# ABSTRACT

Title of Dissertation: COMPUTATIONAL INVESTIGATION OF TRANSCRIPTOMIC AND GENETIC UNDERPINNING OF AGING AND HGPS

Kun Wang, Doctor of Philosophy, 2018

Dissertation directed by: Sridhar Hannenhalli and Kan Cao, Department of Cell Biology and Molecular Genetics

Normal aging is a complex process affecting everyone, and also a major risk factor for many complex diseases. Hutchinson Gilford progeria syndrome (HGPS) is a rare genetic disease with symptoms of aging at a very early age. There are some known and other presumed overlaps between HGPS and normal aging process. My goal in this dissertation is to perform computational investigation in both transcriptomic and genomic level to uncover potential underpinnings of these two models using high throughput genomic data.

Firstly in order to detect the common and distinct gene expression patterns between HGPS and normal aging, which might suggest their potential molecular links, I developed a novel approach that leverages co-expressed gene clusters to identify gene clusters whose expression co-varies with age and/or HGPS with limited sample size. Our results recapitulate previously known processes underlying aging as well as suggest numerous unique processes underlying aging and HGPS. Moreover, it is

known that alternative splicing contributes to phenotypic diversity at multiple biological scales, and its dysregulation is implicated in both aging and age-associated diseases in human. We aim to provide more insight into aging and age related diseases by studying splicing regulation. Then secondly we performed the first comparative investigation on splicing predictability of genomic and epigenomic features using a deep neural network model (DNN). We showed genomic features are the primary driver of splicing, and epigenomics is not contributing extra regulatory information independent to genomics. In addition, cross-tissue variability in splicing further complicates its links to age-associated phenotypes and elucidating these links requires a comprehensive map of age-associated splicing changes across multiple tissues. Thus thirdly we generate such a map by analyzing ~8500 RNA-seq samples across 48 tissues in 544 individuals. Employing a stringent model controlling for multiple confounders, we identify 49,869 tissue-specific age-associated splicing events of 7 distinct types. We find that genome-wide splicing profile is a better predictor of biological age than the gene and transcript expression profiles, and furthermore, age-associated splicing provides an additional independent contribution to age-associated complex diseases. In fact in this specific study we presented the first systematic investigation of age-associated splicing changes across tissues, and further strengthening the links between age-associated splicing and age-associated diseases. Besides aging factor, genetic variations also potentially contribute to age-related disease shown by GWAS studies. However, potential interactions between aging and genomic variations have not been elucidated fully. It is highly likely that phenotypic effect of systemic molecular changes through aging may depend on the genotype of

the individual. Lastly we approximate the environmental changes by age-associated changes in the levels of regulatory proteins, and exploiting the known mechanisms of transcriptional regulation, explore potential causal interaction between genotype and aging toward explaining age-related transcriptional and ultimately, age-related diseases. We detected numerous interactions across 25 tissues and showed they could potentially be associated with hypertension disease. In summary, our investigations in this dissertation provided predictive hallmarks along with implied molecular basis insight about normal aging and HGPS in transcriptomic and genetic level.

# COMPUTATIONAL INVESTIGATION OF TRANSCRIPTOMIC AND GENETIC UNDERPINNING OF AGING AND HGPS

by

Kun Wang

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Sridhar Hannenhalli, Chair
Professor Kan Cao, Co-Chair
Professor Stephen Mount
Professor Max Leiserson
Professor Mihai Pop, Dean’s Rep

# Dedication

This dissertation is dedicated to my family,

my Mom, my Dad, my wife and my son,

for their unlimited love, support and encouragement.

# Acknowledgments

First, I would like to express my gratitude to my two academic advisors— Dr. Sridhar Hannenhalli and Dr. Kan Cao. This dissertation could not be accomplished without their continuous support and guidance. It is my huge pleasure to get the opportunity to join their labs and work with them. I really appreciate the professional training and unlimited encouragement I have received from them, which transformed me to a better scientist and human being.  I am also grateful for everything they have done for my family as friends, which make our life easier.

Second, I would also like to acknowledge my committee members: Dr. Steve Mount, Dr. Hector Bravo and Dr. Mihai Pop. I really appreciate their time and helpful suggestions, which definitely made my dissertation much better. Especially for Dr. Steve Mount, thank you so much for being always patient and helpful when I kept stopping by. I want to thank Dr. Hector Bravo for all his amazing ideas and suggestions, and also thank Dr. Mihai Pop for always promptly responding to my queries and his warmth toward me.

In addition, I would like to thank all my lab-mates, who have already become my “family”. I cannot imagine whether I could survive through my PhD life without their companionship. Especially my two senior brothers – Dr. Justin Malin and Dr. Avinash Das, I cannot achieve this without all their help and support.

And also I am really grateful to the CBCB and BISI community, I appreciate all the resources and opportunities provided to us. Especially I would like to thank Molly Burk, Gwen Warman and Barbara Lewis, who have been always helping us out of the heavy policy work. And of course, I also really appreciate UMIACS staffs' technical support.

Finally, I would like to thank my family. I cannot forget all the efforts and love from my parents. I hope they will be proud of me at this moment. In the meantime, I want to say thanks to my wife, I appreciate her unconditional love, support, understanding, patience and companionship. Last appreciation is for unlimited fun and responsibilities to me brought by my little boy, all of that makes me a better and mature person.

# Table of Contents

# List of Figures

# List of Abbreviations

PSI: percent splicing index

HGPS: Hutchinson Gilford progeria syndrome

TF: Transcription Factor

eQTL: Expression quantitative trait loci

SNP:   Single Nucleotide Polymorphism

PWM: Position Weighted Matrix

sQTL: Splicing quantitative trait loci

GWAS: Whole Genome Association Study

SR protein: protein domain that contain repeats of serine and arginine amino acid

ESE:  Exonic Splicing Enhancer

ESS:  Exonic Splicing Silencer

ISE:   Intronic Splicing Enhancer

ISS:   Intronic Splicing Silencer

NMD: Nonsense-mediated mRNA decay

CHIP: Chromatin Immunoprecipitation

DNase: Deoxyribonuclease

CLIP:  Cross-linking Immunoprecipitation

FPKM:  Fragments per Kilobase per Million reads

RPKM: Reads per Kilobase per Million reads

RBM:  Restricted Boltzmann Machine

LLR: log likelihood ratio

CF: Confounding Factor

# Chapter 1: Introduction

## 1.1 Aging process

### *1.1.1 What is aging?*

Aging is a longitudinal complex process affecting all human beings, during which significant deleterious physiological changes occur correspondingly, including wrinkled skin, increased blood pressure, slower metabolism, loss of muscle etc. Moreover age could significantly increase the chance to develop age related diseases (hypertension, cardiovascular disease, neurodegeneration, heart attack and all types of cancers etc) (Finkel et al., 2007; Niccoli and Partridge, 2012; Sinclair et al., 2012; Wyss-coray, 2015). Understanding the molecular genetic basis of aging process and the mechanisms about age related complex diseases are crucial to improve the health care during aging.

Onsets of Aging in different organs might vary, in other words aging could affect different organs at various speeds, which could attribute to tissue-specific gene regulation during the aging process. Cassano et al observed significant differential age related effect between liver and brain in the rat with respect to the mitochondria content (Cassano et al., 2004). In addition, Ismene Karakasilioti and George A. Garinis suggested age might have specific effects in adipose tissue due to its unique attributes ( ″fats, oil, lipid peroxidation and inherent propensity”) (Ismene Karakasilioti, 2014). Taken together, the tissue-specificity of aging effect could complicate the link between age and phenotypes.

### 1.1.2 Potential mechanisms of aging and age-related diseases

Although significant progress has been made in recent years, the underlying genetic and molecular basis of aging is still unclear, which is crucial for improving the medical health care and increasing lifespan. Two classical theories have been proposed to explain aging mechanism: damage theory, which hypothesizes that the accumulation of damage in cells causes the failure of the biological systems; programmed theory, which assumes an internal program controlling the aging process (Jin, 2010).

For decades, numerous studies have provided insight about hallmarks for aging with either computational or experimental evidence. Those hallmarks include "genomic instability, telomere attrition, epigenetic alterations, loss of proteostasis, deregulated nutrient sensing, mitochondrial dysfunction, cellular senescence, stem cell exhaustion, and altered intercellular communication" (Blasco et al., 2013). Genome instability supports the damage concept, which assumes that accumulation of damage could lead to system failure. The genome damage could refer to nuclear DNA, mitochondria or nuclear architecture (Blasco et al., 2013). It seems that these hallmarks could shed light on the molecular basis of aging process; however, how these factors are involved in the aging process in an integrated manner is still a mystery. In addition, more in vivo evidence supporting the links between those hallmarks and improvement for health care are needed.

### 1.1.3 Age-associated gene expression and alternative splicing changes

As discussed above, normal aging is a major risk for complex diseases. To understand the molecular basis about how aging contributes to those diseases, it is necessary to identify age related changes at the molecular level or potential predictors for the aging process and complex diseases. Gene expression is one crucial measurement reflecting gene function activity since it is supposed to be correlated with protein concentration. Alternative splicing, which significantly contributes to proteomics diversity, is also an important process in gene regulation. Thus age related changes in gene expression and splicing level are expected to provide insights about aging process and complex diseases. Previous related studies are as follows.

**Age related gene expression changes**: de Magalhães et al identified 56 over-expressed and 17 under-expressed genes which mostly are involved in immune response and inflammation process by using a linear regression over 27 microarray datasets (de Magalhães et al., 2009); Azad Kumar et al specifically re-assessed age related gene expression in different regions of brain tissues using microarray datasets, which suggested multiple gene signatures for aging and validated that gene RHBDL3 expression is correlated with age by RT-PCR (Kumar et al., 2014); Daniel Glass et al proposed a linear mixed model to detect age-associated gene across multiple tissues (skin, adipose, blood and brain) using microarray datasets from 865 individuals. They specifically detected more age-associated genes (1672) in skin tissues than others and showed age-associated changes are more likely to be tissue-specific (Glass et al., 2013); Yang et al developed a linear regression model, which explicitly controls for potential hidden variables and known confounding factors (gender, BMI and genotype) to detect age-associated gene expression using RNA-seq data across 9 tissues from GTEx consortium (GTEx Consortium, 2015). They identified numerous tissue-specific age-associated genes and suggested those genes could be related to complex diseases (Yang et al., 2015).

**Age-associated alternative splicing changes**: Tollervey et al detected age-associated splicing events among normal population using microarray datasets in brain tissue and also showed they are present in both frontotemporal lobar degeneration (FTLD) and Alzheimer's disease (AD), which suggested a potential link between aging and complex diseases (Tollervey et al., 2011b); Mazin et al identified age-associated splicing changes using RNA-seq across 35 individuals and in the meantime showed splicing factor expression change could be one potential cause for the splicing changes (Mazin et al., 2013a).

Although age-associated gene expression and splicing changes have been investigated, it is still a challenge to capture the causal factors among observed correlated predictors, and also the link between those predictors and complex diseases is still uncertain. In addition, technically some drawbacks and limitations of previous studies (specifically for age-associated splicing studies) need to be dealt with: (1) small sample size (2) missed control for confounding factors (3) limitation of microarray datasets (4) validation difficulty.

## 1.2 Age related diseases

### 1.2.1 Hutchinson–Gilford progeria syndrome (HGPS)

HGPS (Hutchinson-Gilford progeria syndrome) is an extremely rare (1 of 4 million live births) genetic disease, caused by one de novo point mutation within exon 11 of LMNA gene (Eriksson et al., 2003). Instead of the classical splicing site, the point mutation creates a novel splice site, causing the generation of the mutated protein – progerin, shown in Fig. 1-1 (Eriksson et al., 2003). The LMNA proteins (LMN A/B) are the major nuclear structure components and also interact with various proteins and chromatins (Andrés and González, 2009). The lack of LMN proteins unexpectedly causes cell nuclear blebbing of HGPS patients. HGPS patients usually have pronounced forehead, short stature, hair loss, a "pinched" nose, and extreme lipodystrophy etc (Merideth et al., 2008). What's more, they usually suffer severe organ degeneration and coronary artery disease. Though the cause of HGPS – the point mutation has already been discovered, the mechanisms linking the point mutation to the symptoms are still not clear. To uncover that, a number of hypotheses and studies have been developed and performed, among which gene expression model is one. It is proposed that progerin alters the nuclear structure and subsequently affects gene expression as well as regulates gene expression via interacting with some gene regulator proteins. In other words, pathway signals related to the clinical manifestations of HGPS are altered by the overexpression of progerin or the lack of LMN proteins. To date, both microarray and RNA-seq have been employed to identify differentially expressed genes between control and HGPS patient samples (Ly et al., 2000; Park et al., 2001).

Biological functions like transcription factors, extracellular proteins, and cell cycle regulators consistently show up. In addition, due to similar symptoms, HGPS is also treated as a reasonable aging model to study with the expectation that the mutated protein is also linked to the normal aging process. Progerin was actually observed in normal dermal fibroblast cell lines (Rodriguez et al., 2009). However, there is still no strong evidence for the link between progerin and normal aging. A comparative investigation on pathways between HGPS and normal aging is needed to uncover the mystery.

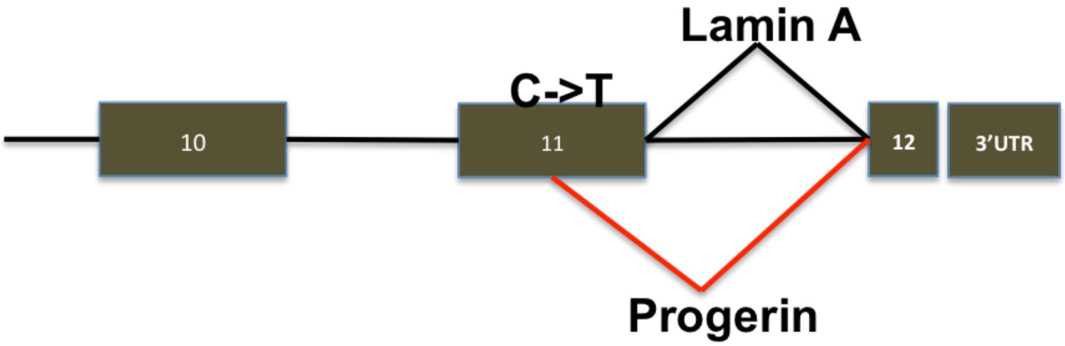


**Figure 1-1: The novel mutation within exon 11 of LMNA gene (C-> T).**

### 1.2.2 Hypertension

Hypertension is a complex disease, which was clinically characterized as having systolic blood pressure persistently higher than 140 mm hg and diastolic blood pressure persistently higher than 90 mm hg (Pinto, 2007). Hypertension is significantly related to biological age and could also potentially contribute to other cardiovascular diseases (Drazner, 2011; Franklin and Wong, 2013; Kokubo, 2014).

Investigation on the molecular basis of hypertension pathology could improve the prevention and therapy strategies of cardiovascular related diseases.

Complex diseases could be caused by the interaction between gene regulation and environmental factors including diet (Kuneš and Zicha, 2009; Olden et al., 2014; Renz et al.). Those hidden confounding factors and variables make the investigations more challenging. Differential gene expression is usually an important type of predictor for diseases. Huan et al identified 34 genes associated with hypertension across 7017 individuals (Huan et al., 2015). Basu et al provided a tissue-specific map of hypertension related genes across dozens of primary tissues using GTEx data (Basu et al., 2017); Chiang et al specifically performed an investigation of hypertension related genes in Han Chinese population of Tai Wan and they showed those genes could be involved in inflammation, visceral fat metabolism and homeostasis (Chiang et al., 2018).

In addition, hypertension exhibits population bias to some extent (Bosu et al., 2017). For example, the prevalence of hypertension among Africa Americans is significantly higher than that of Caucasian (Fuchs, 2006; Ortega et al., 2015), which could be partly attributed to genetic background. GWAS have identified ~120 SNPs significantly related to elevated blood pressure and hypertension (Franceschini et al., 2013; Ortega et al., 2015; Sofer et al., 2017; The UK Biobank Cardio-metabolic Traits Consortium Blood Pressure Working Group et al., 2018; Wain et al., 2017) , and Basu et al also showed that numerous eQTLs that explain the variance of gene expressions are associated with hypertension disease (Basu et al., 2017). However, the precise mechanism exhibiting the regulatory roles of genome variations is not certain. In addition, as a risky "environment" factor, age is significantly related to hypertension, and the interplay between genetic background and age factor is uncertain. Comprehensive studies combining regulation and association investigation is needed to explore the causalities in complex diseases.

## 1.3 How genetic information flows in biological system?

### 1.3.1 Central Dogma

It is known that inherited genetic information is contained in double strand DNA; thus it is crucial to understand how the genetic information flows in the biological system, which guides the investigations on gene regulation. What's more, uncovering the disrupted regulation that causes common diseases could provide better therapeutic strategies for the diseases. Francis Crick first stated central dogma, which assumes the genetic information contained in DNA flows to messenger RNA during transcription process in the nucleus, and then mRNA is translated to protein in the cytoplasm via translation process (Crick, 1970). Many regulatory proteins and complexes are involved in transcription, alternative splicing and translation. Disruptions in any of these regulatory processes by mutations, environmental factors and their interactions may cause diseases.

### 1.3.2 Tissue-specific transcription and alternative splicing

Although all eukaryotic cells in an organism have almost identical genomes, they have to undergo differentiation to generate tissues, which exhibit diverse functions. For example, heart smooth muscle cells could induce heart beatings, which cannot be accomplished by skin fibroblast cells. The functional specificities are governed by tissue-specific gene regulation, which includes both transcription and alternative splicing.

Transcription is a regulatory process, in which DNA is transcribed to mature message RNA (Lee and Young, 2013). The mRNA levels usually referred to as gene expression level are correlated with corresponding protein concentrations (Li et al., 2014). Thus tissue-specific gene expression should be crucial predictors for tissue-specific functions. Tissue-specific regulation could attribute to tissue-specific regulatory pathway and combinatorial interactions among transcription factors (Cheng et al., 2012; Todeschini et al., 2014). In addition, epigenomics also plays important roles in tissue-specific regulation (Eccleston et al., 2013). Unlike tissue-specific genes, housekeeping genes, which are involved in broad biological processes, are expressed ubiquitously across tissues (Eisenberg and Levanon, 2013; Kouadjo et al., 2007) .

Alternative splicing is a regulatory process, in which a gene locus encodes multiple transcripts leading to multiple protein isoforms (Chen and Manley, 2010). Alternative splicing regulation is also tissue-specific (Barash et al., 2010; Streuli and Saito, 1989; Zhang et al., 2008a). Barash et al reported the "splicing code" which can predict splicing trends using motifs involved in splicing regulation in a tissue-specific manner (Barash et al., 2010). Tissue-specific splicing regulatory proteins could play essential roles in splicing regulation; in addition, there is some evidence that epigenomics, which also exhibits tissue-specificity, could also affect splicing regulation (Luco et al., 2011; Wang et al., 2017).

In summary, tissue-specific gene regulation could contribute to cellular function variations across tissues, and moreover it complicates the investigation of complex diseases. It is crucial to uncover context specific regulation modules to solve fundamental problems and improve therapeutic strategies for complex diseases.

## 1.4 Transcription regulation

### 1.4.1 Cis-regulatory elements (enhancers and promoters) for transcription

Transcription is a regulatory process, during which gene is transcribed to mRNA (Lee and Young, 2013). Numerous regulatory elements mediate the transcription process in a tissue-specific manner. Promoters and enhancers are two essential cis-regulatory elements. Promoters are usually defined as DNA segments extending from the transcription start site of genes, and containing short DNA motifs, which could be bound by basal transcription factors. The major transcription complex -- RNA polymerase II usually assembles at promoters to initiate the transcription process. In addition, another unique feature for promoters is TATA-box, which is a DNA region enriched for AT and could be bound by TBP (TATA-binding) protein to specify where transcription starts (Patikoglou et al., 1999; Yang et al., 2007). However only a subset of promoter have a TATA-box. The activity of promoters is significantly related to transcription level, and promoters could be mediation targets for many regulation mechanisms. For example, DNA methylation could completely shut down a gene by masking its promoter (Maurano et al., 2015; Yin et al., 2017; Zhu et al., 2017). In addition, genes could have multiple alternative promoters.

As another crucial cis-regulatory element for transcription regulation, the definition of enhancers is not as clear as that of promoters, however usually they could be any DNA segments falling upstream, downstream and within genes, which could enhance the activation of genes. Enhancers contain DNA motifs, which could be bound by transcription factors. The distance between enhancers and transcription start site varies much and it could be up to 1 Mb (Krivega and Dean, 2013; Marsman and Hors, 2012). The investigations of enhancer activities confront several challenges (Pennacchio et al., 2013): 1) mostly falls within noncoding regions which cover ~98% of human genome; the regulatory role of noncoding regions is still a big mystery; 2) context-specific transcription variations imply the complexity of enhancer activities; 3) difficult to determine the target gene; 4) polymorphisms could induce variations to enhancer activities.

The interaction between enhancer and promoter is a fundamental mechanism to enhance transcription, in which distal enhancer could spatially loop over to interact with promoters through physical interactions between numerous transcription factors (Krivega and Dean, 2013; Marsman and Hors, 2012).

