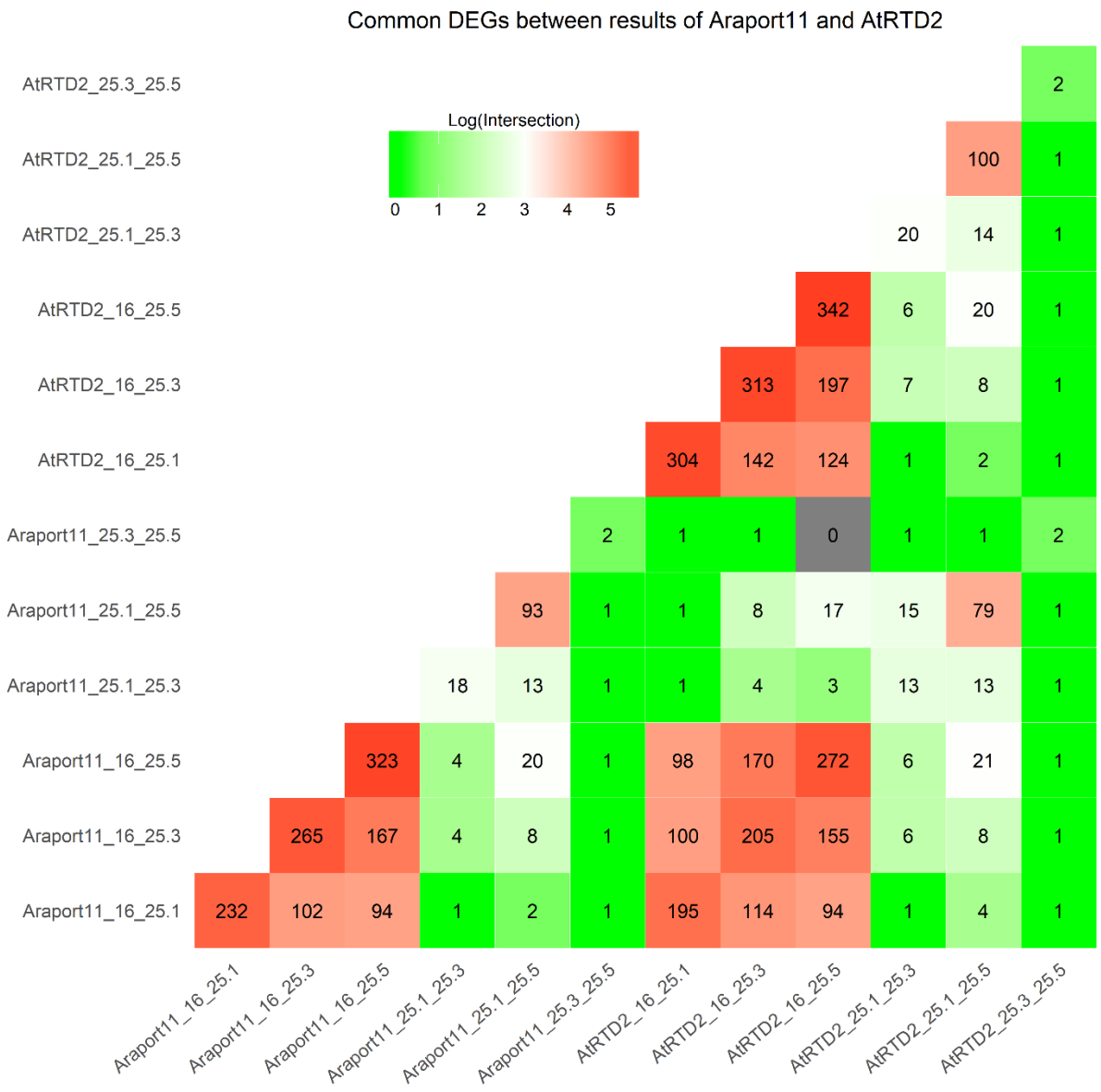


*Figure 2.1* ***Summary of differentially expressed genes****: Number of common differentially expressed genes (DEGs) from Araport11 and AtRTD2. The diagonal represents total genes predicted for the comparison and the color is based on the natural log of the common genes (intersection).*

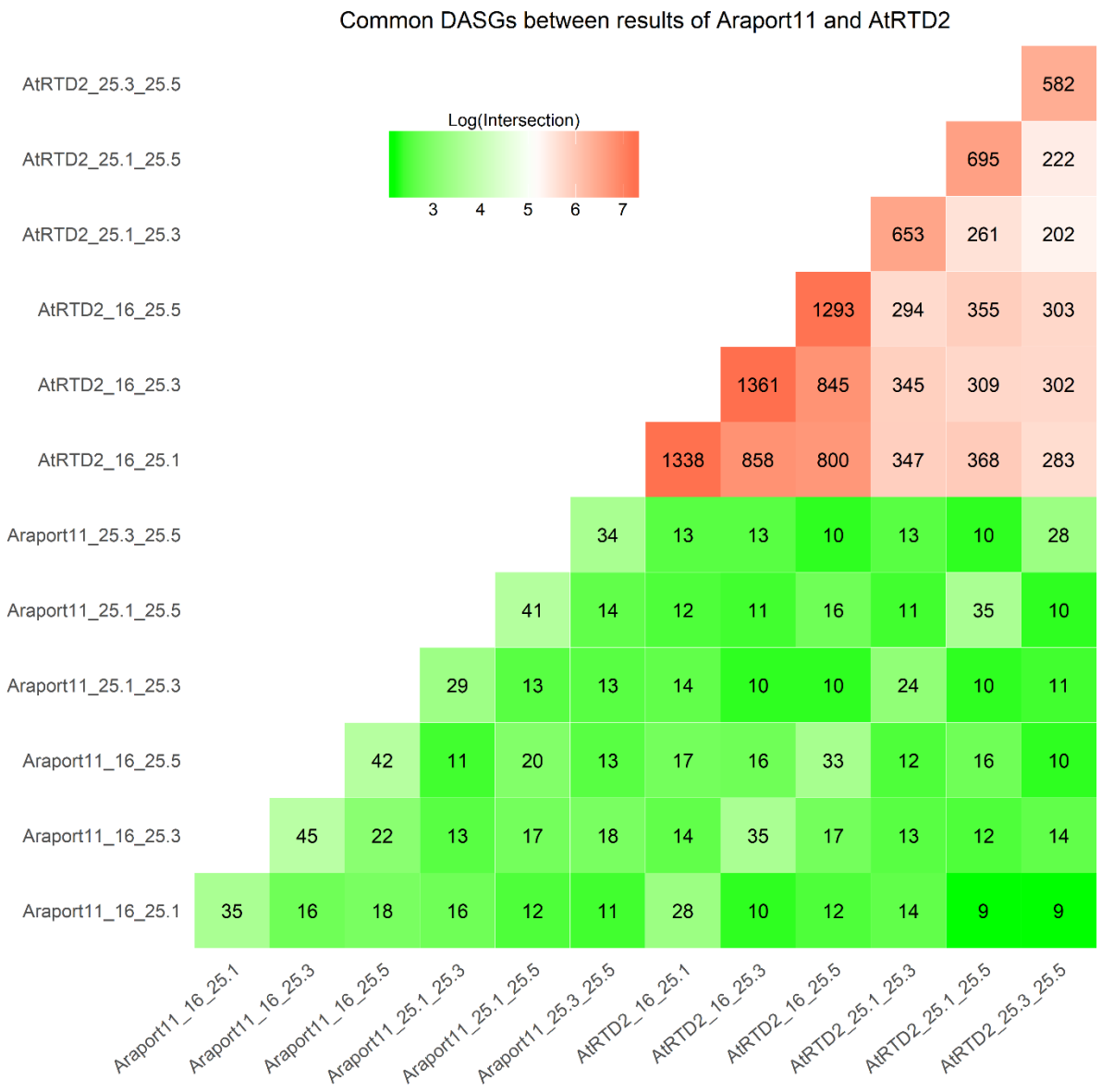


*Figure 2.2* ***Summary of differentially alternatively spliced genes****: Number of common differentially alternatively spliced genes (DASGs) from Araport11 and AtRTD2. The diagonal represents total genes predicted for the comparison and the color is based on the natural log of the common genes (intersection).*

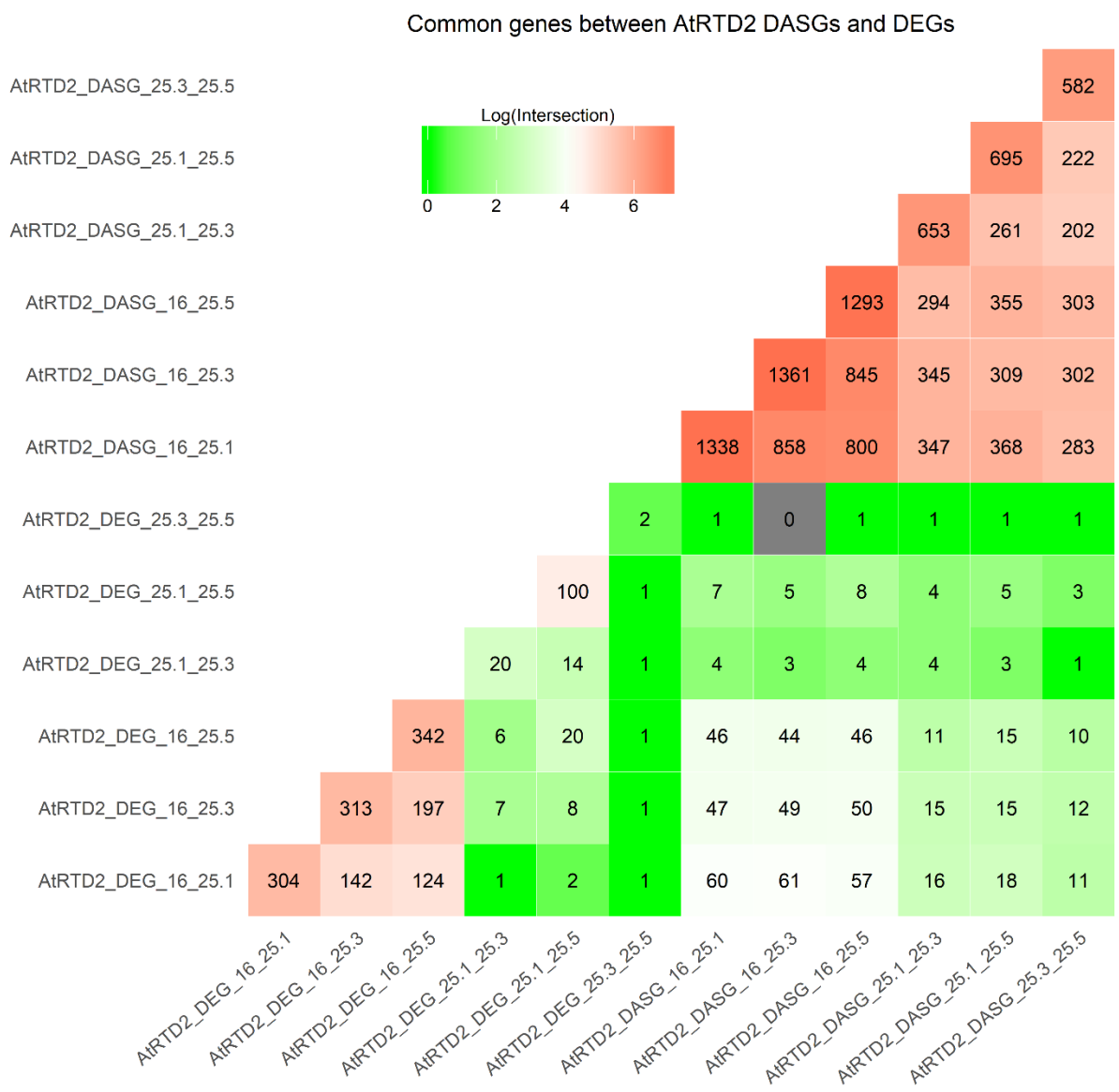


*Figure 2.3* ***Summary of differential genes****: Number of common differentially expressed and differentially alternatively spliced genes from AtRTD2. The diagonal represents total genes predicted for the comparison and the color is based on the natural log of the common genes (intersection).*

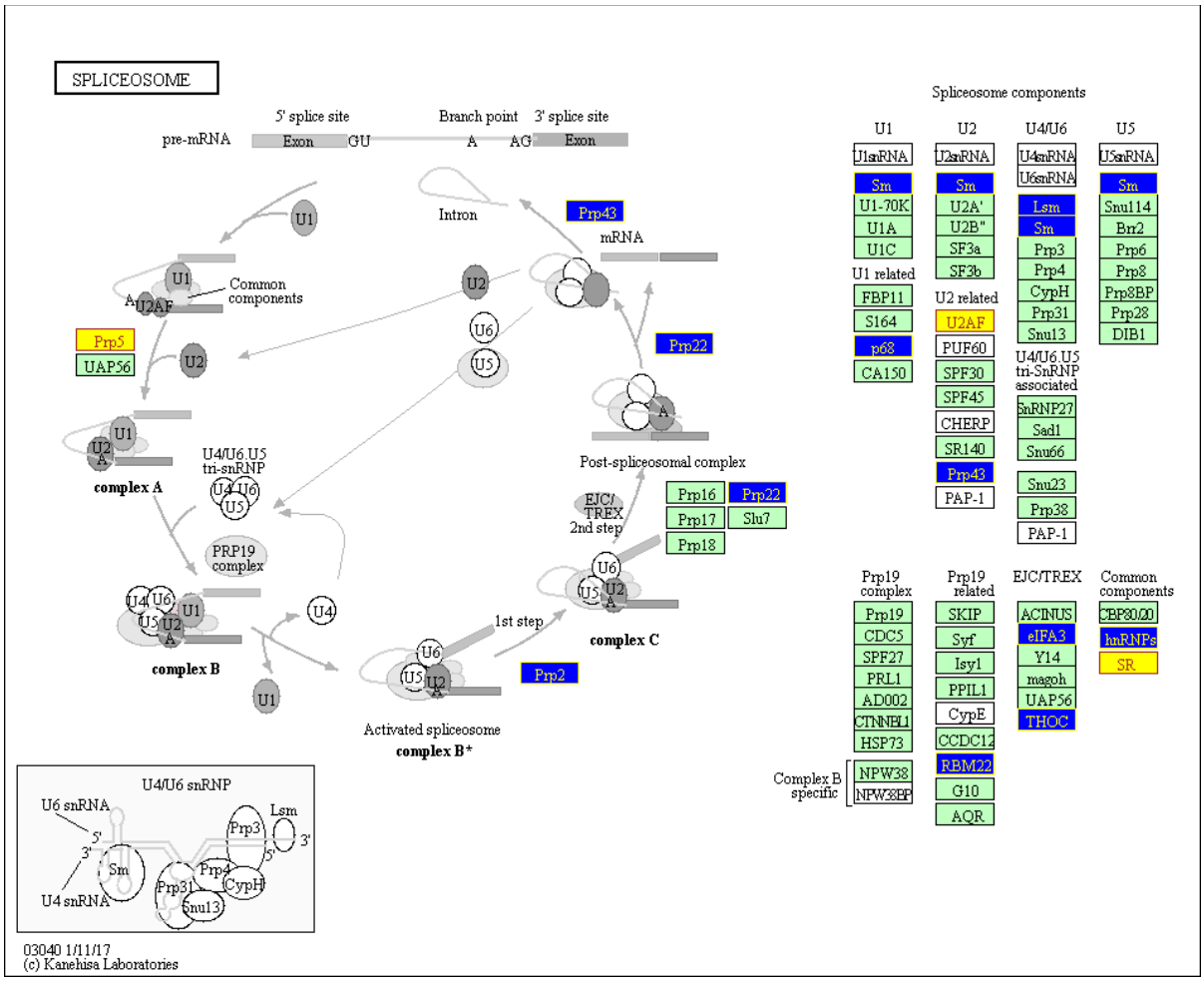


*Figure 2.4* ***Spliceosome pathway****: Differentially alternatively spliced genes (blue) and differentially expressed genes mapped to the spliceosome pathway from the 16C vs 25.1C case. Genes in yellow are both differentially expressed and differentially alternatively spliced.*

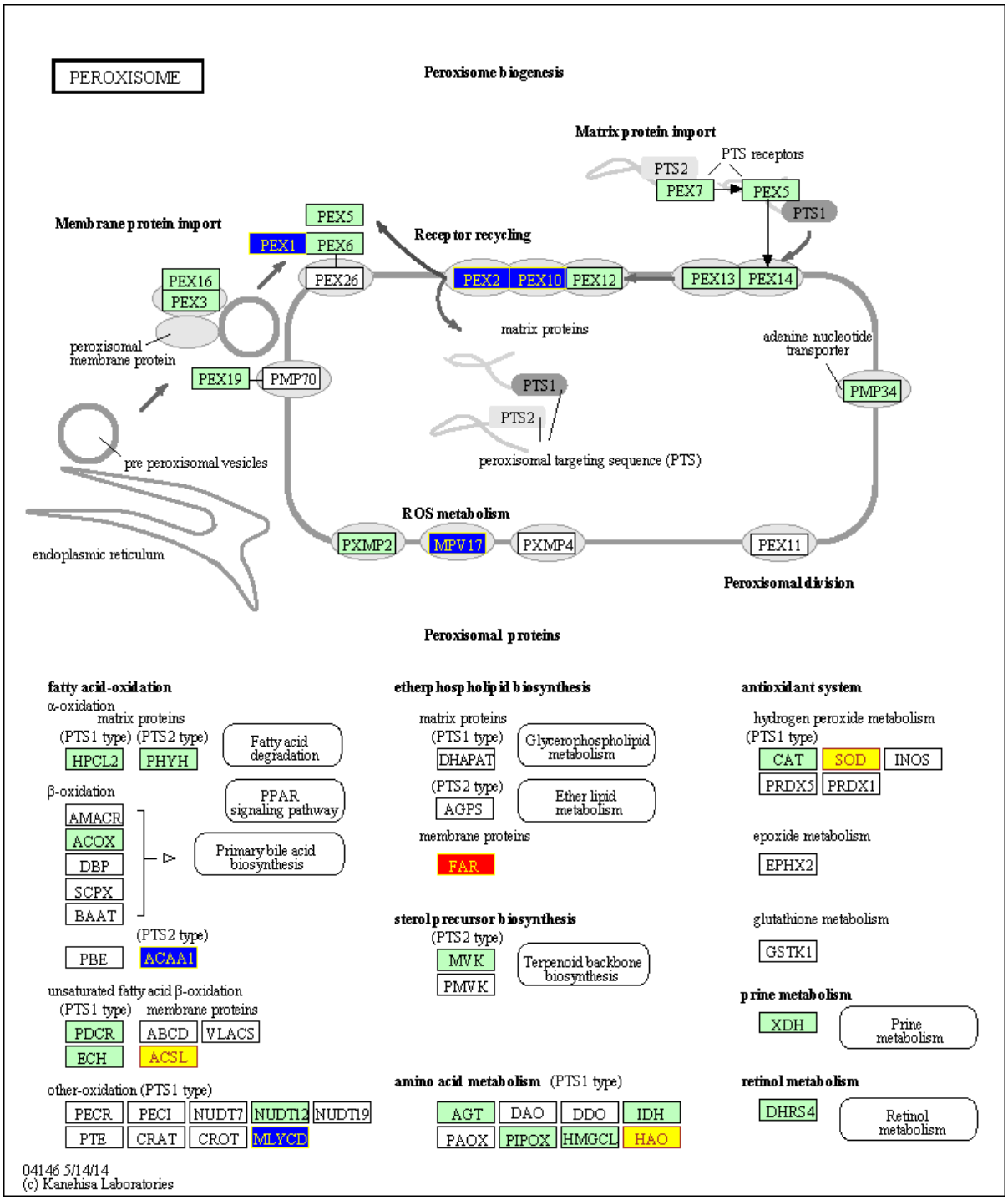


*Figure 2.5* ***Peroxisome pathway****: Differentially alternatively spliced genes (blue) and differentially expressed genes (red) mapped to the peroxisome pathway from the 16C vs 25.1C case. Genes in yellow are both differentially expressed and differentially alternatively spliced.*

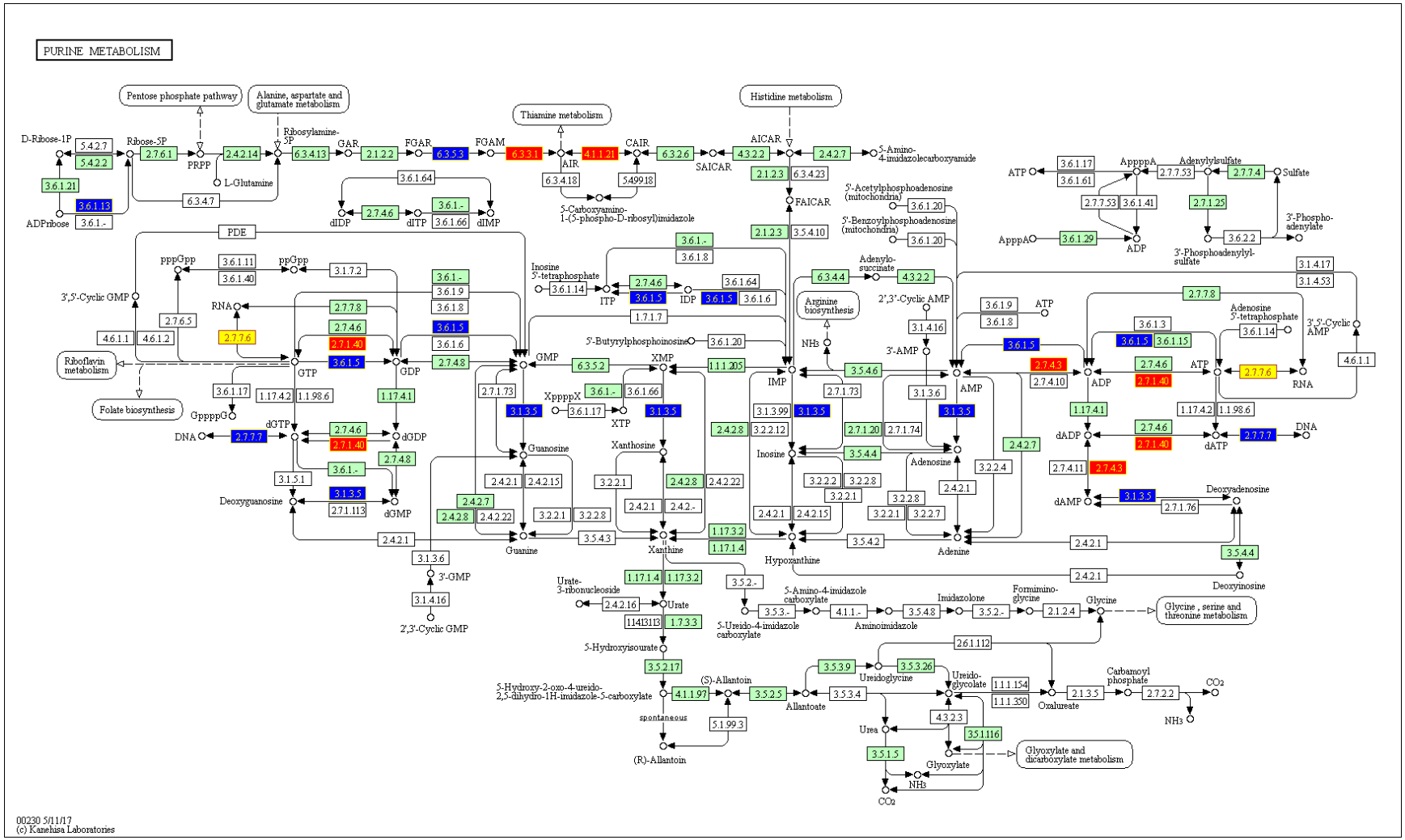


*Figure 2.6* ***Purine metabolism pathway****: Differentially alternatively spliced genes (blue) and differentially expressed genes (red) mapped to the purine metabolism pathway from the 25.1C vs 25.5C case. Enzymes in yellow are encoded by genes found to be both differentially expressed and differentially alternatively spliced.*

*Table 2.1 Summary of significant differentially alternatively spliced and expressed genes*

| **Dataset** | **Araport11** | **AtRTD2** |
|---|---|---|
| **DASGs** | | |
| **16C vs 25.1C** | 35 (43) | 1338 (2014) |
| **16C vs 25.3C** | 45 (55) | 1361 (2091) |
| **16C vs 25.5C** | 42 (47) | 1293 (2005) |
| **25.1C vs 25.3C** | 29 (33) | 653 (803) |
| **25.1C vs 25.5C** | 41 (45) | 695 (872) |
| **25.3C vs 25.5C** | 34 (41) | 582 (695) |
| **DEGs** | | |
| **16C vs 25.1C** | 232 (239) | 304 (324) |
| **16C vs 25.3C** | 265 (273) | 313 (322) |
| **16C vs 25.5C** | 323 (330) | 342 (348) |
| **25.1C vs 25.3C** | 18 (18) | 20 (20) |
| **25.1C vs 25.5C** | 93 (93) | 100 (106) |
| **25.3C vs 25.5C** | 2 (2) | 2 (2) |
| **The numbers within the parenthesis denote the number of transcripts/events while that outside are genes.** | | |

*Table 2.2 KEGG Pathways enriched in the differentially alternatively spliced genes (DASGs) and differentially expressed genes (DEGs) in AtRTD2*

| 16C vs 25.1C | 16C vs 25.3C | 16C vs 25.5C | 25.1C vs 25.3C | 25.1C vs 25.5C | 25.3C vs 25.5C |
|---|---|---|---|---|---|
| **DASG** | | | | | |
| **Spliceosome [03040]** | **Spliceosome [03040]** | **Spliceosome [03040]** | Pantothenate and CoA biosynthesis [00770] | Pyrimidine metabolism [00240] | RNA polymerase [03020] |
| Folate biosynthesis [00790] | **Folate biosynthesis [00790]** | **Folate biosynthesis [00790]** | RNA polymerase [03020] | Folate biosynthesis [00790] | Pyrimidine metabolism [00240] |
| Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | **Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563]** | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | Ubiquinone and other terpenoid-quinone biosynthesis [00130] |
| Ubiquinone and other terpenoid-quinone biosynthesis [00130] | Ubiquinone and other terpenoid-quinone biosynthesis [00130] | Basal transcription factors [03022] | Glycosaminoglycan degradation [00531] | Homologous recombination [03440] | Purine metabolism [00230] |
| Circadian rhythm - plant [04712] | Circadian rhythm - plant [04712] | **Circadian rhythm - plant [04712]** | Riboflavin metabolism [00740] | Nicotinate and nicotinamide metabolism [00760] | Insulin resistance [04931] |
| Sulfur relay system [04122] | Sulfur relay system [04122] | Sulfur relay system [04122] | Sulfur relay system [04122] | Lysine degradation [00310] | |
| Peroxisome [04146] | Peroxisome [04146] | Peroxisome [04146] | Purine metabolism [00230] | Purine metabolism [00230] | |
| | | | | | |
| **16C vs 25.1C** | **16C vs 25.3C** | **16C vs 25.5C** | **25.1C vs 25.3C** | **25.1C vs 25.5C** | **25.3C vs 25.5C** |
| **DEG** | | | | | |
| Peroxisome [04146] | Cutin, suberine and wax biosynthesis | Cutin, suberine and wax biosynthesis [00073] | Ribosome biogenesis in eukaryotes [03008] | **Ribosome biogenesis in eukaryotes [03008]** | |
| Spliceosome [03040] | | Limonene and pinene degradation [00903] | Ribosome [03010] | Ribosome [03010] | |
| Ubiquitin mediated proteolysis [04120] | | Stilbenoid, diarylheptanoid and gingerol biosynthesis [00945] | | Purine metabolism [00230] | |
| Fatty acid metabolism [01212] | | Galactose metabolism [00052] | | | |

*Table 2.2 (continued)*

| 16C vs 25.1C | 16C vs 25.3C | 16C vs 25.5C | 25.1C vs 25.3C | 25.1C vs 25.5C | 25.3C vs 25.5C |
|---|---|---|---|---|---|
| **DASG + DEG** | | | | | |
| **Spliceosome [03040]** | **Spliceosome [03040]** | **Folate biosynthesis [00790]** | Glycosaminoglycan degradation [00531] | **Ribosome biogenesis in eukaryotes [03008]** | RNA polymerase [03020] |
| Peroxisome [04146] | **Folate biosynthesis [00790]** | Spliceosome [03040] | Pantothenate and CoA biosynthesis [00770] | Pyrimidine metabolism [00240] | Pyrimidine metabolism [00240] |
| Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | Circadian rhythm - plant [04712] | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | Ubiquinone and other terpenoid-quinone biosynthesis [00130] |
| Folate biosynthesis [00790] | Peroxisome [04146] | Sulfur relay system [04122] | Riboflavin metabolism [00740] | Purine metabolism [00230] | Purine metabolism [00230] |
| Ubiquinone and other terpenoid-quinone biosynthesis [00130] | Ubiquinone and other terpenoid-quinone biosynthesis [00130] | Glycosylphosphatidylinositol(GPI)-anchor biosynthesis [00563] | RNA polymerase [03020] | Homologous recombination [03440] | Insulin resistance [04931] |
| Circadian rhythm - plant [04712] | Sulfur relay system [04122] | Peroxisome [04146] | Nucleotide excision repair [03420] | Folate biosynthesis [00790] | |
| Sulfur relay system [04122] | Circadian rhythm - plant [04712] | Basal transcription factors [03022] | | Nicotinate and nicotinamide metabolism [00760] | |
| **The KEGG pathways enriched after multiple testing correction using Holm-Bonferroni at p-value < 0.05 are marked in bold. The number in brackets represents the KEGG pathway ID. The analysis was performed on ThaleMine at Araport.** | | | | | |

*Table 2.3 Top 7 significant biological process terms*

| 16C vs 25.1C | 16C vs 25.3C | 16C vs 25.5C | 25.1C vs 25.3C | 25.1C vs 25.5C |
|---|---|---|---|---|
| **DASG** | | | | |
| heterocycle metabolic process | heterocycle metabolic process | heterocycle metabolic process | regulation of circadian rhythm | |
| organic cyclic compound metabolic process | organic cyclic compound metabolic process | organic cyclic compound metabolic process | phosphorus metabolic process | |
| nucleobase-containing compound metabolic process | nucleobase-containing compound metabolic process | circadian rhythm | phosphate-containing compound metabolic process | |
| cellular aromatic compound metabolic process | nucleic acid metabolic process | nucleic acid metabolic process | | |
| vegetative to reproductive phase transition of meristem | RNA splicing | RNA splicing | | |
| circadian rhythm | mRNA processing | mRNA processing | | |
| | mRNA metabolic process | regulation of circadian rhythm | | |
| | | | | |
| **16C vs 25.1C** | **16C vs 25.3C** | **16C vs 25.5C** | **25.1C vs 25.3C** | **25.1C vs 25.5C** |
| **DEG** | | | | |
| cell cycle | cell cycle | cell cycle | | ribosome biogenesis |
| cell cycle process | mitotic cell cycle | cell cycle process | | rRNA methylation |
| mitotic cell cycle | cell cycle process | mitotic cell cycle | | rRNA processing |
| mitotic cell cycle process | mitotic cell cycle process | mitotic cell cycle process | | rRNA metabolic process |
| regulation of cell cycle process | microtubule-based movement | microtubule-based movement | | cellular component biogenesis |
| regulation of cell cycle | microtubule-based process | cell division | | ncRNA processing |
| cellular water homeostasis | cell division | nuclear division | | ncRNA metabolic process |

**Top 7 significant biological process terms enriched in the differentially alternatively spliced genes (DASGs) and differentially expressed genes (DEGs) after multiple testing correction using Holm-Bonferroni at p-value < 0.05. No significant term was found in the 25.3C vs 25.5C comparison. The analysis was performed on ThaleMine at Araport.**

*Table 2.4 Top 7 significant molecular function terms*

| 16C vs 25.1C | 16C vs 25.3C | 16C vs 25.5C | 25.1C vs 25.3C | 25.1C vs 25.5C |
|---|---|---|---|---|
| **DASG** | | | | |
| nucleoside phosphate binding | nucleoside phosphate binding | | transferase activity, transferring phosphorus-containing groups | small molecule binding |
| nucleotide binding | nucleotide binding | | | nucleotide binding |
| small molecule binding | small molecule binding | | | nucleoside phosphate binding |
| | | | | ribonucleotide binding |
| | | | | purine ribonucleotide binding |
| | | | | purine nucleotide binding |
| | | | | carbohydrate derivative binding |
| **16C vs 25.1C** | **16C vs 25.3C** | **16C vs 25.5C** | **25.1C vs 25.3C** | **25.1C vs 25.5C** |
| **DEG** | | | | |
| glycerol channel activity | microtubule binding | microtubule motor activity | | RNA methyltransferase activity |
| glycerol transmembrane transporter activity | microtubule motor activity | glycerol transmembrane transporter activity | | RNA binding |
| organic hydroxy compound transmembrane transporter activity | ATP-dependent microtubule motor activity | cyclin-dependent protein kinase activity | | rRNA methyltransferase activity |
| water channel activity | tubulin binding | tubulin binding | | methyltransferase activity |
| polyol transmembrane transporter activity | cytoskeletal protein binding | cyclin-dependent protein serine/threonine kinase activity | | S-adenosylmethionine-dependent methyltransferase activity |
| alcohol transmembrane transporter activity | histone kinase activity | histone kinase activity | | transferase activity, transferring one-carbon groups |
| water transmembrane transporter activity | motor activity | glycerol channel activity | | |

**Top 7 significant molecular function terms enriched in the differentially alternatively spliced genes (DASGs) and differentially expressed genes (DEGs) after multiple testing correction using Holm-Bonferroni at p-value < 0.05. No significant term was found in the 25.3C vs 25.5C comparison. The analysis was performed on ThaleMine at Araport.**

## References

Black, D. L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. Annual Review of Biochemistry, 72(1), 291–336. https://doi.org/10.1146/annurev.biochem.72.121801.161720

Blencowe, B. J. (2006, July). Alternative Splicing: New Insights from Global Analyses. Cell. https://doi.org/10.1016/j.cell.2006.06.023

Celotto, A. M., & Graveley, B. R. (2001). Alternative splicing of the Drosophila Dscam pre-mRNA is both temporally and spatially regulated. Genetics, 159(2), 599–608.

Cheng, C. Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant Journal, 89(4), 789–804. https://doi.org/10.1111/tpj.13415

Corpas, F. J., Barroso, J. B., & Del Río, L. A. (2001, April). Peroxisomes as a source of reactive oxygen species and nitric oxide signal molecules in plant cells. Trends in Plant Science. https://doi.org/10.1016/S1360-1385(01)01898-2

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Flaishman, M. A., Peles, Y., Dahan, Y., Milo-Cochavi, S., Frieman, A., & Naor, A. (2015). Differential response of cell-cycle and cell-expansion regulators to heat stress in apple (Malus domestica) fruitlets. Plant Science, 233, 82–94. https://doi.org/10.1016/j.plantsci.2015.01.005

Horikawa, I., Tanaka, H., Yuasa, Y., Suzuki, M., & Oshimura, M. (1995). Molecular cloning of a novel human cDNA on chromosome 1q21 and its mouse homolog encoding a nuclear protein with DNA-binding ability. Biochemical and Biophysical Research Communications, 208(3), 999–1007. https://doi.org/10.1006/bbrc.1995.1433

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews. Genetics, 11(5), 345–355. https://doi.org/10.1038/nrg2776

Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M., … Town, C. D. (2015). Araport: The Arabidopsis Information Portal. Nucleic Acids Research, 43(D1), D1003–D1009. https://doi.org/10.1093/nar/gku1200

Mach, J. (2009). Alternative splicing produces a JAZ protein that is not broken down in response to jasmonic acid. The Plant Cell, 21(1), 14. https://doi.org/10.1105/tpc.108.210111

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., … Bryant, S. H. (2015). CDD: NCBI's conserved domain database. Nucleic Acids Research, 43(D1), D222–D226. https://doi.org/10.1093/nar/gku1221

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Research, 22(6), 1184–1195. https://doi.org/10.1101/gr.134106.111

Pajoro, A., Severing, E., Angenent, G. C., & Immink, R. G. H. (2017). Histone H3 lysine 36 methylation affects temperature-induced alternative splicing and flowering in plants. Genome Biology, 18(1), 102. https://doi.org/10.1186/s13059-017-1235-x

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics, 40(12), 1413–1415. https://doi.org/10.1038/ng.259

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., … Blencowe, B. J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Molecular Cell, 16(6), 929–941. https://doi.org/10.1016/j.molcel.2004.12.004

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protocols, 11(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

Reddy, A. S. N., Rogers, M. F., Richardson, D. N., Hamilton, M., & Ben-Hur, A. (2012). Deciphering the Plant Splicing Code: Experimental and Computational Approaches for Predicting Alternative Splicing and Splicing Regulatory Elements. Frontiers in Plant Science, 3, 18. https://doi.org/10.3389/fpls.2012.00018

Shen, S., Park, J. W., Lu, Z., Lin, L., Henry, M. D., Wu, Y. N., … Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences of the United States of America, 111(51), E5593-601. https://doi.org/10.1073/pnas.1419161111

Sheng, Z., Sun, Y., Zhu, R., Jiao, N., Tang, K., Cao, Z., & Ma, C. (2015). Functional cross-talking between differentially expressed and alternatively spliced genes in human liver cancer cells treated with berberine. PLoS ONE, 10(11), e0143742. https://doi.org/10.1371/journal.pone.0143742

Simon Andrews. (2010). FastQC: a quality control tool for high throughput sequence data. Http://Www . Bioinformatics . Babraham . Ac . Uk/Projects/ Fastqc. https://doi.org/citeulike-article-id:11583827

Staiger, D., & Brown, J. W. S. (2013). Alternative Splicing at the Intersection of Biological Timing, Development, and Stress Responses. The Plant Cell, 25(10), 3640–3656. https://doi.org/10.1105/tpc.113.113803

Sugliani, M., Brambilla, V., Clerkx, E. J. M., Koornneef, M., & Soppe, W. J. J. (2010). The conserved splicing factor SUA controls alternative splicing of the developmental regulator ABI3 in Arabidopsis. The Plant Cell, 22(6), 1936–1946. https://doi.org/10.1105/tpc.110.074674

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., & Brown, J. W. S. (2012, October). Alternative splicing in plants - coming of age. Trends in Plant Science. https://doi.org/10.1016/j.tplants.2012.06.001

Zhang, R., Calixto, C. P. G., Marquez, Y., Venhuizen, P., Tzioutziou, N. A., Guo, W., … Brown, J. W. S. (2017). A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic Acids Research, 45(10), 5061–5073. https://doi.org/10.1093/nar/gkx267

# CHAPTER 3. TISSUE-SPECIFIC MOUSE MRNA ISOFORM NETWORKS

Modified from a manuscript under review at Scientific Reports

Gaurav Kandoi, and Julie A. Dickerson

## Author's contributions

GK leads this study. GK and JAD contribute to the design of the study and the interpretation of the results. GK and JAD together wrote the manuscript. GK wrote the programs and performed data analysis. All the authors read and approved the final manuscript.

## Abstract

Alternative Splicing produces multiple mRNA isoforms of genes which have important diverse roles such as regulation of gene expression, human heritable diseases, and response to environmental stresses. However, little has been done to assign functions at the mRNA isoform level. Functional networks, where the interactions are quantified by their probability of being involved in the same biological process are typically generated at the gene level. We use a diverse array of tissue-specific RNA-seq datasets and sequence information to train random forest models that predict the functional networks. Since there is no mRNA isoform-level gold standard, we use single isoform genes co-annotated to Gene Ontology biological process annotations, Kyoto Encyclopedia of Genes and Genomes pathways, BioCyc pathways and protein-protein interactions as functionally related (positive pair). To generate the non-functional pairs (negative pair), we use the Gene Ontology annotations tagged with "NOT" qualifier. We describe 17 Tissue-spEcific mrNa iSoform functIOnal Networks (TENSION) following a leave-one-tissue-out strategy in addition to an organism level reference functional

network for mouse. We validate our predictions by comparing its performance with previous methods, randomized positive and negative class labels, updated Gene Ontology annotations, and by literature evidence. We demonstrate the ability of our networks to reveal tissue-specific functional differences of the isoforms of the same genes. All scripts and data from TENSION are available at: https://doi.org/10.25380/iastate.c.4275191

## Introduction

Recent studies illustrate that genes can have distinct functions with different mRNA isoforms, highlighting the importance of studying mRNA isoforms of a gene (Chen & Crowther, 2012; H. D. Li, Menon, Omenn, & Guan, 2014). This diversity in mRNA isoforms is a result of Alternative Splicing (AS). Many alternatively spliced mRNA isoforms are variably expressed across cell and tissue types (Buljan et al., 2012; Ellis et al., 2012; Raj & Blencowe, 2015; Sun et al., 2018; Vitulo et al., 2014; Wei & Jin, 2016; Wu et al., 2018; Xu, Modrek, & Lee, 2002). AS affects regulation of gene expression, development, human heritable diseases, and response to environmental stresses. This article builds mouse tissue-specific functional networks by integrating heterogeneous expression and sequence datasets at the mRNA isoform level.

In higher organisms such as mouse and human, AS plays a significant role in expanding the variety of protein species (Kelemen et al., 2013; Resch et al., 2004; Suzuki et al., 2011; Yura et al., 2006). As an effect, a gene may produce multiple mRNA isoforms whose protein translations differ in expression, interaction and function (Ellis et al., 2012; Kelemen et al., 2013; H. D. Li et al., 2014; Pan et al., 2004). For example, there are more than 75,000 mRNA isoforms encoded by over 20,000 genes in the Mouse genome annotation (GRCm38.p4). The

fact that a gene is a mixture of mRNA isoforms makes referencing a gene as being "upregulated" or "downregulated", uninformative.

Massively parallel sequencing of mRNA isoforms has led to a rapid accumulation of expression and sequence data at the mRNA isoform level. RNA-Seq has provided evidence confirming the production and expression of distinct mRNA isoforms under different conditions (Marquez et al., 2012; Pan et al., 2008; Raj & Blencowe, 2015). This has led to the improvement and refinement of genome annotations. Functional networks, at the mRNA isoform level are important for understanding gene function but are largely uninvestigated (H.-D. D. Li et al., 2016; Tseng et al., 2015).

Traditionally, functional experiments are performed at the gene level. Therefore, there are very few (few hundreds) functional annotations for alternatively spliced mRNA isoforms. The functional data recorded in databases such as Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), and UniProt Gene Ontology Annotations (UniProt-GOA) are focused on the canonical mRNA isoform and contain only few hundred annotations describing the functions of alternatively spliced mRNA isoforms. These databases do not store tissue specific information either.

The task of mRNA isoform function prediction is a challenging problem. Some mRNA isoforms are non-functional and introduce noise in the data. Many mRNA isoforms are tissue and condition specific. Since a gene can produce multiple mRNA isoforms (Liu, Loraine, & Dickerson, 2014), the direct transfer of function from the gene to its mRNA isoforms doesn't work. Gene function prediction methods consider a gene as a single entity. Therefore, these cannot be directly applied to mRNA isoform function prediction because they ignore the distinct functions of alternatively spliced mRNA isoforms. However, important advancements

have been made by recent studies towards mRNA isoform level understanding of gene functions (Eksi et al., 2013; H.-D. D. Li et al., 2016; H. D. Li et al., 2015; W. Li et al., 2014; Luo et al., 2017; Panwar et al., 2016; Tseng et al., 2015) such as the prediction of more immune related gene ontology terms for the mRNA isoform ADAM15B than isoform ADAM15A of ADAM15 gene, which is involved in B-cell-mediated immune mechanisms.

One such study developed the human isoform-isoform interactions database (IIIDB) using RNA-Seq datasets, domain-domain interactions and protein-protein interactions (PPIs) (Tseng et al., 2015). A logistic regression model was built using physical interaction data from the IntAct database (Orchard et al., 2014). The predicted human isoform-isoform physical interaction network was restricted to the gene pairs already present in IntAct. The problem of mRNA isoform functional network prediction is formulated as a complex multiple instance learning (MIL) problem in (H.-D. D. Li et al., 2016). In MIL, a gene is treated as a "bag" of mRNA isoforms ("instances"). A gene pair is formulated as a bag of multiple instance pairs, each of which has different probabilities to be functionally related. The goal of MIL is to identify the specific instance pairs which are functional and maximize the difference between them and the instance pairs of non-functionally related bags. A Bayesian network based MIL algorithm was developed by (H.-D. D. Li et al., 2016) to predict a mouse mRNA isoform level functional network using RNA-Seq datasets, exon array, pseudo-amino acid composition and isoform-docking data.

The studies (H.-D. D. Li et al., 2016; Tseng et al., 2015) above introduce bias in the training and testing dataset by using random mRNA isoform pairs as non-functional pairs (negative pairs) and do not consider the tissue-specific mRNA isoform functions. Our work is fundamentally different and improves upon the studies (H.-D. D. Li et al., 2016; Tseng et al.,

2015) above both in terms of research content and computational approaches. First, we formulate the problem of mRNA isoform functional network prediction as a simple supervised learning task. Second, our goal is to develop tissue-specific functional networks for mouse. Lastly, like previous methods, we do not introduce bias by assuming that functionally unrelated (negative pair) mRNA isoform pairs can be selected based on the cellular localization (Tseng et al., 2015) or at random (H.-D. D. Li et al., 2016), which is crucial to the selection of training data in a machine learning system.

We have developed 17 tissue-specific mRNA isoform level functional networks in addition to an organism level reference functional network for mouse. Using the leave-one-tissue-out strategy with a diverse array of tissue-specific RNA-Seq datasets and sequence information, we trained a random forest model to predict the functional networks. Because there is no mRNA isoform-level gold standard for testing, we have used the single mRNA isoform genes co-annotated to GO biological process, KEGG pathways, BioCyc pathways and PPIs as functionally related (positive pair). The non-functional pairs (negative pairs) were generated by using the GO annotations tagged with "NOT" qualifier. We have validated our predictions by comparing its performance with previous methods, datasets with randomized positive and negative class labels, updated GO annotations and literature evidence.

## Methods

### mRNA isoform level data processing

This study considers mRNA isoforms annotated in the NCBI *Mus musculus* genome assembly (GRCm38.p4) for which both mRNA and protein sequences are available. All protein (and corresponding mRNA) sequences smaller than 30 amino acids and those

containing non-standard characters are not considered. This resulted in a filtered set of 75,826 mRNA isoforms from 21,813 genes.

To comprehensively characterize mRNA isoform pairs, we have processed 359 RNA-Seq samples from 17 tissues and calculated protein and mRNA sequence properties as described below. Such heterogeneous features have been shown to be useful for predicting several biological properties (Du, Hu, Yao, Sun, & Zhang, 2017; Kandoi, Acencio, & Lemke, 2015; H.-D. D. Li et al., 2016). All calculations and analyses were performed on the Extreme Science and Engineering Discovery Environment (XSEDE) Comet cluster (Towns et al., 2014).

The mRNA and protein level features are summarized in Table 1 and an overview of the workflow is presented in Fig 1. Every feature type resulted in 1 feature (as described in the following sections).

**Preprocessing of RNA-seq datasets.** To capture tissue specific functions, RNA-Seq datasets from multiple tissues are processed to extract the expression values. Starting with the ENCODE mouse RNA-Seq datasets, the following filtering criteria are used to select the datasets: 1) Read length ≥ 50; 2) Mapping percent ≥ 70%; and 3) No error or audit warning flags were generated. For the tissue specific networks, only those tissues with at least 10 samples were used. Based on these filters we retained 359 RNA-Seq samples from around 20 tissues, 17 of which have at least 10 samples (Table S1).

The mouse genome build GRCm38.p4 from NCBI was used to align the RNA-Seq datasets using STAR (version 2.5.3a) (Dobin et al., 2013). Then, the relative abundance of the mRNA isoforms as fragments per kilobase of exon per million fragments mapped (FPKM) is

calculated using StringTie (version 1.3.3b) (Pertea et al., 2016). The GFF3 annotation file corresponding to the GRCm38.p4 build was also used during the alignment and quantification.

**mRNA sequence composition.** mRNA sequences can be represented as the frequencies of *k* neighboring nucleic acids, jointly referred to as *k-mers*. For an mRNA sequence there are $\mathbf{4^k}$ possible *k-mers* in a *k-mer group*, while there are $\mathbf{20^k}$ possible *k-mers* for protein sequences. For a sequence of length *l*,

$$f(kmer_i) = \begin{cases} \frac{N_i}{l} & i \in A, T, C, G \\ \frac{N_i}{(l-1)} & i \in AA, AT, \dots, GC, GG \\ \dots & \\ \dots & \\ \dots & \\ \frac{N_i}{(l-(k-1))} & i \in A\{k\}, A\{k-1\}T, \dots, G\{k-1\}C, G\{K\} \end{cases}$$

where, $f(kmer_i)$ is the frequency of the *ith k-mer* and $N_i$ is the count of the *ith k-mer*. We compute the *k-mer* composition for *k = 3 to 6* for all mRNA isoform sequences using the rDNAse library in R (R Core Team, 2017; Zhu, Dong, & Cao, 2016).

**Protein Sequence Properties.** Each protein sequence can be characterized in multiple ways by exploiting its sequence and order composition. Like the mRNA sequence *k*-mer composition described above, we compute the *k*-mer compositions for *k = 1 and 2* for all protein sequences. We also compute the conjoint triad descriptors (J. Shen et al., 2007) for all protein sequences. For this, the standard 20 amino acids are grouped into 7 classes according to the volume of the side chains and their dipoles. Then, the *k-mer* composition is calculated at *k = 3* for this newly represented protein sequence. For *k = 3*, protein sequences can lead to highly sparse 8000 (20*20*20) features as opposed to only 243 (7*7*7) in case of the conjoint-

triad descriptors. The dramatically reduced feature dimension also results in a reduced variance dimension and may also partially overcome the overfitting problem (J. Shen et al., 2007).

To take the sequential information of the amino acids in a protein sequence into account, we also compute the pseudo-amino acid composition (Chou, 2001) and Moran autocorrelation(Moran, 1950) for all protein sequences. The amino acid composition (*k = 1*) does not contain any of its sequence-order information, whereas pseudo-amino acid composition includes additional position-related features (Chou, 2001). Therefore, the pseudo-amino acid composition reflects both sequential as well as compositional order (Chou, 2001). Moran autocorrelation is a type of topological descriptor which measures the level of spatial correlation between two objects (amino acid residues) in terms of their specific physicochemical or structural property.

All protein sequence properties were computed using the protr library in R (R Core Team, 2017; Xiao, Cao, Zhu, & Xu, 2015).

$$Moran\ autocorrelation\ I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \bar{P}')\,(P_{i+d} - \bar{P}')}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \bar{P}')^2} \qquad d = 1,2,\ldots,30$$

where, d is called the lag of the autocorrelation; $P_i$ and $P_{i+d}$ are the properties of the amino acid $i$ and $i + d$; $\bar{P}'$ is the considered property $P$ along the sequence, i.e.,

$$\bar{P}' = \frac{\sum_{i=1}^{N} P_i}{N}$$

**mRNA isoform level feature calculation**

The goal is to accurately predict a functional network which represents the probability of two mRNA isoforms belonging to the same GO biological process or pathway. Lower edge weights correlate with mRNA isoform pairs' involvement in the same GO biological process or pathway. The weighted functional network is modeled as a graph $G = (V, E)$, where the set $V$ represents the mRNA isoforms (nodes) and the set $E$ represents the mRNA isoform pairs (edges). For an mRNA isoform pair $(E_{ij})$ in the functional network, the class label $(L_{ij})$ is assigned as following:

$$L_{ij} = \begin{cases} 1 \ if \ both \ mRNA \ i \ and \ j \ interact, or \ are \ co-annotated \ to \ the \ same \ GO \ biological \ process \ or \ pathway \\ 0 \ otherwise \end{cases}$$

Many mRNA isoforms have zero FPKM values. The FPKM values were corrected by performing log-transformation and a small constant value of 1 was added to all FPKM values, i.e. $\log_2(FPKM + 1)$. The log-transformation is intended to normalize and re-scale the FPKM values. The addition of a small constant value alleviates the problem where the log of zero FPKM value is not defined, which is not an acceptable input for machine learning methods.

For all mRNA isoform pairs, Fisher's z-transformed Pearson correlation scores are calculated and used as input features for machine learning.

$$z = \frac{1}{2}\log_2\frac{1-\rho}{1+\rho}$$

Pearson correlation coefficient of 1 and -1 leads to z-score of -∞ and ∞ respectively, so these z-scores are replaced with an extreme value of -100 and 100 respectively. In cases where the Pearson correlation coefficient is not defined, we set the z-score to 0.

For every mRNA isoform pair, we calculate one z-score using the samples from one tissue and use this as one feature. For instance, one z-score for heart, one z-score for liver, one z-score for lungs and so on for all 17 tissues. Additionally, one z-score is also calculated using all 359 RNA-Seq samples. This resulted in 18 features, one for each of the 17 tissues and one using all RNA-Seq samples. Similarly, for every mRNA isoform pair, we calculate one z-score for each of $k = 3, 4, 5, \text{and } 6$ for mRNA isoform sequences, $k = 1 \text{ and } 2$ for protein sequences, conjoint-triad descriptors, pseudo-amino acid composition and Moran autocorrelation. This led to 9 further features resulting in a total of 27 features.

**mRNA isoform level functional labels**

The mRNA isoform level functional labels are created by combining the information from GO biological process annotations (downloaded on 23 October 2017), KEGG pathways (downloaded on 25 September 2017), BioCyc pathways (downloaded on 25 September 2017) and PPIs (downloaded on 25 September 2017). We remove all GO biological process annotations with the evidence codes: Inferred from Electronic Annotation (IEA), Non-traceable Author Statement (NAS) and No biological Data available (ND). We utilize the GO hierarchy (gene ontology downloaded on 25 October 2017) and propagate all annotations by following the "true path rule", which means that all genes/proteins annotated to a GO term *T* will also be annotated to all ancestor terms of *T*.

The PPIs were integrated from IntAct (Orchard et al., 2014), Biological General Repository for Interaction Datasets (BioGRID) (Chatr-Aryamontri et al., 2017), Agile Protein Interactomes DataServer (APID) (Alonso-López et al., 2016), Integrated Interactions Database (IID) (Kotlyar, Pastrello, Sheahan, & Jurisica, 2016) and Mentha (Calderone, Castagnoli, & Cesareni, 2013). For APID (Alonso-López et al., 2016), we include interactions with at least 2

experimental evidences (level 2 dataset). For IID (Kotlyar et al., 2016), we remove all interactions for which there is only orthologous evidence. For Mentha (Calderone et al., 2013), we remove interactions with a score less than 0.2. Finally, we consider PPIs only if both interactors are from mouse.

After propagation, we remove the GO biological process terms which are too broad (more than 1000 genes annotated) or too specific (less than 10 genes annotated). A gene is assumed to be functional if it is annotated to a GO biological process or a pathway. Two genes are assumed to be functionally related if both are co-annotated to the same GO biological process or pathway. The information in GO, KEGG, BioCyc and PPI databases usually focus on the canonical form of a gene/protein and doesn't distinguish between the mRNA isoforms resulting from AS. The current biological databases do not explicitly differentiate the functions of different mRNA isoforms of the same gene. This unavailability of mRNA isoform level functional information is the cause for having no mRNA isoform level gold standard datasets. To overcome this challenge for building machine learning methods, there are two ways: 1) Randomly assign the functions of a gene to its mRNA isoforms; and 2) Use only single mRNA isoform producing genes. The first approach introduces large bias in the functional datasets while also losing information from the random assignment of function. In the second approach, we lose information from multiple mRNA isoform producing genes in the functional data, but avoid biasing the functional dataset. Because, we do not randomly select unannotated genes for building the non-functional dataset, we still introduce some complementary information from multiple mRNA isoform producing genes in the training and testing datasets. Both ways have their pros and cons, and we believe that although we lose information by using only single

mRNA isoform producing genes as functional pairs, we reduce a lot of false functional labels by not assigning the functions of a gene to its mRNA isoforms randomly.

Therefore, we construct mRNA isoform level functional labels by utilizing the information from single mRNA producing genes and gene annotations tagged with a "NOT" qualifier. A summary of the mRNA isoform level functional label generation is illustrated in Fig 1.

In our functional networks, if a gene $G_1$ produces only a single mRNA $M_1$, then $M_1$ is assumed to perform the functions of $G_1$and is considered functional. Similarly, if two genes $G_1$ $and$ $G_2$, both of which produce single mRNAs, $M_1$ $and$ $M_2$ respectively, are co-annotated to the same GO biological process or pathway, the pair (edge) $M_1 - M_2$ is assumed to be functionally related (positive pair). Additionally, if $G_1$ $and$ $G_2$ are involved in a PPI, the pair (edge) $M_1 - M_2$ is also assumed to be functionally related (positive pair).

We utilize a more robust way of defining functionally unrelated (negative pair) mRNA isoform pairs by using the GO biological process annotations tagged with "NOT" qualifier. A gene/protein tagged with "NOT" qualifier means that it is not involved in the respective GO biological process and hence can be considered non-functional (negative) for this GO biological process. All such annotations are propagated by the inverse of "true path rule", which means that if a gene/protein is explicitly 'NOT' annotated to a GO term *T*, it will also be 'NOT' annotated to all the children of *T*. Consider a GO biological process term $T_1$ annotated with genes $G_1, G_2, G_3$ and $G_4$ which produce mRNA isoforms $M_1, M_2, M_{31}, M_{32}, M_{41}, M_{42},$ and $M_{43}$. Of these genes, if $G_3$ is tagged with a 'NOT' qualifier (Fig 1), all pairs of $M_{31}$and $M_{32}$ with $M_1, M_2, M_{41}, M_{42},$and $M_{43}$ are assumed to be

functionally unrelated (negative pair). It should be noted that currently there are only few hundred such annotations.

Genes can be annotated to multiple GO biological process terms. In Fig 1, single mRNA isoform producing genes $G_1 and\ G_2$ are annotated to GO biological process terms $T_1\ and\ T_2$. However, the gene $G_2$ is tagged with a "NOT" qualifier for term $T_2$. Consequently, the mRNA isoform pair $M_1 - M_2$ is functionally related for term $T_1$but functionally unrelated for term $T_2$. In cases where an mRNA isoform pair ($M_1 - M_2$) is found to be both functionally related (positive pair) for one term ($T_1$) but functionally unrelated (negative pair) for another term ($T_2$), we consider the mRNA isoform pair ($M_1 - M_2$) as functionally related (positive pair) because $M_1(G_1)$ and $M_2(G_2)$ are involved in at least one common GO biological process.

**Predicting functional networks**

**Generating training and testing datasets.** There are approximately 2.9 billion possible mRNA isoform pairs resulting from the 75,826 annotated mRNA isoforms. Using the method described above (see methods section 'mRNA isoform level functional labels'), we labelled 2,083,679 mRNA isoform pairs as functional pairs (positive) and 818,071 mRNA isoform pairs as non-functional pairs (negative). All the remaining mRNA isoform pairs are considered to be 'unknown', i.e. neither functional nor non-functional pairs. The mRNA isoform pairs in the functional and non-functional groups are mutually exclusive, i.e. an mRNA isoform pair can be either functional or non-functional, but not both.

We generate two types of datasets: training and testing. The training and testing datasets are mutually exclusive, i.e. an mRNA isoform pair can be either in a training or testing

dataset, but not both. The training dataset contains randomly selected 640,000 functional and 640,000 non-functional mRNA isoform pairs. The testing dataset contains randomly selected 160,000 functional and 160,000 non-functional mRNA isoform pairs not included in the training dataset. The functional pairs in the original testing dataset are made up of only single mRNA isoform genes. The non-functional pairs are however not restricted to single mRNA isoform genes. All datasets are balanced.

**Random forest model for the functional networks.** We formulate the task of mRNA isoform functional network prediction as a simple supervised learning problem. In supervised learning, a model capable of distinguishing a pre-defined set of 'positives' (functional mRNA isoform pairs in our case) from a set of 'negatives' (non-functional mRNA isoform pairs in our case) is built using a set of features derived from potential predictors of the property under consideration (mRNA isoform pair function in our case).

Using all 27 features for our training dataset, we train a Scikit-learn (Pedregosa et al., 2011) Random Forest (Breiman, 2001) model to predict the mRNA isoform functional network. Then we evaluate the performance of the random forest model by making predictions on the testing dataset. Commonly used performance evaluation metrics such as Accuracy, Area Under the Receiver Operating Characteristics Curve (AUROC), Area Under the Precision-Recall Curve (AUPRC), Precision, Recall, F1 Score, and Matthews Correlation Coefficient (MCC) are calculated using the predictions for testing dataset to assess the performance of the random forest model. The predictions are only evaluated when all 27 features are used for predictions. Finally, we use the random forest models to make predictions on all 2.9 billion possible mRNA isoform pairs.

**Building tissue-specific mRNA isoform networks.** To build the tissue-specific mRNA isoform networks, we utilize the leave-one-tissue-out strategy. First, using all 27 features, we train an organism-level mRNA isoform functional network prediction random forest model. Then, we generate 17 tissue-specific mRNA isoform functional network prediction random forest models by removing the tissue specific RNA-Seq features, one tissue at a time. The mRNA isoform pairs for which the prediction is unaffected after leave-one-tissue-out are referred to as "reference pairs". The two tissue-specific cases are: 1) mRNA isoform pairs which are predicted to be functional in only one tissue (tissue specific functional mRNA isoform pairs), and 2) mRNA isoform pairs which are predicted to be non-functional in only one tissue (tissue specific non-functional mRNA isoform pairs). These are also summarized in Fig 2.

If the prediction for an mRNA isoform pair changes from functional (positive) to non-functional (negative) after removing a tissue derived RNA-Seq feature, we consider such mRNA isoform pairs as tissue specific functional pairs. Similarly, if the prediction for an mRNA isoform pair changes from non-functional (negative) to functional (positive) after removing a tissue derived RNA-Seq feature, we consider such mRNA isoform pairs as tissue specific non-functional pairs. For instance, consider the case of heart specific mRNA isoform functional network prediction. We train two random forest models, 1) using all 27 features and, 2) after removing the heart derived RNA-Seq feature. Then, the heart specific functional mRNA isoform pairs are those which are predicted as functional by the first model but non-functional by the second model and vice-versa for the non-functional mRNA isoform pairs.

**From mRNA isoform networks to gene networks.** We collapse the tissue-specific mRNA isoform networks to gene networks as illustrated in Fig 3. All mRNA isoform nodes

of the same gene are merged into a single gene node. All direct edges from the mRNA isoforms of the same gene are transferred to the single gene node. This resulted in 17 gene level tissues networks in addition to the 17 tissue-specific mRNA isoform networks.

**Tissue-specific network analysis**

We use igraph (Csardi & Nepusz, 2006) in R (R Core Team, 2017) for analyzing the graph properties of tissue-specific networks. We calculate basic statistics like number of nodes, number of edges, density, number of components, and size of the largest connected component for both mRNA isoform and gene level networks. Using the largest connected component for every network, we find central nodes (top 10%) using betweenness centrality and degree centrality. We also check the overlap between the central nodes as found using both centrality measures. The overlapping central gene nodes are further subjected to functional enrichment analysis.

In addition to calculating the global network properties, we also extract the mRNA isoforms, genes and gene pairs that are specific to a tissue and those that are shared by multiple tissues.