### 1.4.2 Transcription factor binding

Transcription factors are key regulators of transcription; they could bind to both enhancer and promoter regions to mediate transcription (Todeschini et al., 2014). The short DNA segments bound by transcription factors are named TF binding sites.  also contribute to cell and tissue-specific transcription responses (Todeschini et al., 2014; Yu et al., 2006). Multiple transcription factors and co-factor could form a regulatory module complex to enhance transcription (Kaufmann et al., 1998; Reiter et al., 2017). Different TFs have a distinct preference for the binding sites, which are usually conserved (Dermitzakis and Clark, 2002). Disruption of binding sites could cause mis-regulation of transcription and diseases (Lee and Young, 2013). Validated transcription factor binding sites are extremely limited and investigation of transcription factor binding across tissues could help shed light on the underlying context-specific regulatory roles of transcription factors. What's more, robust prediction model could be used to estimate the potential transcription response and phenotypic outcome of mutations falls within TF binding regions (Lee et al., 2015).

However, the experimental techniques (such as Chromatin ImmunoPrecipitation) are limited and expensive for large-scale studies; dozens of computational approaches have been developed to achieve this goal (Bailey et al., 2009; Grant et al., 2011; Levy and Hannenhalli, 2002). The most widely used approach is based on position weighted matrix (Stormo and Schneider, 1982), which is basically 4×N matrices providing the probabilities or likelihoods that respectively 4 types of nucleotides ('A', 'C', 'T', 'G') are observed at position n (n = 1…N). The position weighted matrices are usually derived from validated transcription factor binding sites using experimental techniques (SELEX, PBM, CHIP-Seq, CHIP-CHIP) (Hu et al., 2010; Portales-casamar et al., 2010). The power and accuracies may vary with different approaches. The two most well-known databases of PWMs for human transcription factors are TRANSFAC (commercial) (Wingender et al., 1996) and JASPER (open access) (Portales-casamar et al., 2010).

Most TF binding sites prediction packages (PWM-scan, MEME, FIMO etc) scan the query DNA sequence using position weighted matrices to estimate the likelihood for a potential TF binding hit. However usually the dependencies between neighboring positions (adjacent nucleotides) are not considered (Tomovic and Ã, 2007). Relatively ClusterBuster takes into account of the neighboring dependencies using HMM model, which could predict a cluster of transcription factor binding sites module (Frith et al., 2003). In addition, some methods based on supervised learning methods are also developed (Qin and Feng, 2017; Salekin et al., 2017), which split the validated TF binding sites sequence to k-mers as features to train the model and then perform prediction.

Although consensus information contained in PWM could be powerful for predicting putative binding sites, it cannot take into account tissue-specificity, which actually could be overcome by using tissue-specific epigenomic features – DNase-Seq and ChIP-seq Histone modification data. Combining epigenomic features with putative TF binding sites would provide more accurate tissue-specific binding sites predictions  (Chen et al., 2017a; Ka and La, 2015).

### 1.4.3 Epigenomics related to transcription

As we know from the central dogma, genetic information is contained in DNA sequence, however completely identical nucleotide sequence could not fully explain the tissue/cell specific gene expression variations (Jaenisch and Bird, 2003). “Epigenetics” which refers to any functional heritable chromatin state information instead of DNA sequence, could potentially contribute to gene regulation and be in charge of the expression variations (Gibney and Nolan, 2010; Jaenisch and Bird, 2003). Monozygotic twin studies are playing crucial roles to uncover epigenetic effect and suggest that epigenetic features actually involved in gene regulation (Bell and Spector, 2011). These epigenetic features could be chromatin accessibility state, DNA methylation, Histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me1 et al). Ultimately the interplay between these functional markers and genetic background will determine the gene expression profile.

Although epigenetic markers’ roles in gene regulation might vary, to target transcription factor binding is definitely one of the common mechanisms (Maurano et al., 2015; Moore et al., 2012; Zhu et al., 2017). DNA methylation mostly could shut down the TF binding in enhancers and promoters regions to inactivate transcription (Curradi et al., 2002); Chromatin remodeling process could rearrange chromatin accessibility states which then affect the TF binding landscape accordingly (Luo and Dean, 1999; Steger and Workman, 1996); multiple histone modification markers are observed related to enhancer and promoter activities (Creyghton et al., 2010; Zentner and Scacheri, 2012); for example H3K4me1 is a marker for enhancers, H3K4me3 is a marker for promoters, H3K27ac is a marker for active enhancers.

The advent and development of next-generation sequencing technology accelerated the investigation of epigenetics by providing more accurate high throughput data. ChIP-seq data could provide sensitive histone modification signals, DNase-Seq is a powerful tool to estimate the tissue-specific DNA accessibility and Bisulfite-Seq could provide high resolution of DNA methylation measurement (Sarda and Hannenhalli, 2014). Both ENCODE and Roadmap provides numerous tissue or cell-specific epigenetic high throughput data (Roadmap Epigenomics Consortium et al., 2015; The ENCODE Project Consortium, 2012), which significantly promoted human genetic studies.

### 1.4.4 Expression quantitative trait loci study (eQTL)

Individual specific gene expression variations are observed in the large population, partly determined by genome variability (Zhang et al., 2008b). Expression quantitative trait loci (eQTL) study is the standard approach to detect gene expression associated genome polymorphisms -- more specifically, single nucleotide polymorphism (SNP) (Gilad et al., 2008). Gene expression associated SNPs detected in eQTL studies are also called eSNPs. The distance between eSNPs and their potential target gene transcription start site could vary a lot, usually the eQTLs are divided to cis-eQTLs (similar locations) and trans-eQTL (different locations and chromosome) dependent on the locations (Gilad et al., 2008). eQTL studies are usually performed in a tissue-specific manner and exhibit the tissue-specific effect. cis-eQTL mostly are more consistent across tissues relative to trans eQTLs which show strong tissue-specificity (Westra and Franke, 2014). The detection power is sensitive to the population sample size. In addition, although the individual effect of single SNP is usually small, the accumulation of those effects could significantly affect gene regulation (Petretto et al., 2006).

Numerous computational approaches have been developed to detect eSNPs in a tissue-specific manner, and the basic well known approach is to measure the association between any SNP within a certain distance and the target gene expression in a linear model (Rantalainen et al., 2015). Some hidden variables could be derived over gene expression profile using packages like PEER, SVA and even PCA to control for the hidden confounding factors in the population, which could significantly increase the detection power and accuracy (Leek et al., 2012; Stegle et al., 2012). One of the challenges in eQTL study is linkage disequilibrium regions, where SNPs are correlated with each other and it make the differentiation of real causal polymorphisms difficult (Gilad et al., 2008; Westra and Franke, 2014).

GTEx is a publicly funded effort to gather large-scale genotype and gene expression data in human primary tissues. Currently, it has data for ~52 tissues with transcriptome expression, imputed SNPs and phenotype information. This is meant to support tissue-specific gene regulation investigations and provide more insights into the underlying molecular basis of expression variations (GTEx Consortium, 2015). Along with the raw data, tissues-specific eQTL results are also provided, which were predicted using matrix-eQTL package (Shabalin, 2012), a widely used efficient approach for eQTL analysis. GTEx consortium provided a comprehensive reference of eSNPs and potential corresponding targets genes across 46 tissues.

eQTL studies are expected to provide  molecular insight into the mechanism, by measuring the associations between SNPs and gene expression. One potential mechanism is that SNP could fall within enhancers and promoters to disrupt a transcription factor binding site or create a cryptic do novo TF binding site (Shi et al., 2016).  Since many eSNPs fall within non-coding regions, eQTL studies might be able to provide an understanding of the regulatory role of non-coding regions in the genome. In order to detect potential functional and causal SNPs, a better idea is to employ more biological insight into the model instead of only increasing the detection power by adopted more advanced statistical techniques. Based on the assumption that SNPs can fall within regulatory regions, transcription related epigenetic markers could be used to imply the regulatory potential of the SNPs (Acharya et al., 2017; Das et al., 2015). Das et al developed a Bayesian model to detect causal SNPs by employing epigenetic markers information (Das et al., 2015).

### 1.4.5 Genome Wide Association Study (GWAS)

Whole Genome Association Study is an approach to measure the association between genome variations and phenotypic traits (Bush and Moore, 2012; Visscher et al., 2017). In GWAS, the genome variations are usually single nucleotide polymorphisms which could be identified by performing whole genome sequencing or SNP imputation based on large population whole genome sequencing reference datasets like 1000 genome (Li et al., 2009). The phenotypes could be any traits like height, body mass index (BMI) or any clinical outcomes (blood pressure, diseases etc.). Basically GWAS measured whether some SNP alleles are observed significantly more frequently in the disease group than the control group (Bush and Moore, 2012; Visscher et al., 2017). Large population sample size is needed to perform an accurate analysis. In addition, similar to eQTL studies, potential confounding factors need to be dealt with to improve the detection power (Leek et al., 2012; Stegle et al., 2012).

The first GWAS was performed in 2005 by Haines et al to explore Age related Macular degeneration associated SNPs (Haines et al., 2005). Colorectal cancer is another significant genetic diseases and multiple groups have already identified associated common SNPs using GWAS (Broderick et al., 2007; Cogent Study, 2008; Tomlinson et al.; Wang et al., 2014a). In addition, blood pressure is a trait related to hypertension disease and persistent high blood pressure could potentially induce cardiovascular disease. About 120 SNPs have been identified to be related to elevated blood pressure (Ehret and Teresa Ferreira et al, 2016; Human et al., 2016; Kato et al., 2015; The UK Biobank Cardio-metabolic Traits Consortium Blood Pressure Working Group et al., 2018). At this point, hundreds of GWAS have been published with thousands of associated genomic loci (Fong et al., 2018; Macarthur et al., 2017). However it is still challenge to apply the results to improve therapeutic strategies with the actual mechanisms are largely unknown. But GWAS does provide an effective database to test for enrichment for the potential linked phenotypic outcome for predicted functional SNP candidates. It is expected GWAS and eQTLs could be aligned to imply more causal SNPs and potential mechanisms.

### 1.4.6 Interactions between germline mutation and envirmental factors

In general genome variations have been shown related to complex genetic diseases, for example ~120 SNPs are associated with hypertension diseases and blood pressure (Adeyemo et al., 2009; Wain et al., 2017). In addition, those complex diseases are most significantly associated with age – a critical risk factor (Niccoli and Partridge, 2012; Sinclair et al., 2012). It is expected that the interplay between variations and age factor could also play essential roles in the molecular basis (Kuneš and Zicha, 2009). Interactions should be observed through age dependent effect for SNPs.

Multiple previous studies have incorporated SNP-age interaction terms in eQTL model to measure the association between the interaction and target gene expression; Smino et al detected ~9 SNPs which have age dependent effect in blood pressure using a meta-analysis approach (Simino et al., 2014). Dongen et al showed the interaction between genetics and age factor could be associated with methylome in whole blood (Dongen et al., 2016). However, the molecular basis of how they interact with each other is not clear. Also, simply identifying a statistical interaction between Age and SNPs is not sufficient to provide insight into the interaction mechanism. Specific mechanism hypothesis based on a reasonable biological model should be proposed to dig down.

## 1.5 Alternative splicing regulation

### 1.5.1 What is alternative splicing?

Alternative splicing is a regulatory process during which multiple isoforms are produced (Black, 2003). It was first reported by Berget et al (Berget et al., 1977) in the adenovirus model in 1977.  Actually alternative splicing is prevalent in eukaryotic cells and almost 90% genes with multiple exons undergo alternative splicing (Wang et al., 2008). It is evident that alternative splicing significantly contributes to proteomics complexity and diversity across cell types and species. In general, there are seven different types of alternative splicing event as Fig. 4-1B shows.  The exon skipping event is the most prevalent and the best studied one.

### 1.5.2 Cis-regulatory elements for splicing

Alternative splicing is carried out by Spliceosome complex following RNA Polymerase II. Spliceosome complex is composed of five ribonucleo protein U1, U2, U4, U5, U6 and many auxiliary proteins (Black, 2003; Wang and Burge, 2008; Will and Lührmann, 2011). The spliceosome crucially regulates alternative splicing by recognizing splicing sites and defining exon/intron boundaries (Black, 2003; Wang and Burge, 2008; Will and Lührmann, 2011). The splicing sites include donor sites and acceptor sites, which refer to the 5' and 3' ends of introns respectively. ~95% of donor sites and acceptor sites are "GT" and "AG" respectively in the human genome (Black, 2003; Mount, 2000). The strength of splicing site signals could significantly affect the efficiency of splicing process, in other words stronger splicing sites are more likely to be detected by the spliceosome.

Auxiliary proteins (SR protein or hnRNPs) could bind to splicing cis-regulation elements (splicing enhancers and silencers) to enhance or inhibit the recognition of splicing sites by Spliceosome (Black, 2003; Chen and Manley, 2010). Alternative usage of splicing sites leads to diverse transcriptome. The mechanisms for splicing factors and cis-regulatory elements could vary dependent on involved regulatory protein and the genomic context, for example PTBP1 is a known splicing repressor and it could repress the inclusion of Pbx1 exon7 in neuron cells (Linares et al., 2015). In addition, SRSF1 could enhance the inclusion of DBF4B exon 6 in colon cancer cells (Chen et al., 2017b). Usually the splicing enhancer and silencer are divided into four categories based on their locations and function: ESE (exonic splicing enhancer), ESS (exonis splicing silencer), ISE (intronic splicing enhancer), ISS (intronic splicing silencer). Barash et al suggested a splicing code which could accurately predict relative exon inclusion in a cell type using numerous genomics features (Barash et al., 2010). What’s more, a branch site, which usually resides upstream from the acceptor site, is also crucial for the splicing cutoff since it could be bound by U2 to form the catalytic center (Black, 2003). Nonsense-mediated mRNA decay (NMD) is another significant mechanism affecting alternative splicing level, which could be induced by premature termination codons within the mRNA sequence (Maquat, 2004).

### 1.5.3 Splicing Factor binding

Although it is known that Splicing regulatory proteins play essential roles in splicing regulation, the mechanisms are far from certain. Searching for splicing sites and splicing factor binding sites are important to uncover the mechanisms. As for splicing site prediction, the most widely used approach is based on consensus sequence comparisons, one of which is Chris Burge's MaxEntScan tool which accepts an adjacent sequence and reports strength scores for the potential splicing site candidates (Yeo and Burge, 2004). There are also some other tools based on machine learning model trained over benchmarked k-mer datasets (Dogan et al., 2007). In addition, conservation could be another hallmark of splicing sites, Prichard's group filtered out noisy splicing junctions based on conservation score of splicing sites (Pickrell et al., 2010).

For splicing factor binding sites, specific experiments have been developed: cross-linking immunoprecipitation (CLIP) is one approach to detect protein-RNA interactions by using both UV cross-linking and immunoprecipitation (Jensen and Darnell, 2015). CLIP-seq which combined CLIP experiment and high throughput techniques could promise genome wide sensitive and accurate protein-RNA interaction signals (Stork and Zheng, 2017). There are also some other varied CLIP based approaches (PAR-CLIP, iCLIP and sCLIP) with specific advantages and limitations (Cook et al., 2014). Ray et al published a comprehensive map of RNA-binding motifs related to 205 proteins across 24 eukaryotes (Ray et al., 2013), which could be a powerful reference for splicing factor binding sites prediction. The CLIP-seq data from ENCODE project is also increasing rapidly. As for the methodology aspect, multiple approaches have proposed to predict splicing factor binding sites, most of which (SFmap, SpliceAid, ESEfinder) determine binding sites based on the consensus sequence derived from validated binding sites, genomic environment and conservation (Akerman et al., 2009; Cartegni, 2003; Paz et al., 2010; Piva et al., 2018).

### 1.5.4 Epigenomic regulation of splicing

Since alternative splicing is coupled to transcription (Schor et al., 2013), it is highly likely that epigenomic features could regulate splicing, which has been supported by numerous studies (Enroth et al., 2012; Pradeepa et al., 2012; Zhou et al., 2012). Correlations between several chromatin marks and splicing have been observed by Shindo et al (Shindo et al., 2013), which suggests that histone marks could be involved in the mechanism of splicing regulation. Actually they could affect the recruitment of splicing regulatory protein indirectly to regulate splicing (Luco et al., 2011). For example, MRG15 could recruit PTB to regulate alternative splicing by binding to H3K36me3 (Luco et al., 2010), which implies that some histone modification features may be able to enhance the recruitment of splicing factors to regulate splicing. The RNA polymerase II elongation rate could affect splicing by mediating the splicing sites selection (Shukla et al., 2011), more specifically weak splicing sites might not be detected by spliceosome at high elongation rate. Thus the mediation of elongation rate could modify alternative splicing. One of the potential mechanisms could be related to the nucleosome, which could specifically work as barriers that slow down RNA polymerase II (Schwartz et al., 2009). Multiple previous studies have observed significant association between nucleosome occupancy and splicing level (Brodsky et al., 2005; Schwartz et al., 2009). DNA Methylation is another epigenetic marker, which plays a role in splicing regulation. Bound CTCF is also able to decrease RNA polymerase II elongation rate, which could enhance splicing level with weak splicing sites involved. DNA methylation could avoid the binding of CTCF, which consequently affect splicing in a reverse manner (Shukla et al., 2011).

Although some specific mechanisms and association regarding of epigenetic regulation of splicing have been observed, a genome wide analysis is needed to further strengthen proposed general mechanisms. In addition, it seems that epigenetic markers' regulatory role in splicing could be dependent on cis-regulatory elements. The investigation on the interplay between epigenetic markers and cis elements is needed. Moreover, underlying tissue-specificity in epigenetic regulation of splicing needs to be addressed.

### 1.5.5 Potential links between alternative splicing and complex diseases

Alternative splicing regulates almost all the protein-coding genes and significantly affects the protein structure variation (Wang et al., 2008). Mis-regulation of alternative splicing could cause numerous human diseases (Wang and Cooper, 2007a). As one of the most critical factors, a mutation may happen in splicing sites, spliceosome complex components, splicing factor binding sites (Li et al., 2016; Wang and Cooper, 2007a), which could potentially cause lack of specific functional protein or generation of harmful isoforms.  LMNA is one of the fragile genes, which is related to many diseases induced by mutations (Eriksson et al., 2003; Prince et al., 2001). As mentioned above, HGPS is caused by a de novo mutation within exon 11 of LMNA and creates a cryptic splicing site. Instead of LMNA protein, which could support the nucleus membrane scaffold and interact with numerous regulatory proteins, the mutated dysfunctional protein--progerin is generated. Familial partial lipodystrophy type 2 (FPLD2) is caused by a mutation in one 5' splicing site of LMNA gene, leading to intron retention, which triggers NMD process (Tu and Ara, 2016). Furthermore, alternative splicing could be linked to cancer. Some known hallmarks of cancers like apoptosis and metastasis could be affected by mis-regulation of alternative splicing (Sveen et al., 2015). SRSF1 is a crucial splicing factor that could promote the assembly of spliceosome complex and also play roles in the translation process (Black, 2003; Chen and Manley, 2010). SRSF1 is significantly over expressed in multiple types of tumors and has already been used as an important biomarker (Sveen et al., 2015). At the molecular level, it is known that SRSF1 could interact with many tumorigenesis related genes like MYC (Das and Krainer, 2014; Sveen et al., 2015). In addition, Shen at al showed splicing variation model could predict survival of cancer patients better than gene expression model (Shen et al., 2016), which implies that splicing might play crucial roles in tumorigenesis.

### 1.5.6 Splicing quantitative trait loci study (sQTL)

Alternative splicing is crucial for gene regulation and complex disease and numerous biological factors could affect splicing regulation. Genome variation probably is also one of them. For example, Heinzen et al showed that a polymorphism falling within splicing regulatory cis-elements could affect the ratio of SCN1A isoforms, which are related to drug-responsive (Heinzen et al., 2007). Splicing quantitative trait loci (sQTLs) is a standard approach to identify SNPs which are associated with splicing levels (Monlong et al., 2014; Zhao et al., 2013). Many different computational approaches ranging from the simple linear regression model to more advanced statistical model accounting for over-dispersion have been developed, and numerous sQTLs have been detected. Compared to eQTLs, it seems sQTLs exhibit some unique characteristics. They usually have a smaller effect than eQTL and reside not too far from the target genes/events (Monlong et al., 2014; Zhao et al., 2013).

One of the major steps for sQTL studies is splicing level quantification. Actually splicing complexity is much higher than we expected, and the standard splicing annotation may not be sufficient to capture all the local splicing variations. Li et al developed "LeafCutter", an intron-centric splicing quantification approach, which basically clusters competing splicing junctions as generalized splicing events and in the meantime sQTL could be detected by assessing the association between SNPs and the ratio of splicing junctions within each cluster  (Li et al., 2018). The advantage is the quantification will not be limited to fixed exon-centric annotation, which is far from complete. In addition, regarding tissue-specificity, consistent with eQTL studies that a comprehensive map for sQTL across multiple tissues is needed to fully understand tissue-specific splicing regulation interacting with polymorphisms.