**Functional enrichment analysis.** We use the tissue-specific list of overlapping central gene nodes to perform functional enrichment analysis using the ReactomePA (version 1.26.0) and clusterProfiler (version 3.10.0) packages in R (R Core Team, 2017; Yu & He, 2016; Yu, Wang, Han, & He, 2012). Enrichment is performed for Reactome pathways (version 66), KEGG pathways (release 88.2), GO biological process, GO molecular function and GO cellular components (GO data with a time stamp from the source of 10 October 2018 used by tools). In reactome data model, the core unit is a reaction while KEGG provides information about higher-level systemic functions of the cell and the organism. Due to the differences in

the underlying data model and how pathways are defined, we perform enrichment analysis for both Reactome and KEGG pathways. We use a p-value cutoff of 0.05, false discovery rate control using Benjamini-Hochberg (Benjamini & Hochberg, 1995) with a cutoff of 0.05, minimum term size of 10, and maximum term size of 1000 for the enrichment analyses. We also remove redundant GO terms with a semantic similarity greater than 0.7 using the "Wang" measure (Wang, Du, Payattakool, Yu, & Chen, 2007) and keep the terms with most significant adjusted p-value. We further filter the GO terms to four levels (Yu & He, 2016; Yu et al., 2012) and plot only the top 5 most significant terms for every tissue. Neural tube was removed from the functional enrichment analysis because there was only 1 central gene.

**Model evaluation**

**Randomization experiments.** To test the effect of randomization during the generation of training and testing datasets, we performed 1000 iterations of random training and testing dataset generation. In each iteration, we shuffle the combined functional and non-functional pairs, select 640,000 functional pairs and non-functional pairs respectively for the training dataset, select 160,000 functional and non-functional pairs respectively for the testing dataset, train a random forest model on the training dataset, use the trained model to make predictions on the testing dataset, and compute performance metrics. These datasets are referred to as "randomized datasets".

To examine whether the random forest model learns genomic and sequence features that are predictive of functional mRNA isoform pairs, we perform a control experiment in which the functional and non-functional class labels are randomly shuffled to destroy the feature-class relationship in the original dataset. We perform 500 iterations of random training and testing dataset generation in which the functional and non-functional mRNA

isoform pair class labels are shuffled. We train a random forest model on the class label shuffled training dataset, use the trained model to make predictions on the class label shuffled testing dataset and compute performance metrics. These datasets are referred to as the "class-label shuffled datasets".

We also evaluate the impact of number of trees on the performance of the random forest model. For this, we use the following number of trees: 10, 20, 50, 100, 200, 500, 1000, 2000, and 5000. Again, we train one model with each of these number of trees using the training dataset and then evaluate the performance using the predictions on the testing dataset.

**Using stratified cross-validation.** We evaluate the performance of TENSION using a Stratified 10-Fold cross-validator. In terms of bias and variance, stratification, a sampling technique without replication and where class frequencies are preserved, is generally a better scheme as compared to regular cross-validation (Kohavi, 1995). We use the original training data to create the 10-fold splits using StratifiedKFold function from Scikit-learn (Pedregosa et al., 2011) which preserves the relative class frequency in each training and held out test fold. We then evaluate the performance of each fold by computing the AUROC and AUPRC using the predictions made on the held out test fold.

**Validating predictions using new annotations.** Because there is no gold standard dataset available for mRNA isoform level functions, we validate our predictions using the latest annotations from GO, KEGG pathways, BioCyc pathways, IntAct (Orchard et al., 2014), BioGRID (Chatr-Aryamontri et al., 2017), APID (Alonso-López et al., 2016), IID (Kotlyar et al., 2016) and Mentha (Calderone et al., 2013). The new annotations (downloaded on 5 June 2018) were also processed as described in the "mRNA isoform level functional labels" section. Using our strategy to utilize the single isoform gene annotations for creating functional pairs,

we found 284,916 functional pairs in the new annotations not present in our original functional pairs. Similarly, we found 112,827 non-functional pairs in the new annotations not present in our original non-functional pairs. We refer this new set of functional and non-functional mRNA isoform pairs as the "validation set".

**Validation of literature datasets.** We also validate the predictions made by TENSION using two datasets from the literature: 1) a list of 20 ubiquitously expressed genes (Söllner et al., 2017) and, 2) a list of 5035 genes that are expressed higher (expression fold change greater than 4 relative to all other tissues) in a specific tissue (B. Li et al., 2017). Only the tissues present in both TENSION and the transcriptomic BodyMap of mouse are selected for validation. We merge the three brain regions used in TENSION, forebrain, midbrain and hindbrain into a single brain entity for the analysis. Additionally, we removed the transcriptomic BodyMap of mouse genes that were not included in our initial 21,813 genes. This resulted in a final gene set of 1654 genes for the transcriptomic BodyMap of mouse. It is important to note that the above gene lists are based solely on the gene expression and do not necessarily translate to functionally enriched genes and as such we expect to find interactions involving these genes in multiple tissues.

**Comparison with existing methods.** To demonstrate the utility of using a simple supervised learning framework and improvements over previous methods for mRNA isoform functional network prediction, we compare TENSION with the Bayesian network based multi-instance learning model in (H.-D. D. Li et al., 2016). We use our original training dataset with all 27 features to train the Bayesian network classifier and TENSION and make predictions on our original testing dataset. The output scores for mRNA isoform pairs in the original testing dataset from Bayesian network classifier and TENSION were used to compare the

performance of the methods. We evaluate the performance of both the methods by computing the AUROC and AUPRC.

## Results

### A random forest model for functional network prediction

We use the mouse genome build GRCm38.p4 from NCBI in this study. After filtering the mRNA isoforms containing non-standard characters, less than 30 amino acid protein products and those missing either sequence or expression profile, we retained 2,874,753,225 mRNA isoform pairs. We have calculated 27 heterogeneous genomic and sequence-based features for all the mRNA isoform pairs (Table 1). Of these, we labelled 2,083,679 mRNA isoform pairs as functional pairs (positive) using the single mRNA isoform genes (described in methods section). And 818,071 mRNA isoform pairs as non-functional pairs (negative) by using the "NOT" annotation tag in the GO annotations (described in methods section). These functional and non-functional mRNA isoform pairs are used to train and develop random forest models for predicting mouse mRNA isoform level functional networks. The predictions made by random forest have an associated probability score which measures the strength of mRNA isoform interactions.

**Randomization experiments.** Randomization experiments test the effect of selecting functional and non-functional pairs when generating training and testing datasets. Fig 4 shows that there is very little to no variance in the performance of randomized datasets. Therefore, we generate one final training and testing dataset ("original datasets") by randomly selecting functional and non-functional pairs and use it to generate the final functional network prediction models.

To help us identify if TENSION is actually learning from the data and not just making random predictions, we estimate the performance of the random forest model on the class-label shuffled datasets. The AUROC obtained on the class-label shuffled datasets is 0.5 (as compared with 0.947 on the original testing dataset) indicating that our functional network prediction model performs significantly better than random predictions (Fig 5).

**Performance evaluation.** We evaluate the performance of TENSION when using all 27 features from the predictions on the original testing dataset. We first evaluate the impact of number of trees on the performance of random forest model. It can be seen in Fig S1 that there is very little improvement in the performance of the model after 100 trees. To reduce computational complexity without sacrificing the performance while making predictions for all 2.9 billion mRNA isoform pairs, we use 100 trees in our final models. On the original testing dataset, we obtain a high correlation as seen in Table 2 and Fig S2 suggesting a highly accurate model.

**Evaluation by stratified cross-validation.** In addition to evaluating the performance of our random forest on a held-out test set, we also perform stratified 10-fold cross validation. The AUROC and AUPRC curves for each fold are shown in Fig 6. We see that there is very little variance in the results of each fold. The results are also very close to those obtained on the original testing dataset (S2 Fig). The results of stratified cross-validation emphasize the robustness of TENSION.

**Validating predictions using new annotations.** After processing the new GO annotations, pathway, and PPIs data, we learned a new set of 397,743 previously unknown mRNA isoform pairs. Of these, we labelled 284,916 as functional and 112,827 as non-functional mRNA isoform pairs. Using all 27 features, TENSION correctly classified 315,844

(out of 397,743) mRNA isoforms pairs at an overall accuracy of 79.4%. The true positives, true negatives, false positives, and false negatives collectively represented by a confusion matrix are presented in Table 3. Since the distribution of functional and non-functional mRNA isoform pairs in the validation set is imbalanced, we also access the performance of our classifier by computing the AUPRC and AUROC. We observe an AUPRC of 0.926 and an AUROC of 0.855 (Fig 7). In addition to these curves, we also calculate the Precision (0.885), Recall (0.819), F1 score (0.851) and MCC (0.524). These are much higher than random predictions shown in Fig 5 suggesting that TENSION performs better than random guessing and is also able to predict potential functional and non-functional mRNA isoform pairs accurately.

Of these new mRNA isoform pairs, 8200 are predicted as tissue-specific functional mRNA isoform pairs. However, the annotations in GO, KEGG, BioCyc, and PPI databases do not store tissue information, so we cannot validate the tissue specificity of these predictions.

**Comparison with existing methods.** We compare the performance of TENSION when using all 27 features with that of the Bayesian network based MIL method (H.-D. D. Li et al., 2016). The default parameters are used for the Bayesian network-based MIL method. We use our original training dataset to train the Bayesian network-based MIL method and TENSION and then make predictions on our original testing dataset. We calculate the AUROC and AUPRC using these predictions for both models to compare their performance. The functional mRNA isoform pairs are derived from single mRNA producing genes co-annotated to GO biological process, pathways or PPIs whereas the non-functional mRNA isoform pairs are constructed by using the ‘NOT’ tagged GO biological process annotations.

The Bayesian network based MIL method achieves an AUROC of 0.761 (Fig 8) which is higher than the original AUROC value of 0.656 reported in the original study (H.-D. D. Li et al., 2016). TENSION achieves significantly higher AUROC of 0.947. Similarly, TENSION achieves significantly higher AUPRC of 0.947 as compared to Bayesian network based MIL method's AUPRC of 0.757 (Fig 8). The significantly higher AUROC and AUPRC values of TENSION highlights the importance of using a simple supervised learning framework and improvements over the more complex MIL-based methods for mRNA isoform functional network prediction. It should be noted that the MIL-based method was originally developed using different set of features, however, for the purpose of comparison we have used the same training and testing datasets for both methods. The improved performance of Bayesian network based MIL method on our dataset also highlights the significance of mRNA isoform level functional label and feature generation in TENSION.

**Tissue-specific networks**

**Tissue-specific functional mRNA isoform pair networks.** As shown in Fig 2, to build the tissue-specific mRNA isoform level functional networks, we assume that, for a tissue $\boldsymbol{i}$, if an mRNA isoform pair is predicted to be functional (positive) using all 27 features, but the prediction after removing the tissue $\boldsymbol{i}$- specific feature is non-functional (negative), the mRNA isoform pair is only functional under tissue $\boldsymbol{i}$. The strength of mRNA isoform interactions is measured by the probability score predicted by random forest. To remove noise, low confidence predictions and organism-wide reference mRNA isoform pairs from tissue-specific functional networks, we only consider the mRNA isoform pairs which have a random forest predicted probability score $\geq \mathbf{0.6}$ when using all 27 features and a probability score $\leq \mathbf{0.4}$

after removing the tissue derived RNA-Seq feature. For the tissue specific functional networks, a lower probability score corresponds to higher strength of mRNA isoform pair to be involved in the same GO biological process or pathway. A summary of all 17 tissue-specific mRNA isoform functional networks as obtained after applying the above filtering criteria is provided in Table 4.

The tissue-specific functional networks identify around 10.6 million tissue-specific functional mRNA isoform pairs (0.37% of all possible mRNA isoform pairs). The density of tissue-specific functional networks is in the order of $10^{-2} - 10^{-5}$ and most networks are very sparse. The number of tissue-specific functional mRNA isoform pairs vary greatly across the tissues, from few thousands in limb and neural tube to few million in large intestine and ovary (Table 4). All these mRNA isoform pairs are present in only one tissue. Table 4 shows the number of functional mRNA isoform pairs identified as single tissue-specific in each of the 17 tissues.

All tissues have many connected components (Table 4). Limb, neural tube and kidney have less than 50% mRNA isoform nodes in their largest connected component, whereas some others like hindbrain, large intestine, ovary, and forebrain etc. have over 90%. These differences in the size of networks, mRNA isoforms involved and the network structures highlight the differences in tissue-level biological processes as evident by the differences in the enriched pathways and gene ontology terms (discussed later).

To highlight the differences that arise when analyzing functional networks at the mRNA isoform and gene level and because all functional enrichment tools are built for analyzing genes, we also compress the mRNA isoform level networks to gene level networks. In the gene level networks, all mRNA isoform nodes of the same gene are combined into a

single gene node. Table 5 provides a summary of the gene level networks for all 17 tissues. We identified around 7.79 million unique gene pairs (3.27% of all possible gene pairs) using these tissue level gene functional networks. It was recently observed in mouse and humans that testis and ovary express the highest number of genes whereas brain and liver express the highest number of tissue enriched genes under normal conditions (B. Li et al., 2017; Uhlen et al., 2015). This is also reflected in our gene level networks (Table 5) where ovary, hindbrain and forebrain networks have the largest number of edges (gene pairs) and nodes (genes).

While the majority of gene pairs are present in only one tissue level gene functional networks (98% of identified gene pairs; Table 5), a small fraction (2% of identified gene pairs; Table 5 and Fig 9) is present in at least two tissue level gene functional networks. Although the gene pairs are shared between tissues, the mRNA isoform pairs resulting from these gene pairs are specific to only one tissue. This highlights that different mRNA isoforms of the same gene can have different functional partners across tissues.

Shared gene-pairs may indicate shared processes between tissues. The spleen and embryonic facial prominence share the highest fraction of gene pairs (about 7.6% of gene pairs; Table 5) with other tissues, while ovary shares the lowest fraction (3% of all ovary gene pairs; Table 5). The composition of gene pairs shared between the tissue level functional networks is quite complex and is shown in Fig 9. Upon further investigation, we find that the spleen network shares 4.8% of its gene pairs with ovary network while ovary network shares only 0.1% of its gene pairs with the spleen network. We also find that thymus shares about 3.7% of its gene pairs with ovary, supporting the notion that thymus is necessary for normal ovarian development and function after the neonatal period (Garcia, Hinojosa, Dominguez, Chavira,

& Rosas, 2000; Michael, 1979). These findings further justify the importance of our networks in characterizing tissue level processes.

Like the mRNA isoform networks, the gene-level neural tube network contains only 2.9% of genes in its largest connected components (Tables 4 and 5). All other gene-level tissue networks have a very high fraction of genes and gene pairs in the largest connected components (Table 5).

**Central genes in tissue-specific functional networks have tissue related characteristics.** The central genes identified in our tissue-specific networks are enriched in tissue related GO terms and pathways (Figs 10 and 11). The central genes in the heart specific gene network are significantly enriched in transmembrane transporter activity, vitamin binding, complement and coagulation cascades etc. (Figs 10 and 11). Supplementation of several vitamins such as Vitamin B6, Vitamin D, Vitamin E, and folate etc. are linked to reduced risk of cardiovascular diseases (Rimm et al., 1993, 1998; Schnyder, Roffi, Flammer, Pin, & Hess, 2002; Stephens et al., 1996; Zittermann et al., 2003). The serine proteinase cascades of the coagulation and the complement systems have been associated with functions of the cardiovascular and immune systems (Oikonomopoulou, Ricklin, Ward, & Lambris, 2012).

Several important renal processes such as JAK-STAT signaling pathway, cytokine signaling in immune system, cytokine-cytokine receptor interaction, and signaling by interleukins etc. are enriched in the central genes of the kidney specific gene network (Figs 10 and 11). Defects in these processes and pathways have been linked to several renal disorders and related co-morbidities (Berthier et al., 2009; Brosius & He, 2015; Chuang & He, 2010; Yang et al., 2008). Genes in the kidney network are also enriched for interferon-gamma

production and inflammatory bowel disease (IBD; Figs 10 and 11). In IBD, interferon-gamma negatively regulates the $Na^{+}/Ca^{2+}$ exchanger 1 (NCX1) -mediated renal $Ca^{2+}$ absorption contributing to IBD-associated loss of bone mineral density and altered $Ca^{2+}$ homeostasis (Radhakrishnan et al., 2015).

The large intestine has specific and efficient carrier mediated transporter mechanisms for the absorption of several water soluble vitamins (pantothenic acid, biotin, thiamin, riboflavin and folate) (Said & Mohammed, 2006). These vitamins are essential for several biological processes and their enrichment in large intestine specific gene network only seems natural (Figs 10 and 11). The *brain-in-the-gut* or the enteric nervous system (ENS) is the largest component of the autonomous nervous system (Nezami & Srinivasan, 2010; Rao & Gershon, 2016; Wood, 2016). The small intestine ENS is equipped to perform functions relating to inflammation, digestion, secretion and motility among others (Nezami & Srinivasan, 2010; Rao & Gershon, 2016; Wood, 2016). The identification of several neuronal terms for central genes in the small intestine network is in line with such literature findings (Figs 10 and 11) (Nezami & Srinivasan, 2010; Rao & Gershon, 2016; Wood, 2016).

Fertility and energy metabolism are reciprocally regulated and tightly linked in female animals and this relation has been conserved throughout evolution (Della Torre et al., 2011; Fontana & Della Torre, 2016; Torre, Benedusi, Fontana, & Maggi, 2014). Metabolic disorders such as those of the liver can lead to changes in reproductive functions and vice-versa (Della Torre et al., 2011; Fontana & Della Torre, 2016; Torre et al., 2014). It was recently proposed that in case of protein scarcity, the estrous cycle is blocked and the liver acts as a critical mediator of reproductive and energetic functions (Della Torre et al., 2011; Fontana & Della

Torre, 2016; Torre et al., 2014). The enrichment of several reproduction and fertility related terms in our liver specific network also point towards such observations (Figs 10 and 11).

We also find significantly enriched tissue related process terms for other tissues such as spleen, ovary, adrenal glands and limb etc. (Figs 10 and 11). However, the tissue specific central genes do not always lead to significantly enriched terms.

The identification of tissue related biological processes via the central genes highlights that TENSION can correctly capture the tissue-specific functional mRNA isoform pairs produced by genes involved in tissue related functions. We can identify the specific mRNA isoforms of these genes by looking back at the mRNA isoform level tissue networks. Finding the specific mRNA isoforms responsible for these processes should provide a significant clue towards understanding of developmental and molecular processes of diseases and biological functions.

**Tissue-specific non-functional mRNA isoform pair networks.** To build the tissue-specific mRNA isoform level non-functional networks, we assume that, for a tissue $\boldsymbol{i}$, if an mRNA isoform pair is predicted to be non-functional (negative) using all 27 features but the prediction after removing the tissue $\boldsymbol{i}$- specific feature changes to functional (positive), the mRNA isoform pair is only non-functional under tissue $\boldsymbol{i}$ (Fig 2). To remove noise and low confidence predictions in tissue-specific non-functional mRNA isoform networks, we only consider the mRNA isoform pairs which have a random forest predicted probability score of $\leq \mathbf{0.4}$ when using all 27 features and a probability score of $\geq \mathbf{0.6}$ after removing the tissue derived RNA-Seq feature. Higher probability score reflects stronger tissue-specific non-functional mRNA isoform pair. A summary of all 17 tissue-specific mRNA isoform level non-

functional networks as obtained after applying the above filtering criteria is provided in Tables 6.

Using these tissue-specific mRNA isoform level non-functional networks we identified around 3.5 million tissue-specific non-functional mRNA isoform pairs (0.12% of all possible mRNA isoform pairs). The tissue-specific non-functional networks are also sparse with density in the order of $10^{-3} - 10^{-5}$. The number of tissue-specific non-functional mRNA isoform pairs also vary greatly across the tissues. For instance, forebrain has a very high number of 1.4 million (40% of all tissue-specific non-functional mRNA isoform pairs) non-functional mRNA isoform pairs. All these mRNA isoform pairs are specifically non-functional in only one tissue.

Similar to the functional networks, we also compress the non-functional mRNA isoform networks to gene level non-functional networks. In the gene level networks, all mRNA isoform nodes of the same gene and their edges are combined into a single gene node. Many gene pair (but no mRNA isoform pair) are present in at least two tissue level gene non-functional networks.

**Different mRNA isoforms of the same gene are functional in different tissues and have tissue preferred functional partners**

The tissue level functional mRNA isoform networks along with the identification of gene pairs that are shared across tissues provide us an opportunity to distinguish the tissue-specific functional mRNA isoforms of a gene. We have identified around 164,000 functional gene pairs with different mRNA isoform pairs that are shared by multiple tissues. This points to the tissue specific expression and function of different mRNA isoforms of a gene.

The fraction of gene pairs shared between tissues is presented in Fig 9. We see that several pairs of tissues such as limb and forebrain, heart and large intestine, midbrain and

forebrain, thymus and ovary, spleen and ovary etc. share a large number of gene pairs. This suggests that while these gene pairs are functional in multiple tissues, the actual mRNA isoform pairs can differ and our networks are capable of identifying such differential relationships between mRNA isoform pairs of the same gene pair.

The gene pair Fundc2 (FUN14 domain containing 2) and Necab1 (N-terminal EF-hand calcium binding protein 1) is present in both ovary and heart. The Fundc2 gene produces a single mRNA isoform NM_026126.4 while Necab1 gene produces two mRNA isoforms, XM_006538234.1 and NM_178617.4. The interaction between Fundc2 and Necab1 can be dissected into two interactions corresponding to the two mRNA isoform pairs (Fig 12A). Among the two mRNA isoform pairs, the pair involving XM_006538234.1 is heart specific functional mRNA isoform pair while the other pair involving NM_178617.4 is functional in ovary. This reveals the tissue preferred interaction partners of Fundc2 mRNA isoform NM_026126.4. Further investigation of all tissue specific functional mRNA isoform pairs involving Necab1 mRNA isoform XM_006538234.1 revealed that most of its interactions are found in heart (366 out of 391). Similarly, most of the interactions involving Necab1 mRNA isoform NM_178617.4 are found in ovary (836 out of 859). This highlights the expression and functional preference of Necab1 mRNA isoforms.

Another such gene pair involves two mRNA isoform producing genes, Apoc2 (apolipoprotein C-II) and Nts (neurotensin). The gene pair involving Apoc2 and Nts is found in the networks of ovary and forebrain and can be dissected into four interactions corresponding to the four mRNA isoform pairs. Three of these mRNA isoform interactions are found to be tissue-specific functional mRNA isoform pairs (Fig 12B). Interactions involving the Apoc2 mRNA isoform NM_001309795.1 are preferred in forebrain (1310 out of 1903) and

NM_001277944.1 are preferred in ovary (355 out of 586). The NM_024435.2 mRNA isoform of Nts is enriched in ovary (1314 out of 1358) and interacts with the ovary enriched Apoc2 mRNA isoform NM_001277944.1 in ovary, suggesting a tissue preferred interaction pattern.

TENSION is also able to distinguish the tissue-specificity of mRNA isoforms of a gene between closely related tissues. For example, the gene Olfr994 (olfactory receptor 994) produces two mRNA isoforms, XM_006499549.1 and NM_146433.1. The mRNA isoform NM_146433.1 is preferred in hindbrain (223 out of 309 interactions) while XM_006499549.1 is preferred in midbrain (57 out of 65 interactions). There are several cases in which the mRNA isoforms of the same gene exhibit tissue preferred interactions. However, this is not true for all multi-isoform genes. The mRNA isoforms of many multi-isoform genes are not involved in tissue preferred interactions.

**Some mRNA isoform pairs are functional while other mRNA isoform pairs of the same gene pair are non-functional**

We find about 660,000 instances where an mRNA isoform pair is functional while other mRNA isoform pairs of the same gene pair are non-functional. Around 143,000 of such cases are within the same tissue. For example, the mRNA isoforms of genes Agrp (agouti related neuropeptide) and Olfr1152 (olfactory receptor 1152) result in two mRNA isoform pairs (Fig 12C). The pair involving NM_001011834.1 (Olfr1152) and NM_001271806.1 (Agrp) is predicted to be functional in hindbrain while the other pair involving Agrp mRNA isoform NM_007427.3.1 is non-functional in hindbrain (Fig 12C). The NM_007427.3.1 mRNA isoform of Agrp is functionally enriched in the forebrain but has most of its non-functional interactions in hindbrain (362/447 functional interactions in forebrain vs 324/343 non-functional interactions in hindbrain), but the opposite is true for the isoform

NM_001271806.1. The NM_001271806.1 mRNA isoform of Agrp contains an alternate 5' exon, although both Agrp mRNA isoforms produce the same protein.

Similarly, for the gene pair involving Iqcf6 (IQ motif containing F6) and Gstcd (glutathione S-transferase, C-terminal domain containing), only one mRNA isoform pair is functional in adrenal glands while two other pairs are non-functional (Fig 12D). The remaining mRNA isoform pair could be functional or non-functional in multiple tissues.

The remaining 520,000 instances are across tissues, i.e., one mRNA isoform pair is tissue-specific functional in one tissue while other mRNA isoform pairs of the same gene pair are tissue-specific non-functional in other tissue.

**Validation of super-conserved and transcriptomic BodyMap of mouse tissue-specific genes**

The first gene set contains 20 genes that are known to be widely expressed (Söllner et al., 2017). These genes have tissue-specific functional interactions in most of our 17 tissue-specific networks validating their ubiquitous expression and function (Fig. 13). The second gene set contains 1654 genes from the transcriptomic BodyMap of mouse that have a very high expression in one tissue (relative to all other tissues) and thus a higher propensity to have more tissue-specific functions. For every gene, we compute the top $n = \{1, 3, 5, 7, 9, All\}$ tissues for its mRNA isoforms based on the number of functional interactions in the tissue.

We find that the top tissue ($n = 1$) among our tissue-specific networks and that in the transcriptomic BodyMap of mouse matches for 503 genes (30%; Table 7). However, a gene can be involved in multiple functions across multiple tissues due to different mRNA isoforms.

Therefore, when we consider the top 3 (52% match) or top 5 (68% match) tissues, we find a much higher correlation with the transcriptomic BodyMap of mouse (Table 7). Overall, we find 1245 (75%) genes to have at least one tissue specific interaction in the same tissue as described in the transcriptomic BodyMap of mouse.

It is interesting to note that if we consider the tissue-specificity of only the genes, ignoring the tissue-specificity of different mRNA isoforms of the same gene, we find a weaker correlation with the transcriptomic BodyMap of mouse (15% and 41% respectively for n = 1 and 3). Most studies including the transcriptomic BodyMap of mouse focus only on the gene expression and function, completely ignoring the effects of alternatively spliced mRNAs. Our study further illustrates the importance of distinguishing the functions of different mRNA isoforms of the same gene.

**Similar tissues have similar mRNA isoform expression profile**

Tissues that are functionally and morphologically similar tend to have more consistent gene expression profile than other tissues (B. Li et al., 2017). We also observe that similar tissues such as midbrain, forebrain, hindbrain and neural tube have a very high Pearson correlation coefficient ($\rho \geq 0.97$; Fig 13) based on the median mRNA isoform expression profile. Likewise, adrenal gland is most highly correlated with ovary ($\rho = 0.87$), large intestine with small intestine ($\rho = 0.84$) and thymus with spleen ($\rho = 0.88$) among others, and are consistent with previous findings (B. Li et al., 2017).

## Discussions

We have developed tissue-level functional networks to study mRNA isoform functional relationships, providing a higher resolution view of biological processes as

compared to traditional gene-level networks. Learning the differences in the functional connections of mRNA isoforms of the same gene are crucial for functional genomics, and helps us in deepening our understanding of gene functions. Determining the functional interaction patterns of mRNA-isoforms of the same gene also provides useful information about biological regulation, diseases, and stress response caused by AS.

It is widely believed that the fate of biological processes and pathways varies with different mRNA isoforms of the same gene. Many pathways and molecular processes differ across cell and tissue-types. These mechanisms are also altered by external conditions such as abiotic and biotic stress. Understanding of such deviations in cell, tissue and condition specific functional relationships would be of interest to understand the perturbed mechanisms.

Based on the analysis of 359 mouse tissue-specific RNA-Seq samples along with 9 diverse sequence properties, we have constructed 17 tissue-specific mRNA isoform level functional networks. These networks constitute ~10.6 million unique functional and ~3.5 million non-functional mRNA isoform interactions across 17 tissues. In addition to these tissue-specific networks, we have also developed an organism-wide reference network. We show that TENSION is highly accurate with very high precision and recall by comparing our predictions with class label shuffled datasets, ten-fold stratified cross validation, previous method, and updated annotations from gene ontology, pathway databases and PPIs. In addition to these, we also validate our predictions by using a gene set of 20 ubiquitously expressed genes and 1654 genes with a very high expression in one tissue from the transcriptomic BodyMap of mouse. The improvement in the performance (compared to the original study) of Bayesian network based MIL method on our dataset also prove the utility of TENSION in generating better mRNA isoform level datasets.

Our tissue-specific networks capture the differences in functional relationships of mRNA isoforms of the same gene across multiple tissues highlighting the importance of tissue-specific changes in biological processes and pathways. We are also able to distinguish the tissue-specific functional mRNA isoforms of a gene. We also find that different mRNA isoforms of the same gene are enriched in different tissues, suggesting differential tissue-level activity of mRNA isoforms of the same gene. Furthermore, we also see that morphologically and functionally similar tissues tend to have more consistent mRNA isoform expression profile.

By studying the gene level networks in conjunction with mRNA-isoform level functional networks, we are able to gain different insights into the molecular mechanisms of biological processes. Diving down further into the tissue-specific networks sheds more light on the tissue-level activities of a gene and its mRNA isoforms. The central genes identified in these tissue-level networks are enriched in tissue related processes.

Despite all the efforts to reduce bias and account other variables that can impact the results, there are few shortcomings. Like similar studies, we do not distinguish between the co-variates such as sex and age, but rather build generic mouse functional networks. A very important and common assumption of all machine learning studies in biological sciences is the fact that the current biological databases are accurate and complete to-date. And like previous studies, our study will also suffer from the loss of information not present in biological databases such as Gene Ontology, Pathway and PPI databases.

In summary, we provide the research community with a comprehensive characterization of mRNA isoform level tissue-specific functional networks for mouse. TENSION is simple and generic, making it easily applicable to other organisms. We expect

that these networks will allow further in-depth investigations of the impact of alternatively spliced mRNA isoforms on biological processes. We anticipate that tissue-specific mRNA-isoform functional networks will find wide applications in genomics, agriculture and biomedical sciences.

## Data availability

All data and scripts have been deposited and is available at DataShare: Iowa State University's Open Research Data Repository through doi: https://doi.org/10.25380/iastate.c.4275191 (Dickerson & Kandoi, 2019).

## Acknowledgement

This material is based upon work supported by the National Science Foundation under Grant IOS-1062546. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. This work used the XSEDE Comet cluster at San Diego Supercomputer Center (SDSC) through allocation TG-BIO170049. We would like to thank Dr. Mahidhar Tatineni and Dr. Martin Kandes from SDSC User Services group for providing technical support at Comet cluster. We also thank Megan O'Donnell and Levi Baber for setting up the data repository at DataShare: Iowa State University's Open Research Data Repository. We would also like to thank Dr. Yuanfang Guan and Dr. Hongdong Li for providing some data and the code for the comparison with Bayesian network-based MIL method.

## Competing interests

The author(s) declare no competing interests.

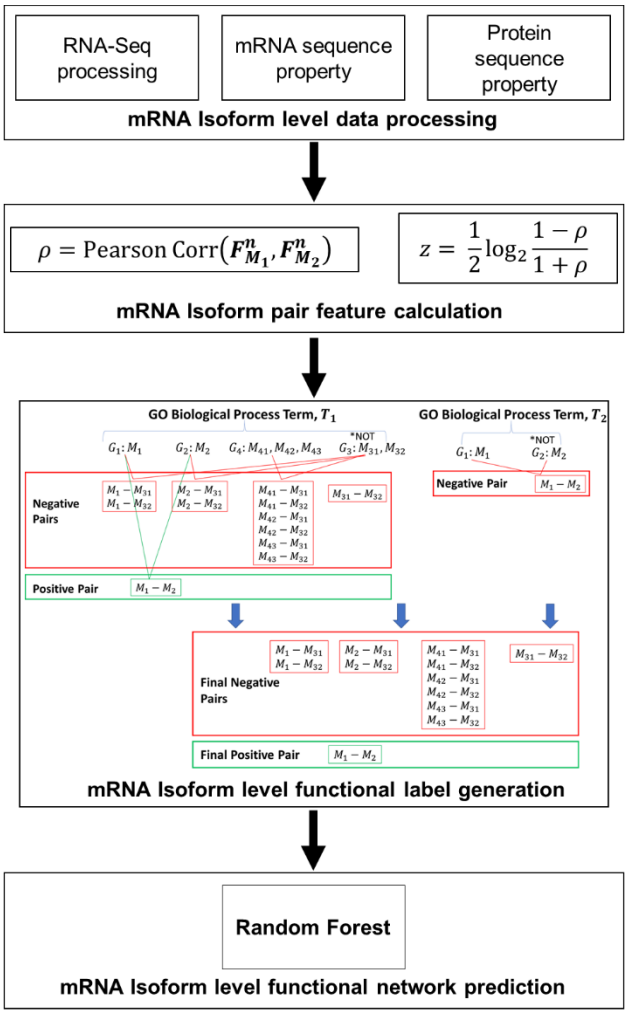


Figure 3.1 **Overview of our workflow.** A brief overview of TENSION is provided. We also illustrate the process of generating the mRNA isoform level labels using two dummy gene ontology biological process terms, T1 and T2. Functional mRNA isoform pairs (positive pairs) are shown in green and non-functional pairs (negative pairs) are shown in red.

| | Using all features | Removing any other tissue | Removing tissue (i) |
|---|---|---|---|
| Tissue (i) specific functional pairs | **Functional** | **Functional** | **Non-functional** |
| Tissue (i) specific non-functional pairs | **Non-functional** | **Non-functional** | **Functional** |
| Reference pairs | **Functional** | **Functional** | **Functional** |

Figure 3.2 **Defining tissue specific functional and non-functional mRNA isoform pairs.** Here we illustrate the process of classifying the mRNA isoforms as tissue specific functional, tissue specific non-functional or organism wide reference pairs. If the prediction is functional (positive) when using all 27 features but changes to non-functional (negative) after removing the tissue derived RNA-Seq feature, we assume such mRNA isoform pairs as tissue-specific functional pairs. Contrary to tissue-specific functional pairs, if the prediction changes from non-functional (negative) to functional (positive) after removing the tissue derived RNA-Seq feature, we assume such pairs as tissue-specific non-functional pairs. For the reference pairs, the prediction is constant after removing any tissue derived RNA-Seq feature.

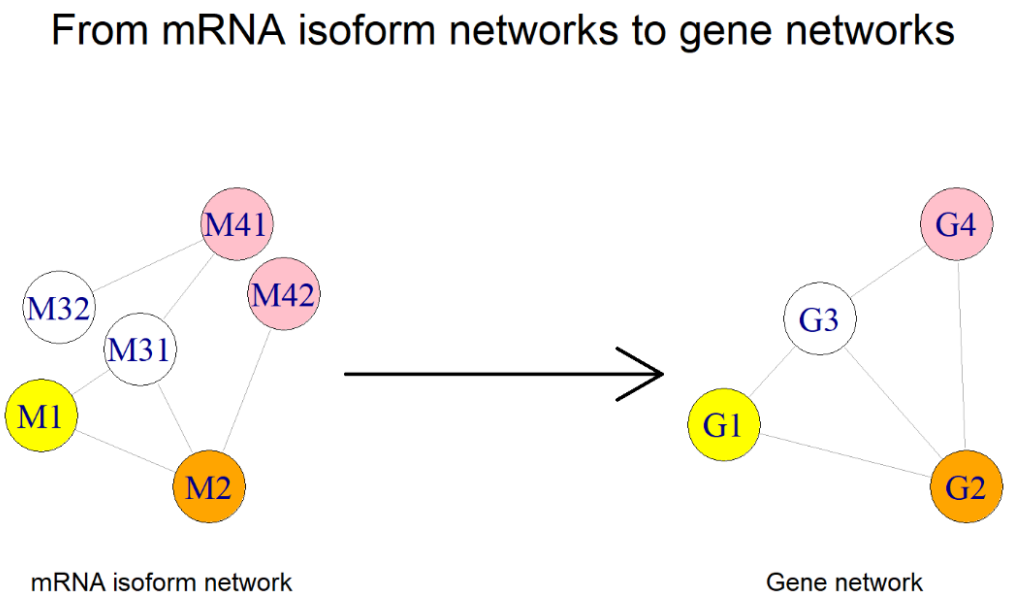


Figure 3.3 **Constructing gene level networks from mRNA isoform networks.** Shown here is the process by which we construct gene level networks using the tissue-specific functional mRNA isoform pair networks. All edges from the mRNA isoforms of the same gene in the mRNA isoform network are transferred to the single gene node in the gene level network. The gene and its mRNA isoforms have the same color.

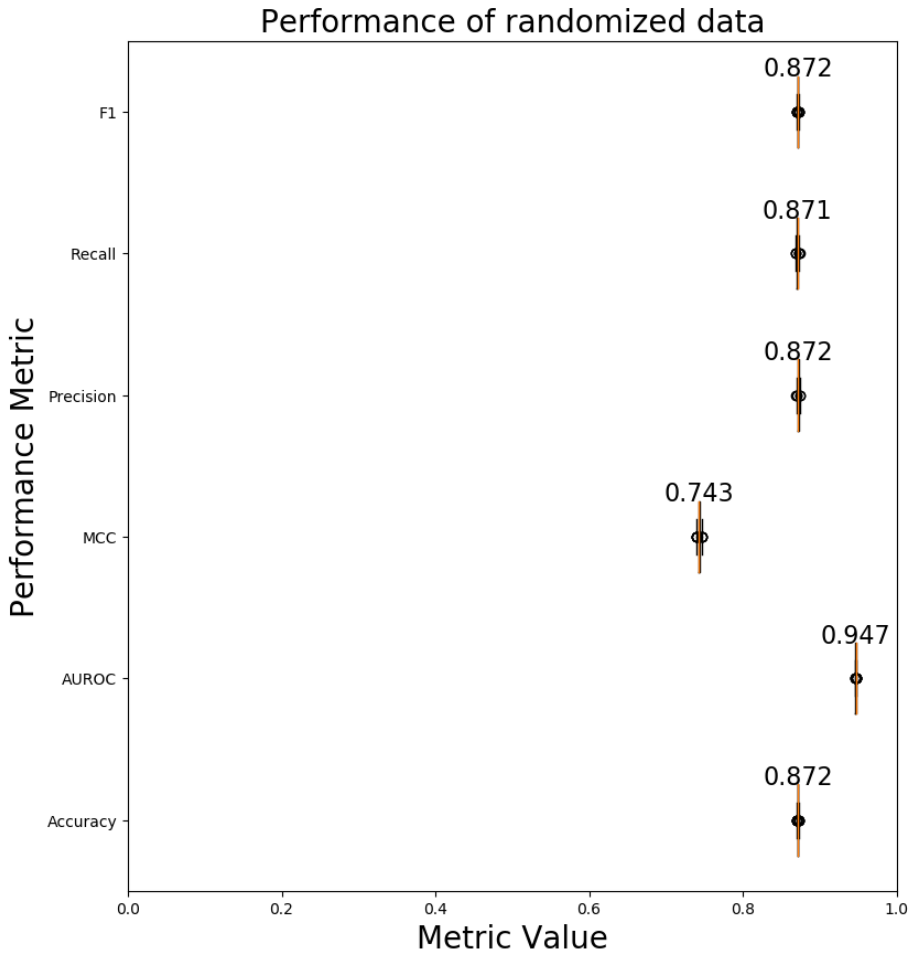


Figure 3.4 **Performance evaluation on randomized datasets.** A boxplot of various performance evaluation metrics calculated using 1000 randomized datasets. The median value is shown for the performance metrics. The width of the boxes along the x-axis represent the variability in the value of the performance metric across 1000 randomized datasets. Higher metric value and smaller box width is better. Abbreviations - AUROC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient.

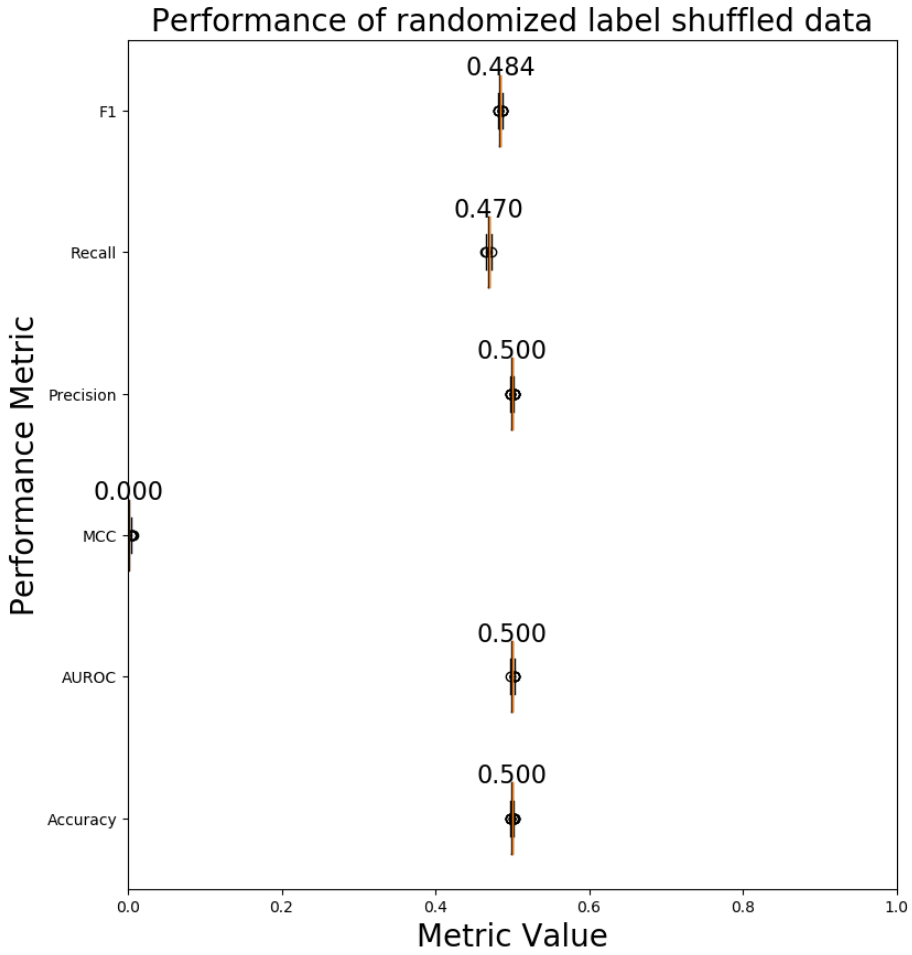


Figure 3.5 **Performance evaluation on label shuffled datasets.** A boxplot of performance evaluation metrics calculated using 1000 label shuffled datasets. The functional and non-functional labels for mRNA isoform pairs are randomly shuffled while still maintaining the class distribution (equal functional/non-functional pairs). The median value is shown for the performance metrics. The width of the boxes along the x-axis represent the variability in the value of the performance metric across 1000 label shuffled datasets. Higher metric value and smaller box width is better. The performance of a model which makes random guesses is about 0.5 (or 0 for MCC because it ranges from -1 to 1). Abbreviations - AUROC: Area Under the Receiver Operating Characteristic Curve; MCC: Matthews Correlation Coefficient.

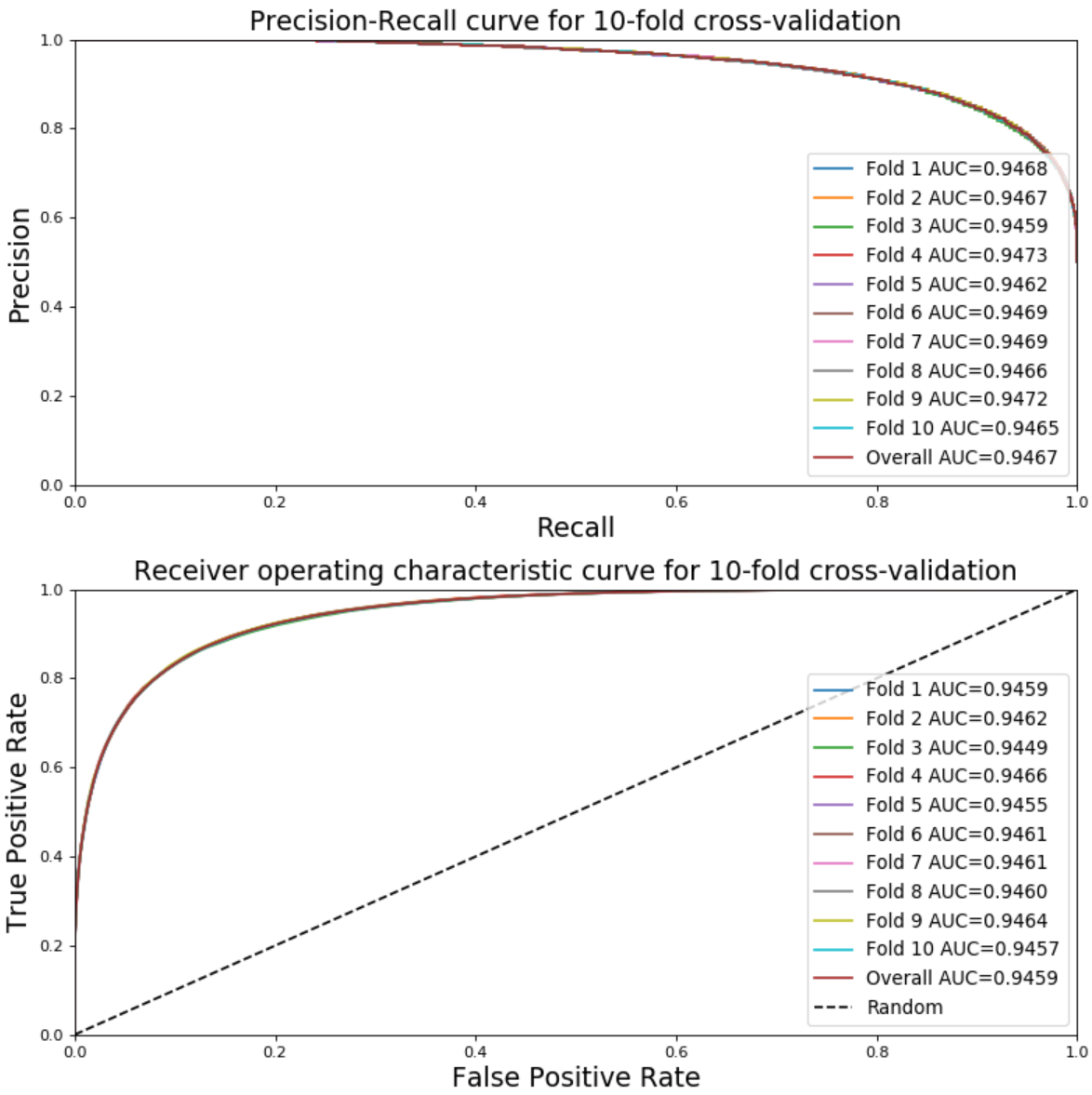


Figure 3.6 **Performance evaluation by 10-fold stratified cross-validation.** The precision-recall and receiver operating characteristic curve for all 10 folds of the stratified cross-validation. Note that the performance is virtually identical for all folds suggesting the robustness of TENSION. A model with area under the curve closer to 1 is better while a model with an area under the curve of 0.5 is equivalent to making random guess. Abbreviations - AUC: Area Under the Curve.

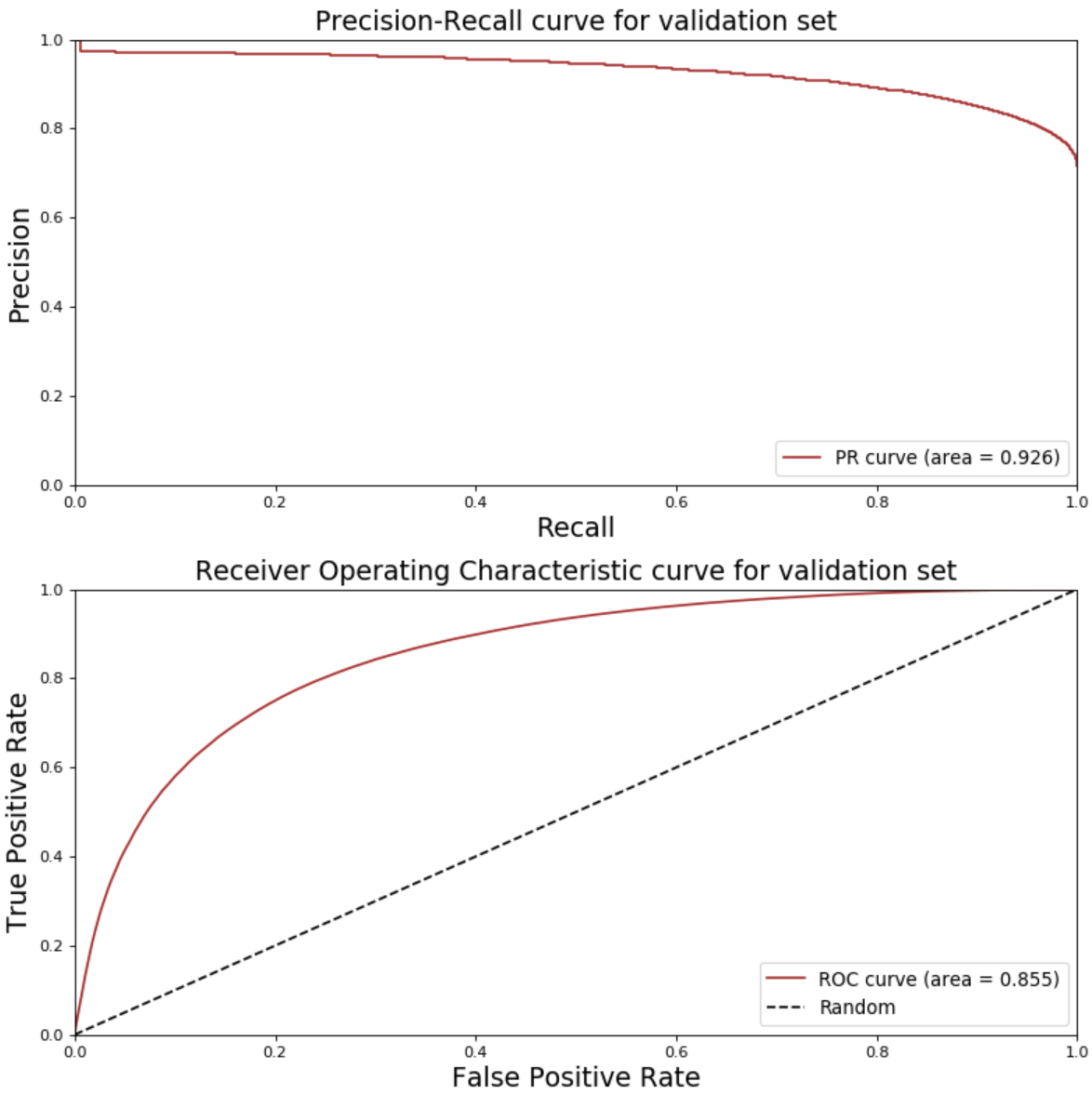


Figure 3.7 **Performance evaluation on validation dataset.** The precision-recall and receiver operating characteristic curve for predictions on the validation dataset. The validation dataset is constructed by using the later version of gene ontology annotations, pathways and protein-protein interactions than those used for our original mRNA isoform level label generation. A model with area under the curve closer to 1 is better while a model with an area under the curve of 0.5 is equivalent to making random guess. Abbreviations - PR: Precision-Recall; ROC: Receiver Operating Characteristic.

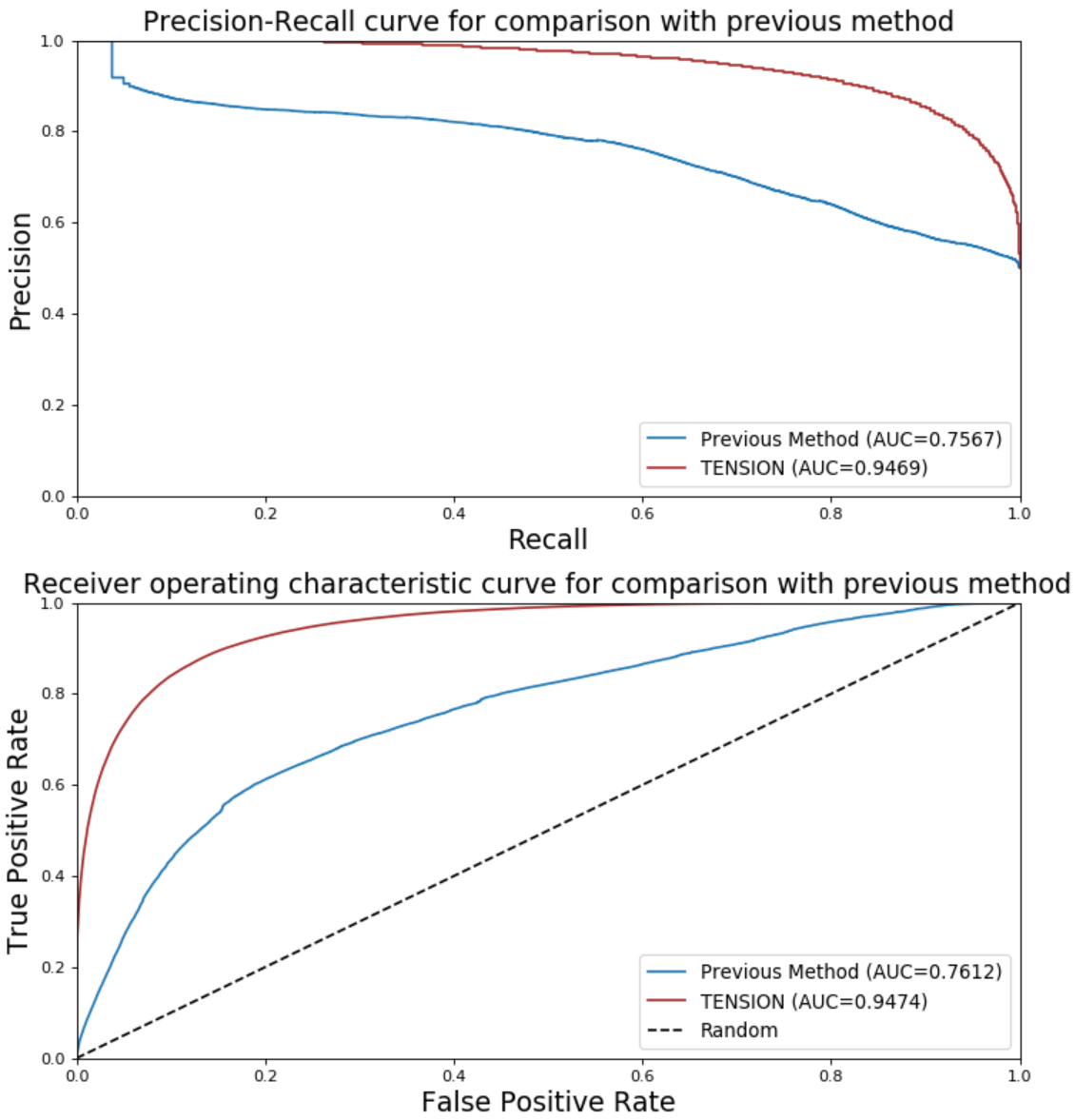


Figure 3.8 **Performance comparison with Bayesian network based multi-instance learning method.** The precision-recall and receiver operating characteristic curve for performance comparison of TENSION with previously published Bayesian network based multi-instance learning method. The original training dataset was used to train both models and performance was calculated using the predictions made on the original testing dataset. Abbreviations - AUC: Area Under the Curve.

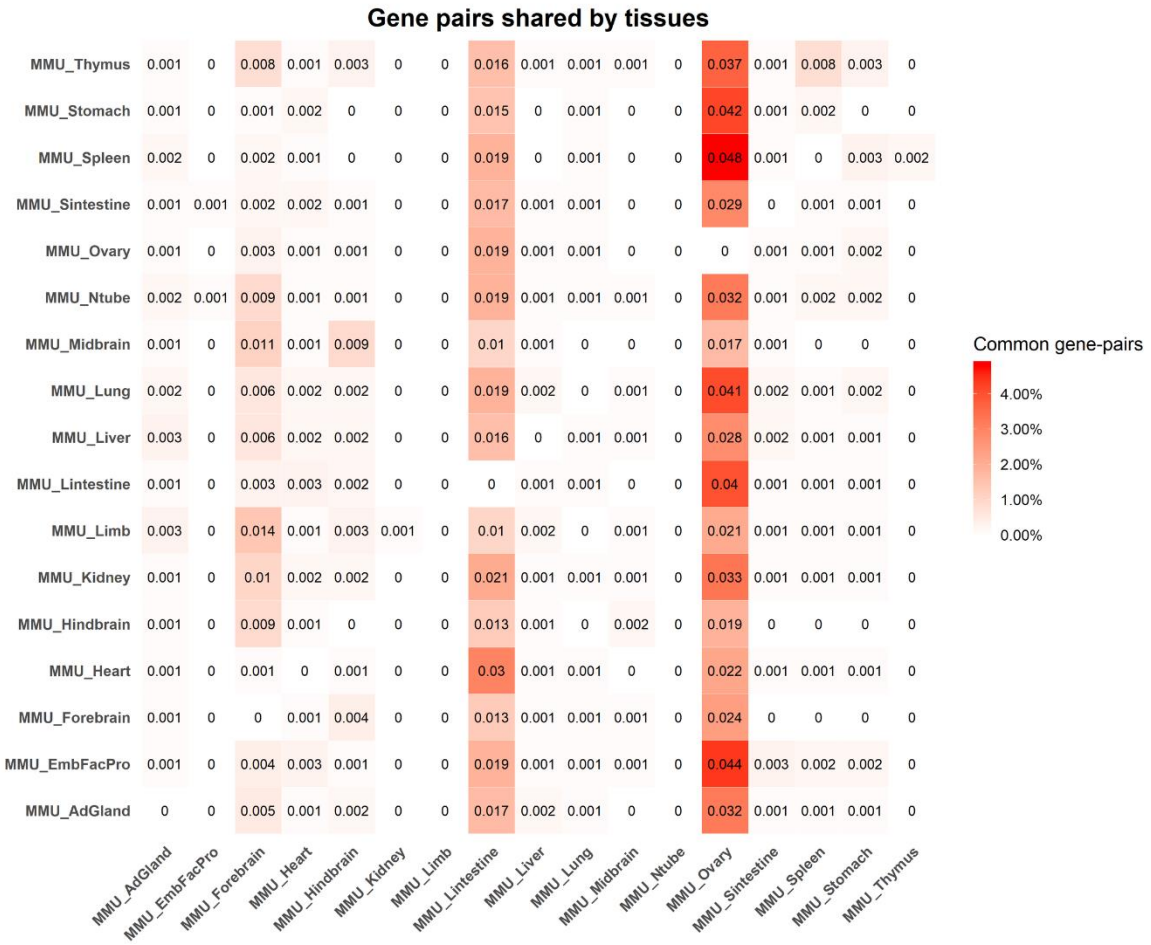


Figure 3.9 **Fraction of gene pairs shared between tissues.** The heatmap represents the fraction of gene pairs shared between two tissues. The numbers shown in the heatmap are not symmetric because the fraction is weighted by total gene pairs in that row's tissue. The fraction is weighted by the total number of pairs in the tissue specified on row. For instance, Midbrain shares 2.9% of all gene pairs present in the midbrain network with hindbrain. Darker shades refer to higher fractions of shared gene pairs. The numbers in the heatmap should be interpreted as reading a matrix rowwise. Abbreviations - AdGland: Adrenal glands; EmbFacPro: Embryonic facial prominence; Sintestine: Small intestine; Lintestine: Large intestine.

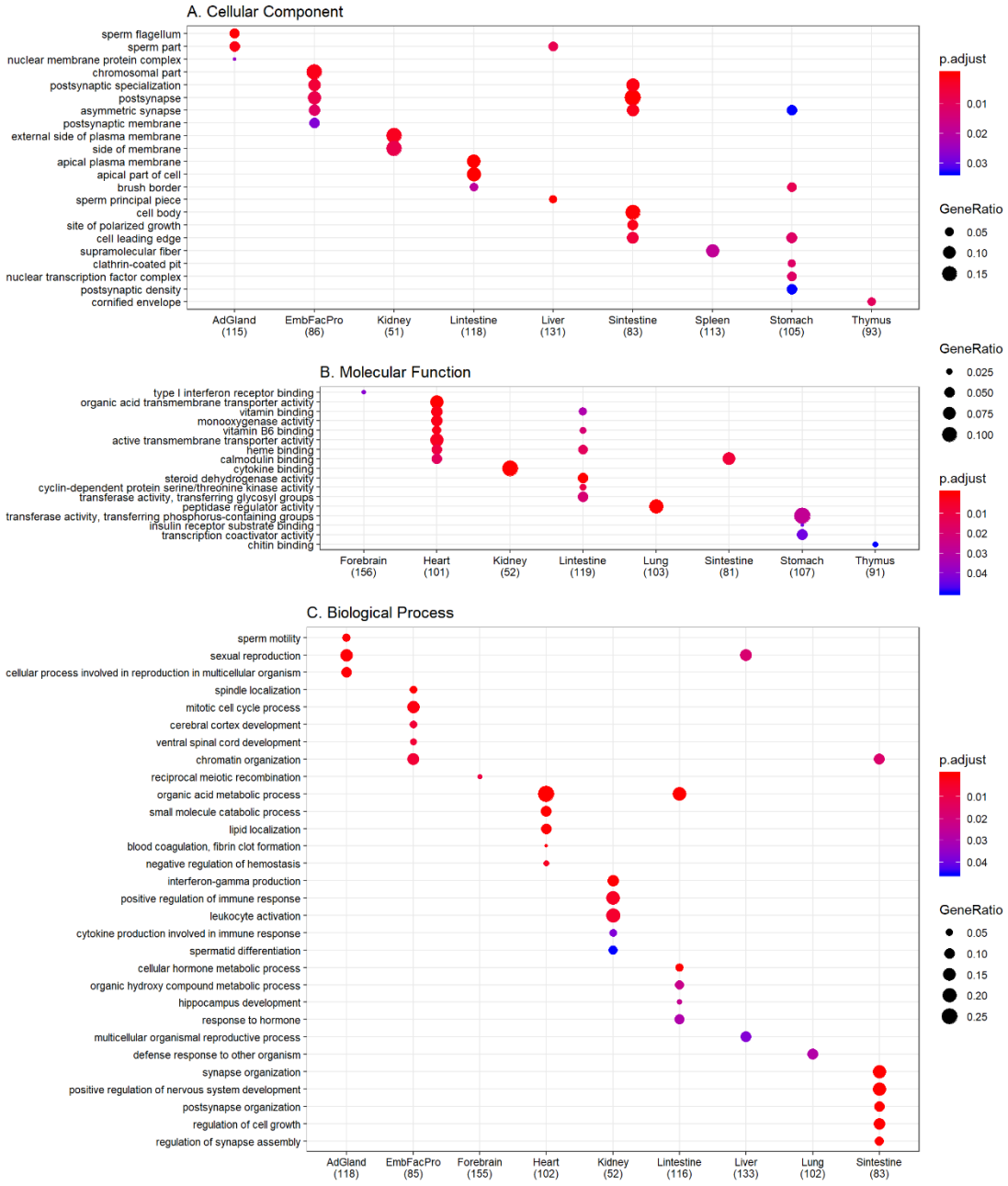


Figure 3.10 **Gene ontology functional enrichment.** Since the functional annotations are at the gene level, we use the central genes identified by both betweenness centrality (top 10%) and degree centrality (top 10%) to perform gene ontology enrichment. Only the top 5 terms for every tissue are shown here. The dot size represents the ratio of genes present in our central genes annotated to a gene ontology term to genes present in our central genes. The color signifies the value of adjusted p-value from false discovery rate control using Benjamini-Hochberg, with lower adjusted p-values shown in darker intensities of red. **A.** Enrichment for cellular component aspect of gene ontology. **B.** Enrichment for molecular function aspect of gene ontology. **C.** Enrichment for biological process aspect of gene ontology. Abbreviations - EmbFacPro: Embryonic facial prominence; Sintestine: Small intestine; Lintestine: Large intestine.