### 1.6 Significance of this study

HGPS is a rare premature genetic disease. Although the causal de novo mutation has already been identified, its molecular basis is still not certain. What are more, common symptoms between HGPS patients and seniors imply potential connections in molecular basis between the two models, which raises the difficult and interesting question. Gene expression study is expected to be able to shed light on both processes. However, the limited RNA-seq sample size in the context of HGPS further complicates the challenge. In chapter 2, we provided the first set of matched HGPS and aging RNA-seq samples and proposed a novel approach, effective with limited sample size, to identify gene clusters whose expression co-varies with age and/or HGPS. We have applied our approach to our novel RNA-seq profiles at three different cellular ages, both from HGPS patients and normal samples. Our results suggest novel insights into biological processes underlying aging and HGPS. In addition, our novel approach could be applied to any general problems with limited sample size.

It is well known that alternative splicing significantly contributes to proteomics diversity and mis-regulation of splicing can cause diseases in human. Although both genomic and chromatin features have been shown to associate with splicing, the corresponding specific mechanism is not clear. Moreover, it is not known whether the regulatory effect is context dependent. In chapter 3 we predict exon skipping level using a deep neural network model and performed the first comparative investigation on splicing regulatory effect of genomic and epigenomic features. Our analysis showed genomic features are the primary drivers of splicing, and chromatin features probably are not contributing extra regulatory information independent to genomic features.

As we mentioned above, Alternative splicing is one of the key drivers for phenotypic diversity and its dysregulation could underlie diseases in human. Tissue-specificity in splicing further complicates its links to phenotypes. To elucidate these links we generated a comprehensive map of age-associated splicing changes across 48 tissues in Chapter 4. More specifically, we identified 49,869 tissue-specific age-associated splicing events of 7 distinct types using a stringent model controlling for multiple hidden confounders by analyzing ~8500 RNA-seq samples across 48 tissues. Moreover, we also showed those age-associated splicing could be linked to complex diseases. In addition, we performed a comparative investigation on predictive ability of age among gene expression, transcript expression and splicing level, showing that splicing level is most correlated with age.

It is known that both normal aging and genomic variation could be linked to age related complex diseases. However, a potential interplay between the two factors is not understood fully. It is highly likely that genomic variations exhibit age dependent effect on phenotypes. Previous studies have incorporated SNP-Environment term in a regression model to study such interactions. However, those interaction terms do not provide insights about the interaction mechanisms. We instead incorporate well understood transcriptional regulation mechanism in our association model. More specifically, we approximate the age factor as age-associated TF concentration changes and explore potential causal interaction effect between genotype and age-associated TFs. In addition, we performed analysis in 25 tissues to account for tissue-specificity. Numerous SNP-TFs interactions are identified across 25 tissues and enriched epigenomic signals associated with regulatory elements further validate their functionality.

Overall, my dissertation focuses on fundamental questions in gene regulation and in the meantime specifically provides new insights into aging process and age related complex diseases.

# Chapter 2: Phenotype-Dependent Co-expression Gene Clusters: Application to Normal and Premature Aging

## 2.1 Introduction

Although the causal of HGPS is already known, the mechanisms leading to the clinical manifestations are still mysteries. Among numerous hypotheses, the ”gene expression” model, which proposes that progerin alters the nuclear structure and subsequently affects gene expression, has been supported by various lines of evidence (Mounkes et al., 2003). A general loss of heterochromatin and dislocation of epigenetic marks have been observed in HGPS cells (Goldman et al., 2004; Mccord et al., 2013; Shumaker et al., 2006). In addition, it has been shown that lamin A interacts with transcription regulatory proteins (e.g., retinoblastoma protein pRb), signaling molecules (e.g., protein kinase C), and chromatin proteins (e.g., histones and barrier-to-autointegration factor (BAF)), implicating its direct involvement in gene expression and signaling (Gotzmann and Foisner, 2006; Somech et al., 2005; Wilson and Foisner, 2010).

Accordingly, changing expression levels of various genes have been observed in HGPS cells. To date, four independent HGPS microarray studies have been published. Park et al. examined 384 known genes and reported four genes with more than twofold changes (Park et al., 2001). Ly et al. monitored the expression of approximately 6,000 genes and found 61 altered in HGPS (Ly et al., 2000). Csoka et al. analyzed approximately 33,000 genes and found 361 genes that showed statistically significant change (Csoka et al., 2004), and more recently, Marji et al. compared 4 HGPS fibroblast lines with four age-matched controls, and suggested that a lamin A-Rb signaling is a major defective signaling pathway in HGPS cells (Marji et al., 2010). While these microarray studies are not in complete agreement with each other, transcription factors, extracellular proteins, and cell cycle regulators appear to be the largest affected functional category.

As the relationship between nuclear lamins and gene expression is continued to be explored, we are optimistic that the gene expression model may help to shed light on the causes of the premature aging phenotypes associated with HGPS. On the other hand, it is of great interests to determine how the gene expression pattern in this disease resembles and is distinct from the pattern observed in normal aging. A detailed comparative investigation of genome-wide gene expression patterns associated with HGPS and normal cellular aging has not yet been reported and may reveal common and distinctive biological pathways underlying these two conditions. Comparative exploration of gene expression changes in normal aging and HGPS has not been possible thus far due to unavailability of genome-wide expression profiles in HGPS samples at different cellular ages. Thus, in this study, we have collected RNA samples from a HGPS primary fibroblast cell line and from a genetic background matched normal control at early, middle and late cellular passages, and conducted genome-wide RNA-seq.

Although, we have generated the first whole genome RNA-seq based transcriptomic profile in cell cultures at three different ”ages” in both normal and HGPS samples, the number of samples (n = 6) is not sufficient to assess individual genes with respect to their co-variation with age or HGPS using standard regression approaches, such as those used for eQTL studies, with hundreds of samples (Stranger et al., 2007). At the same time genes are known to form co-expression clusters reflecting common or interdependent regulatory mechanisms, and the traditional gene-centric regression approach does not leverage this fact. To address the limitations in the sample size, we have developed an iterative procedure that leverages co-expressed gene clusters while iteratively refining the cluster based on a cluster-centric multivariate regression's goodness of fit criteria (Wang et al., 2014b). We have performed a number of tests to show the robustness and efficacy of the approach.

## 2.2 Results

### 2.2.1 Method Performance and Efficacy

We first cluster the 9,453 genes into 200 clusters (see METHODS) and applied the regression model within each cluster independently. Fig. 2-1 shows (”Initial FG” plot) the goodness of fit as represented by Adj-$R^2$. We then iteratively refine the clustering with the explicit goal to improve the cluster ”tightness” which interestingly has an indirect effect on the Adj-$R^2$. As shown in ”Final FG” plot in Fig. 2-1, the Adj-$R^2$distribution shifts to higher values. As a control when we randomly permuted the initial expression data and repeat the entire procedure, the final refined clusters show a much inferior distribution as shown in ”Final BG” plot in Fig. 2-1. This supports the efficacy of the refinement step and indicates a substantial pattern in the expression data. The refinement step took 143 iterations until convergence.

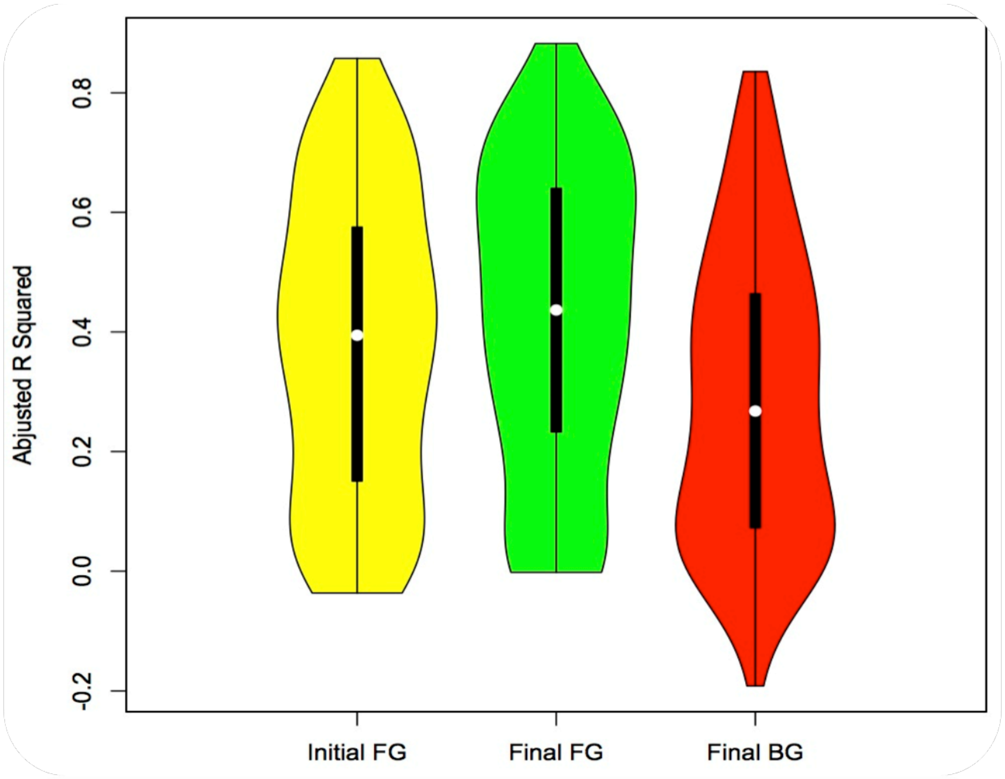


**Figure 2-1. Goodness of fit (Adj-$R^2$) distributions for gene clusters.** Yellow plot: initial clustering, Green plot: Final refined clustering, Red plot: Re-fined clustering for randomly permuted gene expression data.

### 2.2.2 Convergence

The maximal matching clustering refinement (see METHODS) monotonically decreases the value of $\tilde{\mathrm{F}}^2$. Intuitively, the algorithm will converge because if $\tilde{\mathrm{F}}^2$ is minimized the cluster will contain co-expressing genes which will have similar regression coefficient vectors. Fig. 2-2 shows changes in $\tilde{\mathrm{F}}^2$ and concomitant changes in total squared residuals due to regression through successive iterations.

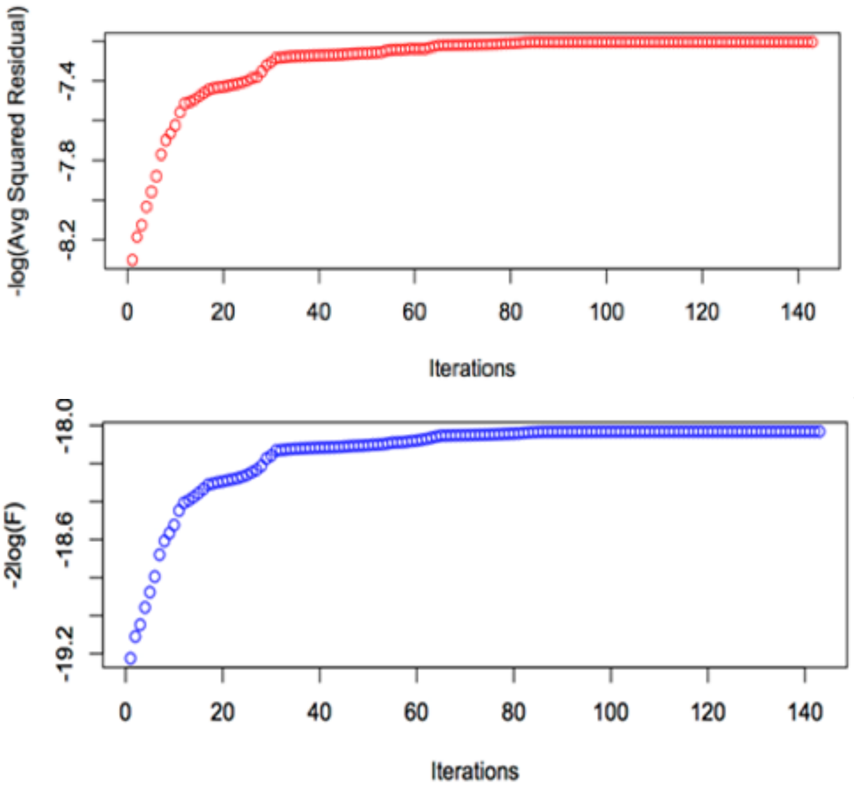


**Figure 2-2: Convergence of $\tilde{F}^2$ and total average residual of linear regression.**

### 2.2.3 Method Robustness

Next we assessed the extent to which the quality of final clustering depends on the initial clustering step. To do so, starting with initial clustering, we perturb the clustering to various extent (defined by parameter $\alpha$) and quantify the quality of final clustering. For instance, we randomly select a fraction of genes and randomly assign to existing clusters. We first noticed that as we increase $\alpha$, it takes longer for the clustering to converge—roughly an eight-fold increase in real run time when using a random clustering compared with co-expression cluster as described in

Methods. We compared the Adj-$R^2$ distributions for increasing values of α. As shown in Fig. 2-3, the overall Adj-$R^2$ distribution shows modest reduction in quality (compare plots for α > 0 with α = 0), which nevertheless is better than initial clustering (compare plots for α > 0 with initial clustering). A direct comparison of Adj-$R^2$ between each of the perturbed data and the unperturbed data using Wilcoxon test shows no significant difference, thus supporting robustness of the iterative procedure.

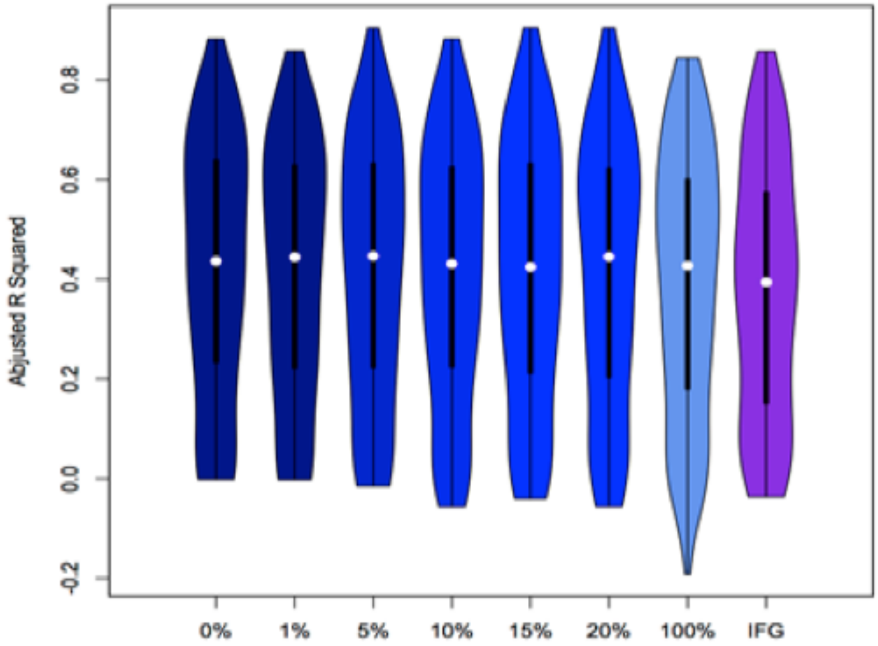


**Figure 2-3. Robustness of the iterative refinement.** For varying degree (α) of perturbation of the initial clustering the plots show the Adj-$R^2$ distribution of the refined clustering. IFG represents initial clustering.

### 2.2.4 Performance

To illustrate the advantage of the proposed RegressionClust model in small sample size, we compared its p-values of regression coefficients with those of single gene regression model. Figs. 2-4 and 2-5 show the p-value distributions of

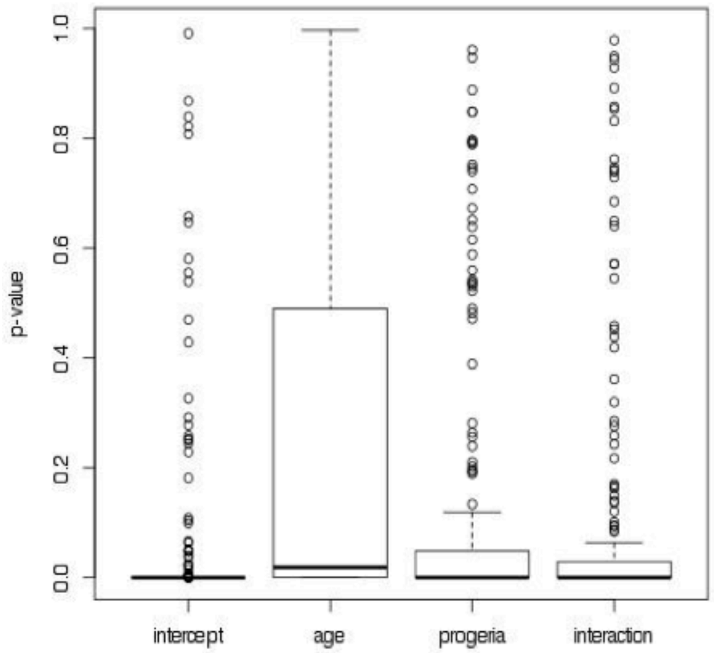


**Figure 2-4. Distribution of p-values of the four parameters in the single gene fitting model.**

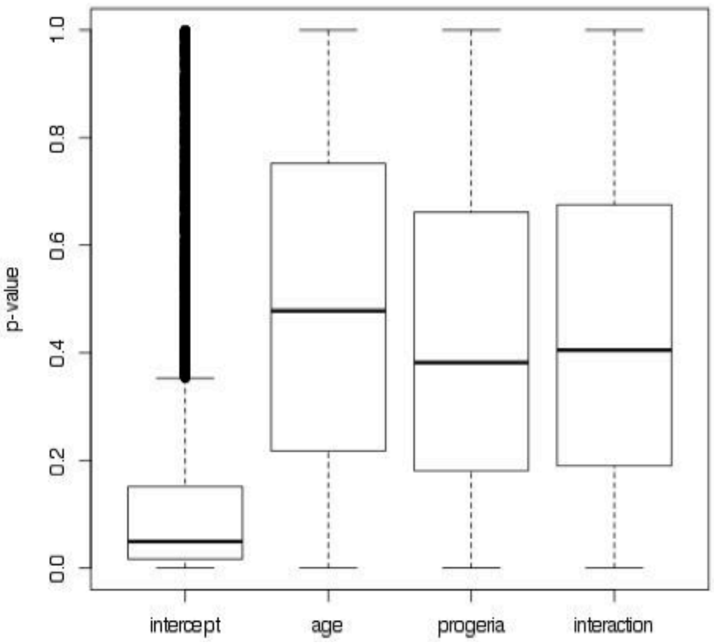


**Figure 2-5. Distribution of p-values of the four parameters in the gene cluster fitting model.**

regression coefficients of both models for six samples of age-progeria data. We observed coefficients estimated from single gene model are not significant. On the other hand most of coefficients generated from the RegressionClust model are significant.

### 2.2.5 Identification of Gene Clusters Whose Expression Co-Vary with Age and/or HGPS

We applied our approach to our in house RNA-seq gene expression data for six fibroblast samples—three normal at different cellular ages and three from HGPS at different ages. Using 200 initial co-expression clusters, we iteratively refined the clusters based on tightness of the cluster criterion (see Methods) while estimating cluster specific regression coefficients $\beta^1$, $\beta^2$, and $\beta^3$ for each final cluster along with the p-value for the null hypothesis that the coefficient is zero. We corrected all p values thus obtained using Benjamini-Hochberg procedure. Next we examined the normalized coefficient values (effect size).

In clusters where only the age and interaction coefficients were significant, the age alone tended to have positive effect on gene expression relative to interaction (Fig. 2-6). The gene expressions increased with age while the interaction terms in general had negative effect on gene expressions (Fig. 2-6). Likewise in clusters where only progeria and interaction coefficients were significant, the interaction terms had negative effect compared to progeria (Fig. 2-7).

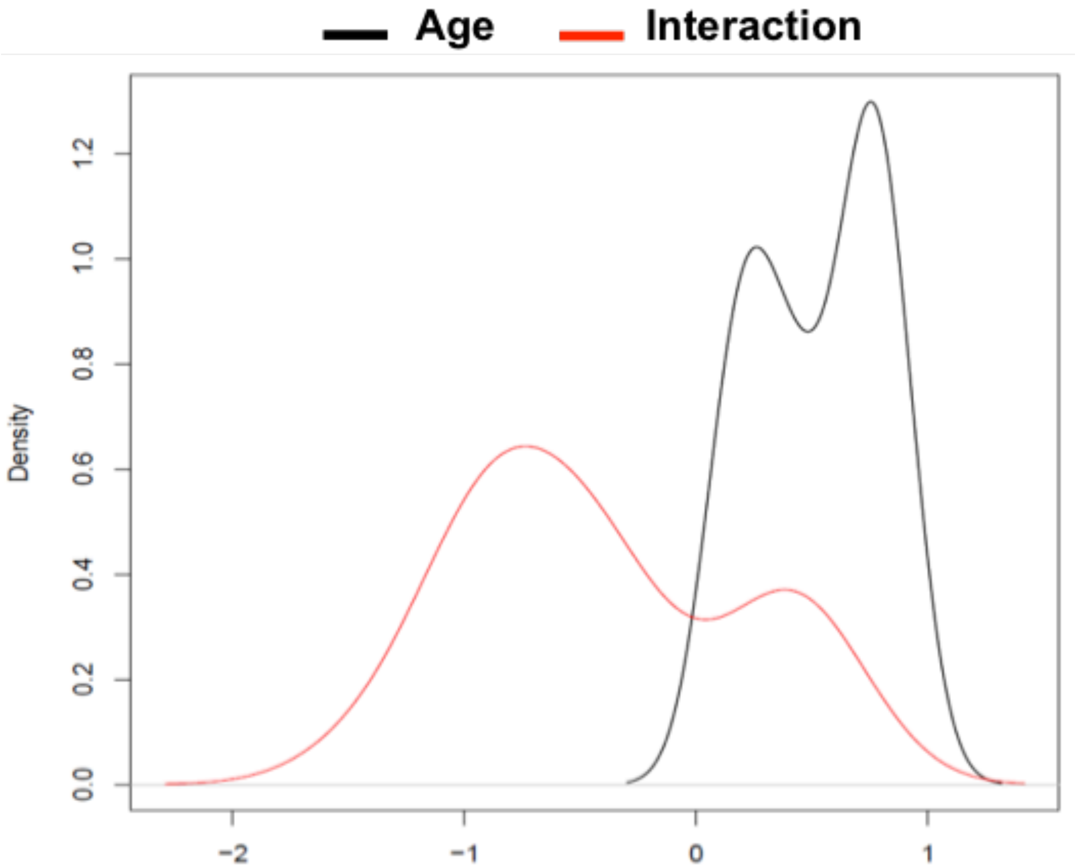


**Figure 2-6. Distribution of normalized coefficients $\beta_1$, and $\beta_3$ specifically for the clusters for which only these two coefficients were significant.**

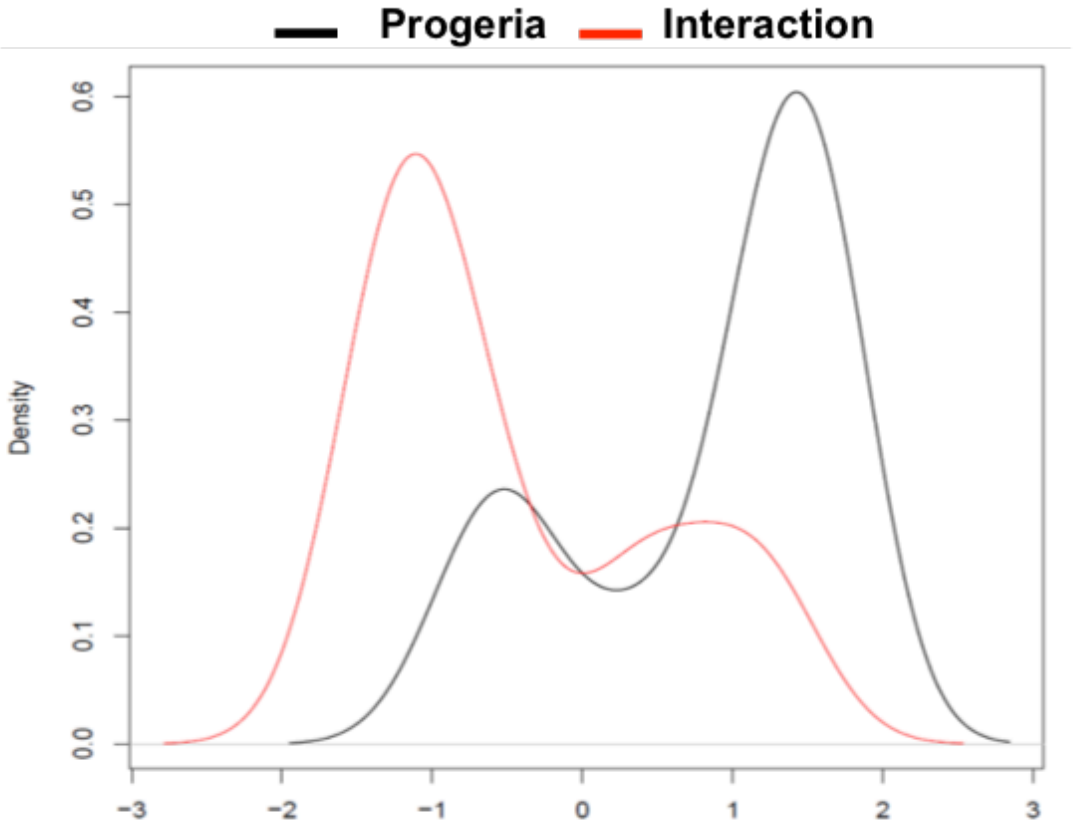


**Figure 2-7. Distribution of normalized coefficients β₂, and β₃ specifically for the clusters for which only three two coefficients were significant.**

### 2.2.6 Functional Analysis of Specific Gene Clusters Whose Expression Co-Vary with Age and/or HGPS

We selected the significant coefficients whose absolute value was at least 0.5, to exclude extremely small effect sizes. Considering three coefficients $\beta^1$, $\beta^2$ , and $\beta^3$, and their signs, there are multiple possibilities for various combinations of these coefficients being significant (FDR <= 0.05 and effect size >= 0.5). For instance, 1+2- represents the clusters for which both $\beta^1$and $\beta^2$ were significant

(FDR<= 0.05 and effect size >= 0.5) and $\beta^3$ was not. Table 2-1 shows the number of clusters with different combinations of significant coefficients with relative large effect sizes.

**TABLE 2-1**
**Number of clusters in various categories based on which combination of coefficients were significant and with large effect size (absolute value ≥ 0.5).**

| Category | Number of Clusters |
|---|---|
| 1+ | 4 |
| 1- | 3 |
| 2+ | 9 |
| 2- | 15 |
| 3+ | 24 |
| 1-2- | 1 |
| 1+3- | 11 |
| 2+3- | 42 |
| 2-3+ | 9 |
| 1+2+3- | 6 |
| 1-2-3+ | 6 |

We performed functional enrichment analysis (Huang et al., 2008; Huang et al., 2009) using GO Biological processes and KEGG pathways based on FDR threshold of 0.05. For ease of interpretation, we consider only the clusters in six categories: 1+: the expression increases with age (and no other significant large effect), 1-: the expression decreases with age (and no other significant large effect), 2+: the expression increases with HGPS (and no other significant large effect), 2-: the expression decreases with HGPS (and no other significant large effect), 1-2-: the expression decreases both with age and HGPS (no significant large interaction), and finally 3+: the expression significantly and largely increases with age only in HGPS patients (there were no significant clusters in 3- category). To underscore the relative advantage of our specific approach that can potentially distinguish between mechanisms mediated by increase or decrease in gene expression, and also age related and HGPS-related mechanisms, we directly compare the functions enriched in specific categories. Functions enriched in 1+ (73 terms enriched) and 1- (39 terms enriched) category are significantly distinct. Terms unique to 1+ included "extracellular matrix organization", "regulation of cell proliferation", "response to oxidative stress", "regulation of cellular metabolic process" and "immune response" etc. Interestingly, all terms unique to 1- referred to "cell cycle", "regulation of growth", "cytoskeleton/spindle organization", and "chromosome segregation", which according to our analysis are suppressed with age. The number of enriched terms in 2+ and 2- clusters were 185 and 251 respectively. To reduce this to a manageable list we only considered GO terms with at most 20 genes annotated, bringing the numbers down to 63 and 152. The common enriched terms referred to "protein localization/transportation", "signal transduction" and "metabolic process". The terms unique to 2+ referred to "negative regulation of molecular function", "negative regulation of signaling", "negative regulation of response to stimulus", and "metabolic process". The terms unique to 2- were clearly different and referred to "biosynthetic process", "DNA repair/double-strand break repair", "transcription","RNA splicing", "mRNA processing" and "chromatin/histone modification". Interesting we did not detect any gene cluster in "1+2+" category, but we did detect clusters in "1-2-" category and the main theme among the enriched terms in these clusters were “metabolic process”, “apoptosis process”, “signal transduction”. Finally, category 3+ which refers to clusters whose expression increases with age, particularly in HGPS population has several enriched terms in common with 1+ category, which include “signal transduction”, “regulation of cell proliferation”, “metabolic process”. The 3+ category includes numerous enriched terms not detected for 1+ category. These include “cell migration”, “development growth”, “muscle contraction”, “cell adhesion”, “heart growth” and “protein folding”.

| Functional Terms | FDR |
|---|---|
| Transcription | 1.5e-6 |
| Regulation of transcription | 4.9e-6 |
| DNA binding | 3.7e-4 |
| Transcription co-activator activity | 1.1e-4 |
| Transcription factor binding | 2.5e-2 |
| Transcription co-factor activity | 6e-3 |

**Table 2-2. Cluster of enriched terms relating to transcriptions are enriched among gene clusters in 2- category.**

### 2.2.7 Functional Comparison of the Initial Clusters and Refined Clusters

To assess whether our joint regression clustering approach improves functional enrichment of the clusters, we compared the enrichment of functional GO terms in the initial and final clusters for a few selected genes previously known to be involved in aging or Progeria. For some genes, we found their initial cluster and final assigned cluster have different biological functions. As an illustrative example when we compared the clusters containing FOXM1 gene, a key transcriptional regulator of cell cycle progression, the final refines cluster containing FOXM1 was specifically enriched for Cell Cycle while the initial cluster containing FOXM1 was enriched in general terms such as transcription activity and DNA binding but not for Cell Cycle. However we acknowledge that it is difficult to quantify the functional enrichment difference between the initial clusters and the final reassigned clusters because the clusters compared may have different size that will influence the enrichment score.

## 2.3 Materials and Methods

### 2.3.1 Cell Culture, RNA Preparation, and RNA-Seq Experiment

Normal (HGADFN168, Father) and HGPS (HGADFN167, son) primary fibroblasts were obtained from the Progeria research foundation. All fibroblast cell lines were cultured in MEM (Life Sciences) supplemented with 15 percent FBS (Gemini Bio-Products) and 2 mM L-glutamine (Life Sciences) at 37°C with 5 percent $CO^2$. RNA samples were collected from these two cell lines at early (passage 11), middle (passage 16) and late stages (passage 20 for HGPS, and passage 23 for normal) during replicative senescence. Total RNA from different cell lines was extracted with Trizol (Life Sciences) and purified using the RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. The RNA yield was determined using the NanoDrop 2000 spectrophotometer. The RNA-seq sample preparation and sequencing were conducted according to the illumina

Truseq RNA sample preparation V2 guide by the IBBR sequencing Core facility at the University of Maryland.

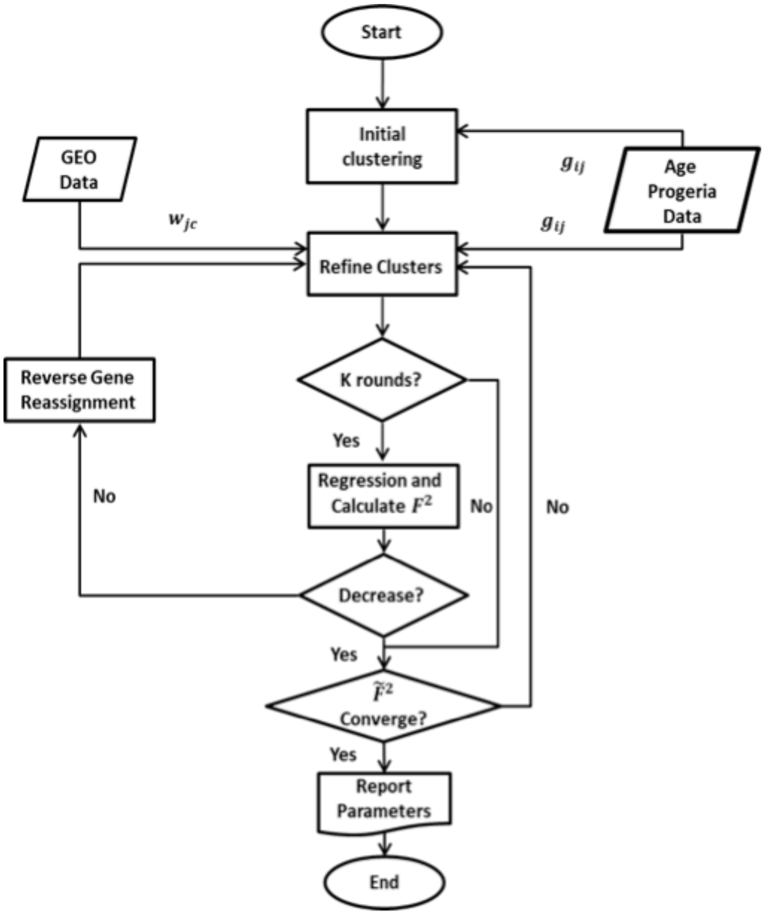


**Figure 2-8. Method workflow.** See text for details.

### 2.3.2 RNA-Seq Data processing

We processed each of the six samples identically using the Cufflinks suite of tools following the recommended protocol (Trapnell et al., 2012) yielding RNA expression value (FPKM) for ~14,000 human genes in each of the six samples. In addition, to guide the iterative cluster refinement procedure (see below), we obtained the RNA-seq profiles for 15 independent tissue types from Gene

Expression Omnibus (GEO). There were 9,453 genes in common between our samples and the GEO samples, which were ultimately used for all follow up analyses.

### 2.3.3 Joint Regression Clustering

*Workflow*. Fig. 2-8 shows the complete workflow of proposed joint regression clustering method. We start with initial clustering of genes (k-mean clustering) based on age-progeria data. We refine the clusters iteratively to minimize the average error of predicted (from cluster regression) gene expression till convergence. It is computationally expensive to calculate the average error for all possible refinements. Therefore, we approximated the average error by its maximum likelihood (ML) estimate. After every k rounds if actual average error increases, we randomly reverse some of the gene reassignments.

*Linear regression model*. Linear regression is widely used method to study the effect of covariates on expression variance between samples. The linear regression model for gene expression with aging and HGPS as covariates can be expressed as:

$$g_{ij} = \mu_j + \beta_j^1 a_i + \beta_j^2 d_i + \beta_j^3 a_i d_i \text{ (1)}$$

Where, $g_{ij}$ is expression of jth gene in ith sample. $\mu_j$ is the basal expression of jth gene. $\beta_j$ is vector of the regression coefficients of jth gene for covariates age $a_i$ (1: young, 2: middle age, 3: old) and HGPS state $d_i$ (0: normal, 1: HGPS), and interaction term $a_i d_i$. A vector of coefficients must be estimated for each gene separately for model (1). This is clearly limiting as we have only six samples. Moreover, there are thousands of genes that will be differentially expressed in different sample. To learn regression coefficient separately for each gene is not an effective approach because small sample size will have low statistical power (see results) and many genes are expected to vary in a similar manner with respect to the covariates. Additionally, it will be hard to extract meaningful result and visualize the effect of covariates from separate regression coefficients for thousands of genes. We can cluster genes based on its expression variance w.r.t its covariates and estimate coefficients jointly for a cluster of co-varying genes.

*RegressionClust model*. To overcome above limitations and to leverage clusters of potentially co-varying genes we propose the following model, RegressionClust:

$$g_{ijc} = \mu_{ic} + \beta_c^1 a_i + \beta_c^2 d_i + \beta_c^3 a_i d_i$$

$g_{ij} - w_{jc} = g_{ijc}$ (2)

Where, $g_{ijc}$ is imputed expression of jth gene belonging to cth gene-cluster in ith sample. $\mu_{ic}$ is the basal expression of genes in cth gene-cluster in ith sample. $\beta_c$ is cluster specific vector of the regression coefficients of the covariates. $w_{jc}$ is distance of jth gene from its cluster center and is computed as the difference between the mean expression value of the gene and that of all genes in the cluster that the gene is assigned to. $w_{jc}$ is updated after each reassignment.

This is a dual optimization problem—fit a regression model for each cluster and refine clusters to maximize overall explained variance. The objective functions are: model for each cluster and refine clusters to maximize overall explained variance. The objective functions are:

Regression: find optimal regression coefficients such that gene expression variance within cluster explained by covariates is maximized.

i.e. $argmin_{\beta_c} Q_c^2 = \sum_j \sum_i \left(g_i(\beta_c) - g_{ijc}\right)^2$.

Where, $g_i(\beta_c) = \mu_{ic} + \beta_c^1 a_i + \beta_c^2 d_i + \beta_c^3 a_i d_i$. This is equivalent to

$I = argmax_{\beta_c} R_c^2 = 1 - \frac{\sum_j \sum_i \left(g_i(\beta_c) - g_{ijc}\right)^2}{\sum_j \sum_i \left(\bar{g} - g_{ijc}\right)^2}$ (3).

$\bar{g}$ is the mean gene expression value in cluster c.

Cluster refinement: find optimal set of clusters (or, clustering) such that each cluster is tight (maximize overall explained variance), i.e. minimize

$F^2 = \sum_c F_c^2 = \sum_c \sum_{j\in c} \sum_i \left|g_{ij} - w_{jc} - g_i(\beta_c)\right|^2$ (4)

### 2.3.4 Greedy maximal matching cluster

**Inference**: Independent maximization of $R^2$ is equivalent to linear regression, while independent minimization of $F^2$ is clustering. We estimate the parameters of RegressionClust model by iteratively optimizing the two objective functions. It is important to note that $w_{jc}$ should be independent of the expression variance due to the covariates, therefore we estimate them from gene expression of

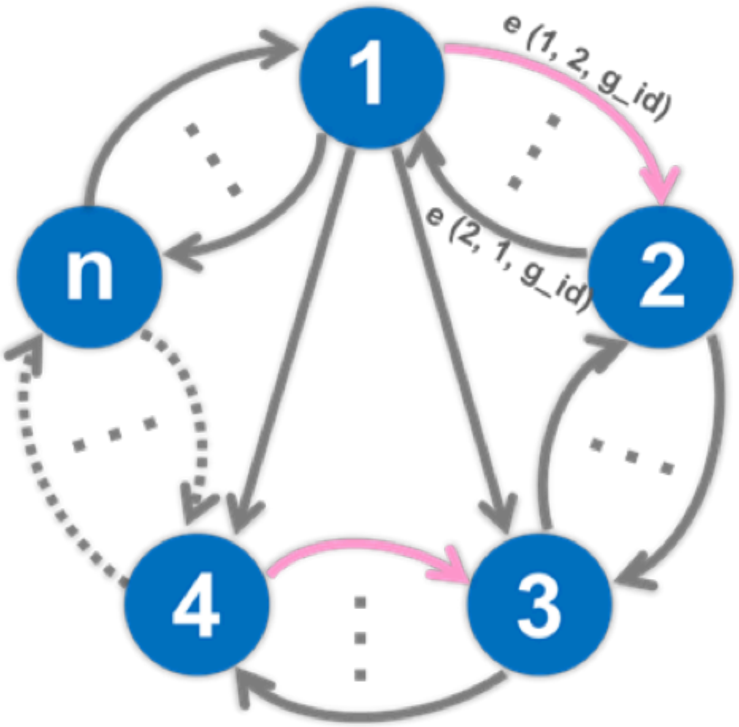


**Fig. 2-9 The Graph constructed for greedy maximum matching cluster refinement.** There are n clusters, we construct a node for each cluster, and the edge is added if the possible reassignment will decrease the objective function score. The Edge e(1,2,g_id) is the gene reassignment that moves gene gene_id from cluster 1 to cluster 2. The red edges are greedily selected edges, based on which we perform the final genes reassignment.

**Box. 2-1 This is the greedy maximum matching clusters refinement algorithm.**

```
Greedy maximal matching cluster refinement algorithm:

Greedy_Cluster_Refinement(c_vector,real_gene_exp,geo_gene_exp,num_c)
    G= Initialize_graph(c_vector,real_gene_exp, geo_gene_exp, num_c);
    while |unmarked_E| > 0
        max(w(e(i, j, gene_id) ∈ unmarked_E ));
        marked_E ← e(i, j, gene_id);
        marked_V ← i, j;
    for each e(m, n, gene_id') ∈ unmarked_E
        if m or n ∈ marked_V||gene_id=gene_id'
            delete e(m,n,gene_id');
   for each e(m, n, gene_id)
                    c_vector[gene_id] = n;
   return c_vector;

Initialize_Graph(c_vector, real_gene_exp, geo_gene_exp, num_c)
   Graph G = (V,E);
   for each cluster c(i)
       V ← n(i);
  for each gene gene_id
       for t = 1 to num_c - 1
           update w_jc;
           calculate obj_diff;
           if var_diff < 0
               E ← e(c_vector[gene_id], t, gene_id)
               w(e(c_vector[gene_id], t, gene_id) = obj_diff;
   return Graph G;
```

15 independent normal expression samples collected from GEO database. As a side note, this iterative inference is similar to Expectation Maximization (EM) algorithm. In particular, if instead of hard assignment of gene cluster, fuzzy assignment to cluster is used, it can be proved that it is equivalent to EM algorithm. However, we chose to use hard assignment because we found that fuzzy clustering increases computational cost without significant gain in the overall performance. Maximization of $R^2$ is explained next.