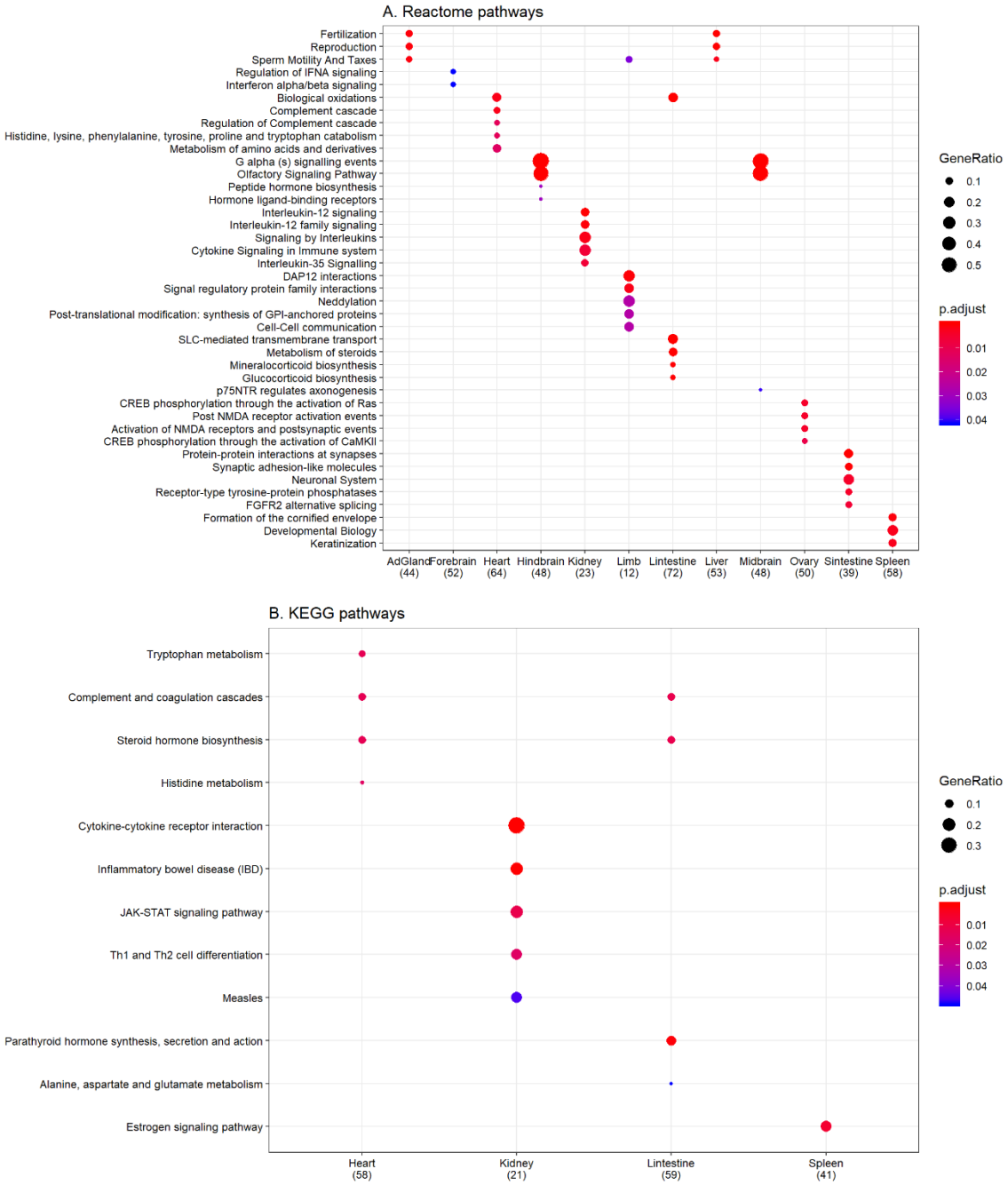


Figure 3.11 **Pathway enrichment analysis.** We use the central genes identified by both betweenness centrality (top 10%) and degree centrality (top 10%) to perform pathway enrichment. Only the top 5 pathways for every tissue are shown here. The dot size represents the ratio of genes present in our central genes annotated to a pathway to genes present in out central genes. The color signifies the value of adjusted p-value from false discovery rate control using Benjamini-Hochberg, with lower adjusted p-values shown in darker intensities of red. **A.** Enrichment for reactome pathways. **B.** Enrichment for KEGG pathways. Abbreviations - KEGG: Kyoto Encyclopedia of Genes and Genomes; AdGland: Adrenal glands; EmbFacPro: Embryonic facial prominence; Sintestine: Small intestine; Lintestine: Large intestine.

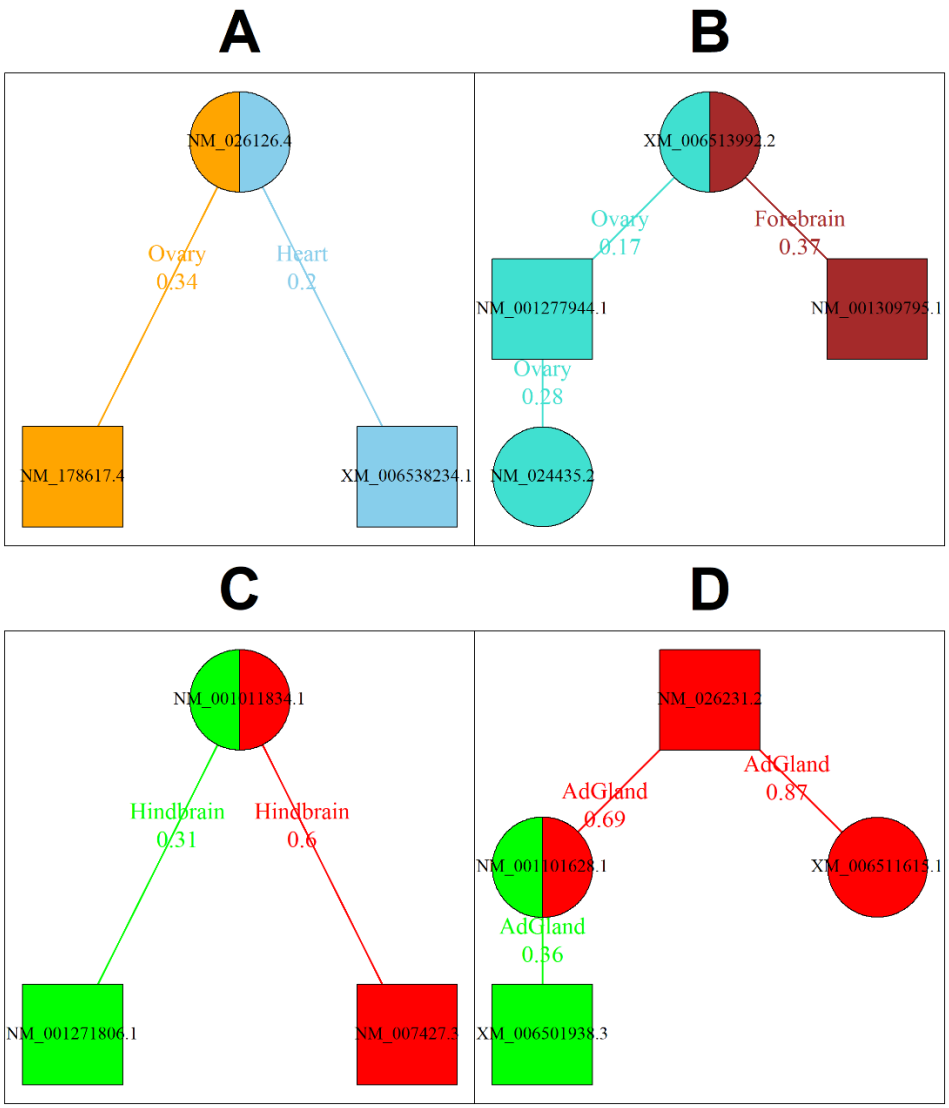


Figure 3.12 **mRNA isoforms of the same gene have different functional partners across tissues.** Examples where the mRNA isoforms of the same gene have different functional/non-functional partners in specific tissues. The mRNA isoforms of the same gene are represented in same shape. The node color, edge color and the edge label color are encoded based on the tissue for part A and B. Functional pairs have green, while non-functional pairs have red node color, edge color and edge label color in parts C and D. Lower edge weight reflects higher strength of functional mRNA isoform pair. **A.** The mRNA isoform NM_030678.3 of gene Gys1 forms a functional pair with different mRNA isoforms of Wap gene in hindbrain and midbrain. **B.** The ovary enriched mRNA isoform NM_001327860.1 of gene Magohb forms a functional pair with another ovary enriched Tbcb mRNA isoform NM_025548.3 in ovary. Other Magohb mRNA isoform NM_025564.2 is preferred in large intestine. **C.** The Chchd2 mRNA isoform NM_024166.6 forms a functional pair with Tktl2 mRNA isoform NM_001271574.1 in hindbrain while the other pair involving Tktl2 mRNA isoform NM_028927.3 is non-functional in hindbrain. **D.** The gene pair Scgb1b30 and Pou4f1 result in four mRNA isoform pairs of which two pairs are functional within hindbrain and one is non-functional in hindbrain.

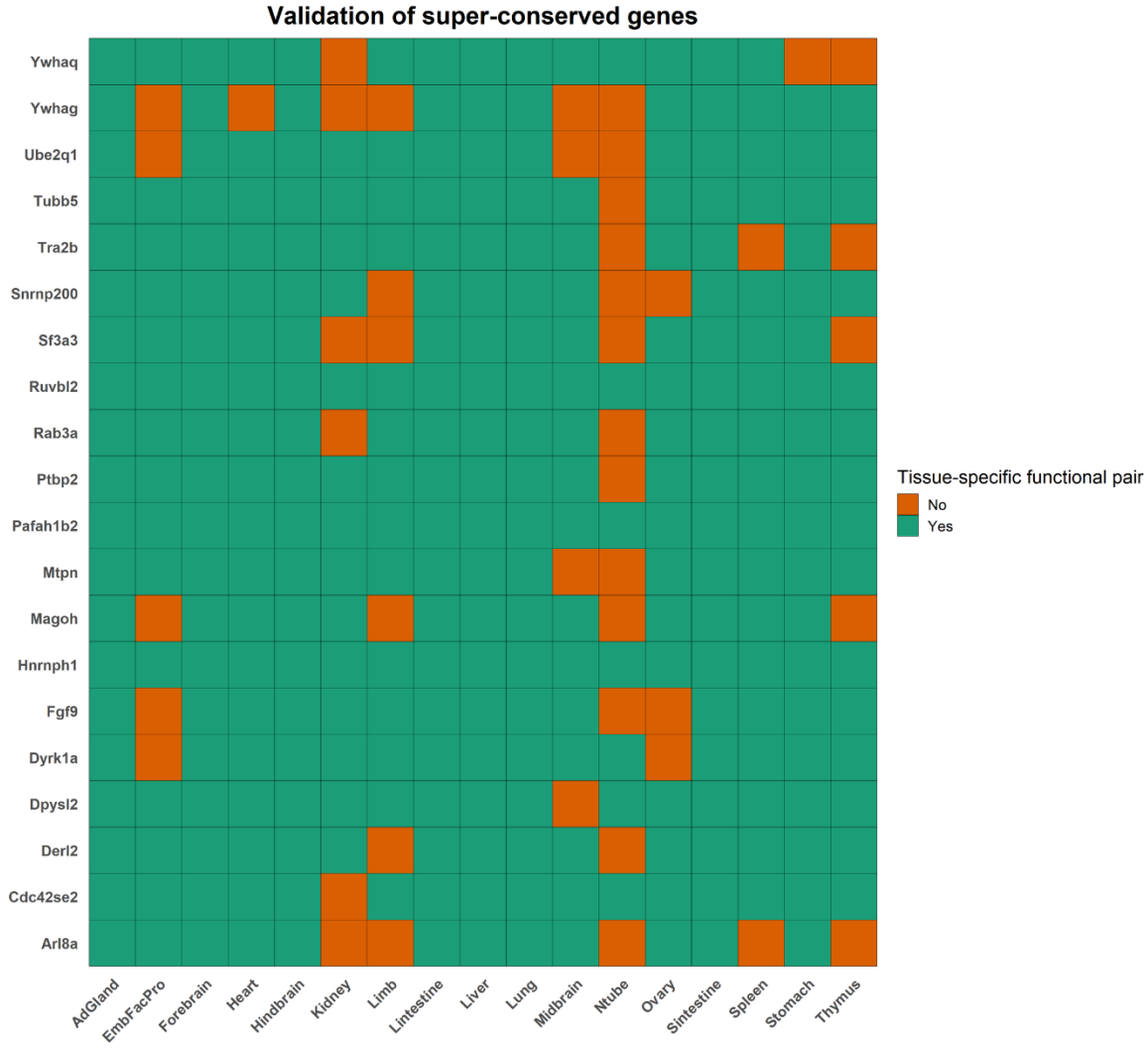


Figure 3.13 **Validation of super-conserved genes.** A heatmap showing the presence or absence of a tissue-specific functional interaction for the 20 super-conserved genes. The genes are on the y-axis and the tissues are on the x-axis. If a gene has a tissue-specific functional interaction, the corresponding block is filled green, or orange otherwise. Abbreviations - AdGland: Adrenal glands; EmbFacPro: Embryonic Facial Prominence; Lintestine: Large intestine; Ntube: Neural tube; Sintestine: Small intestine.

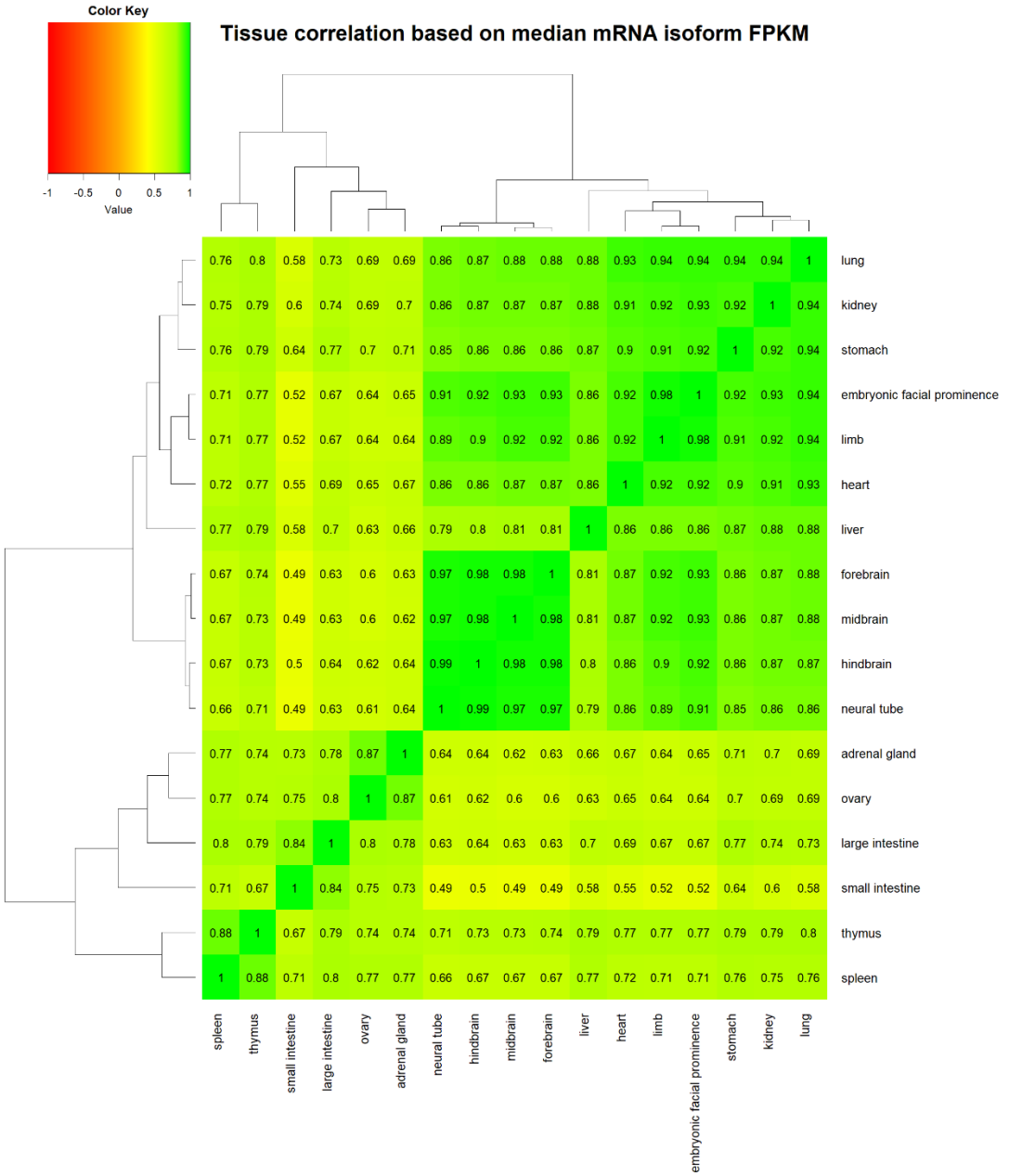


Figure 3.14 **Similar tissues have similar mRNA isoform expression profile.** A heatmap showing the Pearson correlation coefficient between pairs of tissue based on the median mRNA isoform expression values. The dendrogram on the rows and columns reflects the clustering of tissues. Green represents higher positive correlation between a pair of tissue while red reflects higher negative correlation. Similar tissues can be seen being clustered together.

Table 3.1 A list of all mRNA and protein level feature types used in this study.

| Level | Entity | Feature Type |
|---|---|---|
| Sequence | mRNA | 3-mers |
| | | 4-mers |
| | | 5-mers |
| | | 6-mers |
| | Protein | Amino acid composition (1-mer) |
| | | Di-amino acid composition (2-mer) |
| | | Conjoint Triad Descriptors |
| | | Pseudo-amino acid composition |
| | | Moran autocorrelation |
| Expression | mRNA | Heart |
| | | Liver |
| | | Kidney |
| | | Adrenal Glands (AdGland) |
| | | Forebrain |
| | | Midbrain |
| | | Hindbrain |
| | | Embryonic facial prominence (EmbFacPro) |
| | | Large intestine (Lintestine) |
| | | Small intestine (Sintestine) |
| | | Lung |
| | | Limb |
| | | Neural tube (Ntube) |
| | | Ovary |
| | | Spleen |
| | | Stomach |
| | | Thymus |
| | | Organism-wide |

Table 3.2 Prediction performance metrics for TENSION on the original testing dataset with all 27 features

| **Metric** | **Value** |
|---|---|
| Accuracy | 0.802 |
| Area Under the Receiver Operating Characteristic Curve (AUROC) | 0.888 |
| Area Under Precision-Recall Curve (AUPRC) | 0.892 |
| Precision | 0.814 |
| Recall | 0.783 |
| F1 score | 0.798 |
| Matthews Correlation Coefficient (MCC) | 0.604 |

Table 3.3 Confusion matrix for predictions on validation set

| | **Predicted Label** | | |
|---|---|---|---|
| **True Label ↓** | Functional | Non-Functional | |
| Functional | 52515 | 29055 | **81570 (64.4%)** |
| Non-Functional | 16263 | 36634 | **52897 (69.3%)** |
| | **68778** | **65689** | |

Table 3.4 Summary statistics for mRNA isoform level single tissue functional networks

| **Tissue** | **mRNA isoforms (Nodes)** | **mRNA isoform pairs (Edges)** | **Density** | **Clusters (Connected Component)** | **Largest connected component size** |
|---|---|---|---|---|---|
| Neural tube | 9929 | 5546 | 0.000113 | 4412 | 16 |
| Limb | 10130 | 5714 | 0.000111 | 4418 | 18 |
| Kidney | 15763 | 9666 | 7.78E-05 | 6102 | 35 |
| Embryonic facial prominence | 17864 | 12456 | 7.81E-05 | 5427 | 2328 |
| Stomach | 20546 | 15001 | 7.11E-05 | 5732 | 4678 |
| Heart | 19392 | 15294 | 8.13E-05 | 4819 | 6755 |
| Lung | 20994 | 15803 | 7.17E-05 | 5311 | 5724 |
| Spleen | 22152 | 18487 | 7.54E-05 | 4329 | 10140 |
| Small intestine | 33000 | 33649 | 6.18E-05 | 3923 | 22806 |
| Thymus | 35129 | 48056 | 7.79E-05 | 1798 | 31104 |
| Adrenal gland | 36883 | 59421 | 8.74E-05 | 1949 | 32469 |
| Forebrain | 43340 | 72749 | 7.75E-05 | 1705 | 39498 |
| Midbrain | 43179 | 95159 | 0.000102 | 954 | 41146 |
| Liver | 46613 | 145619 | 0.000134 | 1075 | 44262 |
| Large intestine | 49709 | 383029 | 0.00031 | 302 | 49075 |
| Ovary | 45441 | 429726 | 0.000416 | 363 | 44702 |
| Hindbrain | 75286 | 1145680 | 0.000404 | 1 | 75286 |

Table 3.5 Summary statistics for gene level functional networks

| Tissue | Genes (Edges) | Gene pairs (Edges) | Tissue specific gene pairs | Density | Clusters | Largest connected component size | Gene pairs shared with other tissues |
|---|---|---|---|---|---|---|---|
| Neural tube | 7316 | 5523 | 5339 | 2.00E-04 | 2006 | 190 | 184 |
| Limb | 7373 | 5691 | 5474 | 2.01E-04 | 1907 | 1268 | 217 |
| Kidney | 10157 | 9638 | 9364 | 1.82E-04 | 1336 | 6742 | 274 |
| Embryonic facial prominence | 10969 | 12389 | 12003 | 2.00E-04 | 943 | 8827 | 386 |
| Stomach | 11789 | 14929 | 14491 | 2.09E-04 | 699 | 10212 | 438 |
| Heart | 11582 | 15161 | 14653 | 2.18E-04 | 743 | 9932 | 508 |
| Lung | 11958 | 15712 | 15220 | 2.13E-04 | 670 | 10531 | 492 |
| Spleen | 12255 | 18345 | 17780 | 2.37E-04 | 546 | 11124 | 565 |
| Small intestine | 15281 | 33322 | 32354 | 2.77E-04 | 257 | 14789 | 968 |
| Thymus | 15389 | 47307 | 45990 | 3.88E-04 | 150 | 15127 | 1317 |
| Adrenal gland | 16287 | 57805 | 56249 | 4.24E-04 | 160 | 16002 | 1556 |
| Forebrain | 17377 | 69504 | 67469 | 4.47E-04 | 87 | 17226 | 2035 |
| Midbrain | 17209 | 88300 | 84820 | 5.73E-04 | 72 | 17094 | 3480 |
| Liver | 18370 | 135667 | 132669 | 7.86E-04 | 45 | 18302 | 2998 |
| Large intestine | 17877 | 367275 | 359311 | 2.25E-03 | 43 | 17803 | 7964 |
| Ovary | 17413 | 404513 | 396133 | 2.61E-03 | 54 | 17322 | 8380 |
| Hindbrain | 21697 | 975020 | 961759 | 4.09E-03 | 2 | 21696 | 13261 |

Table 3.6 Summary statistics for single tissue mRNA isoform level non-functional networks

| **Tissue** | **mRNA isoforms (Nodes)** | **mRNA isoform pairs (Edges)** | **Density** | **Clusters** | **Largest connected component size** |
|---|---|---|---|---|---|
| Limb | 17614 | 10337 | 6.66E-05 | 7290 | 18 |
| Embryonic facial prominence | 18796 | 11176 | 6.33E-05 | 7623 | 21 |
| Heart | 20313 | 12259 | 5.94E-05 | 8104 | 24 |
| Lung | 20318 | 12347 | 5.98E-05 | 7989 | 17 |
| Neural tube | 21200 | 12925 | 5.75E-05 | 8321 | 22 |
| Kidney | 22762 | 14194 | 5.48E-05 | 8581 | 30 |
| Thymus | 24273 | 15815 | 5.37E-05 | 8503 | 1306 |
| Midbrain | 25537 | 16291 | 5.00E-05 | 9263 | 35 |
| Stomach | 27744 | 19624 | 5.10E-05 | 8124 | 137 |
| Ovary | 31642 | 23926 | 4.78E-05 | 7752 | 5845 |
| Large intestine | 34276 | 26208 | 4.46E-05 | 8081 | 4116 |
| Liver | 35905 | 30592 | 4.75E-05 | 6456 | 17265 |
| Spleen | 28989 | 30877 | 7.35E-05 | 5668 | 14706 |
| Adrenal gland | 35319 | 41155 | 6.60E-05 | 4981 | 22738 |
| Forebrain | 45509 | 59838 | 5.78E-05 | 2462 | 39862 |
| Small intestine | 53086 | 119726 | 8.50E-05 | 861 | 51262 |
| Hindbrain | 72220 | 3380521 | 0.001296 | 8 | 72206 |

Table 3.7 Summary statistics for single tissue gene level non-functional networks

| Tissue | Genes (Nodes) | Gene pairs (Edges) | Tissue specific gene pairs | Density | Clusters | Largest connected component size | Gene pairs shared with other tissues |
|---|---|---|---|---|---|---|---|
| Limb | 10389 | 10294 | 9794 | 1.82E-04 | 1252 | 7210 | 500 |
| Embryonic facial prominence | 10716 | 11137 | 10641 | 1.85E-04 | 1112 | 8088 | 496 |
| Heart | 11235 | 12194 | 11620 | 1.84E-04 | 1003 | 8913 | 574 |
| Lung | 11147 | 12280 | 11711 | 1.89E-04 | 973 | 8890 | 569 |
| Neural tube | 11484 | 12871 | 12245 | 1.86E-04 | 997 | 9255 | 626 |
| Kidney | 11864 | 14129 | 13444 | 1.91E-04 | 794 | 10109 | 685 |
| Thymus | 12449 | 15738 | 14956 | 1.93E-04 | 760 | 10777 | 782 |
| Midbrain | 12813 | 16214 | 15433 | 1.88E-04 | 738 | 11226 | 781 |
| Stomach | 12960 | 19516 | 18508 | 2.20E-04 | 502 | 11937 | 1008 |
| Ovary | 14219 | 23834 | 22868 | 2.26E-04 | 409 | 13420 | 966 |
| Large intestine | 14870 | 26102 | 25004 | 2.26E-04 | 340 | 14228 | 1098 |
| Liver | 15298 | 30030 | 28755 | 2.46E-04 | 281 | 14789 | 1275 |
| Spleen | 13170 | 30341 | 28790 | 3.32E-04 | 448 | 12270 | 1551 |
| Adrenal gland | 15008 | 40312 | 38528 | 3.42E-04 | 264 | 14513 | 1784 |
| Forebrain | 16726 | 57735 | 53353 | 3.81E-04 | 124 | 16538 | 4382 |
| Small intestine | 17532 | 114283 | 105327 | 6.85E-04 | 82 | 17417 | 8956 |
| Hindbrain | 20900 | 2458032 | 2435986 | 1.12E-02 | 2 | 20898 | 22046 |

## References

Alonso-López, D., Gutiérrez, M. A., Lopes, K. P., Prieto, C., Santamaría, R., & De Las Rivas, J. (2016). APID interactomes: Providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Research, 44(W1), W529–W535. https://doi.org/10.1093/nar/gkw363

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57(1), 289–300. https://doi.org/10.2307/2346101

Berthier, C. C., Zhang, H., Schin, M., Henger, A., Nelson, R. G., Yee, B., … Kretzler, M. (2009). Enhanced expression of janus kinase-signal transducer and activator of transcription pathway members in human diabetic nephropathy. Diabetes, 58(2), 469–477. https://doi.org/10.2337/db08-1328

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brosius, F. C., & He, J. C. (2015, January). JAK inhibition and progressive kidney disease. Current Opinion in Nephrology and Hypertension. https://doi.org/10.1097/MNH.0000000000000079

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. Molecular Cell, 46(6), 871–883. https://doi.org/10.1016/j.molcel.2012.05.039

Calderone, A., Castagnoli, L., & Cesareni, G. (2013, August 1). Mentha: A resource for browsing integrated protein-interaction networks. Nature Methods. Nature Publishing Group. https://doi.org/10.1038/nmeth.2561

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., … Tyers, M. (2017). The BioGRID interaction database: 2017 update. Nucleic Acids Research, 45(D1), D369–D379. https://doi.org/10.1093/nar/gkw1102

Chen, K.-F., & Crowther, D. C. (2012). Functional genomics in Drosophila models of human disease. Briefings in Functional Genomics, 11(5), 405–415. https://doi.org/10.1093/bfgp/els038

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function and Genetics, 43(3), 246–255. https://doi.org/10.1002/prot.1035

Chuang, P. Y., & He, J. C. (2010, August 1). JAK/STAT signaling in renal diseases. Kidney International. Elsevier. https://doi.org/10.1038/ki.2010.158

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695(5), 1–9.