**Initializing the cluster set**: We initialized the clusters using k-means clustering on the sample data, for different number of clusters (200,300,400,500,600).

**Greedy maximal matching cluster refinement**: To greedily refine the clusters, a change in $F^2$ should be calculated for each possible reassignment (move of gene from each cluster to another). Each possible reassignment changes $\beta_c$ and $g_i(\beta_c)$. Running linear regression for each possible reassignment is clearly computationally limited. We therefore use maximum likelihood estimate of $g_i(\beta_c)$ to estimate change of $F^2$. The Maximum Likelihood (ML) estimate of $g_i(\beta_c)$ can be determined by differentiating equation (3) w.r.t to $g_i(\beta_c)$:

$$\frac{dI}{dg_i(\beta_c)} = -\frac{2\left(\sum_j \sum_i \left(g_i(\beta_c) - g_{ijc}\right)\right)}{\sum_j \sum_i \left(\bar{g} - g_{ijc}\right)^2} = 0$$

$$g_{iML}(\beta_c) = \frac{1}{J_c} \sum_j g_{ijc} \qquad (5)$$

Where $J_c$ is total number of genes in cluster c. Now, we can replace this in equation (4) to obtain its ML estimate:

$$\tilde{F}^2 = \sum_c \sum_{j \in c} \sum_i \left| g_{ij} - w_{jc} - g_{iML}(\beta_c) \right|^2 \qquad (6)$$

$g_{iML}(\beta_c)$ can be locally updated efficiently for each possible single gene moves.

To allow at most (a single gene) change to a cluster in an iteration, we construct a graph with all clusters of current clustering as nodes and each single gene move as a directed edge between originating and destination cluster and change in $\tilde{F}^2$ due to the move as the edge weight. We then performed maximal matching on this graph to minimize $F^2$and allowed single change to a cluster. We then proceed with single gene moves corresponding to maximally matched edge as our cluster refinement. The maximal matching also ensures that same cluster will not be source of one gene and destination for some other gene as shown in Fig. 2-9. Box 2-1 shows pseudo code for this greedy cluster refinement. c_vector, real_gene_exp, geo_gene_exp, num_c respectively are cluster membership vector, age-progeria data and GEO data. Although we use $\tilde{F}^2$ for cluster refinement, but after every k iterations if $F^2$ increases for the selected gene moves, we reverse those moves. We chose k=10 for all analysis.

Note that $w_{jc}$ is re-calculated after every cluster update and multiple changes to a cluster can in fact result in overall increase in $F_c^2$. Our matching strategy involving at most one change to each cluster ensures overall reduction of square errors. In addition, by virtue to selecting a maximal matching we maximize the improvement in square errors. The steps of calculating $w_{jc}$, $F_c^2$ and maximal matching cluster refinement are repeated until convergence.

**Adjusted R$^2$:** The quality of regression fit is generally estimated using the R$^2$ statistic as defined above. However, to account for the varying number of clusters and the number of parameters, we instead use adjusted R$^2$ (Adj-R$^2$) for a cluster computed as:

$$R_{adj}^2 = 1 - \left(\frac{(1 - R_c^2)(n - 1)}{n - k - 1}\right)$$

Where *n* is the cluster size, and *k* is the number of coefficients in the multiple linear regressions. $R_c^2$ is defined in equation (3).

### 2.3.5 GO Analysis

We assessed enrichment of GO biological processes and KEGG pathways in co-expressed gene clusters whose expression co-varied with age and/or HGPS using R’s GOstats package. The significance was corrected for multiple testing using the Benjamini-Hochberg procedure. An FDR threshold of 0.05 was used.

### 2.4 Summary and Discussion

Our works make three main contributions which pertain to data, method, and application to make new biological discoveries. With regards to data we have generated the first RNA-seq data in a controlled fashion for aging HGPS primary cells and passage and genetic background matched normal control cells. Methodologically, here we have presented a regression based approach that leverages clusters of genes with co-varying expression to robustly estimate regression coefficients representing dependence on age, HGPS and the interaction between the two. Our approach iteratively refines the clusters using a cluster ”tightness” criterion which, as we show analytically, simultaneously improves the goodness of fit while increasing the computational efficiency substantially. The proposed method should be useful in several other contexts with limited number of samples. Finally, application of our method to the data recapitulates previous discovery of age-dependent gene expression changes as well as makes several important observations in a comparison between age and HGPS. Previous microarray studies of HGPS and aged normal fibroblasts have revealed some insights into the gene expression changes during the normal and the premature aging. Ly et al. used fibroblast cells from young, middle and aged normal donors as well as from a HGPS patient, and identified 61 differentially expressed genes out of the 6,000 genes monitored, among which there are two major functional groups: (1) genes involved in cell cycle progression and (2) genes involved in maintenance and remodeling of the extracellular matrix (ECM) (Ly et al., 2000). Interestingly, most of the cell cycle genes showed down-regulation in aged cells and HGPS cells, and the ECM genes are affected in both directions in aged and HGPS cells. Using genome-wide affymatrix microarrays, Csoka et al. defined 361 differentially expressed genes in HGPS fibroblast cells (out of 33,000 genes on the array), and found that the two most prominent categories encoded transcriptional factors and ECM proteins (Csoka et al., 2004). Because of our specific methodology we were able to identify gene clusters whose expressions co-vary exclusively with age, or disease, or in specific combinations of age and disease. We identified several predominant gene clusters, whose expressions were altered either under the disease condition HGPS and/or during the normal cellular senescence. Of particular note, our analysis indicated that the HGPS gene expression profiles show important differences from the profiles of normal fibroblast passaged into cellular senescence. In the ”1-” clusters, we found that the majority of genes are related to cell cycle regulation, which is in highly significant agreement with the results from Ly et al. despite major differences in samples, methods, and data analysis. For example, Forkhead box protein M1 (FOXM1), a key transcriptional regulator of cell cycle progression, was found to be down regulated in both studies. This gene has been shown to regulate a large group of G2-M specific genes (Myatt and Lam, 2007), including a key mitotic cyclin, cyclin B1, which was also identified by both our analysis and Ly et al. In addition, consistent with previous reports on ECM proteins, we found that the ”1+” clusters are enriched with functional categories of ECM organization. However, the responses in HGPS cells differ: In the ”2+”

clusters that positively associate with HGPS disease condition, the prominent categories are related to metabolic functions, implying an activation (or at least an attempted activation) of the biological processes involved on various cellular metabolic activities in HGPS cells. To date, the metabolism in HGPS cells have not been systematically examined, nor any metabolite profiling in HGPS cells have been reported. Our study provides the first genome-wide evidence of the affected metabolisms in HGPS cells, and points to a potentially important direction for future HGPS research. Interestingly in the ”2-” clusters whose expression is reduced in HGPS, we found gene clusters including transcription, mRNA processing, splicing and protein biosynthesis, reflecting an overall slow down in cellular growth in the prematurely aged HGPS cells. Table 2-2 shows the cluster of terms related to transcription significantly enriched among these gene clusters computed by NIH DAVID tool (Huang et al., 2008; Huang et al., 2009). Because lamin A/progerin resides in the inner nuclear rim, and plays a role in organizing chromatin, it is not surprising to identify wide spreading changes in gene expression in HGPS cells. The challenge is to determine the specificity of progerin-related changes and of the age-related changes, and illuminate their potential interplays. In an attempt, we examined the functional groups in the genes whose expression increases with age specifically in HGPS cells (the ”3+” clusters). Interestingly, a prominent functional gene group is related to signal transduction, including trans-membrane receptors (e.g. insulin-like family peptide receptor 1 and stannin) and protein kinases (e.g. membrane associated guanylate kinase and protein kinase B). Additional studies, especially those conducted in cell types other than fibroblasts, are required before we can understand the contributions of progerin/lamin A and cellular aging to gene expression in complex organisms. The study reported here provides a first genome-wide, multi stage RNA-seq experiment with a novel iterative multiple regression approach to examine this important mechanism.

# Chapter 3: Epigenomic and Genomic determinants of alternative splicing

### 3.1 Introduction

Recent availability of RNA-seq data has spurred several computational investigations into the determinants of alternative splicing (Barash et al., 2010; Flores et al., 2012; Shindo et al., 2013; Xiong et al., 2011; Zhou et al., 2012). While several different types of alternative splicing events have been documented, due to ease and robustness of inference, most investigations have focused on alternative exon inclusion/exclusion events, specifically cassette exons, where an alternative internal exon is flanked by ubiquitously included exons. A previous work has suggested existence of a 'splicing code' composed of numerous genomics features such as splice sites signals, conservation score, ESE, ESS, ISE, ISS, etc, that can accurately predict relative exon inclusion in a cell type (Barash et al., 2010; Xiong et al., 2011). On the other hand, Shindo et al have shown correlation between several chromatin marks and splicing (Shindo et al., 2013); they found H3K36me3 and H3K79me1 around the exon-intron boundaries to be strongly correlated with splicing. Zhou et al have shown a correlation between H3K36me3 and splicing (Zhou et al., 2012). Another report suggested that DNA methylation within exon body may have positive effect on exon inclusion (Flores et al., 2012). Finally, a linear regression model to predict exon inclusion based on multiple chromatin features showed several chromatin features, especially H3K36me3 and H4K20me1 to be correlated with exon expression (Zhu et al., 2013). However, this previous approach does not separate exon expression and gene expression, and an alternative measure – percent splicing index (PSI), better quantifies alternative. What's more, chromatin's ability to affect splicing can be overestimated without using genomics features as control since genomics features can, presumably causally, predict chromatin features (Whitaker et al., 2014). Overall, the relative contributions of genomic and epigenomic contributions to alternative splicing, specifically, alternative exon inclusion event, is not known. Here, based on deep neural network model, we carefully analyzed the relative contribution of genomic and epigenomic features on exon inclusion levels, in multiple cell lines.

Our analyses showed that genomics features could predict exon inclusion much more accurately than chromatin features. Integrating the two types of features does not improve the prediction accuracy. We specifically assessed the contributions made by either genomic or epigenomic feature in addition to the other type of feature using multiple approaches, and found that, while genomic feature make a significant additional contributions to predictability of exon inclusion, the converse is not true, suggesting that genomic features encode most information relevant to exon inclusion. Besides the assessment of predictability, we specifically model the position-specific contribution of each feature. Finally, even though epigenomic features do not make substantial contribution independent of the genomic features, our model detected specific interactions between genomic and epigenomic features, suggesting that the effect of specific genomic features may be sensitive to the epigenomic context.

Overall, we provide a first direct comparative assessment of genomic and epigenomic features, and interaction thereof, in predicting cell type specific alternative splicing (Wang et al., 2015).

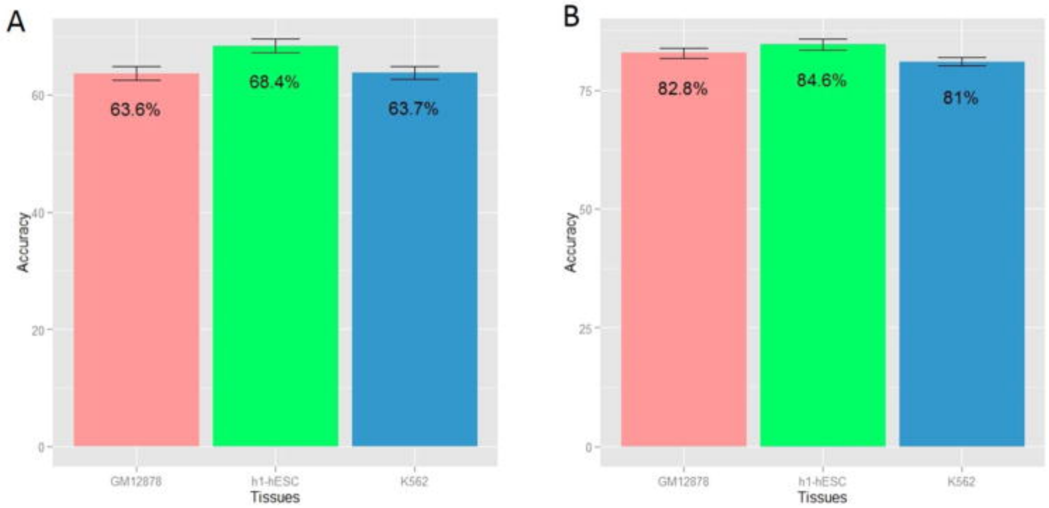


**Figure 3-1: Cross-validation prediction accuracy of exon inclusion using chromatin features for three cell types GM12878 (blood), h1-hESC (human embryonic stem cell) and K562 (leukemia). The accuracy is the mean accuracy of 8-fold cross validation.** (A) Prediction accuracy using chromatin features; (B) Prediction accuracy using genomic features.

## 3.2 Results

### 3.2.1 Chromatin Features are Weak Predictors of Exon Inclusion

Previous studies have reported correlations between various types of splicing events and proximal chromatin features (Whitaker et al., 2014; Zhu et al., 2013). We directly assessed the cross-validation predictability of exon inclusion using chromatin features alone. Fig. 3-1A shows the 8-fold cross-validation classification accuracy in three different cell lines. The prediction accuracies in all cell types are significantly higher than the random expectation of 50%, albeit, modest. Notably, prediction accuracy is much higher in h1-ESC relative to the other two cell lines. This may be either because the chromatin state is indeed more closely associated with splicing in pluripotent cells or alternatively, because of better quality of chromatin modification data in h1-ESC cell line; these need to be explored in future.

### 3.2.2 Genomic Features are Robust Predictors of Exon Inclusion

Previous studies have shown that genomic features can accurately predict change in exon inclusion propensity in a cell line relative to other cell lines (Barash et al., 2010; Xiong et al., 2011). We emphasize that our goal here is not necessarily to improve exon inclusion predictability, but rather to contrast the independent and synergistic contributions of chromatin and genomic features and also to assess location specificity of various features relative to splice sites. Nevertheless, we first establish a baseline for predictability of exon inclusion using genomic features in our datasets and using tissue-specific performance metric. Also, in contrast to previous genomic element-based relative exon inclusion prediction approach, here we only employ genomic features with a potential mechanistic link to splicing machinery and excluded features such as 'exon translatability' that was shown to be the single-most powerful predictor but is not linked to the splicing mechanisms per se. We used only the cis-elements discussed above to predict splicing. However, we note that by excluding translatability as a feature, our approach does not account for nonsense mediated decay of the mRNA caused by pre-mature stop codon (Cusack et al., 2011).

Similar to chromatin features, we employed 8-fold cross validation to estimate prediction accuracy. As shown in Fig. 3-1B, genomic features can predict exon inclusion very accurately, consistent with previous studies (Barash et al., 2010; Xiong et al., 2011), and importantly, much more accurately than chromatin features. This suggests that exon inclusion, even in a specific context, is largely determined by genomic sequences.

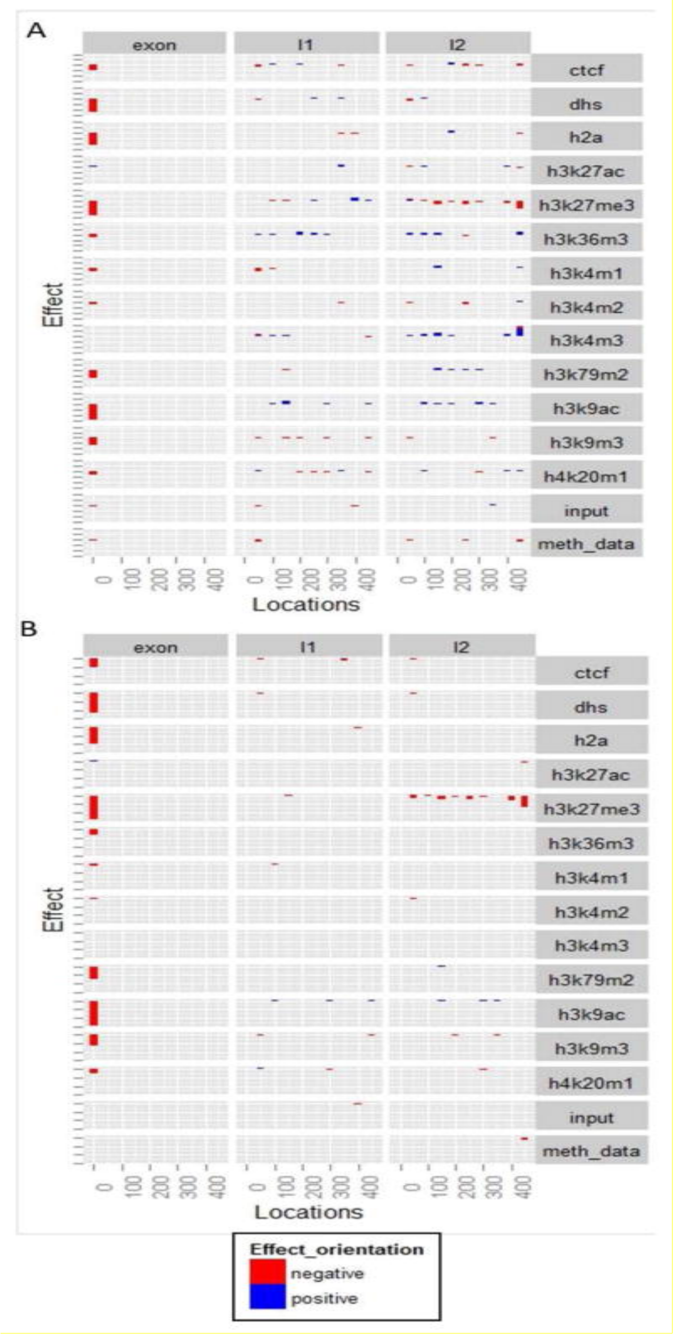


**Figure 3-2: (A) The effect size of chromatin features at different genome locations in h1-hESC cell line; (B) The relevance of chromatin features at different genome locations in h1-hESC cell line.**

### 3.2.3 Location-Specific Map of Chromatin Features

In our deep learning model, we assessed the effect of each chromatin mark in 3 regions (multiple sub-regions in each broad region) relative to the cassette exon; each mark-locus combination is a distinct feature in our model. Here we report the locus-specific effect size of various chromatin marks. Fig. 3-2A, supplementary Fig. 3-S1, and 3-S2, show, respectively for h1-hESC, GM12878, and K562, the most significant chromatin features (Methods) in all locations considered – the cassette exon, the 5' flanking intron and in the 3' flanking intron. Interestingly, by and large, almost all features in exonic regions have negative effect on exon skipping, i.e., their presence in specific exonic regions is associated with higher inclusion levels, discussed later. Also the trends are largely consistent across cell types, particularly across h1-hESC and K562.

We further ascertained the importance of features selected above as follows. We partitioned the entire set of exons into two sets based on the feature values (top and bottom half). We then randomly sampled (100 times) 1000 exons from each of the two groups and compared their inclusion levels using Wilcoxon test. We noted the fraction of tests (out of 100 tests) that yielded significant results consistent with the directionality of the feature's effect according to the model above. To rank the features in terms of their overall relevance, which captures both significance and effect size, we multiplied each feature's effect size (obtained from the model) with the fraction of Wilcoxon tests that were significant (significance). This procedure yields a view of every single feature's independent contribution (without considering interactions). Fig. 3-2B, supplementary Fig. 3-S3 and 3-S4 shows the relevance for all the three tissues. We ranked the features based on their relevance as estimated above. Our results suggest that H3K36me3 is one of the most relevant features consistent with previous reports (Shindo et al., 2013; Zhu et al., 2013). The analysis also reveals H3K79me2, H4K20me1, H3K27me3, H3K9ac to be highly relevant to exon inclusion. Interestingly, we found that leukemia and stem cell lines have more and stronger feature signals for enhancing inclusion, however, blood cell lines have more features associated with repression of exon inclusion.

### 3.2.4 Epigenomic features are not contributing extra regulatory information in addition to genomics features

We have shown in section 2.3 that chromatin features are modestly predictive of exon inclusion. Even though specific mechanisms linking histone modifications to splicing have been reported (Luco et al., 2010; Luco et al., 2011), it is not clear to what extent the predictive power of chromatin features are independent of genomic features. To specifically investigate this, we assessed the extent to which chromatin features can explain the variance in exon inclusion that is unexplained by genomic features. We used two approaches to assess this: (i) we trained a model using chromatin (genomics respectively) features and then assessed the prediction accuracy using genomic (chromatin respectively) features with an additional feature representing the prediction score using the chromatin-based (genomic-based respectively) model; an improvement in prediction accuracy associated with the added feature represents additional contribution of that feature.

(ii) we quantified the extent to which chromatin (genomics respectively) features could explain the residuals of a linear model based on genomic-based (chromatin-based respectively) model. A high explanatory power of the residual is consistent with an independent contribution.

Fig. 3-3A and 3-3B show the results for the first analysis, which suggest that while adding chromatin-based model score to genomic features does not improve prediction accuracy, adding genomic-based model score to chromatin features substantially improves the prediction accuracy. Analogously, Fig. 3-4A and 3-4B show the result of the residual analysis, consistent with the first analysis, namely, chromatin features explain very small fraction of variance of the residual from the genomics-based model, as opposed to the converse. Overall, these analyses strongly suggest that that genomics features provide robust prediction of exon inclusion, largely independent of chromatin features and that the previous observed associations between chromatin features and splicing can largely be explained by the links between genomics and chromatin features, also noted previously (Whitaker et al., 2014).

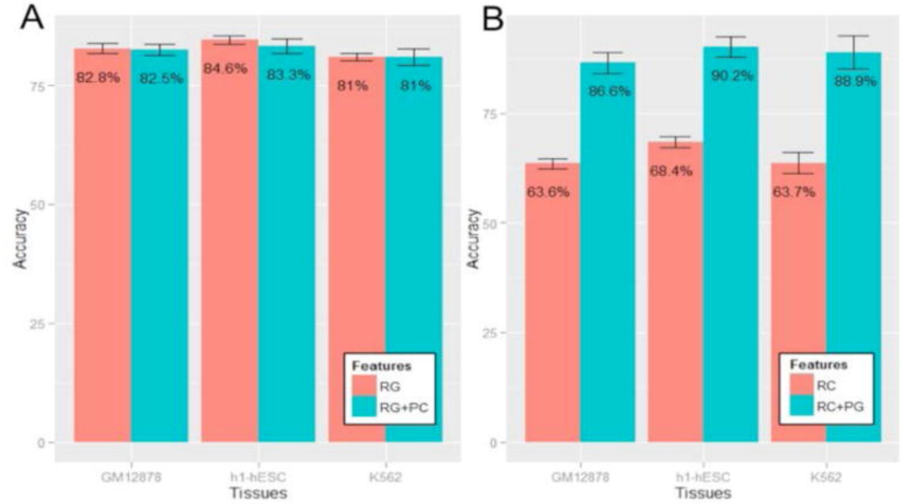


**Figure 3-3: Cross-validation prediction accuracy using raw genomics (chromatin respectively) features and chromatin (genomics respectively) feature prediction score as an additional feature, for three cell types, GM12878 (blood), h1-hESC (human embryonic stem cell) and K562 (leukemia).** The accuracy is the mean accuracy of 8-fold cross validation. RG indicates the raw genomics features, PC indicates prediction score using chromatin features. RC indicates raw chromatin features, PG means prediction score using genomics features. (A) Comparison between accuracy using RG + PC and only RG; (B) Comparison between accuracy using RC + PG and only RC.

### 3.2.5 Cross Tissue Generalization of Chromatin and Genomics Predictors

Next, to assess the extent to which similar rules govern exon inclusion in different cell lines, for each pair of tissues we trained the model on one tissue and tested on the other. First, for chromatin-based modeling, as shown in Table 3-1, GM12878 model cannot predict exon inclusion in the other cell types and conversely, model trained on other cell types cannot predict exon inclusion in GM12878. However, cross-tissue predictability is much higher than random expectation between h1-hESC and K562. As shown in Table 3-2, genomics-based model exhibits a similar

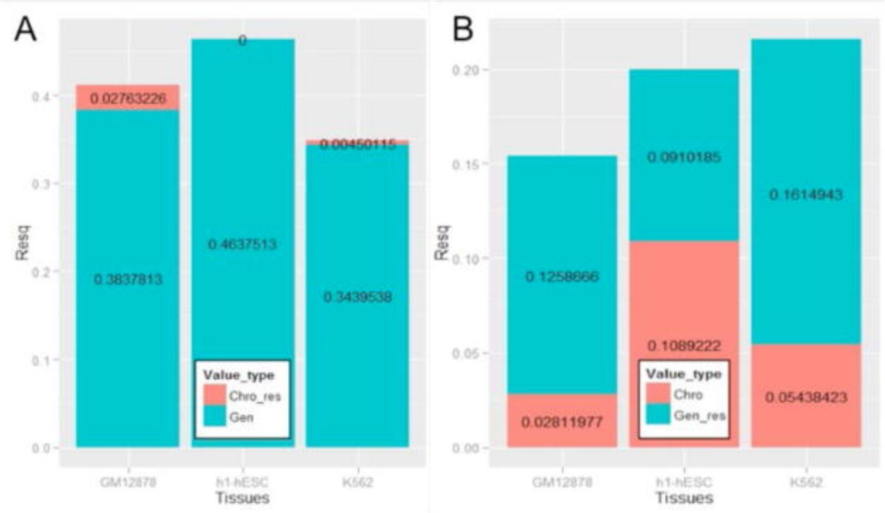


**Figure 3-4: R-squared for explaining residuals of genomics feature prediction using chromatin features and residuals of chromatin feature prediction using genomics features, in three cell lines, GM12878 (blood), h1-hESC (human embryonic stem cell) and K562 (leukemia).**
Chro-res: chromatin feature explain residuals of genomics model. Gen: genomics model. Gen_res: genomics feature explain residuals of chromatin model. Chro: chromatin model. (A) R-squared of Chro-res and Gen; (B) R-squared of Gen-res and Chro.

trend, however the absolute prediction is much greater for the genomics-based models. These results suggest that, even though a large portion of cis elements contribute to exon inclusion across cell types, exon inclusion also depends on specific cis elements that are recognized by cell type specific splicing factors (including splice enhancers and repressors). And the cross-cell type predictability for chromatin feature follow similar trend likely because chromatin features are largely encoded in cis elements (Whitaker et al., 2014).

|  |  |  |
|---|---|---|
| GM12878 | 49.2% | 51.6% |
| 46.8% | h1-hESC | 60.2% |
| 50.1% | 64.4% | K562 |

**Table 3-1: Cross tissue test using chromatin model.** In each row, we used one tissue model to predict exon inclusion of the rest. Red accuracy means not significant, green ones means significant.

|  |  |  |
|---|---|---|
| GM12878 | 80.56% | 78.9% |
| 75.82% | h1-hESC | 80.27% |
| 75.14% | 80.79% | K562 |

**Table 3-2: Cross tissue test using genomics model.** In each row, we used one tissue model to predict exon inclusion of the rest. Red accuracy means not significant, green ones means significant.

### 3.2.6 Potential interactions between Chromatin Features and Genomic Features

Our results thus far suggest that previously observed links between chromatin features and splicing may be largely explained by their correlations with genomic features, which are more directly and likely mechanistically linked with splicing.

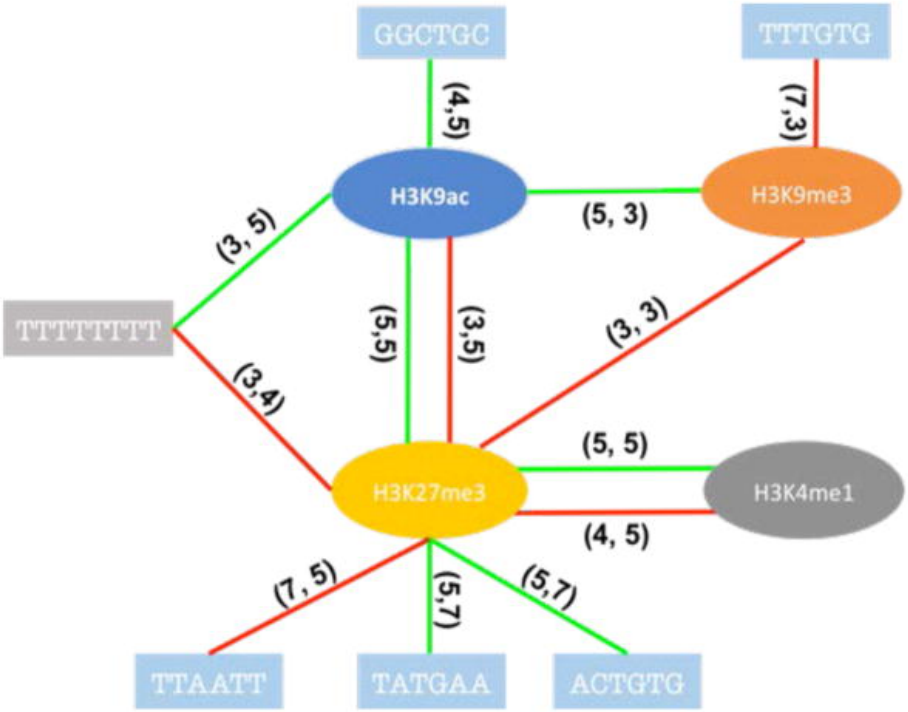


**Figure 3-5: Potential interactions for chromatins-genomics, chromatins-chromatins in GM12878.** The red line means negative to exon exclusion, green line means positive to that. The numbers on the line indicate feature location (Fig. 3-8).

Nevertheless, it is possible that the effect of some of the location-specific genomic features may be modulated by chromatin context. In other words, there may be interactions between specific genomic and chromatin features. However, these interactions cannot be directly quantified in our DNN model. Therefore, we applied L1 norm to the first layer of the DNN model to make the connections sparse, then explicitly assessed the interactions among the selected features based on a linear regression model, using the model selection package “stepwiselm” in Matlab (Create linear regression model using stepwise regression - MATLAB stepwiselm). We investigated both chromatin-genomic and chromatin-chromatin interactions. The results are summarized in Fig. 3-5 – 3-7 for each cell line respectively. Our results suggest that chromatin context can potentially modulate the effect of genomic features on splicing. Moreover, both chromatin-genomic and chromatin-chromatin interactions are position specific, which is consistent with a mechanisms that relies on specific genomic and RNA conformation and binding of splicing factors. What's more, many cis-elements within the skipped exons tend to interact with chromatin features. However, interactions between chromatin and genomic features in the context of splicing has not been studied before making it difficult to directly assess the observed interactions based on existing literature and more experimental work is needed to further investigate our findings.

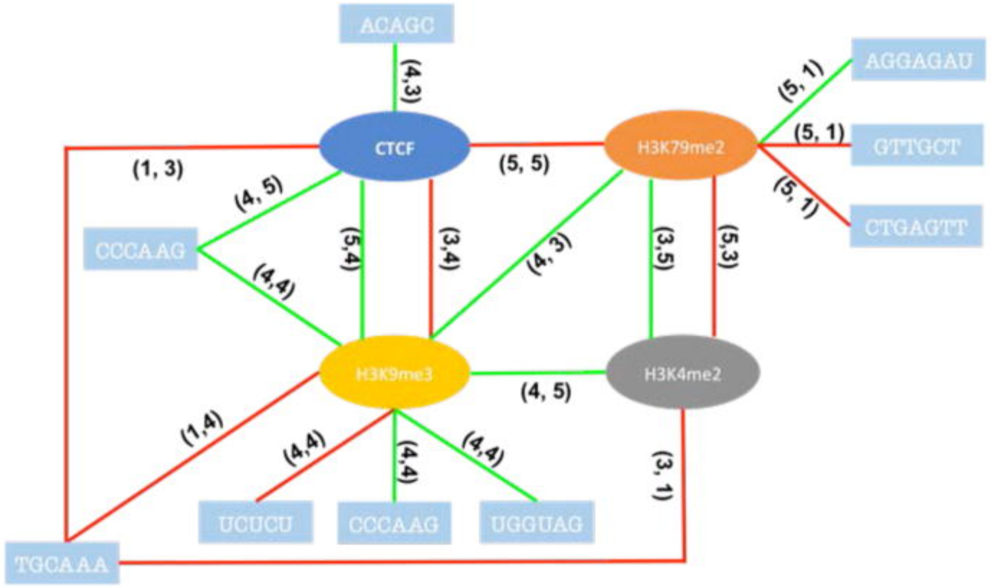


**Figure 3-6:Potential interactions for chromatins-genomics, chromatins-chromatins in h1-hESC.** The red line means negative to exon exclusion, green line means positive to that. The numbers on the line indicate feature location (Fig. 9).

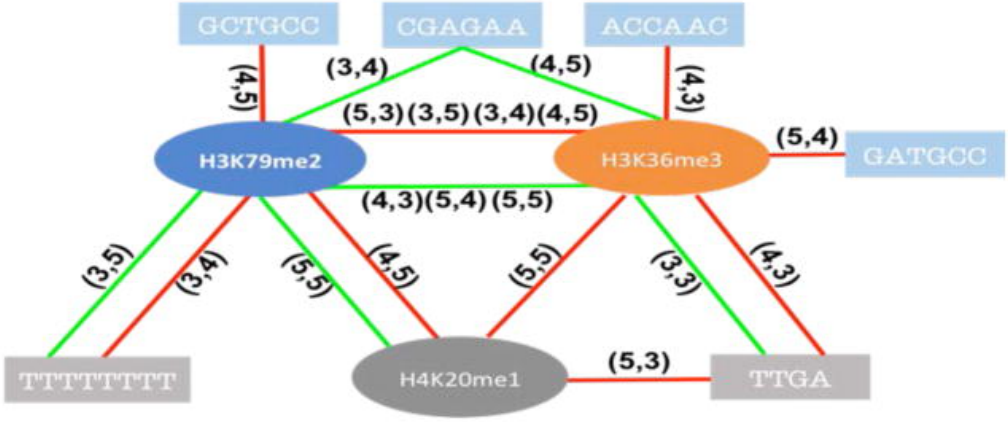


**Figure 3-7: Potential interactions for chromatins-genomics, chromatins-chromatins in K562.** The red line means negative to exon exclusion, green line means positive to that. The numbers on the line indicate feature location (Fig. 9).

## 3.3 Materials and Methods

### 3.3.1 Approach Overview

We downloaded MISO skipped exon splicing events annotations (Katz et al., 2010). Based on the MISO skipped exon splicing events annotations and the RNA-seq data in three cell lines, we used MISO package (Katz et al., 2010) to estimate the sample-specific exon inclusion fractions for all annotated cassette exons. We excluded the genes which are not expressed in any given cell type based on expression data from Gene Expression Omnibus as an independently ascertained expression data. The distributions of exon inclusion levels are bimodal as supplementary figure 3-S5 shows, suggesting that most exons tend to be either included or excluded in a given context. Moreover, these extreme cases are more likely to be robust. We therefore considered 40% of events from each end (total of 80% of data) of the distributions for three tissues for the investigation of determinants of exon inclusion. Exons whose inclusion levels are closer to 0 represent excluded exons, whereas the exons to the right of the spectrum are considered included exons. We thus formalize the problem as a two-class classification problem.

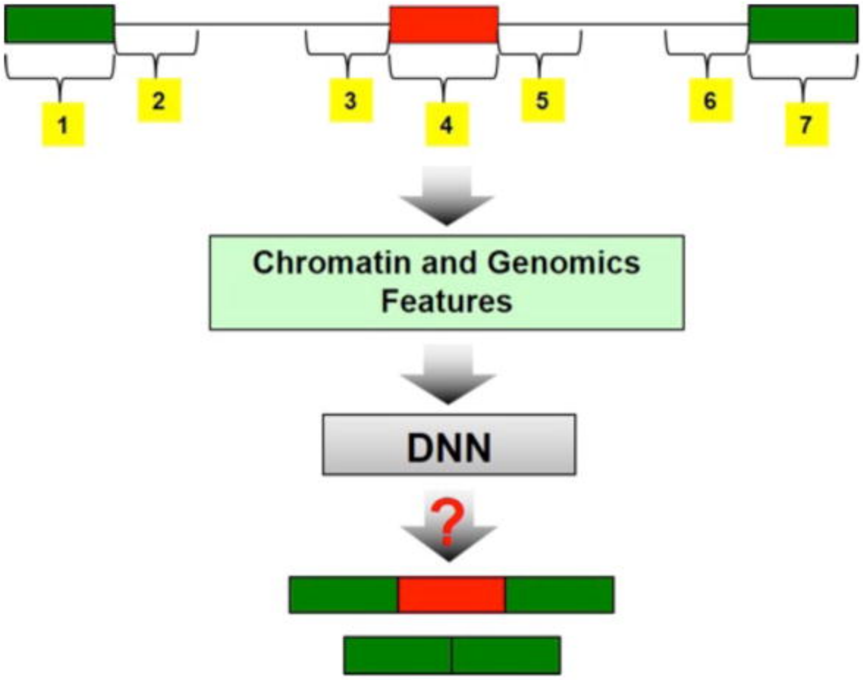


**Figure 3-8: Predictive model for exon inclusion prediction.** We extracted features from the 7 regions in yellow in the skipping exon event structure. We employed deep neural network model to perform supervised learning to predict exon inclusion.

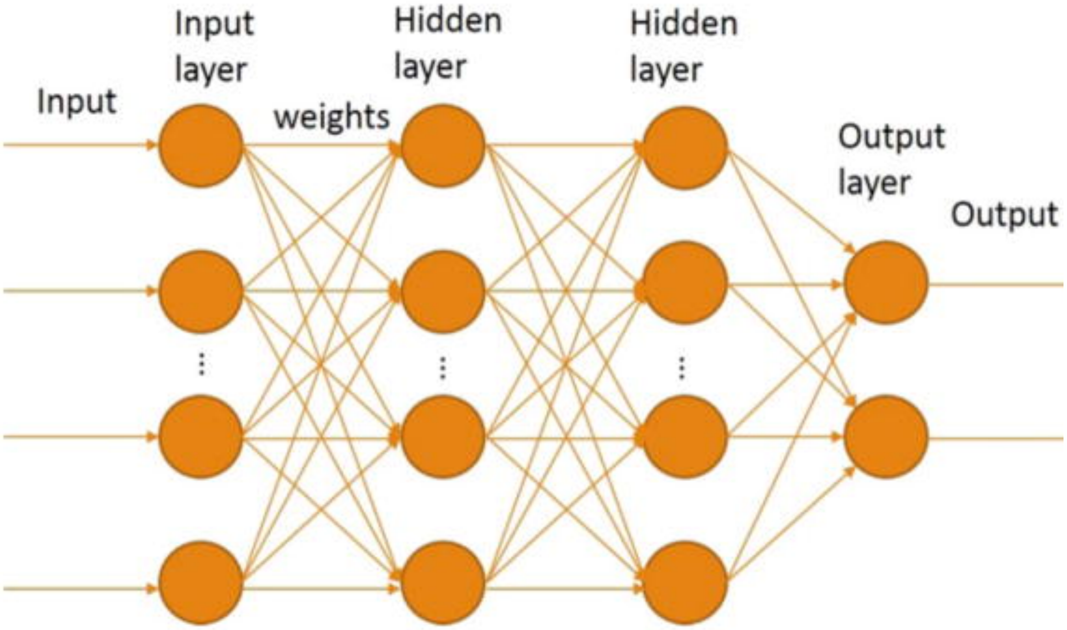


**Figure 3-9: The deep neural network architecture we used.**

We obtained processed Histone modification (Chip-seq), CTCF (ChIP-seq), RNA-seq, DNA Methylation (Methyl RRBA), Dnase-seq data for each other for Blood tissue (GM12878), Embryonic Stem cell tissue (H1-hESC), Leukemia tissue (K562) respectively from ENCODE project (www.encode.org). The chromatin features include histone modifications (H2AFZ, H3K36me3, H3K27ac, h3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9me3, H3K9ac, H4K20me1), DNA methylations, DNase hypersensitivity (DHS), and CTCF binding. The genomics features include a total of 560 motifs (which include validated known splicing motifs (Barash et al., 2010; Cartegni, 2003) and new potential splicing motifs (Fairbrother et al., 2002; Yeo et al., 2004), splice sites scores, exon length, intron length and conservation scores for the 7 regions.

For each cassette exon, given the flanking exons and the introns, we selected seven regions for feature extraction as shown in Fig. 3-8. Regions 1, 4, 7 are exons whose length varied. Regions 2, 3, 5, 6 are 450 bps intron regions proximal to the 3 exons. A previous work used the regions 1–7 for genomics features (Barash et al., 2010). For chromatin features we only used regions 3, 4, and 5 because we found signals from region 1, 2, 6, 7 not to be effective by comparing the prediction accuracies before and after we include these regions (results not shown). For each region, we divided it into 9 windows and used as the chromatin feature value, the fraction of the windows overlapping a broad peak for each feature.

### 3.3.2 Deep Neural Network model

As mentioned earlier we treat the exon inclusion prediction problem as a two-class classification problem. We applied the deep neural network model, which has been widely used in computer vision and nature language processing field. DNNs are probabilistic generative network models with multiple hidden layers (Deep Neural Networks for Acoustic Modeling in Speech Recognition - Microsoft Research - http://research.microsoft.com/apps/pubs/default.aspx?id=171498). All nodes at a layer have complete directed connections to the nodes in the next layer. Each layer includes multiple neural units, which contain a transfer function. Transfer function can be customized based on the application.

The activation ac of each node depends on the input features f, connection weights w, the bias b and transfer function T:

$$ac = T(\sum f_{ij} w_{ij} + b_j)$$

We used logistic function as the transfer function:

$$T(x) = \frac{1}{1+ e^{-x}}$$

The DNN architecture is shown in Fig. 3-9. There is a one-to-one mapping between features and the nodes in the input layer. The number of nodes for the output layer is two since this is a two-class classification problem; one of the nodes outputs the probability pe for an exon to be excluded, and the other node outputs the probability of being included pi. The predicted class c is based on maximum of the two probabilities:

$$c = \max(p_e, p_i)\ ?\ (excluded, included)$$

We utilized previously developed convenient deep neural network toolbox (Prediction as a candidate for learning deep hierarchical models of data).