Della Torre, S., Rando, G., Meda, C., Stell, A., Chambon, P., Krust, A., … Maggi, A. (2011). Amino acid-dependent activation of liver estrogen receptor alpha integrates metabolic and reproductive functions via IGF-1. Cell Metabolism, 13(2), 205–214. https://doi.org/10.1016/j.cmet.2011.01.002

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Du, X., Hu, C., Yao, Y., Sun, S., & Zhang, Y. (2017). Analysis and prediction of exon skipping events from RNA-seq with sequence information using rotation forest. International Journal of Molecular Sciences, 18(12). https://doi.org/10.3390/ijms18122691

Eksi, R., Li, H. D., Menon, R., Wen, Y., Omenn, G. S., Kretzler, M., & Guan, Y. (2013). Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. PLoS Computational Biology, 9(11). https://doi.org/10.1371/journal.pcbi.1003314

Ellis, J. D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T. H., Calarco, J. A., … Blencowe, B. J. (2012). Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. Molecular Cell, 46(6), 884–892. https://doi.org/10.1016/j.molcel.2012.05.037

Fontana, R., & Della Torre, S. (2016, February 11). The deep correlation between energy metabolism and reproduction: A view on the effects of nutrition for women fertility. Nutrients. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/nu8020087

Garcia, L., Hinojosa, L., Dominguez, R., Chavira, R., & Rosas, P. (2000). Effects of infantile thymectomy on ovarian functions and gonadotrophin-induced ovulation in prepubertal mice: Role of thymulin. Journal of Endocrinology, 166(2), 381–387. https://doi.org/10.1677/joe.0.1660381

Kandoi, G., Acencio, M. L., & Lemke, N. (2015). Prediction of druggable proteins using machine learning and systems biology: A mini-review. Frontiers in Physiology. https://doi.org/10.3389/fphys.2015.00366

Kandoi, G., & Dickerson, J. A. (2018). Tissue-spEcific mrNa iSoform functIOnal Networks (TENSION) Collection. figshare. https://doi.org/10.25380/iastate.c.4275191

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013, February 1). Function of alternative splicing. Gene. https://doi.org/10.1016/j.gene.2012.07.083

Kohavi, R. (1995). A study of Cross validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conference on Neural Networks (Vol. 2, pp. 1137–1143). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.529

Kotlyar, M., Pastrello, C., Sheahan, N., & Jurisica, I. (2016). Integrated interactions database: Tissue-specific view of the human and model organism interactomes. Nucleic Acids Research, 44(D1), D536–D541. https://doi.org/10.1093/nar/gkv1115

Li, B., Qing, T., Zhu, J., Wen, Z., Yu, Y., Fukumura, R., … Shi, L. (2017). A Comprehensive Mouse Transcriptomic BodyMap across 17 Tissues by RNA-seq. Scientific Reports, 7(1), 4200. https://doi.org/10.1038/s41598-017-04520-z

Li, H.-D. D., Menon, R., Eksi, R., Guerler, A., Zhang, Y., Omenn, G. S., & Guan, Y. (2016). A Network of Splice Isoforms for the Mouse. Scientific Reports, 6(April), 1–11. https://doi.org/10.1038/srep24507

Li, H. D., Menon, R., Govindarajoo, B., Panwar, B., Zhang, Y., Omenn, G. S., & Guan, Y. (2015). Functional networks of highest-connected splice isoforms: From the chromosome 17 human proteome project. Journal of Proteome Research, 14(9), 3484–3491. https://doi.org/10.1021/acs.jproteome.5b00494

Li, H. D., Menon, R., Omenn, G. S., & Guan, Y. (2014). The emerging era of genomic data integration for analyzing splice isoform function. Trends in Genetics, 30(8), 340–347. https://doi.org/10.1016/j.tig.2014.05.005

Li, W., Kang, S., Liu, C. C., Zhang, S., Shi, Y., Liu, Y., & Zhou, X. J. (2014). High-resolution functional annotation of human transcriptome: Predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Research, 42(6), e39–e39. https://doi.org/10.1093/nar/gkt1362

Liu, R., Loraine, A. E., & Dickerson, J. A. (2014). Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. BMC Bioinformatics, 15(1). https://doi.org/10.1186/s12859-014-0364-4

Luo, T., Zhang, W., Qiu, S., Yang, Y., Yi, D., Wang, G., … Wang, J. (2017). Functional Annotation of Human Protein Coding Isoforms via Non-convex Multi-Instance Learning. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17, 345–354. https://doi.org/10.1145/3097983.3097984

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Research, 22(6), 1184–1195. https://doi.org/10.1101/gr.134106.111

Michael, S. D. (1979). The role of the endocrine thymus in female reproduction. Arthritis and Rheumatism, 22(11), 1241–1245. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/508374

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika, 37(1), 17–23. https://doi.org/10.2307/2332142

Nezami, B. G., & Srinivasan, S. (2010, October). Enteric nervous system in the small intestine: Pathophysiology and clinical implications. Current Gastroenterology Reports. NIH Public Access. https://doi.org/10.1007/s11894-010-0129-9

Oikonomopoulou, K., Ricklin, D., Ward, P. A., & Lambris, J. D. (2012, January). Interactions between coagulation and complement - Their role in inflammation. Seminars in Immunopathology. NIH Public Access. https://doi.org/10.1007/s00281-011-0280-x

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., … Hermjakob, H. (2014). The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Research, 42(D1), D358–D363. https://doi.org/10.1093/nar/gkt1115

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics, 40(12), 1413–1415. https://doi.org/10.1038/ng.259

Pan, Q., Shai, O., Misquitta, C., Zhang, W., Saltzman, A. L., Mohammad, N., … Blencowe, B. J. (2004). Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. Molecular Cell, 16(6), 929–941. https://doi.org/10.1016/j.molcel.2004.12.004

Panwar, B., Menon, R., Eksi, R., Li, H.-D., Omenn, G. S., & Guan, Y. (2016). Genome-Wide Functional Annotation of Human Protein-Coding Splice Variants Using Multiple Instance Learning. Journal of Proteome Research, 15(6), 1747–1753. https://doi.org/10.1021/acs.jproteome.5b00883

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res., 12, 2825–2830. Retrieved from http://dl.acm.org/citation.cfm?id=1953048.2078195

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protocols, 11(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Radhakrishnan, V. M., Kojs, P., Ramalingam, R., Midura-Kiela, M. T., Angeli, P., Kiela, P. R., & Ghishan, F. K. (2015). Experimental colitis is associated with transcriptional inhibition of Na+/Ca2+ exchanger isoform 1 (NCX1) expression by interferon γ in the renal distal convoluted tubules. Journal of Biological Chemistry, 290(14), 8964–8974. https://doi.org/10.1074/jbc.M114.616516

Raj, B., & Blencowe, B. J. (2015, July 1). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron. Cell Press. https://doi.org/10.1016/j.neuron.2015.05.004

Rao, M., & Gershon, M. D. (2016, September 20). The bowel and beyond: The enteric nervous system in neurological disorders. Nature Reviews Gastroenterology and Hepatology. Nature Publishing Group. https://doi.org/10.1038/nrgastro.2016.107

Resch, A., Xing, Y., Modrek, B., Gorlick, M., Riley, R., & Lee, C. (2004). Assessing the Impact of Alternative Splicing on Domain Interactions in the Human Proteome. Journal of Proteome Research, 3(1), 76–83. https://doi.org/10.1021/pr034064v

Rimm, E. B., Stampfer, M. J., Ascherio, A., Giovannucci, E., Colditz, G. A., & Willett, W. C. (1993). Vitamin E Consumption and the Risk of Coronary Heart Disease in Men. New England Journal of Medicine, 328(20), 1450–1456. https://doi.org/10.1056/NEJM199305203282004

Rimm, E. B., Willett, W. C., Hu, F. B., Sampson, L., Colditz, G. A., Manson, J. A. E., … Stampfer, M. J. (1998). Folate and vitamin B6 from diet and supplements in relation to risk of coronary heart disease among women. Journal of the American Medical Association, 279(5), 359–364. https://doi.org/10.1001/jama.279.5.359

Said, H. M., & Mohammed, Z. M. (2006, March). Intestinal absorption of water-soluble vitamins: An update. Current Opinion in Gastroenterology. https://doi.org/10.1097/01.mog.0000203870.22706.52

Schnyder, G., Roffi, M., Flammer, Y., Pin, R., & Hess, O. M. (2002). Effect of homocysteine-lowering therapy with folic acid, vitamin B12, and vitamin B6 on clinical outcome after percutaneous coronary intervention: The Swiss heart study: A randomized controlled trial. Journal of the American Medical Association, 288(8), 973–979. https://doi.org/10.1001/jama.288.8.973

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., … Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America, 104(11), 4337–4341. https://doi.org/10.1073/pnas.0607879104

Söllner, J. F., Leparc, G., Hildebrandt, T., Klein, H., Thomas, L., Stupka, E., & Simon, E. (2017). An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Scientific Data, 4, 170185. https://doi.org/10.1038/sdata.2017.185

Stephens, N. G., Parsons, A., Schofield, P. M., Kelly, F., Cheeseman, K., Mitchinson, M. J., & Brown, M. J. (1996). Randomised controlled trial of vitamin E in patients with coronary disease: Cambridge Heart Antioxidant Study (CHAOS). Lancet, 347(9004), 781–786. https://doi.org/10.1016/S0140-6736(96)90866-1

Sun, Y., Hou, H., Song, H., Lin, K., Zhang, Z., Hu, J., & Pang, E. (2018). The comparison of alternative splicing among the multiple tissues in cucumber. BMC Plant Biology, 18(1), 5. https://doi.org/10.1186/s12870-017-1217-x

Suzuki, H., Osaki, K., Sano, K., Alam, A. H. M. K., Nakamura, Y., Ishigaki, Y., … Tsukahara, T. (2011). Comprehensive analysis of alternative splicing and functionality in neuronal differentiation of P19 cells. PLoS ONE, 6(2), e16880. https://doi.org/10.1371/journal.pone.0016880

Torre, S. Della, Benedusi, V., Fontana, R., & Maggi, A. (2014, January 22). Energy metabolism and fertility - A balance preserved for female health. Nature Reviews Endocrinology. Nature Publishing Group. https://doi.org/10.1038/nrendo.2013.203

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., … Wilkens-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. Computing in Science and Engineering, 16(5), 62–74. https://doi.org/10.1109/MCSE.2014.80

Tseng, Y. T., Li, W., Chen, C. H., Zhang, S., Chen, J. J. W., Zhou, X. J., & Liu, C. C. (2015). IIIDB: A database for isoform-isoform interactions and isoform network modules. BMC Genomics, 16(Suppl 2), 1–7. https://doi.org/10.1186/1471-2164-16-S2-S10

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., … Ponten, F. (2015). Tissue-based map of the human proteome. Science, 347(6220), 1260419–1260419. https://doi.org/10.1126/science.1260419

Vitulo, N., Forcato, C., Carpinelli, E. C., Telatin, A., Campagna, D., D'Angelo, M., … Valle, G. (2014). A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biology, 14(1), 99. https://doi.org/10.1186/1471-2229-14-99

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics, 23(10), 1274–1281. https://doi.org/10.1093/bioinformatics/btm087

Wei, B., & Jin, J. P. (2016, May 10). TNNT1, TNNT2, and TNNT3: Isoform genes, regulation, and structure-function relationships. Gene. Elsevier. https://doi.org/10.1016/j.gene.2016.01.006

Wood, J. D. (2016, July 3). Enteric Nervous System: Neuropathic Gastrointestinal Motility. Digestive Diseases and Sciences, 61(7), 1803–1816. https://doi.org/10.1007/s10620-016-4183-5

Wu, P., Zhou, D., Lin, W., Li, Y., Wei, H., Qian, X., … He, F. (2018). Cell-type-resolved alternative splicing patterns in mouse liver. DNA Research. https://doi.org/10.1093/dnares/dsx055

Xiao, N., Cao, D. S., Zhu, M. F., & Xu, Q. S. (2015). Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. In Bioinformatics (Vol. 31, pp. 1857–1859). https://doi.org/10.1093/bioinformatics/btv042

Xu, Q., Modrek, B., & Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Research, 30(17), 3754–3766. https://doi.org/10.1093/nar/gkf492

Yang, N., Luo, M., Li, R., Huang, Y., Zhang, R., Wu, Q., … Yu, X. (2008). Blockage of JAK/STAT signalling attenuates renal ischaemia-reperfusion injury in rat. Nephrology Dialysis Transplantation, 23(1), 91–100. https://doi.org/10.1093/ndt/gfm509

Yu, G., & He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. Mol. BioSyst., 12(2), 477–479. https://doi.org/10.1039/C5MB00663E

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS: A Journal of Integrative Biology, 16(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Yura, K., Shionyu, M., Hagino, K., Hijikata, A., Hirashima, Y., Nakahara, T., … Go, M. (2006). Alternative splicing in human transcriptome: Functional and structural influence on proteins. Gene, 380(2), 63–71. https://doi.org/10.1016/j.gene.2006.05.015

Zhu, M., Dong, J., & Cao, D. (2016). rDNAse: Generating Various Numerical Representation Schemes of DNA Sequences. Retrieved from https://cran.r-project.org/package=rDNAse

Zittermann, A., Schulze Schleithoff, S., Tenderich, G., Berthold, H. K., Körfer, R., & Stehle, P. (2003). Low vitamin D status: A contributing factor in the pathogenesis of congestive heart failure? Journal of the American College of Cardiology. https://doi.org/10.1016/S0735-1097(02)02624-4

**Appendix. Supplementary material for Chapter 3**

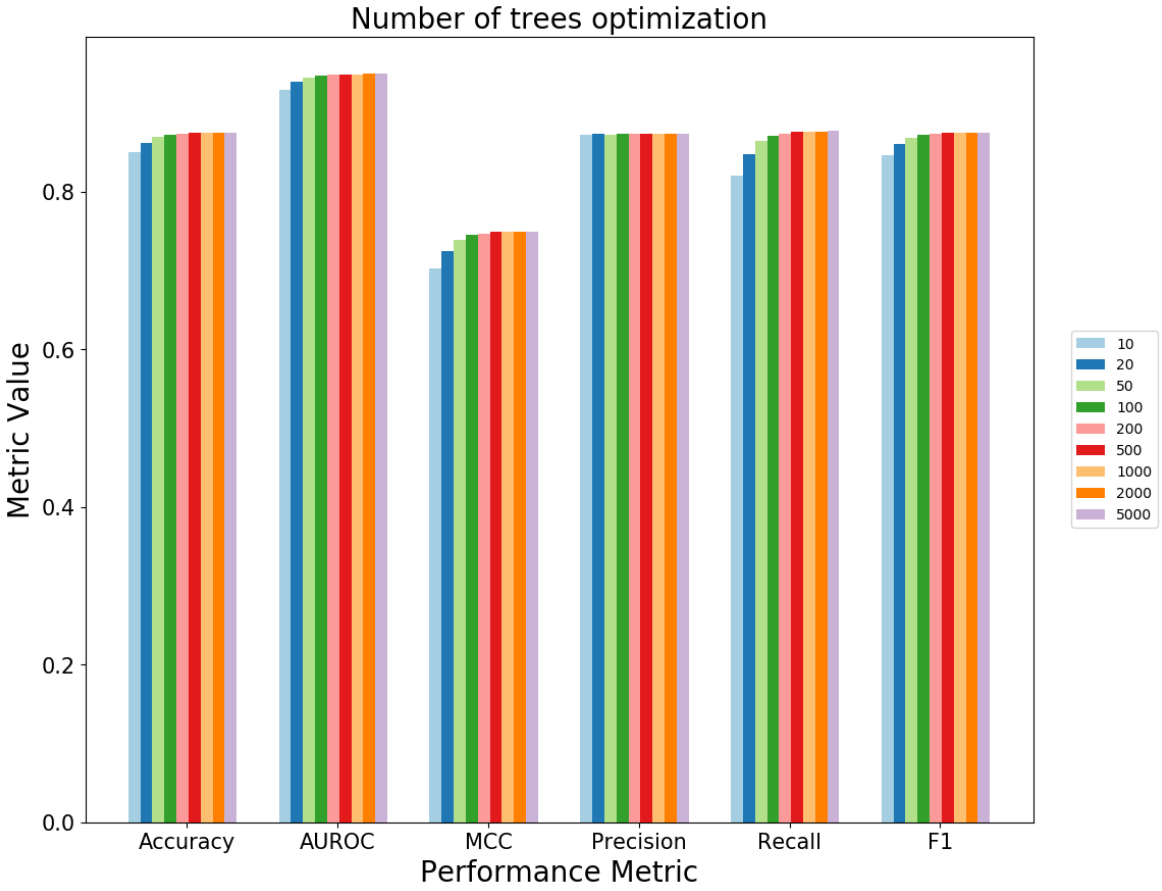


Figure A.1 **Optimization of number of trees in random forest.** A bar plot of performance metrics computed at different number of trees used in random forest. There is very little improvement in the performance beyond 100 trees, therefore, we have used 100 trees while developing TENSION. Abbreviations - ROC: Receiver Operating Characteristic; MCC: Matthews Correlation Coefficient.

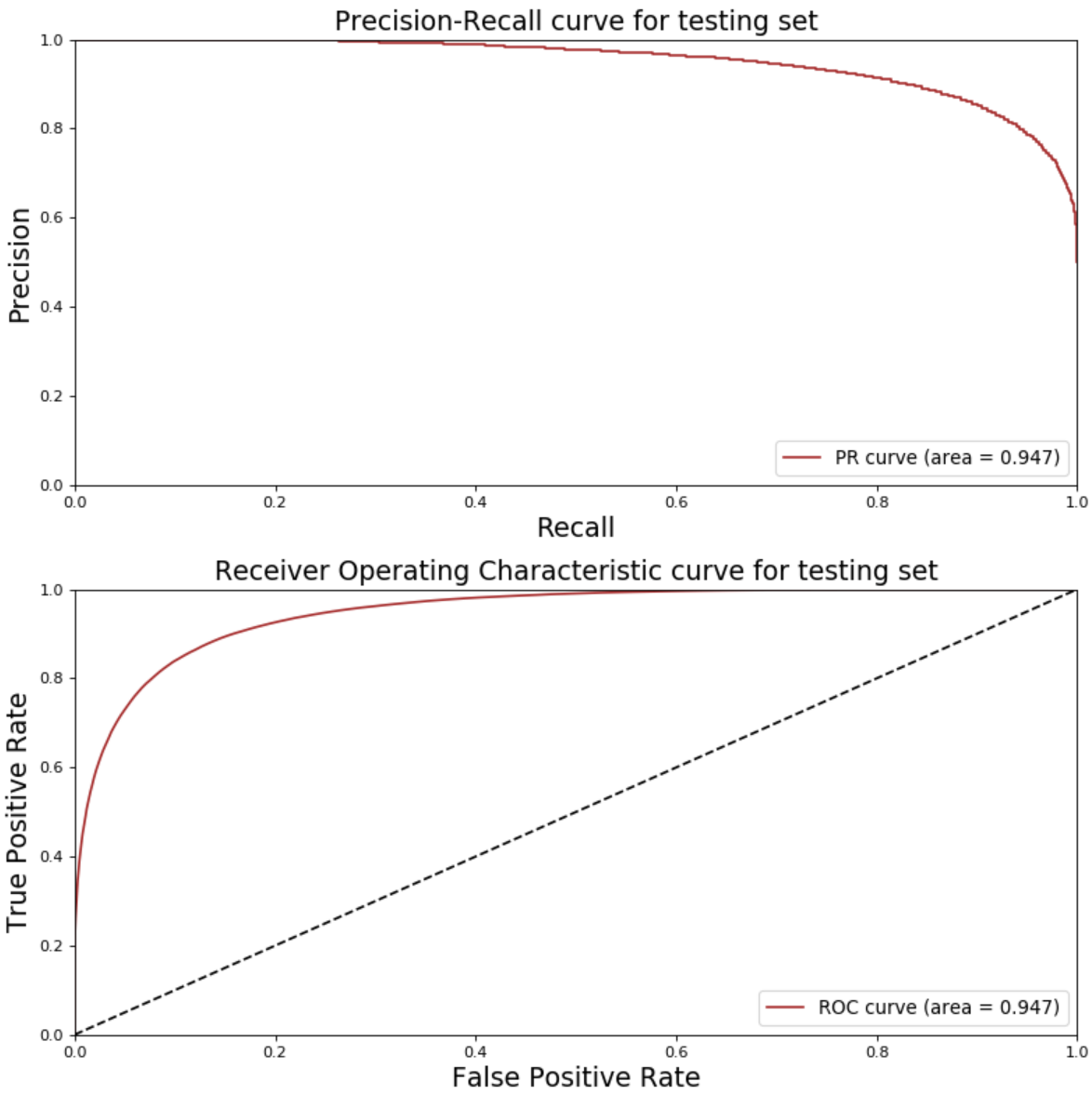


Figure A.2 **Performance evaluation on original testing dataset.** The precision-recall and receiver operating characteristic curve for predictions on the original testing dataset. A model with area under the curve closer to 1 is better while a model with an area under the curve of 0.5 is equivalent to making random guess. Abbreviations - PR: Precision-Recall; ROC: Receiver Operating Characteristic.

Table A.1 **List of RNA-Seq experiments and their tissue.**

| Sample | Tissue Type |
|---|---|
| ENCFF016KLR | adrenal gland |
| ENCFF360XMZ | adrenal gland |
| ENCFF694UNH | adrenal gland |
| ENCFF867HND | adrenal gland |
| SRR5047957 | adrenal gland |
| SRR5047958 | adrenal gland |
| SRR5047959 | adrenal gland |
| SRR5047960 | adrenal gland |
| SRR5047961 | adrenal gland |
| SRR5047962 | adrenal gland |
| SRR5048019 | brain |
| SRR5048020 | brain |
| SRR5048015 | central nervous system |
| SRR5048016 | central nervous system |
| SRR5048023 | central nervous system |
| SRR5048024 | central nervous system |
| SRR5048027 | central nervous system |
| SRR5048028 | central nervous system |
| SRR5048025 | hindbrain |
| SRR5048026 | hindbrain |
| SRR5047913 | large intestine |
| SRR5047914 | large intestine |
| SRR5047915 | large intestine |
| SRR5047916 | large intestine |
| SRR5047917 | large intestine |
| SRR5047918 | large intestine |
| SRR5048041 | forebrain |
| SRR5048042 | forebrain |
| ENCFF053CRD | embryonic facial prominence |
| ENCFF061AVT | embryonic facial prominence |
| ENCFF249ZZI | embryonic facial prominence |
| ENCFF252QAP | embryonic facial prominence |
| ENCFF316SZZ | embryonic facial prominence |
| ENCFF360GSG | embryonic facial prominence |
| ENCFF427WPK | embryonic facial prominence |
| ENCFF500GLW | embryonic facial prominence |
| ENCFF528UUE | embryonic facial prominence |
| ENCFF551UCM | embryonic facial prominence |
| ENCFF557LMN | embryonic facial prominence |
| ENCFF576MIX | embryonic facial prominence |

| | |
|---|---|
| ENCFF599OTY | embryonic facial prominence |
| ENCFF709FPA | embryonic facial prominence |
| ENCFF714ZDW | embryonic facial prominence |
| ENCFF744XBD | embryonic facial prominence |
| ENCFF771GDS | embryonic facial prominence |
| ENCFF781WVF | embryonic facial prominence |
| ENCFF839MMS | embryonic facial prominence |
| ENCFF839UKS | embryonic facial prominence |
| ENCFF896QPV | embryonic facial prominence |
| ENCFF917VEZ | embryonic facial prominence |
| ENCFF037JQC | forebrain |
| ENCFF114DRT | forebrain |
| ENCFF126IRS | forebrain |
| ENCFF179JEC | forebrain |
| ENCFF203BWA | forebrain |
| ENCFF235DNM | forebrain |
| ENCFF251LNG | forebrain |
| ENCFF270GKY | forebrain |
| ENCFF294JRP | forebrain |
| ENCFF320FJX | forebrain |
| ENCFF329ACL | forebrain |
| ENCFF358MFI | forebrain |
| ENCFF447EXU | forebrain |
| ENCFF458NWF | forebrain |
| ENCFF460TCF | forebrain |
| ENCFF528EVC | forebrain |
| ENCFF663SNC | forebrain |
| ENCFF700OLU | forebrain |
| ENCFF748SRJ | forebrain |
| ENCFF891HIX | forebrain |
| ENCFF896COV | forebrain |
| ENCFF920CNZ | forebrain |
| ENCFF920QAY | forebrain |
| ENCFF931IVO | forebrain |
| ENCFF959PSX | forebrain |
| SRR3192667 | forebrain |
| SRR3192668 | forebrain |
| SRR5047970 | gonadal fat pad |
| SRR5047971 | gonadal fat pad |
| SRR5047972 | gonadal fat pad |
| SRR5047973 | gonadal fat pad |
| ENCFF007SHF | heart |
| ENCFF019VGV | heart |

| | |
|---|---|
| ENCFF034XQS | heart |
| ENCFF063CEM | heart |
| ENCFF070OPW | heart |
| ENCFF136TII | heart |
| ENCFF220DJN | heart |
| ENCFF228OAV | heart |
| ENCFF229IFK | heart |
| ENCFF236BUE | heart |
| ENCFF358SEO | heart |
| ENCFF360NRX | heart |
| ENCFF381RLD | heart |
| ENCFF405TRT | heart |
| ENCFF418MSC | heart |
| ENCFF445AIZ | heart |
| ENCFF477PIC | heart |
| ENCFF478ZKL | heart |
| ENCFF485OXT | heart |
| ENCFF490QEC | heart |
| ENCFF500NCE | heart |
| ENCFF503TOH | heart |
| ENCFF586VFP | heart |
| ENCFF637ZYL | heart |
| ENCFF646JUX | heart |
| ENCFF676BDY | heart |
| ENCFF676PFC | heart |
| ENCFF878OGG | heart |
| SRR5047921 | heart |
| SRR5047922 | heart |
| SRR5047923 | heart |
| SRR5047924 | heart |
| ENCFF032XDZ | hindbrain |
| ENCFF060NTC | hindbrain |
| ENCFF094ZGK | hindbrain |
| ENCFF104CQE | hindbrain |
| ENCFF160LUK | hindbrain |
| ENCFF167NXN | hindbrain |
| ENCFF172XPK | hindbrain |
| ENCFF275SGM | hindbrain |
| ENCFF282QML | hindbrain |
| ENCFF336BOI | hindbrain |
| ENCFF372TIL | hindbrain |
| ENCFF377BWR | hindbrain |
| ENCFF378HXV | hindbrain |

| | |
|---|---|
| ENCFF416EWW | hindbrain |
| ENCFF548GUA | hindbrain |
| ENCFF645LDN | hindbrain |
| ENCFF672PAZ | hindbrain |
| ENCFF700EWM | hindbrain |
| ENCFF706XOL | hindbrain |
| ENCFF738FUB | hindbrain |
| ENCFF786VDJ | hindbrain |
| ENCFF863UFD | hindbrain |
| ENCFF874AXO | hindbrain |
| ENCFF876NSY | hindbrain |
| ENCFF913XMQ | hindbrain |
| ENCFF926BFE | hindbrain |
| SRR3192647 | hindbrain |
| SRR3192648 | hindbrain |
| ENCFF014ZBI | intestine |
| ENCFF039KJW | intestine |
| ENCFF093ZAR | intestine |
| ENCFF107WVN | intestine |
| ENCFF113UJZ | intestine |
| ENCFF235BLS | intestine |
| ENCFF316QLU | intestine |
| ENCFF379JZS | intestine |
| ENCFF499WIP | intestine |
| ENCFF553BVK | intestine |
| ENCFF758VQQ | intestine |
| ENCFF904JAW | intestine |
| ENCFF021BPG | kidney |
| ENCFF070BUP | kidney |
| ENCFF140MLD | kidney |
| ENCFF143NRY | kidney |
| ENCFF266SYA | kidney |
| ENCFF301QEB | kidney |
| ENCFF367OPA | kidney |
| ENCFF652UDJ | kidney |
| ENCFF654QAE | kidney |
| ENCFF798JOI | kidney |
| ENCFF901TLF | kidney |
| ENCFF929PSZ | kidney |
| SRR5047925 | kidney |
| SRR5047926 | kidney |
| SRR5047927 | kidney |
| SRR5047928 | kidney |

| | |
|---|---|
| SRR5047929 | kidney |
| SRR5047930 | kidney |
| SRR5047975 | large intestine |
| SRR5047976 | large intestine |
| SRR5047977 | large intestine |
| SRR5047978 | large intestine |
| ENCFF184ELK | limb |
| ENCFF235PJS | limb |
| ENCFF237DCF | limb |
| ENCFF237SXT | limb |
| ENCFF246JLP | limb |
| ENCFF249AZE | limb |
| ENCFF262CIY | limb |
| ENCFF291NWK | limb |
| ENCFF409ZNA | limb |
| ENCFF419QRX | limb |
| ENCFF479HKB | limb |
| ENCFF565KTC | limb |
| ENCFF654OBR | limb |
| ENCFF678XFK | limb |
| ENCFF679RDZ | limb |
| ENCFF682WAX | limb |
| ENCFF775HDI | limb |
| ENCFF780HRS | limb |
| ENCFF820NAK | limb |
| ENCFF959ZAX | limb |
| SRR5048029 | limb |
| SRR5048030 | limb |
| ENCFF085URY | liver |
| ENCFF130DKL | liver |
| ENCFF155LJD | liver |
| ENCFF245OTN | liver |
| ENCFF276ENR | liver |
| ENCFF377KCE | liver |
| ENCFF473WMT | liver |
| ENCFF510RXX | liver |
| ENCFF526QHV | liver |
| ENCFF528MAS | liver |
| ENCFF536HIT | liver |
| ENCFF584CMS | liver |
| ENCFF635VBU | liver |
| ENCFF635YMK | liver |
| ENCFF677ULG | liver |

| | |
|---|---|
| ENCFF746XUK | liver |
| ENCFF810MMJ | liver |
| ENCFF854WTE | liver |
| ENCFF932YNB | liver |
| ENCFF956HCY | liver |
| ENCFF985XAR | liver |
| ENCFF986WFE | liver |
| SRR3192469 | liver |
| SRR3192470 | liver |
| SRR5047931 | liver |
| SRR5047932 | liver |
| SRR5047933 | liver |
| SRR5047934 | liver |
| SRR5047935 | liver |
| SRR5047936 | liver |
| SRR5048017 | liver |
| SRR5048018 | liver |
| SRR5048021 | liver |
| SRR5048022 | liver |
| SRR5048031 | liver |
| SRR5048032 | liver |
| ENCFF289EZB | lung |
| ENCFF503BOB | lung |
| ENCFF618MYZ | lung |
| ENCFF657LQI | lung |
| ENCFF728LAM | lung |
| ENCFF778KZE | lung |
| ENCFF800SJE | lung |
| ENCFF910RNP | lung |
| ENCFF916XIQ | lung |
| ENCFF919VVI | lung |
| SRR5047937 | lung |
| SRR5047938 | lung |
| SRR5047939 | lung |
| SRR5047940 | lung |
| ENCFF051VXS | midbrain |
| ENCFF052VGB | midbrain |
| ENCFF059FUK | midbrain |
| ENCFF062HAD | midbrain |
| ENCFF091YUW | midbrain |
| ENCFF093YSD | midbrain |
| ENCFF099UUS | midbrain |
| ENCFF156BSL | midbrain |

| | |
|---|---|
| ENCFF327YJQ | midbrain |
| ENCFF348BYM | midbrain |
| ENCFF421QJA | midbrain |
| ENCFF476UXE | midbrain |
| ENCFF499UQZ | midbrain |
| ENCFF727ACE | midbrain |
| ENCFF739QUZ | midbrain |
| ENCFF810ZDM | midbrain |
| ENCFF819IDW | midbrain |
| ENCFF819ZTA | midbrain |
| ENCFF839VKV | midbrain |
| ENCFF853SOX | midbrain |
| ENCFF889DNO | midbrain |
| ENCFF938RVG | midbrain |
| SRR3192588 | midbrain |
| SRR3192589 | midbrain |
| ENCFF003CSR | neural tube |
| ENCFF046EJC | neural tube |
| ENCFF064MCV | neural tube |
| ENCFF078SPI | neural tube |
| ENCFF085MBO | neural tube |
| ENCFF090KDU | neural tube |
| ENCFF198HBZ | neural tube |
| ENCFF216XBD | neural tube |
| ENCFF241GLU | neural tube |
| ENCFF321ZGR | neural tube |
| ENCFF378CNA | neural tube |
| ENCFF405VKS | neural tube |
| ENCFF447CLP | neural tube |
| ENCFF489RVW | neural tube |
| ENCFF528JFG | neural tube |
| ENCFF555MUK | neural tube |
| ENCFF581SPK | neural tube |
| ENCFF739BEA | neural tube |
| ENCFF758NAG | neural tube |
| ENCFF834WRE | neural tube |
| ENCFF895DPO | neural tube |
| ENCFF957SPL | neural tube |
| SRR5047985 | ovary |
| SRR5047986 | ovary |
| SRR5047987 | ovary |
| SRR5047988 | ovary |
| SRR5047989 | ovary |

| | |
|---|---|
| SRR5047990 | ovary |
| SRR5047991 | ovary |
| SRR5047992 | ovary |
| SRR5047993 | ovary |
| SRR5047994 | ovary |
| SRR5171100 | ovary |
| ENCFF300QQW | skeletal muscle tissue |
| ENCFF494GEO | skeletal muscle tissue |
| ENCFF562GFS | skeletal muscle tissue |
| ENCFF642EVR | skeletal muscle tissue |
| SRR5048001 | small intestine |
| SRR5048002 | small intestine |
| SRR5048003 | small intestine |
| SRR5048004 | small intestine |
| SRR5048005 | small intestine |
| SRR5048006 | small intestine |
| SRR5048007 | small intestine |
| SRR5048008 | small intestine |
| SRR5048009 | small intestine |
| SRR5048010 | small intestine |
| SRR5171080 | small intestine |
| ENCFF014ITK | spleen |
| ENCFF063LXW | spleen |
| ENCFF082DGE | spleen |
| ENCFF671GFL | spleen |
| SRR5047941 | spleen |
| SRR5047942 | spleen |
| SRR5047943 | spleen |
| SRR5047944 | spleen |
| SRR5047945 | spleen |
| SRR5047946 | spleen |
| ENCFF180LOF | stomach |
| ENCFF263TIR | stomach |
| ENCFF273IHW | stomach |
| ENCFF352EXD | stomach |
| ENCFF417NAF | stomach |
| ENCFF494GQK | stomach |
| ENCFF553BSP | stomach |
| ENCFF775CBB | stomach |
| ENCFF850KIG | stomach |
| SRR5047995 | stomach |
| SRR5047996 | stomach |
| SRR5047997 | stomach |

| | |
|---|---|
| SRR5047998 | stomach |
| SRR5047999 | stomach |
| SRR5048000 | stomach |
| SRR5047953 | testis |
| SRR5047954 | testis |
| SRR5047955 | testis |
| SRR5047956 | testis |
| ENCFF453JXA | thymus |
| ENCFF530SXN | thymus |
| ENCFF638TAT | thymus |
| ENCFF718DNJ | thymus |
| SRR5047947 | thymus |
| SRR5047948 | thymus |
| SRR5047949 | thymus |
| SRR5047950 | thymus |
| SRR5047951 | thymus |
| SRR5047952 | thymus |
| SRR5048035 | urinary bladder |

# CHAPTER 4. MFRECSYS: MRNA FUNCTION RECOMMENDATION SYSTEM

Modified from a manuscript to be submitted to a peer-reviewed journal

Gaurav Kandoi and Julie A. Dickerson

## Author's contributions

GK leads this study. GK and JAD contribute to the design of the study and the interpretation of the results. GK and JAD together wrote the manuscript. GK wrote the programs and performed data analysis. All the authors read and approved the final manuscript.