### 3.3.3 Restricted Boltzmann Machine Pre-training

A Restricted Boltzmann Machine (RBM) is an undirected stochastic neural network model (Hinton et al., 2006), composed of a visible layer and a hidden layer. In neural network architecture this model could efficiently provide better initialization compared to random initialization based on maximum likelihood approach (Hinton et al., 2006). We treat each pair of adjacent layers as RBMs to perform supervised learning pre-training greedily. Hinton et al. have shown that

RBM pre-training substantially improves the training with a backward fine-tuning phase.

### 3.3.4 Dropout

Overfitting is a potential concern in supervised learning, especially for complex model with numerous parameters. Dropout technique introduced by Hinton et al. Dropout could be used to significantly reduce overfitting (Srivastava et al., 2014). Essentially, it tries to randomly drop some nodes along with all their connections in every round, followed by a fine-tuning phase. In this way, dropout can randomly sample diverse network structures and combine them in prediction step.

### 3.3.5 Feature Selection

We counted the occurrences of motifs within the regions of interest as motif features. For both chromatin states and sequence conservation, we determined the average signal within our regions of interest as feature values. We performed greedy feature selection based on the feature contributions derived from the first model. We used all the features to build the model and then employed Milne’s method (Milne, 1995) to calculate all the features’ contributions by making use of the connection weights from the model as follows:

$$rc(i) = \sum_{k=1}^{2h} \left( \frac{w_{ko}}{\sum_{m=1}^{1h} |w_{mk}|} * \sum_{j=1}^{1h} \frac{w_{jk} * w_{ij}}{\sum_{l=1}^{fea} |w_{lj}|} \right)$$

$$nc(i) = \frac{rc(i)}{\sum_{n=1}^{fea} rc(n)}$$

Where 1h is the number of units in the first hidden layer, 2h indicates the number of units in the second hidden layer, rc(i) is the raw contribution of the ith feature, nc(i) is the normalized contribution of the ith feature, w is the connection weight, fea is the number of input features.

Then we ranked all features based on their contributions, and greedily added features to the feature set till convergence of prediction accuracy.

### 3.4 Summary and discussion

In this study, we formalized the exon inclusion prediction problem as a 2-class supervised learning problem. Our primary goals here were to assess the relative contributions of chromatin and genomic features and specifically test the possibility that the previous reported associations between chromatin features and exon skipping might be largely due to their correlations with genomic features (Whitaker et al., 2014). Our additional goal was to test whether the effect of specific cis elements may be modulated by the chromatin context. Based on a comprehensive set of genomic and chromatin features in 7 and 3 regions respectively, and using deep learning model, we first verified that the genomics features are robust predictors of exon inclusion consistent with previous studies (Barash et al., 2010; Xiong et al., 2011). At the same time we found that, not only chromatin features can only modestly predict exon inclusion, they do not lend substantial information beyond what is captured by genomic features. However, our analysis reveals specific significant interactions between chromatin and genomic features suggesting that the effect of latter on exon inclusion may depend on the context provided by the former.

We employed DNN with pre-training and dropout methods, which have been widely used and proved effective in computer vision and natural language processing domains, relative to other machine learning approaches. Essentially, DNN, as a model with greater number of hidden layers, can represent higher level of abstract features, which should contribute to modeling of the association between splicing inclusion and features, in a situation such as splicing where the precise mechanisms are not known and there are likely to be several interactions among features and stepwise decision being made,. However, complex model are more vulnerable to overfitting. Pre-training and dropout algorithms are meant to reduce overfitting (Srivastava et al., 2014). Finally, it is not easy to quantify interactions within this complicated network model. While we rely on DNN to rank features by significance, to assess interaction we employed a simpler model. In our study we used standard linear regression to model interactions because of their high interpretability.

Previous works relying entirely on genomic features have proposed a highly accurate context-specific splicing code (Barash et al., 2010). We rely on the dictionary of cis elements compiled in these previous works. However, there are some notable differences between our work and these previous works. Our focus is not to optimize the prediction accuracy, but rather to explore relative contributions of genomic and chromatin features. We have therefore explicitly excluded operational, but non-mechanistic, features such as ‘translatability’. Moreover, while we estimate prediction accuracy within a cell line independent of other cell lines, these previous works in fact predict increase/decrease in exon inclusion in a cell line relative to other cell lines, and as such they rely on data from all cell lines simultaneously and boost the accuracy through information shared across cell lines. Therefore the absolute prediction accuracy reported here are not directly comparable to the previous reports.

Even though, by and large, the chromatin features are not highly predictive of exon inclusion, we found specific features to be highly significant. H3k36me3 is one of the most significant features and is consistent with previous report. For each feature revealed by our model as significant, we also directly verified the association between that specific feature and splicing inclusion, and examined the joint effect-size and significance as the feature relevance (Fig 3-2B, 3-S3, 3-S4). We found that most of the detected relevant features are consistent with previous correlation study (Shindo et al., 2013; Zhou et al., 2012; Zhu et al., 2013). In both GM12878 and h1-hESC, H3K36me3 is one of the most significant chromatin marks contributing to splicing, consistent with previous reports (Shindo et al., 2013; Zhou et al., 2012; Zhu et al., 2013). While previous computational association studies suggest that H3K36me3 at exon-intron boundary and exon has a positive effect on exon inclusion, in contrast, our analyses suggest that this mark can have both positive and negative effect in GM12878, depending on its precise location, which is consistent with various potential mechanisms based on experimental studies (Zhou et al., 2014). H4k20me1 is significant in all three tissues, consistent with (Zhu et al., 2013). Moreover, H3K79me2, H3K9ac, H3K27me3, H3K9me3 also showed varying degrees of significance. In addition, in stem cell most chromatin features within skipped exon have strong positive correlation with exon inclusion, which may imply that they can contribute to define exon or recruit SR proteins during splicing.

Our finding that genomic signals carry almost all of the information predictive of exon inclusion, and that predictive power of chromatin features is not independent of genomic elements should not come as a surprise. Despite previously shown associations between chromatin marks and splicing, it is likely that the chromatin signals themselves may be ultimately governed by the underlying genomic elements and the proteins binding to them. This could be true even in the rare cases where a direct mechanistic link has been inferred from a specific chromatin feature and splicing (Luco et al., 2010; Luco et al., 2011). Recent reports showing highly accurate predictability of chromatin features by genomic sequence strongly suggests that not just for splicing, but, unsurprisingly, numerous other cellular processes, such as transcription initiation, poly-Adenylation, etc., even when there strong association and mechanistic links with chromatin features, the ultimate drivers are likely to be the underlying genomic elements.

We performed cross tissue test using genomics and chromatin model respectively. We found that the rules learnt from one cell type are reasonably applicable to a different cell type. The differences can be attributed to cell type specific splicing factors. We expect that chromatin features, after being largely determined by genomic features, should have conserved rules governing exon inclusion across cell types. We found high cross-cell type predictability for stem cell and leukemia. This specific observation is consistent with known broad similarities in active cellular processes between stem cell and cancers (Epigenetic similarities between Wilms tumor cells and normal kidney stem cells found -- ScienceDaily; Li and Neaves, 2006; Spike and Wahl, 2011).

Even though our analyses suggest that chromatin features are not likely to be the primary drivers of alternative splicing, they might still be able to affect splicing at the molecular level, as suggested by our interaction study (Fig. 5-7). First, chromatin features may serve to provide the recognition specificity for specific factors, similar to genomics features. At the molecular level, in most of the reported potential mechanism, chromatin features interact with many other molecules to affect splicing, such as chromatin remodeling protein and SR protein. We speculate that the spatial position of those chromatin marks may influence their protein recruitment or conformational changes after recruiting other factors. Moreover, recruitment of different protein factors can have different effect on splicing. For example, H3K36me3 can both facilitate or suppress splicing by recruiting MRG15 or Psip1 respectively (Luco et al., 2010; Luco et al., 2011). In GM12878 cell line, we observed interaction between H3K27me3 and H3K4me1, which have been suggested to together mark poised enhancer (Creyghton et al., 2010; Rada-Iglesias et al., 2011); In h1-hESC cell line we identified interaction between CTCF and H3K9me3, which have been shown to co-localize (Thomas et al., 2011). In K562, we observed interactions between H3K79me2 and H3K36me3, which are both markers of gene bodies, that is likely to be important for exon definition process in splicing regulation. While we do observe interactions between chromatin and genomic features, very little is known in the literature to reasonably corroborate our findings. Moreover, the mapping between specific cis element and corresponding splicing factor is not known for the most part, making it difficult to interpret the results pertaining to cis element interactions. In GM12878 sample, we detected a potential interaction between H3K9ac and motif "GGCTGC". Even though the protein interacting with the cis elements is not known, we speculate that a splicing repressor like hnRNP binds to the motif to repress inclusion when H3K9ac is present. In h1-hESC, we identified an interaction between SRSF9 protein and H3K79me2. However, the specific locations of the two features are genomically distal from each other (Fig. 6). Such distal interactions are entirely possible due to looping at both DNA and RNA level (Matthews, 1992; Paek et al., 2015; Rueda et al., 2009). In K562 sample, the interactions of H3K36me3 with different motifs have different effects on exon inclusion suggesting diverse potential mechanisms discussed earlier.

# Chapter 4: Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases

## 4.1 Introduction

Almost all multi-exon genes in human exhibit alternative splicing (Black, 2003; Wahl et al., 2009), which alongside transcriptional regulation, significantly contribute to the transcriptomic as well as phenotypic diversity at multiple biological scales (Nilsen and Graveley, 2010). Much like transcription, splicing is highly regulated, by both genetic and environmental factors, and its dysregulation is implicated in, among other things, normal aging as well as age-associated diseases (Deschênes and Chabot, 2017; Eriksson et al., 2003; Wang and Cooper, 2007b; Watson et al., 2013).

Normal aging is associated with systemic changes in cellular processes involving both transcriptional and post-transcriptional controls (Johnson et al., 1999). While some of the changes in molecular processes are caused by age-related changes in the cellular environment, it is possible that molecular changes may further contribute to the aging process, and to age-related diseases such as hypertension and cardiovascular diseases. Moreover, such age-associated changes in the transcriptional and post-transcriptional regulation are likely to vary across tissues and organs. While age-associated gene expression changes across several tissues have been previously reported (Glass et al., 2013; Yang et al., 2015), similar investigations of age-associated splicing changes are limited.

Mazin et al. have previously reported age-associated splicing changes in two brain regions (Mazin et al., 2013b), and Tollervey et al. (Tollervey et al., 2011b) have investigated age-associated splicing and transcript expression across normal and

Alzheimer's disease samples. However, these few previous studies: (1) focused only on a single or very few tissues in contrast to 48 primary tissues included in our study, (2) investigated only exon skipping events while we have studied 7 types of splicing events (exon skipping, alternative 5’, alternative 3’, mutually exclusive exon, alternative first exon, alternative last exon, and intron retention), (3) are based on very few individuals (around 35), in contrast to 177 individuals on average per tissue in our study, and highly importantly, (4) in contrast to our study, do not explicitly control for batch effect and potential hidden confounding factors, which may lead to false positives. Our study addresses these limitations in the previous studies toward a comprehensive investigation of age-associated splicing changes across human tissues, which may provide insights into age-related diseases mediated by splicing changes.

Based on ~8500 RNA-seq samples from 544 donors across 48 tissues in the Genotype-Tissue Expression dataset (GTEx version 6) (GTEx Consortium, 2015), here we report a comprehensive detection of age-associated splicing changes across tissues in human. Using a stringent model, we identified 49,869 age-associated splicing events of 7 distinct types (Keren et al., 2010), including 17,447 exon-skipping events, across the 48 tissues.

Overall, we report the first systematic genome-wide analysis of age-associated splicing events spanning 7 types of splicing events (Alamancos et al., 2015) across 48 primary tissues, paving the way for future investigations of links between alternative splicing and aging, and age-related diseases (Wang et al., 2018).

## 4.2 Results

### 4.2.1 Age-associated splicing events are prevalent in most tissues and are largely tissue-specific

Our overall pipeline is illustrated in Fig. 4-1A, and the details are provided in the Methods section. Briefly, we obtained a total of ~8500 expression samples across 48 tissues (the ones having at least 50 donor samples) and a total of 544 donors from GTEx (GTEx Consortium, 2015). The number of samples for each tissue and distributions of age and gender are shown in supplementary Fig. 4-S1. Based on GENCODE annotations (Harrow et al., 2012), we compiled 163,505 alternative splicing events of 7 different types (Fig. 4-1B) (Alamancos et al., 2015), and estimated sample-specific PSI values (Percent Splicing Index) for each event in each sample. A linear regression model was used to identify age-associated splicing

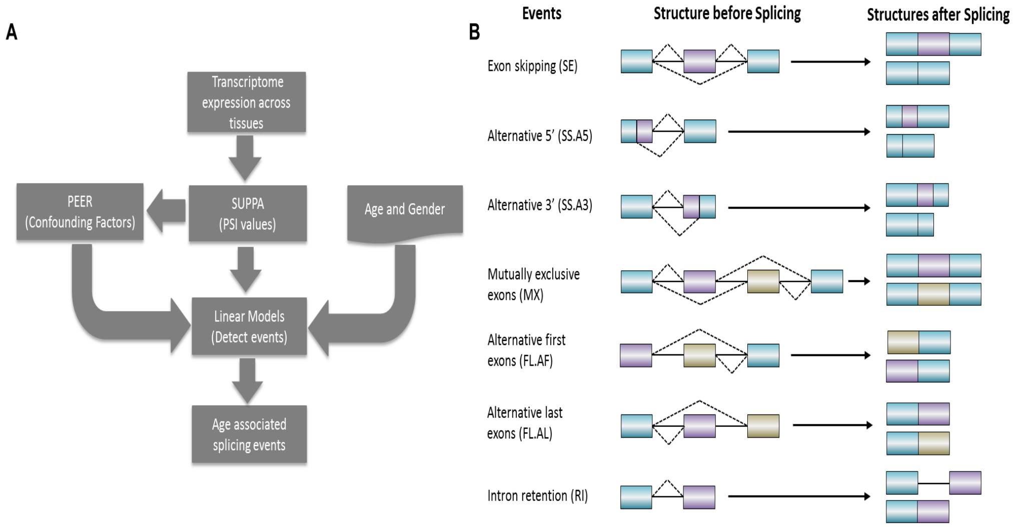

A
Transcriptome expression across tissues
SUPPA (PSI values)
PEER (Confounding Factors)
Age and Gender
Linear Models (Detect events)
Age associated splicing events
B
Events
Structure before Splicing
Structures after Splicing
Exon skipping (SE)
Alternative 5' (SS.A5)
Alternative 3' (SS.A3)
Mutually exclusive exons (MX)
Alternative first exons (FL.AF)
Alternative last exons (FL.AL)
Intron retention (RI)

**Figure 4-1: Overall pipeline and seven types of splicing events.** (A) Overall pipeline to detect age-associated splicing events. (B) Seven types of alternative splicing events before and after splicing. Blue rectangles represent constitutive portion of the exons. Purple and beige rectangles represent the alternatively used portion of the exons. Solid lines represent introns. Dashed lines connect the ends of alternatively spliced out portions of the gene. The splicing event structures before splicing and the generated isoform structures after splicing are shown on left and right side respectively.

events. To ensure sufficient statistical power, we only analyzed 48 tissues with at least 50 samples. Moreover, in a particular tissue, we only analyzed the events that could be quantified in at least 50 samples. More specifically, in a given tissue, we only analyzed the genes that are expressed in at least 50 individuals in the tissue. We thus analyzed a total of 163,505 splicing events spanning 7 types of events across the 48 tissues; a total of 3,723,596 tests.

Summarized in Fig. 4-2A, overall 49,869 events (1.3%) were found to be significantly associated with aging (FDR <= 0.05 and permutation p-value <= 0.05; see Methods) in at least one tissue; on average 1,018 events were detected in each tissue. We ascertained that the number of significant events detected in a tissue is not correlated with sample size (Supplementary Note 4-1). In addition, we show that our results are robust to the potential confounding by human ancestry (detailed in Supplementary Note 2). Fig. 4-2B specifically summarizes the exon skipping event. In Supplementary Table 4-S1 and Supplementary Fig. 4-S2 we provide a detailed summary of significant age-associated events across 48 tissues for each type of splicing event. Age-associated splicing changes are found to be most abundant in Skin (Sun exposed) and Esophagus-Mucosa; interestingly, both these tissues are composed of epithelial cells and are most exposed to external environment, have a well-established effect of aging (Berdyyeva et al., 2005).

To further illustrate age-associated splicing events, Fig. 4-2C shows clustering of samples using the sample-specific PSI values of significant age-associated exon skipping in uterus, revealing subgroups with distinct age distributions (Fig. 4-2D; cluster 2 > cluster 1: Wilcoxon p-value = 3.7e-04; cluster 3 > cluster 2: p-value = 7.8e-3). While inter-cluster differences in age distribution are expected, interestingly, the three splicing-based clusters reflect three important reproductive/hormonal stages in females (Menopause, 2015): the median age of individuals in cluster 1 is 33 years which roughly corresponds to age of first child birth, and the median age of individuals in cluster 2 is 50 years which roughly corresponds to the onset of menopause, while the individuals in cluster 3 are 60 years old on average

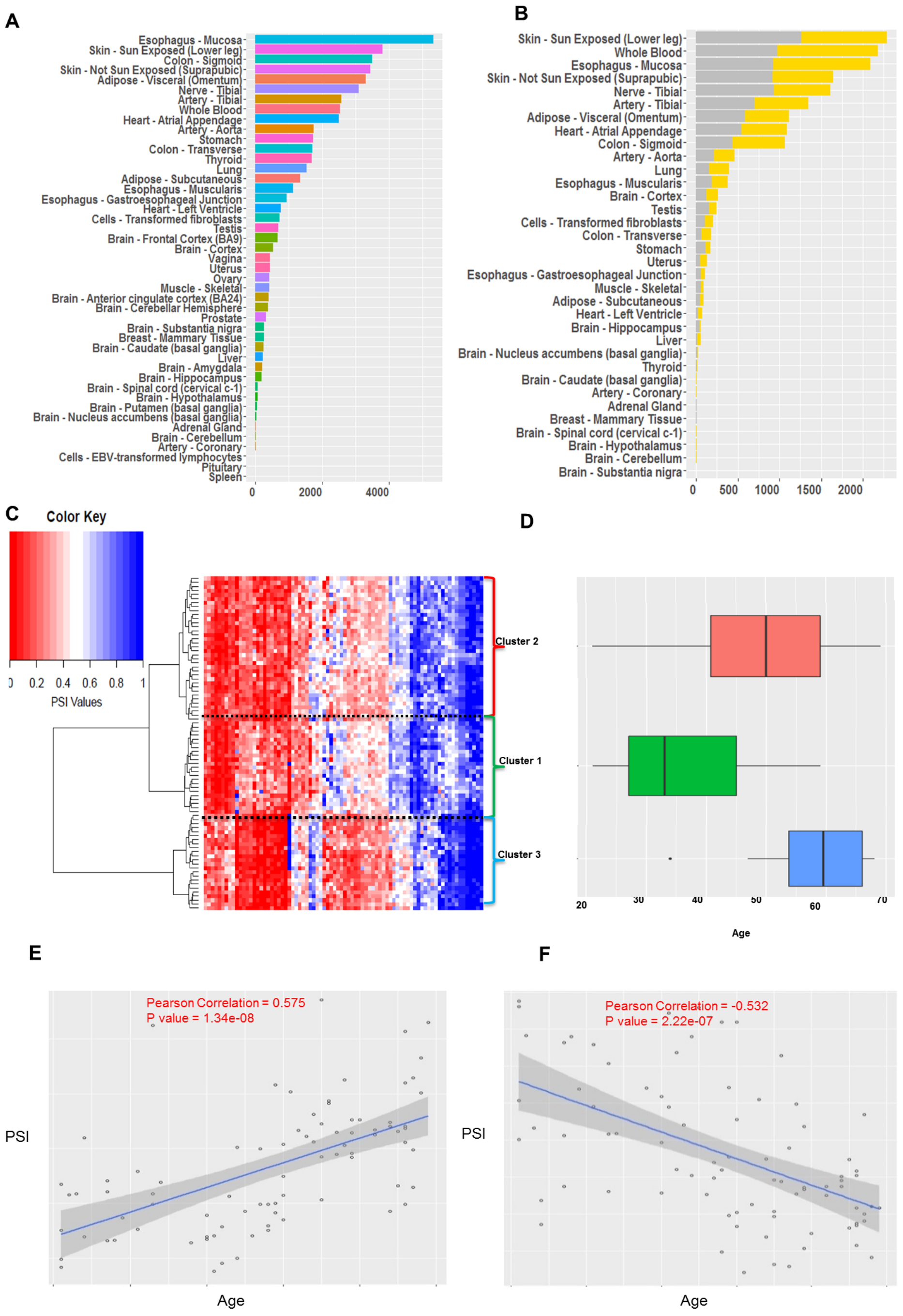


**Figure 4-2: Summary of significant age-associated splicing events.** (A) Number of significant age-associated splicing events across 48 tissues. (B) Number of significant up-regulated (increased with aging) and down-regulated (decreased with aging) exon skipping events across 48 tissues in gray and yellow color respectively. (C) Hierarchical biclustering of top age-associated exon skipping events in Uterus across individuals, based on PSI value of each event; the columns represent events and the rows represent 83 individuals with age information; blue indicates higher PSI and red indicates lower PSI. (D) Box plot of age distributions of the three identified clusters of individuals. (E) Scatter plot illustrating an up-regulated cassette exon event (COL6A3: chr2:238285987-238287279:238287878-238289558) in Uterus. (F) Scatter plot illustrating a down-regulated cassette exon event (NFE2L1:chr17:46133960-46134394:46134483-46134706) in Uterus.

corresponding to post-menopause. Fig. 4-2E and 4-2F illustrate two examples of significantly age-associated events in the uterus. Fig. 4-2E shows an exon skipping event in COL6A3, a procollagen gene important in the extracellular matrix organization, previously shown to be linked to aging in rat muscle tissue (Chaves et al., 2013). In addition, COL6A3 is related to different stages of pregnancy in mouse uterus tissue (Diao et al., 2011). Fig. 4-2F shows an exon skipping event in a nuclear factor gene NFE2L1, whose worm ortholog SKN-1 has been linked to lifespan extension (Tullet et al., 2008). In addition SKN-1 is significantly related to collagen expression (Ewald et al., 2015), which is critical for uterus. Interestingly, these two genes are also age-dependent in the other 6 and 7 tissues respectively.