## Introduction

In higher eukaryotes, more than 95% genes undergo alternative splicing (Kingsmore et al., 2008; Pan et al., 2008), a mechanism that increases protein diversity without increasing genome size. The splicing machinery generates multiple different mRNA isoforms from the same gene that can result in different protein products. Although the sequences of mRNA isoforms of the same gene are very similar, they can have a profound impact on cell regulation and function (Gallego-Paez et al., 2017). These mRNA isoforms of the same gene can have dramatically different functions (Himeji et al., 2002; Melamud & Moult, 2009; Toutant et al., 2007; Vázquez et al., 2011; Végran et al., 2006). The gene *CASP3* is involved in apoptosis and produces two alternative mRNA isoforms. The longer mRNA isoform CASP3-L promotes apoptosis while the shorter mRNA isoform CASP3-S inhibits apoptosis (Végran et al., 2006). Similarly, there are several other genes whose mRNA isoforms perform different or completely opposite functions (Chang et al., 1999; Giorgetti et al., 2007; Himeji et al., 2002; Oberwinkler, Lis, Giehl, Flockerzi, & Philipp, 2005; Rafalska et al., 2004; Végran et al., 2006). In many

cases, such mRNA isoforms of a gene have cell or tissue preferred expression patterns (Buljan et al., 2012; Ellis et al., 2012; Raj & Blencowe, 2015; Sun et al., 2018; Vitulo et al., 2014; Wei & Jin, 2016; Wu et al., 2018; Xu et al., 2002). This article describes mFRecSys, mRNA Function Recommendation System, a tri-factorization based recommender system that uses heterogeneous mRNA isoform properties to make tissue-specific mRNA isoform function recommendations.

The examples show mRNA isoforms of the same gene performing different functions. Traditionally, experiments were mostly performed to identify the functions of the canonical mRNA isoform of a gene. This has resulted in a dearth of functional information about alternative mRNA isoforms. This complexity in understanding of the functions of mRNA isoforms is also reflected in the data stored in biological databases such as Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Carbon et al., 2017; Kandoi & Dickerson, 2017, 2019; Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017; H.-D. Li, Omenn, & Guan, 2016; Shaw, Chen, & Jiang, 2018). Fueled by the advancements in massively parallel sequencing of mRNA isoforms, several computational methods have been developed in the recent years to predict mRNA isoform function (Eksi et al., 2013; W. Li et al., 2014; Luo et al., 2017; Panwar et al., 2016; Shaw et al., 2018).

The task of transcript isoform function prediction is a challenging problem. Some transcript isoforms are non-functional and introduce noise in the data. Many transcript isoforms are tissue and condition specific. Since a gene can produce multiple mRNA isoforms, the direct transfer of function from the gene to its mRNA isoforms doesn't work. Gene function prediction methods cannot be directly applied to mRNA isoform function prediction because these consider a gene as a single entity, ignoring the distinct functions of alternatively spliced

isoforms. However, important advancements have been made by recent studies towards mRNA isoform level understanding of gene functions (Eksi et al., 2013; W. Li et al., 2014; Luo et al., 2017; Panwar et al., 2016; Shaw et al., 2018) such as the prediction of the mRNA isoform ADAM15B of gene ADAM15 to be much more involved in B-cell-mediated immune mechanisms than isoform ADAM15A.

In previous work, the challenge of isoform function prediction has been formulated as Multiple Instance Learning (MIL) (Eksi et al., 2013; W. Li et al., 2014; Luo et al., 2017; Panwar et al., 2016) or deep learning problem (Shaw et al., 2018). IsoPred (Eksi et al., 2013) used isoform level expression data from mouse RNA-Seq to train an SVM model for transcript isoform function prediction. They maintained all evidence codes and selected 1792 biological process terms each annotated with 20 to 300 genes in their method. IsoFunc (Panwar et al., 2016) followed a strategy similar to IsoPred (Eksi et al., 2013) and used mRNA isoform level expression data from human RNA-Seq to train a SVM model for protein-coding splice variant function prediction. The iMILP (W. Li et al., 2014) applied instance-oriented multiple-instance label propagation on a set of isoform co-expression networks. The Weighted Logistic Regression Method (WLRM) (Luo et al., 2017) used a non-convex multiple learning approach using RNA-Seq datasets for predicting the functions of human protein coding isoforms. A deep learning and domain adaptation approach was employed by DeepIsoFun (Shaw et al., 2018) using RNA-Seq datasets.

While the described methods have improved the transcript isoform function prediction, they don't infer the pathways in which these transcript isoforms are involved. These methods don't incorporate the relations between the GO terms apart from the obvious hierarchical relations. The studies introduce bias in the training and testing datasets by using random

mRNA isoforms as non-functional (negative instances) and do not consider the tissue-specific mRNA isoform functions. In this study, we try to overcome these issues and describe mFRecSys, a novel tool for recommending mRNA isoform function. First, we formulate the task of mRNA isoform function prediction as a recommendation problem. This allows for an mRNA isoform to be associated with multiple GO terms and also alleviates the need of generating one model for every GO term. Second, we make tissue-specific function predictions for 17 mouse tissues. Lastly, we do not select random mRNA isoforms as non-functional (negative instances), which is crucial to the selection of training data in a machine learning system.

mFRecSys is a recommender system that uses information from known mRNA isoform- biological process association to make novel association recommendations. A brief overview of mFRecSys is presented in Figure 1. mFRecSys is based on the principles of matrix factorization (MF) for collaborative filtering (Koren, Bell, & Volinsky, 2009). In its simplest form, MF would map mRNA isoforms and GO biological process terms to a latent feature space where their dot product predicts the mRNA isoform – GO biological process term association. The basic matrix factorization method has been useful in building movie recommendation systems (Koren et al., 2009) but it is not ideal for GO biological process term prediction for few reasons. First, the difference in the number of mRNAs and GO biological process terms is large (about 4-fold) making it difficult to project them into same latent feature space. Second, the basic MF approach doesn't allow us to incorporate explicit features (biological context) for mRNA isoforms or GO biological process terms. Third, most mRNA isoforms have none or few known GO biological process term associations, therefore creating the cold-start problem for test mRNAs, i.e. insufficient information to make relevant

recommendations. Therefore, we will use the tri-factorization approach (proposed for predicting multi-relational dyadic data) (Nickel, Tresp, & Kriegel, 2011). The difference in our approach is that we introduce explicit features and use non-linear mappings. In our case, the mRNA isoforms and GO biological process terms play the roles of 'users' and 'items', respectively.

## Methods

### mRNA isoform level feature calculation

The NCBI *Mus musculus* genome assembly (GRCm38.p4) annotated mRNA isoforms were considered. Only mRNA isoforms for which both mRNA and protein sequences are available are used. We remove all protein (and corresponding mRNA) sequences that contain non-standard characters. mRNA isoforms that produce a protein smaller than 30 amino acids are also not considered. This resulted in a filtered set of 75,826 mRNA isoforms from 21,813 genes.

Heterogeneous features such as those derived from RNA-Seq, protein sequences and mRNA sequences have been effective in predicting several biological properties (Du et al., 2017; Kandoi et al., 2015; Kandoi & Dickerson, 2019; H.-D. D. Li et al., 2016). To include a systems level landscape of the mRNA isoforms, we calculated several mRNA and protein sequence properties and processed 359 RNA-Seq samples from 17 tissues. A summary of all the features used for the development of mFRecSys is summarized in Table 1. An overview of the workflow is presented in Fig 1. All analyses were performed on the Extreme Science and Engineering Discovery Environment (XSEDE) Comet cluster (Towns et al., 2014).

**RNA-Seq data processing.** We use mouse RNA-Seq datasets from ENCODE for multiple tissues to extract the tissue-specific expression profile to capture tissue specific

functions. We select datasets which have a read length ≥ 50, a mapping rate of 70% or more, and for which no audit or error warning flags are present in ENCODE. To include sufficient information for tissue-specific function prediction, we select tissues which have at least 10 samples. After applying these filtering criteria we retained 359 RNA-Seq samples from around 20 tissues. There are 17 tissues which have at least 10 samples (Tables 1 and S1).

We use STAR (version 2.5.3a) (Dobin et al., 2013) to align the RNA-Seq datasets. The quantification of mRNA isoform levels in terms of fragments per kilobase of exon per million fragments mapped (FPKM) is performed using StringTie (version 1.3.3b) (Pertea et al., 2016). We use the GRCm38.p4 NCBI genome build (and corresponding GFF3 annotations) during alignment and quantification.

**mRNA sequence composition.** We can represent mRNA sequences as k-mers: the frequencies of k neighboring nucleic acids. Since the mRNA sequences are usually represented by 4 nucleic acids (A, T, C, G), there are $\mathbf{4^k}$ possible k-mers in a k-mer group. For an mRNA sequence of length l,

$$f(kmer_i) = \begin{cases} \dfrac{N_i}{l} & i \in A, T, C, G \\ \dfrac{N_i}{(l-1)} & i \in AA, AT, \dots, GC, GG \\ \dots & \\ \dots & \\ \dots & \\ \dfrac{N_i}{(l-(k-1))} & i \in A\{k\}, A\{k-1\}T, \dots, G\{k-1\}C, G\{K\} \end{cases}$$

where, $f(kmer_i)$ is the frequency of the *ith k-mer* and $N_i$ is the count of the *ith k-mer*. We use the rDNAse library in R (R Core Team, 2017; Zhu et al., 2016) to compute the *k-mer* composition for *k = 3 to 6* for all mRNA isoform sequences.

**Protein Sequence Properties.** Similar to the mRNA sequence, the protein sequence can also be characterized by exploiting its sequence and order composition. Since the protein

sequences are usually represented by 20 amino acids, there are $\mathbf{20}^{\boldsymbol{k}}$ possible k-mers in a k-mer group. We compute the k-mer compositions for k = 1 and 2 for all protein sequences. We also compute the conjoint triad descriptors (J. Shen et al., 2007) for all protein sequences. While these properties are good at capturing the information in the linear vicinity of an amino acid, they don't capture any spatial information such as that obtained from analyzing the protein structures. To take that into account, we also compute the Moran autocorrelation (Moran, 1950) pseudo-amino acid composition (Chou, 2001) for all protein sequences. We use the protr library in R (R Core Team, 2017; Xiao et al., 2015) to compute the protein sequence properties.

$$Moran\ autocorrelation\ I(d) = \frac{\frac{1}{N-d}\sum_{i=1}^{N-d}(P_i - \bar{P}')\,(P_{i+d} - \bar{P}')}{\frac{1}{N}\sum_{i=1}^{N}(P_i - \bar{P}')^2} \qquad d = 1,2,\ldots,30$$

where, d is called the lag of the autocorrelation; $P_i$ and $P_{i+d}$ are the properties of the amino acid $i$ and $i + d$; $\bar{P}'$ is the considered property $P$ along the sequence, i.e.,

$$\bar{P}' = \frac{\sum_{i=1}^{N} P_i}{N}$$

**Biological process level feature calculation**

In addition to mRNA isoform level features, we also compute a GO biological process level feature matrix. We use only mouse specific GO biological process terms. Very specific GO biological process terms (less than 10 genes annotated) and very broad GO biological process terms that are very broad (more than 1000 genes annotated) are removed. This leaves 18,869 GO biological process terms after filtering. We calculate the pairwise semantic similarity between all 18869 GO biological process terms using the graph-based "Wang" method (Wang et al., 2007) with the GOSemSim package in R (R Core Team, 2017; Yu et al., 2010).

**mRNA isoform level functional labels**

To generate the mRNA isoform level functional labels, we use a strategy similar to what we used for the development of TENSION (Kandoi & Dickerson, 2019). We use the GO biological process annotations (downloaded on 23 October 2017) and remove all Inferred from Electronic Annotation (IEA), Non-traceable Author Statement (NAS) and No biological Data available (ND) annotations. The GO hierarchy (gene ontology downloaded on 25 October 2017) allows us to propagate the annotations of a GO term *T* to all its ancestor terms by following the "true path rule".

We generate functional labels associating all mRNA isoforms (75,826) and GO biological process terms that remain after the above filtering (18,869 terms). For the construction of the mRNA isoform level positive labels, we assume a gene to be functional (positive) for a GO biological process term if it is annotated to it. Similarly, if a gene is tagged with a "NOT" qualifier for a GO biological process term, it is considered non-functional (negative) for that term. All such "NOT" tagged annotations are propagated by the inverse of "true path rule", which means that if a gene is explicitly 'NOT' annotated to a GO term *T*, it will also be 'NOT' annotated to all the children of *T*.

Functional database such as GO has very limited information for mRNA/protein isoforms (Kandoi & Dickerson, 2019; H.-D. Li et al., 2016; Luo et al., 2017; Shaw et al., 2018). It usually focusses on the canonical form of a gene/protein and ignore the alternative mRNA isoforms. So, to generate mRNA isoform level functional labels, we exploit the single mRNA producing genes and annotations tagged with a "NOT" qualifier, a method validated using data from mouse (Kandoi & Dickerson, 2019). A summary of the mRNA isoform level functional label generation is illustrated in Fig 1.

For mFRecSys, we assume that if a gene $G_1$ produces only a single mRNA $M_{11}$, then $M_{11}$ is considered functional (positive) for all GO biological process terms (and their ancestors) annotated with $G_1$. In the same way, if $G_2$ produces mRNA isoforms $M_{21}, M_{22}, M_{23}$ and is tagged with a "NOT" qualifier for GO biological process terms, then all mRNA isoforms $M_{21}, M_{22}, M_{23}$ are considered non-functional (negative) for such GO biological process terms and all their child terms.

The GO database doesn't store any information about tissues in which the function is performed. Therefore, to generate tissue-specific mRNA isoform level functional labels, we use data from FANTOM5 (Forrest et al., 2014) for 9 tissues to filter the isoforms based on their tissue expressions. If an mRNA isoform has an expression level below 1 TPM (tags per million) in more than half of the tissue samples, but has a functional label, we exclude such labels from the tissue-specific mRNA isoform level functional labels. This helps us filter out the tissue-specific false positive mRNA isoform level functional labels.

**Generating training and testing datasets**

In our study, we include 75,826 mRNA isoforms and make recommendations for 18869 GO biological process terms. Using the method described above (methods: mRNA isoform level functional labels), we identified 138,786 positive mRNA isoform – GO biological process term associations. We also labelled 26,591 mRNA isoform – GO biological process term negative associations. We label the positive association as 1 and negative associations as -1. All the remaining mRNA isoform – GO biological process term associations are considered 'unknown' and labelled as 0.

To develop an unbiased recommender system, we generate two types of datasets: training and testing. The two datasets are mutually exclusive, i.e. an mRNA isoform is included in only one dataset. We use 70% of the mRNA isoforms (53,078) in the training dataset and the remaining 30% (22,748) are in the testing dataset. The proportion of positive to negative labels (about 5:1) in the training and testing datasets is similar to that of the overall data. The positive labels have been generated using single mRNA producing genes while the negative labels make use of the "NOT" tagged annotations.

**Recommender system for mRNA isoform function prediction**

A workflow of how the recommender system works is shown in Fig. X. To build a recommender system capable of recommending tissue specific mRNA functions, we need to characterize mRNA isoforms and GO biological process terms. The explicit features for the mRNA isoforms include tissue-specific expression profile, mRNA sequence properties and protein sequence properties. For the GO biological process terms, we calculate the semantic similarity between all terms. Let, $F_{RNA} \in \mathbb{R}^{n \times l_{RNA}}$ be the explicit feature matrix associated with $n$ mRNA isoforms and $l$ mRNA isoform level features. Similarly, let $F_{BP} \in \mathbb{R}^{m \times m}$ be the explicit feature matrix associated with $m$ GO biological process terms. These feature-based representations of mRNA isoforms and GO biological process terms are non-linearly projected into latent spaces of different sizes respectively, where a third mapping will associate them. The parameters of the three mappings will be jointly tuned.

Let $A_{RNA} \in \mathbb{R}^{k_{RNA} \times l_{RNA}}$, $A_{BP} \in \mathbb{R}^{k_{BP} \times m}$, and $S \in \mathbb{R}^{k_{RNA} \times k_{BP}}$ denote the three factors in the decomposition. Here, $k_{RNA}$ and $k_{BP}$ are the number of latent features for mRNA isoforms and GO biological process terms, respectively. Then, our model is defined by:

$$R = \sigma\left(F_{RNA}{A_{RNA}}^{T}\right) \in \mathbb{R}^{n \times k_{RNA}}$$

$$B = \sigma\left(F_{BP}{A_{BP}}^{T}\right) \in \mathbb{R}^{m \times k_{BP}}$$

$$\hat{Y} = \sigma(RSB^{T}) \in \mathbb{R}^{n \times m}$$

where $\sigma$ is the logistic function defined by:

$$\sigma = \frac{1}{1 + e^{-x}}$$

Here, $R \in \mathbb{R}^{n \times k_{RNA}}$ is the decomposition of the mRNA isoform feature matrix $F_{RNA}$. Similarly, $B \in \mathbb{R}^{n \times k_{BP}}$ is the decomposition of the biological process feature matrix $F_{BP}$. The $Y \in \mathbb{R}^{n \times m}$ is the interaction matrix defining the true mRNA-biological process labels that we have generated and $\hat{Y}$ is its estimate. We use Adam optimizer (Kingma & Ba, 2014) to train the factorization model to optimize the regularized mean squared error:

$$\min_{A_{BP}, A_{RNA}, S} \frac{\sum_{i=0}^{m} \sum_{j=0}^{n} \left(Y_{ij} - \hat{Y}_{ij}\right)^2}{m \cdot n} + \lambda \cdot r(A_{BP}, A_{RNA}, S)$$

where the regularizer $r(A_{BP}, A_{RNA}, S)$ is the normalized Frobenius norm of the model weights:

$$r(A_{BP}, A_{RNA}, S) = \frac{\|A_{RNA}\|_F}{k_{RNA} \cdot l_{RNA}} + \frac{\|A_{BP}\|_F}{k_{BP} \cdot m} + \frac{\|S\|_F}{k_{RNA} \cdot K_{BP}}$$

Since the three factors of decomposition, $A_{RNA}$, $A_{BP}$, and $S$ have very different sizes, we use the normalized Frobenius norm to cancel out such dependencies.

**Training tissue-specific recommendation systems**

For developing tissue-specific recommendation systems we use a completely different set of RNA-Seq samples from FANTOM5 project (Forrest et al., 2014). We use these RNA-Seq samples from 9 tissues (Adrenal Glands, Heart, Kidney, Liver, Lung, Ovary, Spleen,

Stomach and Thymus) to create tissue-specific mRNA isoform level functional labels. We use these new RNA-Seq samples to control tissue-specific false positive functional labels. For every tissue, only those mRNA isoform level functional labels are retained that contain mRNA isoforms expressed in at least half samples. The remaining mRNA isoform level functional labels are considered false positive and considered as unknowns.

We develop 9 new mFRecSys models, by using all mRNA isoform sequence features, protein sequences features and tissue-specific RNA-Seq features

**Performance evaluation of recommender system**

We calculate multiple performance metrics such as accuracy, regularized mean square error, Matthews Correlation Coefficient (MCC), Area Under the Precision-Recall Curve (AUPRC) and Area Under the Receiver Operating Characteristic Curve (AUROC) to evaluate the performance of mFRecSys. We generate and use several different types of datasets to comprehensively evaluate the performance of mFRecSys.

First, we perform randomization tests to check the impact of randomly selecting the data for training and testing datasets. We perform 50 instances of random training and testing dataset generation. In each instance, we randomly select 70% of the data as training data and the remaining 30% as testing data. Then, we train and optimize the model using the training data alone until 500 iterations. We calculate the performance evaluation metrics for both training and testing dataset after each iteration.

Since there is no gold standard mRNA isoform level functions dataset, we validate the predictions made by mFRecSys using the latest annotations from GO. We process the new GO annotations (downloaded on 13 March 2019) as described above (methods: mRNA isoform level functional labels). We found 145,446 positive mRNA isoform – GO biological process

associations using our strategy to utilize the single isoform gene annotations. Similarly, we found 28661 negative mRNA isoform – GO biological process associations using our strategy to utilize the "NOT" tagged GO annotations. Of these, 21,971 positive and 3,245 negative mRNA isoform – GO biological process associations are new and not present in our original functional labels. We refer this new data as the "validation dataset" and evaluate how the predictions made by mFRecSys compare to these newly discovered associations.

**Feature importance and selection**

Due to computational and time limitations, it is not possible to individually test the importance of all 6582 features used for developing mFRecSys. Therefore, we train recommendation systems using the features groups, namely, mRNA isoform sequence features, protein sequence features, all sequence features, RNA-Seq expression features, and all features. We calculate performance evaluation metrics for both training and testing datasets after every iteration, for up to 3000 iterations. After that, we identify the values of $k_{RNA}$ and $k_{BP}$ for every feature group that results in the highest MCC values on the testing dataset.

## Results

The number of latent features for mRNA isoforms and GO biological process terms, $k_{RNA}$ and $k_{BP}$ respectively, are the two main parameters in mFRecSys. We use grid search (possible values: 10, 20, 50, 100, 200, 500, 1000, and 2000) over both parameters to obtain the optimal values. The three factors in the decomposition, $A_{RNA} \in \mathbb{R}^{k_{RNA} \times l_{RNA}}$, $A_{BP} \in \mathbb{R}^{k_{BP} \times m}$, and $S \in \mathbb{R}^{k_{RNA} \times k_{BP}}$ are initialized as random samples from a uniform distribution between 0 and 1. The three factors in the decomposition are updated after every iteration to minimize the

regularized mean squared error on the training dataset. In a recommendation system or deep neural networks, it is very easy to learn the exact representation of training data. This leads to the problem of overfitting, where the model is unable to generalize. To control this, the model configuration with the highest MCC value on the testing dataset as opposed to the training dataset is selected as the final model (note that the error minimization is done using the training dataset alone).

We compute multiple metrics such MCC, Accuracy, AUPRC and AUROC after every iteration to evaluate the performance of recommendation systems. Using single mRNA isoform producing genes and GO annotations tagged with "NOT", we labelled about 165,000 mRNA isoform – GO biological process term associations as either functional or non-functional. There is a large difference in the number of functional and non-functional labels. This imbalance in the labels results in a baseline AUPRC of 0.839.

We can see from the randomization test that the variance in the model performance among different instances is low (Figure 2). Although there is some variation in MCC values for different instances, the variation in accuracy, AUPRC and AUROC is very low. This suggests there is very little bias in the process of randomly selecting training and testing datasets. Therefore, we generate one random training and testing dataset and use that to develop the final model.

**mRNA sequence properties are most predictive of mRNA isoform functions**

We evaluate which mRNA isoform feature group (mRNA sequence properties, protein sequence properties or mRNA expression profile) results in the best performing model by using a subset of $F_{RNA} \in \mathbb{R}^{n \times l_{RNA}}$ during initialization. We find that the model with only mRNA sequence properties performs better at predicting the known mRNA isoform – GO biological

process term associations (Figure 3; Table 2). The protein sequence properties and the combination of mRNA isoform and protein sequence based properties also have very similar performance. The performance when using only mRNA expression profile is lowest.

**A recommendation system for mRNA isoform function recommendation**

We use the best performing recommendation system using mRNA isoform sequence properties alone with $k_{RNA} = 20$ and $k_{BP} = 200$ (Table 2) to make recommendations for all 75,826 mRNA isoforms and 18,869 GO biological process terms. The recommendations include both functional and non-functional recommendations for mRNA isoforms. These recommendations are at the organism level, and do not necessarily reflect the tissue-specific functions of mRNA isoforms.

**Tissue-specific mRNA isoform function recommendation systems**

We use a completely different dataset for 9 tissues from the FANTOM5 project (Forrest et al., 2014) to generate our tissue-specific mRNA isoform level functional labels. We filter the mRNA isoforms based on their tissue expressions in order to control false positives in our mRNA isoform level functional labels. Only those mRNA isoforms that are expressed in at least half samples for a specific tissue are retained in the mRNA isoform level functional labels. The mRNA isoform level functional labels which contain the remaining mRNA isoforms are discarded for the tissue-specific mRNA isoform level functional labels.

The details of best performing tissue-specific mRNA isoform function recommendation systems is provided in Table 2. We see that the performance of different tissue-specific mRNA isoform function recommendation systems differs. This highlight the complexity of predicting mRNA isoform function at tissue level. Additionally, the

performance of organism-level recommendation system using mRNA sequence properties alone is better than all tissue-specific recommendation systems. This highlights the importance of using mRNA sequence features and points to the noise present in the RNA-Seq expression data. Additionally, many mRNA isoforms are known to be expressed only under certain tissues which introduce bias and error in their function prediction.

## Discussions

The alternatively spliced mRNA isoforms of a gene encode proteins of different function. It is highly beneficial that the investigation of functions is carried out at the mRNA isoform level. Because the paradigm of gene function prediction considers a gene as a single entity without differentiating between its mRNA isoforms, it has a major drawback from the mRNA isoform point of view. There is a rich resource of data at the mRNA isoform level in the form of mRNA isoform expression profile, mRNA isoform and protein sequences that can be used to address this drawback.

However, the prediction of mRNA isoform functions is challenging for multiple reasons. First, because of the lack of labeled training data at the mRNA isoform level. Second, the GO annotations are noisy and most GO biological process terms are only annotated with a small number of genes making the data very imbalanced. Additionally, functions of most genes are yet to be discovered. This results in a high number of false positives leading to a low precision. We overcome this problem by not considering very specific (less than 10 genes annotated) or very broad (more than 1000 genes annotated) GO terms and using GO annotations tagged with “NOT” to create a smaller but high quality non-functional (negative) mRNA isoform label dataset. Third, the heterogeneity of the mRNA isoform expression data from multiple tissues while useful, also contains a lot of noise (W. Li et al., 2014).

Our method is also limited to the incomplete mRNA isoform catalog currently available and maintained by NCBI, but it can be readily updated to incorporate the new genome annotations of mRNA isoform. Our method is further limited by the current technology to assemble and quantify differences in the expression of mRNA isoform of the same gene across multiple tissues.

We present a generic and novel strategy to study gene regulation and functions at a higher resolution. Although our method obtains significant performance in computational evaluations, to validate and characterize the functional dynamics of mRNA isoforms at the scale of entire genome, experimental studies are required. Further integration of other omics data such as Ribo-Seq, proteomics and metabolomics will be useful for improving the performance of mRNA isoform function prediction methods.

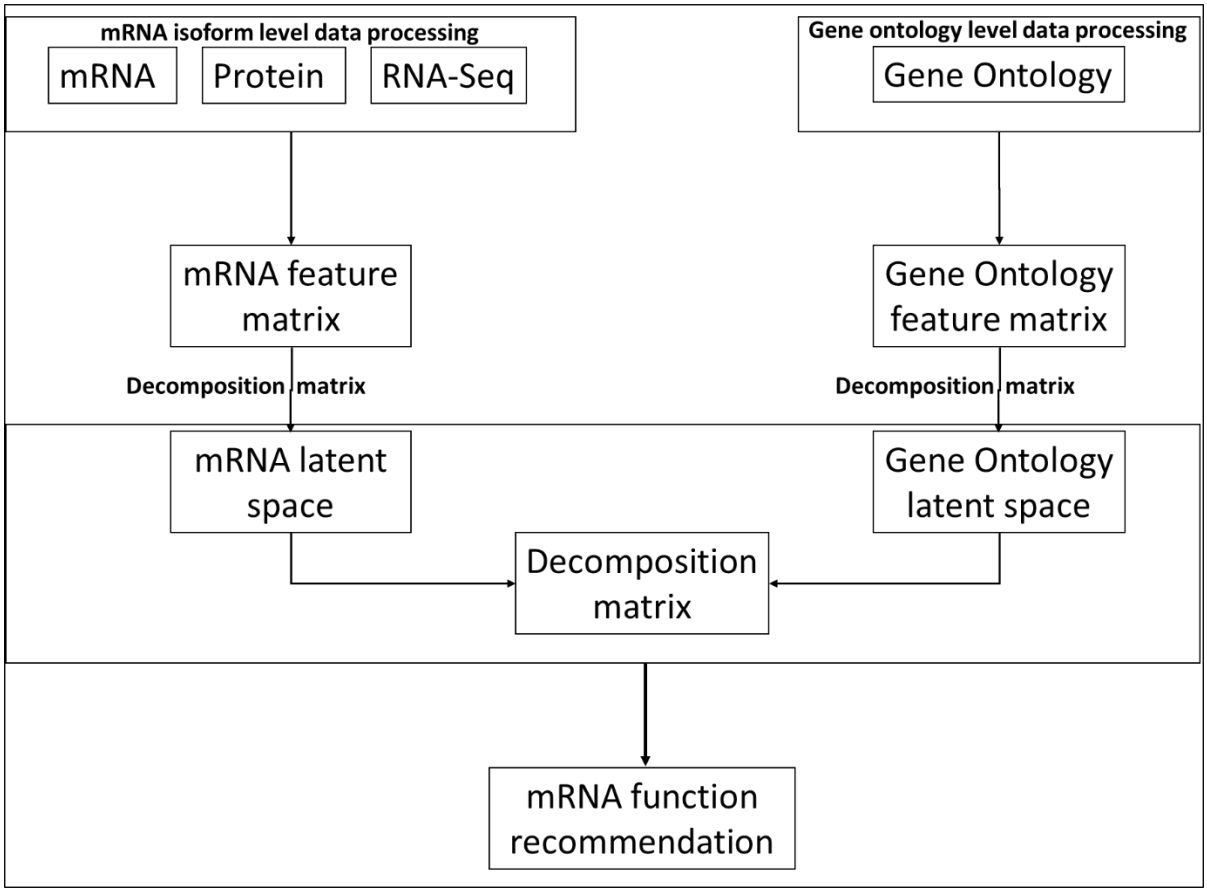


Figure 4.1 **Overview of how mFRecSys works.** We calculate mRNA isoform feature matrix using features calculated from mRNA isoform sequences, protein sequences and RNA-Seq samples from multiple tissues. The elements in the square GO biological process term feature matrix represents the semantic similarity between the GO terms. The mRNA and GO feature matrices are non-linearly projected into latent spaces of different sizes respectively, where a third mapping will associate them, resulting in the mRNA function recommendations.

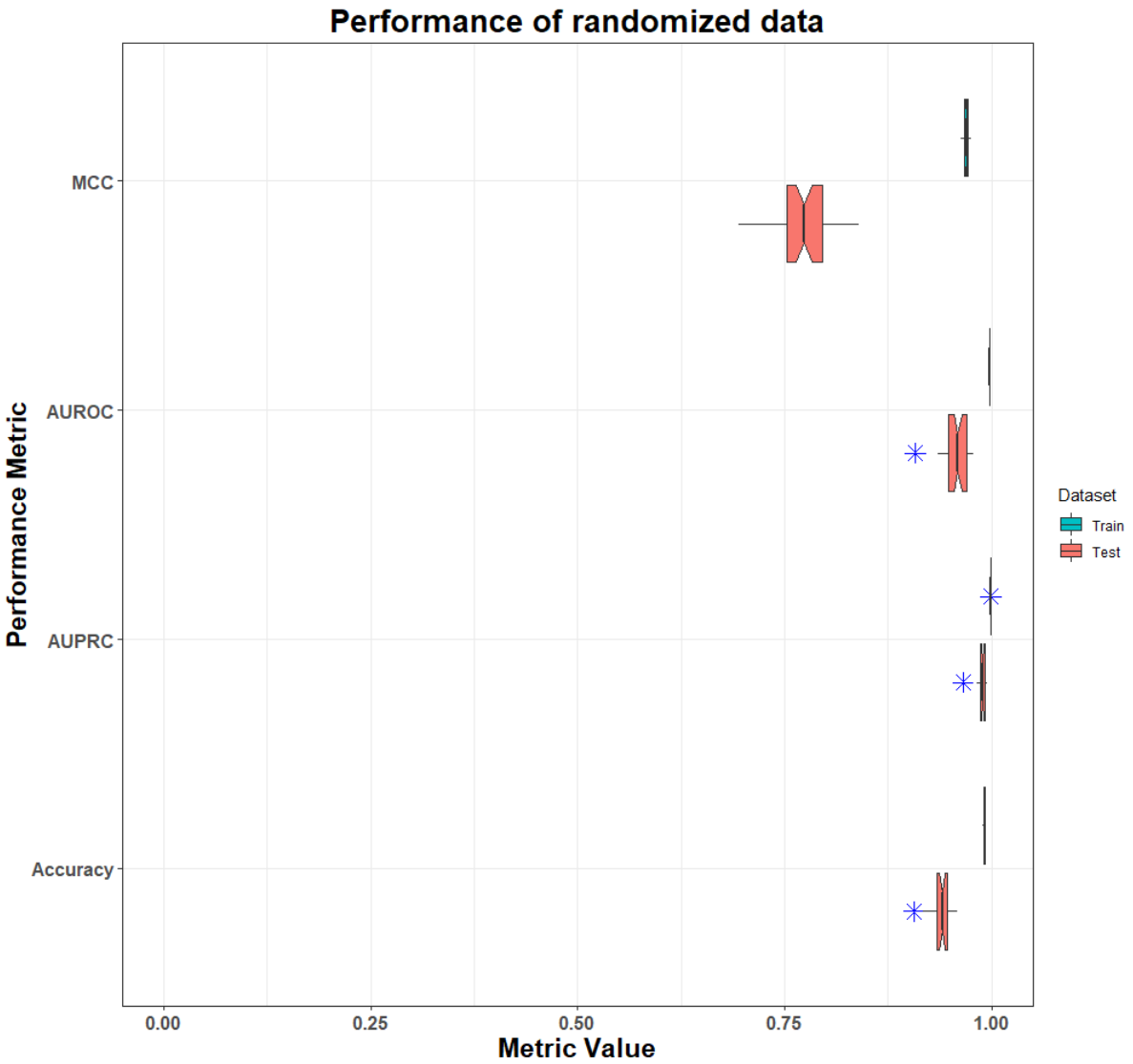


Figure 4.2 **Performance evaluation on randomized datasets.** A boxplot of various performance evaluation metrics calculated at 500$^{th}$ iteration for 50 randomized datasets. The variability in the values of the performance metric across 50 randomized datasets is represented by the width of the boxes along the x-axis. Smaller box widths and higher metric values are better. Abbreviations - AUROC: Area Under the Receiver Operating Characteristic Curve; AUPRC: Area Under the Precision-Recall Curve; MCC: Matthews Correlation Coefficient

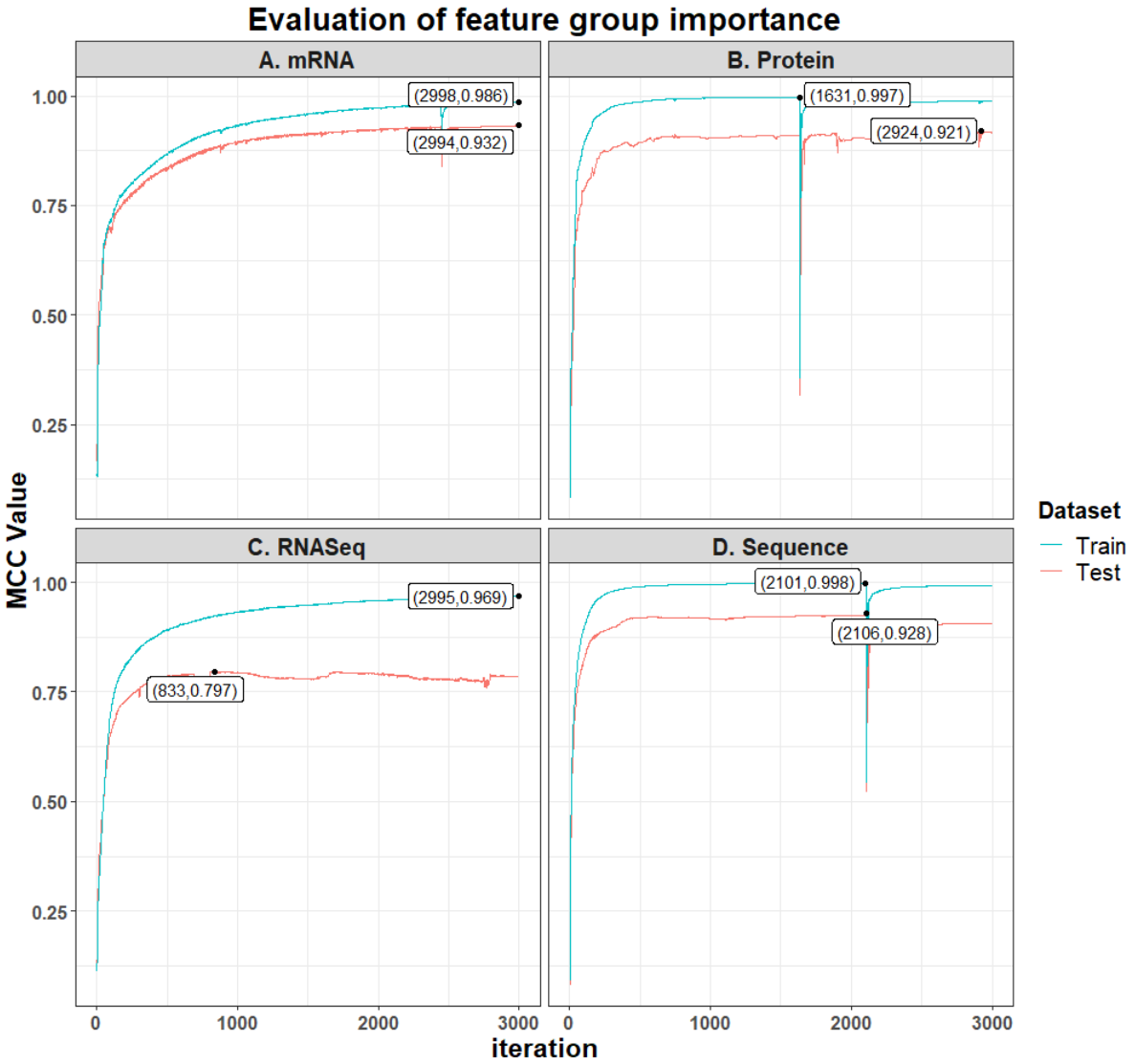


Figure 4.3 **Evaluation of feature group importance**. The plot shows the improvement of MCC for both training and testing datasets over iterations for the best performing models for different feature groups. The best test dataset performance is obtained when using only mRNA isoform features. The number of latent mRNA isoform features is 20 and the number of latent GO biological process term features is 200. The highest MCC along with the iterations at which it occurs is labelled for every feature group. Abbreviations - MCC: Matthews Correlation Coefficient

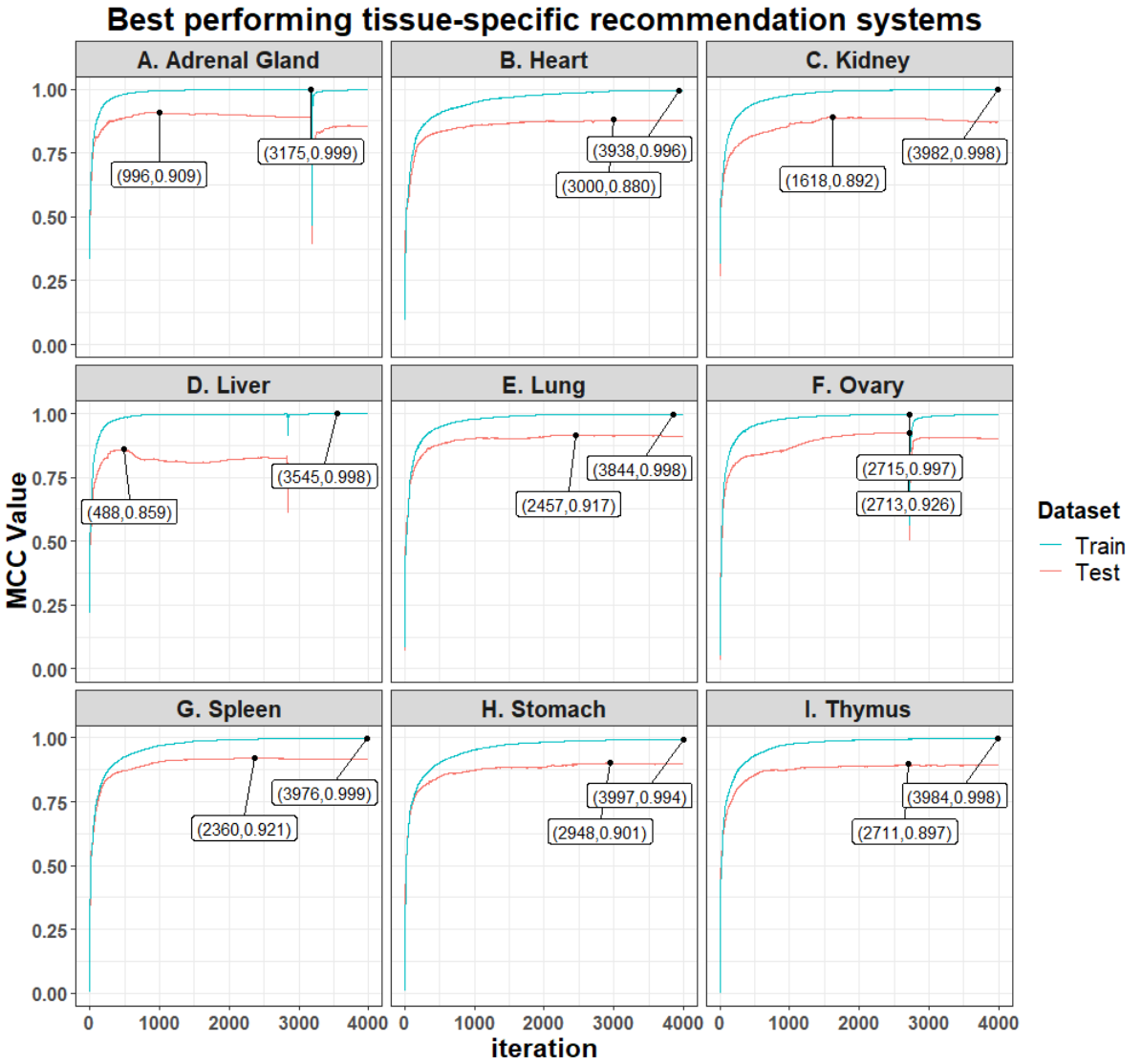


Figure 4.4 **Improvement in MCC over iterations for best tissue-specific recommendation systems.** The plot shows how MCC improves over iterations for both training and testing datasets for the best performing tissue-specific recommendation models. The highest MCC along with the iteration at which it occurs is labelled for all tissues. Abbreviations - MCC: Matthews Correlation Coefficient

Table 4.1 Summary of all the features used for the development of mFRecSys

| Level | Entity | Feature Type | No. of Features |
|---|---|---|---|
| Sequence | mRNA | 3-mers | 64 |
| | | 4-mers | 256 |
| | | 5-mers | 1024 |
| | | 6-mers | 4096 |
| | Protein | Amino acid composition (1-mer) | 20 |
| | | Di-amino acid composition (2-mer) | 400 |
| | | Conjoint Triad Descriptors | 343 |
| | | Pseudo-amino acid composition | 50 |
| | | Moran autocorrelation | 240 |
| Expression | mRNA | Heart | 32 |
| | | Liver | 36 |
| | | Kidney | 18 |
| | | Adrenal Glands (AdGland) | 10 |
| | | Forebrain | 29 |
| | | Midbrain | 24 |
| | | Hindbrain | 30 |
| | | Embryonic facial prominence (EmbFacPro) | 22 |
| | | Large intestine (Lintestine) | 10 |
| | | Small intestine (Sintestine) | 11 |
| | | Lung | 14 |
| | | Limb | 22 |
| | | Neural tube (Ntube) | 22 |
| | | Ovary | 11 |
| | | Spleen | 10 |
| | | Stomach | 15 |
| | | Thymus | 10 |
| | | Other tissues | 33 |

Table 4.2 Summary of best performing recommendation systems.

| **Dataset** | **Krna** | **Kbp** | **MCC** | **Accuracy** | **AUPRC** | **AUROC** |
|---|---|---|---|---|---|---|
| mRNA isoform sequence | 20 | 200 | 0.932 | 0.981 | 0.998 | 0.993 |
| mRNA isoform and Protein sequence | 100 | 20 | 0.928 | 0.980 | 0.996 | 0.988 |
| Protein sequence | 500 | 10 | 0.921 | 0.978 | 0.995 | 0.983 |
| RNASeq | 10 | 20 | 0.797 | 0.944 | 0.995 | 0.976 |
| All features | 20 | 20 | 0.803 | 0.947 | 0.993 | 0.971 |
| Adrenal Glands | 20 | 50 | 0.909 | 0.968 | 0.996 | 0.989 |
| Heart | 10 | 10 | 0.880 | 0.956 | 0.995 | 0.986 |
| Kidney | 10 | 10 | 0.892 | 0.960 | 0.996 | 0.988 |
| Liver | 20 | 50 | 0.859 | 0.948 | 0.991 | 0.979 |
| Lung | 10 | 20 | 0.917 | 0.969 | 0.996 | 0.990 |
| Ovary | 10 | 100 | 0.926 | 0.974 | 0.997 | 0.990 |
| Spleen | 10 | 10 | 0.921 | 0.969 | 0.996 | 0.991 |
| Stomach | 10 | 10 | 0.901 | 0.965 | 0.996 | 0.988 |
| Thymus | 10 | 10 | 0.897 | 0.961 | 0.996 | 0.990 |

## References

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. Molecular Cell, 46(6), 871–883. https://doi.org/10.1016/j.molcel.2012.05.039

Carbon, S., Dietze, H., Lewis, S. E., Mungall, C. J., Munoz-Torres, M. C., Basu, S., … Westerfield, M. (2017). Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. Nucleic Acids Research, 45(D1), D331–D338. https://doi.org/10.1093/nar/gkw1108

Chang, B. S., Kelekar, A., Harris, M. H., Harlan, J. E., Fesik, S. W., & Thompson, C. B. (1999). The BH3 Domain of Bcl-xS Is Required for Inhibition of the Antiapoptotic Function of Bcl-xL. Molecular and Cellular Biology, 19(10), 6673–6681. https://doi.org/10.1128/MCB.19.10.6673

Chou, K. C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Structure, Function and Genetics, 43(3), 246–255. https://doi.org/10.1002/prot.1035

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Du, X., Hu, C., Yao, Y., Sun, S., & Zhang, Y. (2017). Analysis and prediction of exon skipping events from RNA-seq with sequence information using rotation forest. International Journal of Molecular Sciences, 18(12). https://doi.org/10.3390/ijms18122691

Eksi, R., Li, H. D., Menon, R., Wen, Y., Omenn, G. S., Kretzler, M., & Guan, Y. (2013). Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. PLoS Computational Biology, 9(11). https://doi.org/10.1371/journal.pcbi.1003314

Ellis, J. D., Barrios-Rodiles, M., Çolak, R., Irimia, M., Kim, T. H., Calarco, J. A., … Blencowe, B. J. (2012). Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. Molecular Cell, 46(6), 884–892. https://doi.org/10.1016/j.molcel.2012.05.037

Forrest, A. R. R., Kawaji, H., Rehli, M., Baillie, J. K., De Hoon, M. J. L., Haberle, V., … Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. Nature, 507(7493), 462–470. https://doi.org/10.1038/nature13182

Gallego-Paez, L. M., Bordone, M. C., Leote, A. C., Saraiva-Agostinho, N., Ascensão-Ferreira, M., & Barbosa-Morais, N. L. (2017). Alternative splicing: the pledge, the turn, and the prestige: The key role of alternative splicing in human biological systems. Human Genetics. https://doi.org/10.1007/s00439-017-1790-y

Giorgetti, A., Patthy, L., Guigo, R., Jones, P., Schlicker, A., Jones, D. T., … Nagy, A. (2007). The implications of alternative splicing in the ENCODE protein complement. Proceedings of the National Academy of Sciences, 104(13), 5495–5500. https://doi.org/10.1073/pnas.0700800104

Himeji, D., Horiuchi, T., Tsukamoto, H., Hayashi, K., Watanabe, T., & Harada, M. (2002). Characterization of caspase-8L: A novel isoform of caspase-8 that behaves as an inhibitor of the caspase cascade. Blood, 99(11), 4070–4078. https://doi.org/10.1182/blood.V99.11.4070

Kandoi, G., Acencio, M. L., & Lemke, N. (2015). Prediction of druggable proteins using machine learning and systems biology: A mini-review. Frontiers in Physiology. https://doi.org/10.3389/fphys.2015.00366

Kandoi, G., & Dickerson, J. A. (2017). Differential alternative splicing patterns with differential expression to computationally extract plant molecular pathways. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2017-Janua, 2144–2151. https://doi.org/10.1109/BIBM.2017.8217993

Kandoi, G., & Dickerson, J. A. (2019). Tissue-specific mouse mRNA isoform networks. BioRxiv Bioinformatics, 558361. https://doi.org/10.1101/558361

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research, 45(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. Retrieved from http://arxiv.org/abs/1412.6980

Kingsmore, S. F., Wang, E. T., Khrebtukova, I., Zhang, L., Luo, S., Mayr, C., … Schroth, G. P. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature, 456(7221), 470–476. https://doi.org/10.1038/nature07509

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30–37. https://doi.org/10.1109/MC.2009.263

Li, H.-D. D., Menon, R., Eksi, R., Guerler, A., Zhang, Y., Omenn, G. S., & Guan, Y. (2016). A Network of Splice Isoforms for the Mouse. Scientific Reports, 6(April), 1–11. https://doi.org/10.1038/srep24507

Li, H.-D., Omenn, G. S., & Guan, Y. (2016). A proteogenomic approach to understand splice isoform functions through sequence and expression-based computational modeling. Briefings in Bioinformatics, 17(February), bbv109. https://doi.org/10.1093/bib/bbv109

Li, W., Kang, S., Liu, C. C., Zhang, S., Shi, Y., Liu, Y., & Zhou, X. J. (2014). High-resolution functional annotation of human transcriptome: Predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Research, 42(6), e39–e39. https://doi.org/10.1093/nar/gkt1362

Luo, T., Zhang, W., Qiu, S., Yang, Y., Yi, D., Wang, G., … Wang, J. (2017). Functional Annotation of Human Protein Coding Isoforms via Non-convex Multi-Instance Learning. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17, 345–354. https://doi.org/10.1145/3097983.3097984

Melamud, E., & Moult, J. (2009). Stochastic noise in splicing machinery. Nucleic Acids Research, 37(14), 4873–4886. https://doi.org/10.1093/nar/gkp471

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika, 37(1), 17–23. https://doi.org/10.2307/2332142

Nickel, M., Tresp, V., & Kriegel, H.-P. (2011). A Three-Way Model for Collective Learning on Multi-Relational Data. In ICML (pp. 809--816).

Oberwinkler, J., Lis, A., Giehl, K. M., Flockerzi, V., & Philipp, S. E. (2005). Alternative splicing switches the divalent cation selectivity of TRPM3 channels. Journal of Biological Chemistry, 280(23), 22540–22548. https://doi.org/10.1074/jbc.M503092200

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics, 40(12), 1413–1415. https://doi.org/10.1038/ng.259

Panwar, B., Menon, R., Eksi, R., Li, H.-D., Omenn, G. S., & Guan, Y. (2016). Genome-Wide Functional Annotation of Human Protein-Coding Splice Variants Using Multiple Instance Learning. Journal of Proteome Research, 15(6), 1747–1753. https://doi.org/10.1021/acs.jproteome.5b00883

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protocols, 11(9), 1650–1667. https://doi.org/10.1038/nprot.2016.095

R Core Team. (2017). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rafalska, I., Zhang, Z., Ben-Ari, S., Stamm, S., Thanaraj, T. A., Toiber, D., … Soreq, H. (2004). Function of alternative splicing. Gene, 344, 1–20. https://doi.org/10.1016/j.gene.2004.10.022

Raj, B., & Blencowe, B. J. (2015, July 1). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron. Cell Press. https://doi.org/10.1016/j.neuron.2015.05.004

Shaw, D., Chen, H., & Jiang, T. (2018). DeepIsoFun: a deep domain adaptation approach to predict isoform functions. Bioinformatics. https://doi.org/10.1093/bioinformatics/bty1017

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., … Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. Proceedings of the National Academy of Sciences of the United States of America, 104(11), 4337–4341. https://doi.org/10.1073/pnas.0607879104

Sun, Y., Hou, H., Song, H., Lin, K., Zhang, Z., Hu, J., & Pang, E. (2018). The comparison of alternative splicing among the multiple tissues in cucumber. BMC Plant Biology, 18(1), 5. https://doi.org/10.1186/s12870-017-1217-x

Toutant, J., Garneau, D., Cloutier, P., Revil, T., Shkreta, L., & Chabot, B. (2007). Protein Kinase C-Dependent Control of Bcl-x Alternative Splicing. Molecular and Cellular Biology, 27(24), 8431–8441. https://doi.org/10.1128/mcb.00565-07

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., … Wilkens-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. Computing in Science and Engineering, 16(5), 62–74. https://doi.org/10.1109/MCSE.2014.80

Vázquez, Á. V., Blanco, M., Zaborowska, J., Soengas, P., González-Siso, M. I., Becerra, M., … Cerdán, M. E. (2011). Two Proteins with Different Functions are Derived from the KlHEM13 Gene. Eukaryotic Cell, 10(10), 1331–1339. https://doi.org/10.1128/EC.05108-11

Végran, F., Boidot, R., Oudin, C., Riedinger, J. M., Bonnetain, F., & Lizard-Nacol, S. (2006). Overexpression of caspase-3s splice variant in locally advanced breast carcinoma is associated with poor response to neoadjuvant chemotherapy. Clinical Cancer Research, 12(19), 5794–5800. https://doi.org/10.1158/1078-0432.CCR-06-0725

Vitulo, N., Forcato, C., Carpinelli, E. C., Telatin, A., Campagna, D., D'Angelo, M., … Valle, G. (2014). A deep survey of alternative splicing in grape reveals changes in the splicing machinery related to tissue, stress condition and genotype. BMC Plant Biology, 14(1), 99. https://doi.org/10.1186/1471-2229-14-99

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics, 23(10), 1274–1281. https://doi.org/10.1093/bioinformatics/btm087

Wei, B., & Jin, J. P. (2016, May 10). TNNT1, TNNT2, and TNNT3: Isoform genes, regulation, and structure-function relationships. Gene. Elsevier. https://doi.org/10.1016/j.gene.2016.01.006

Wu, P., Zhou, D., Lin, W., Li, Y., Wei, H., Qian, X., … He, F. (2018). Cell-type-resolved alternative splicing patterns in mouse liver. DNA Research. https://doi.org/10.1093/dnares/dsx055

Xiao, N., Cao, D. S., Zhu, M. F., & Xu, Q. S. (2015). Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. In Bioinformatics (Vol. 31, pp. 1857–1859). https://doi.org/10.1093/bioinformatics/btv042

Xu, Q., Modrek, B., & Lee, C. (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Research, 30(17), 3754–3766. https://doi.org/10.1093/nar/gkf492

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., & Wang, S. (2010). GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. Bioinformatics, 26(7), 976–978. https://doi.org/10.1093/bioinformatics/btq064

Zhu, M., Dong, J., & Cao, D. (2016). rDNAse: Generating Various Numerical Representation Schemes of DNA Sequences. Retrieved from https://cran.r-project.org/package=rDNAse

# CHAPTER 5. GENERAL CONCLUSIONS

## 5.1. General Discussions

Advances in high-throughput technologies, computational resources and techniques provide an opportunity, for bioinformatics and computational biology research, to incorporate more biological context when designing tools for biology. With information available at multiple levels of central dogma, we have a chance to study holistically, the biological processes and regulations. Being able to integrate and feed more biological context to statistical and predictive models means better, more accurate, and biologically relevant predictions. Such improved systems will help us better understand how biology works. In this dissertation, we present research that incorporates more biological context from mRNA and protein sequences, mRNA expression profile, tissue specificity and similarity beyond hierarchical relationships between GO terms to better understand biological regulation.

In **Chapter 2**, with the help of multiple use cases from heat or cold stressed *Arabidopsis thaliana*, we argue that differential alternative splicing should be used in conjunction with differential gene expression. We show that several important pathways and processes are missed when considering only DEGs to study biological regulation. Including DASGs along with DEGs provides a more complete picture of the complex biological regulatory machinery. Several pathways have a significant amount of differential alternative splicing when subjected to heat or cold stress, but very little to none differential gene expression. As such, it is important to include the information provided by both differential alternative splicing and differential gene expression when studying regulation.

Only the functions of genes (or canonical protein product) have been primarily studied in the context of biological networks. Such networks, which infer the connections between the

genes (or canonical protein products) leading to identification of novel gene functions are a powerful tool for studying regulation. While these gene-level networks have made important discoveries, lot more can be gained if we work at the mRNA isoform levels, taking into account, the alternate mRNA isoforms. A gene can produce multiple different mRNA isoforms, many of which are functional. A lot of these mRNA isoforms are functional under specific conditions or tissues only (Buljan et al., 2012; Ellis et al., 2012; Raj & Blencowe, 2015; Sun et al., 2018; Vitulo et al., 2014; Wei & Jin, 2016; Wu et al., 2018; Xu et al., 2002). There are several documented examples where the mRNA isoforms of the same gene perform dramatically different functions (Himeji et al., 2002; Melamud & Moult, 2009; Toutant et al., 2007; Vázquez et al., 2011; Végran et al., 2006). In **Chapter 3**, we present TENSION, a computational framework for predicting tissue-specific mRNA isoform level functional networks.

In TENSION, we incorporate heterogeneous data coming from mRNA sequences, protein sequences, and tissue-specific mRNA expression profiles. We also exploit several aspects of GO annotations, pathway databases and protein-protein interactions to create high quality mRNA isoform level functional labels. These labels define whether mRNA isoform pairs are involved in same biological function (co-functional) or not for about 3 million mRNA isoform pairs. We evaluate the performance, robustness and usefulness of TENSION using several tests and case studies.

We identified about 10.6 million mRNA isoform pairs that are co-functional in specific tissues only. Additionally, we identified about 3.5 million mRNA isoform pairs that are not co-functional in specific tissues. This highlights and supports the notion that many alternatively spliced mRNA isoforms are functional under certain tissues and conditions only (Buljan et al.,

2012; Ellis et al., 2012; Raj & Blencowe, 2015; Sun et al., 2018; Vitulo et al., 2014; Wei & Jin, 2016; Wu et al., 2018; Xu et al., 2002). By mapping the mRNA isoform level networks to gene level networks, we also show that the central genes in our tissue-specific functional networks are enriched in biological functions characteristic of the tissues.

Our analysis also identifies about 164,000 functional gene pairs with different mRNA isoform pairs that are shared by multiple tissues. This finding points to the tissue specific expression and function of different mRNA isoforms of the same gene. Additionally, we also identified 660,000 instances where one mRNA isoform pair is functional while other mRNA isoform pairs of the same gene pair are non-functional. We highlight the importance of tissue-specific changes in biological processes and pathways by capturing the differences in functional relationships of mRNA isoforms of the same gene across multiple tissues in mouse.

In **Chapter 4**, we describe mFRecSys, a recommendation system for predicting tissue-specific mRNA isoform functions. Recent methods (Eksi et al., 2013; W. Li et al., 2014; Luo et al., 2017; Panwar et al., 2016; Shaw et al., 2018) have made great progress in developing computational tools for mRNA isoform function prediction. However, these methods have several shortcomings such as biased training and testing label generation, lack of biological context by using only mRNA expression profile for characterizing mRNA isoforms, not considering tissue-specific functions and using only hierarchical relationships between GO terms. This dissertation overcomes many such shortcomings by using a more robust strategy to generate training and testing labels, introducing explicit biological context by using mRNA sequence and protein sequences in addition to tissue-specific mRNA expression profile for characterizing mRNA isoforms and using semantic similarity between GO terms to

characterize GO biological process terms, and developing recommendation systems capable of making tissue-specific recommendations.

We introduce explicit biological context in our system which is missing in previous methods by formulating the problem of mRNA isoform function prediction as a tri-factorization matrix-based recommendation problem. We use semantic similarity between GO terms as part of our recommendation model and information from GO hierarchy to generate our mRNA isoform – GO biological process associations.

Previous methods use unannotated genes as non-functional (negative). Additionally, these either select random mRNA isoform of a gene or initialize all mRNA isoforms of a gene as functional (positive). However, we use a stricter criteria to select functional (positive) and non-functional (negative) instances, therefore limiting the bias in our training and testing labels. We select our non-functional (negative) instances by utilizing the GO annotations tagged with "NOT" after propagating these using the inverse of "true-path rule". Similarly, we use GO annotations involving single mRNA isoform producing genes for selecting our functional (positive) instances. This results in low bias, high quality data labels.

Previous methods have only used mRNA expression profile data to characterize the mRNA isoforms. This limits the amount of information available to distinguish the function of different mRNA isoforms of the same gene. Due to limitations of mRNA isoform expression quantification, the expression profile of several mRNA isoforms of the same gene is highly similar. This makes distinguishing the function of such mRNA isoforms very difficult. To best characterize the mRNA isoforms, we include additional information derived from mRNA isoforms and their corresponding protein sequences.

Furthermore, none of the previous methods takes into account the tissue-specific functions of mRNA isoforms. A primary goal of this dissertation is to develop systems capable of predicting tissue-specific mRNA isoform functions. We introduce tissue-specific context in mFRecSys at two levels. At the first level, we use tissue-specific mRNA isoform expression profile as predictors. In the second level, we use a completely different set of tissue-specific mRNA isoform expression profile to generate labels that are tissue-specific.

### 5.2. Future Works

This work uses the labels obtained from GO annotations, pathway databases, and protein-protein interactions, without any tissue-specific information. However, it might be useful to integrate tissue-specific information when generating the labels. From my analysis in chapter 4, I have found that incorporating tissue-specific information during the label generation stage generally improves the system performance with respect to using organismal level global information. Therefore, the method for generating labels can be improved for chapter 3.

I have not utilized the power of mRNA isoform and protein structures, largely due to limited availability of data. However, it might be worth trying to use predicted secondary or tertiary structures of mRNA isoforms and proteins as more predictors. While the sequence information alone can be used to infer functions, the additional knowledge gained from structures can aid in improvement of performance.

In terms of implementation and availability of data, we have made all data, scripts and models freely available. Additionally, the material for TENSION and mFRecSys are available independently of each other. However, this limits the utility and might not be very user-friendly for those not comfortable working on command-line. In the future, I will develop a single unified singularity container and an R shiny web app to make the tools more accessible.