Alternative splicing has been shown to be tissues-specific (Barash et al., 2010; Chen et al., 1999; Xu et al., 2002). Here we assessed the extent to which this is true of age-associated splicing changes. Toward this, we quantified tissue-pair similarity in age-related splicing as the Jaccard index based on the genes involving age-related splicing in the two tissues. As evident in Fig. 4-3A, most tissues do not share age-associated alternatively spliced genes, implying that such events are tissue-specific, similar to the alternative splicing itself.  Hierarchical clustering of tissues based on their pairwise Jaccard index revealed three clusters (Fig. 4-3A). As expected, Skin – Not Sun Exposed (Suprapubic) and Skin – Sun Exposed (Lower leg) have a high Jaccard index, and so do Colon – Sigmoid and Colon – Transverse. The cluster of 9 tissues (Fig. 4-3A top right corner), share 19 age-associated alternatively spliced genes (Supplementary Table 4-S2).  Interestingly, 15 out of the 19 genes are involved in regulating macromolecule interactions, including binding to proteins, lipids, or nucleic acids.

### 4.2.2 Age-related splicing events are potentially functionally linked to aging process

The 49,869 significant age-associated events across 48 tissues correspond to 9,884 genes. We performed functional term enrichment analysis in these genes in a tissue-specific fashion, using NIH's David online tool (Huang et al., 2008; Huang et al., 2009). We selected the terms that were enriched (FDR <= 5%) in at least 3 tissues and performed hierarchical biclustering based on those terms' enrichment levels (fold change) across tissues (Supplementary Fig. 4-S3). We also provide a TreeMap view (Supek et al., 2011) of GO terms that are enriched among age-associated spliced genes in at least one tissue in supplementary Fig. 4-S9. GO annotation is far from complete, noisy, and lacks resolution, and tissue context, making functional interpretation of enriched process in a specific context challenging. Therefore, even though we identified enriched processes in a tissue-specific way, we chose to take a broad look at the enriched processed across tissues, for a more robust interpretation. Supplementary Table 4-S3 lists the top 15 biological functions ranked according to either the number of affected tissues or fold change. These top functional terms include some well-studied processes linked to aging. For example, mitochondrion and peroxisome and their associated processes are implicated in balancing the levels of reactive oxygen species in the cell (Wallace, 2005), and cell-cell adhesion is essential for mediating tissue integrity and stem cell niche (Geiger et al., 2007).  Ribosome and ribosomal ribonucleoprotein were ranked among the top by both measures, which is in agreement with the emerging view that the ability of cells to maintain a healthy and relatively stable pool of proteins under continuous stresses that accumulate over time is a major determinant of lifespan (Andrew Dillin Cell Meta 2016). Interestingly, genes with age-associated splicing in Muscle – Skeletal (358 genes), Whole Blood (2,073 genes) and Adipose-Subcutaneous tissues (1,143 genes) are linked with all top enriched processes, with very little overlap among the respective gene sets. The interaction between aging process and alternative splicing can be bi-directional, that is, many age-related events may be downstream effects, rather than causes, of aging. The genes that are involved in typical aging-related biological functions, as well as exhibit age-associated splicing patterns may contribute to aging and aging related phenotypes, while other genes that although exhibit age-associated splicing pattern but otherwise are not involved in aging-related processes may represent downstream effects. Overall, these results show that in some tissues age-related splicing events may be functionally linked to the phenotypic changes associated with aging, while they may be the downstream effect of aging process in others. In addition, 78.6% of overall biological processes (include most aging related processes) are recaptured by only performing gene ontology analysis on genes uniquely associated with splicing changes, which implies that these reported aging related process may be related to age-associated splicing instead of gene expression.

### 4.2.3 A focus on splicing uniquely reveals numerous age-associated genes

Both transcriptional and splicing processes can change with age. With regards to splicing regulation, while the age-associate changes must be mediated at the level of individual splicing events, the downstream effects of these changes on the age-related phenotypes are mediated by changes in the levels of specific transcripts. Toward obtaining a global view of age-associated changes in these various aspects of the transcriptome, analogous to splicing event-based model above, we implemented linear models to detect age-associated changes in gene expression, transcript expression, and relative transcript ratios (Methods), and compared the genes corresponding to the significant age-associated events in the four categories – individual splicing events, gene expression, transcript expression, and relative transcript usage, shown in Fig. 4-3B, for 4 select tissues (all tissue results are provided in Supplementary Fig. 4-S4). It is apparent from this result that a focus on splicing and transcripts uniquely reveals numerous age-associated genes. Specifically, for instance, 18% of the metabolic genes, known to be significant aging markers, are revealed as age-associated only at the transcript level, and not at the level of overall gene expression, e.g., Phosphofructokinase gene locus (PFK), which has previously been targeted in cancer therapy (Yi et al., 2012), exhibits age-associated changes at the transcript level in 21 tissues but not at the level of gene expression (Supplementary Fig. 4-S5 illustrates the known isoforms of PFK and their age-associations) (Liu et al., 2015). These results suggest that aging process has substantial association with post-transcriptional regulation beyond its known associations with transcriptional processes.

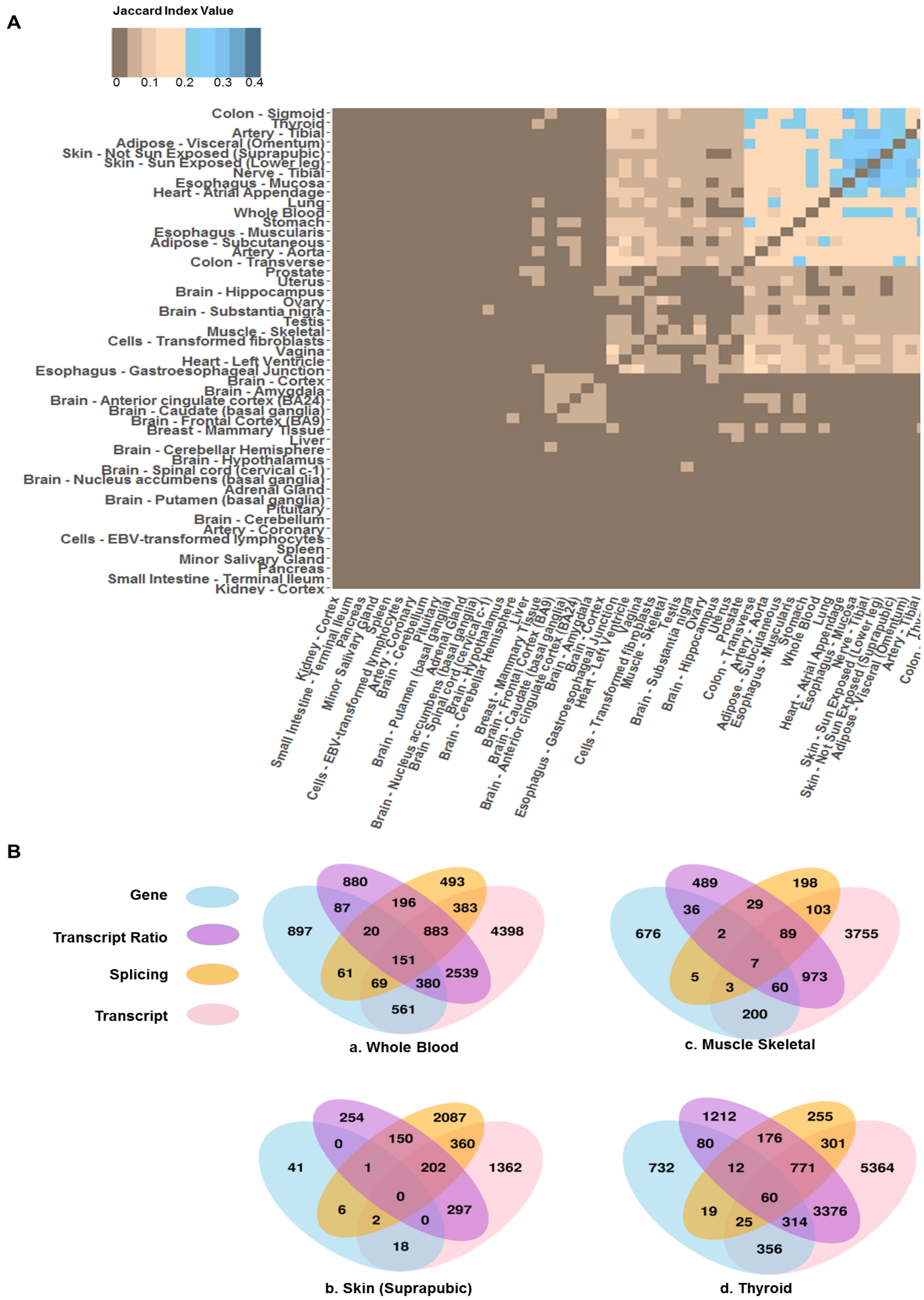


**Figure 4-3: Comparison of age-associated genic changes across tissues and across approaches to capture genic changes.** (A) Clustering of tissues based on pair-wise similarity of genes affected by age-associated splicing, based on Jaccard Index. The darker blue color indicates higher similarity and darker brown indicates lower similarity. (B) Overlap of gene sets affected by age-associated changes detected at the level of gene expression (sky blue), transcript ratio (medium orchid), splicing events (orange) or transcript expression (pink). (a) – (d) show the data for whole blood, Skin, Muscle-skeletal, and Thyroid tissues respectively, as examples. Data for all tissues are provided in Supplementary Fig. 4-S4.

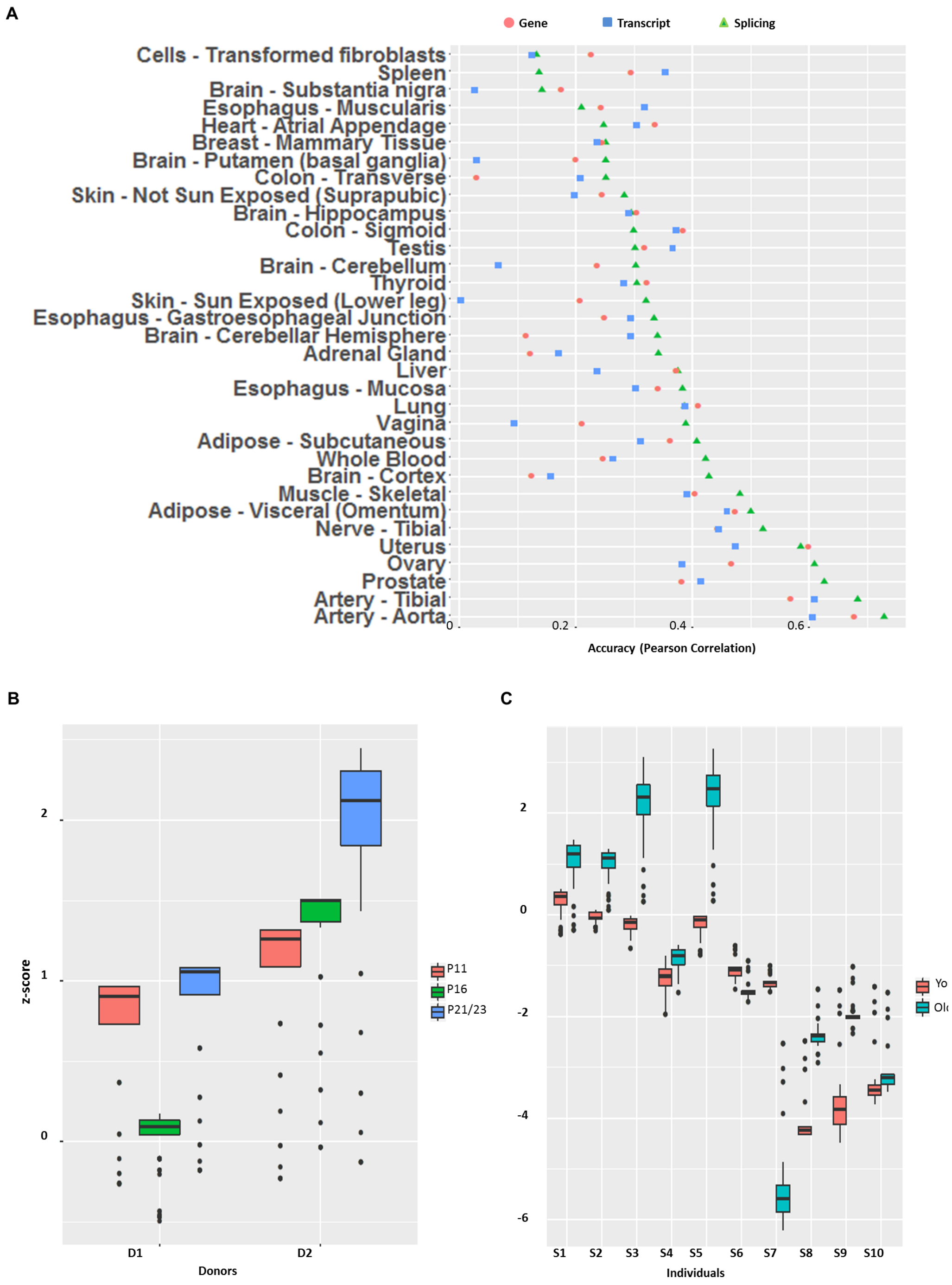


**Figure 4-4: Prediction of age and computational validation.** (A) Accuracies of prediction of age using models based on gene expression (pink circle), splicing events (green triangle) and isoform expression (blue square) across tissues. The accuracy is measured by Pearson correlation between predicted and true ages based on cross-validation.(B) The figure shows predicted relative ages for two sets of skin fibroblast cell culture (D1 and D2) across three passages (PX: passage X). (C) The figure shows predicted relative ages of ten pairs (from individuals S1 through S10) of longitudinal skin fibroblast samples. The matched young and old samples were derived from the same individual over time.  The y-axis in (B) and (C) denote normalized predicted age.

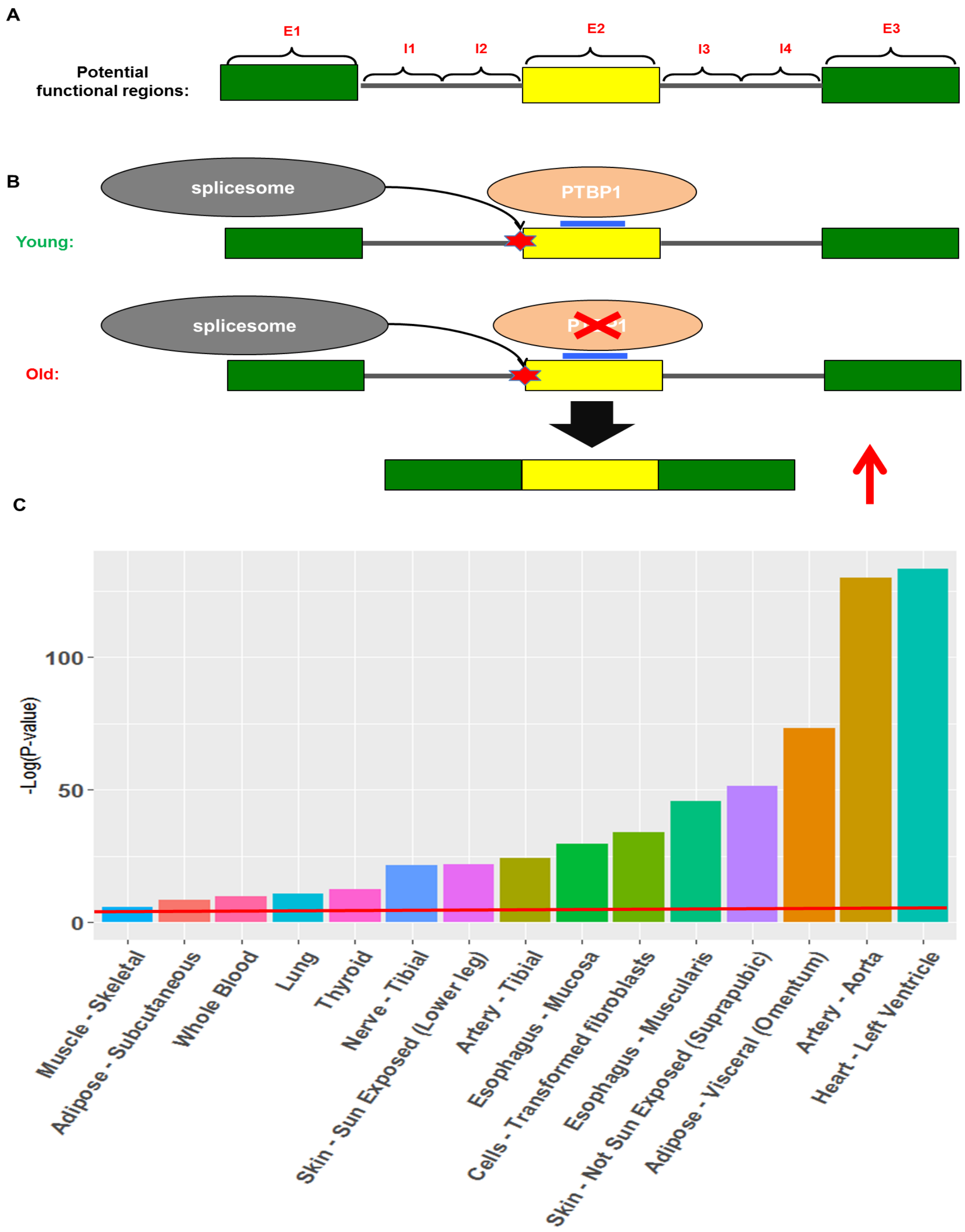


**Figure 4-5: Potential mechanism and contributions to hypertension.** (A) The 7 functional regions relative to an exon (yellow rectangle) potentially impacting the splicing event. (B) Illustration of a potential mechanism of age-associated change in exon inclusion whereby the expression of a splicing factor PTBP1, which suppresses the inclusion of the cassette exon, decreases with age, thereby resulting in an increase in exon inclusion with age. (C) Additional contribution of splicing events to the explained variance of Hypertension, in multiple tissues shown on x-axis. The y-axis denotes the significance (-log(p-value)) based on a log-likelihood ratio test. The red line indicates p-value = 0.05.

### 4.2.4 A splicing-based model is informative of biological and cellular age

We first assessed the extent to which genome-wide splicing profile in an individual is reflective of the individual's biological age. Furthermore, to compare the merits of splicing profile relative to gene expression and transcription expression profiles, we constructed three analogous models of age based on splicing profile, gene expression profile, and transcript expression profile (Methods). The accuracy was quantified as the Pearson correlation between predicted ages and true ages in cross-validation samples. Fig. 4-4A shows the 10-fold cross-validation prediction accuracies of the three models across 33 tissues; only the tissues in which all three models yielded positive predictive accuracy are shown. A direct comparison of model accuracies based on paired Wilcoxon test across tissues reveals that splicing-based model outperforms the other two models (p-values <= 1.1e-3). Surprisingly, the isoform-based model is not significantly better than the gene expression-based model, which may be due to incompleteness and inaccuracies in isoform annotations and noisy quantification of isoform expression. In addition, we implemented an alternative approach to estimate accuracy. We partitioned the individuals into two classes of old and young (Old class: the oldest 25% and Young class: the youngest 25%) and performed a standard classification based on Lasso regression. The results are consistent, in that the splicing events results in better prediction accuracy than the other two modalities, and on average the prediction accuracy is 71% (supplementary Fig. 4-S6). Overall our results suggest that global splicing profile is more predictive of age compared to gene and isoform expression.

We also compared the 7 types of splicing events regarding their individual ability to predict age following an analogous procedure as above. The results are shown for the 25 tissues in which all seven models yielded positive predictive accuracies (Supplementary Fig. 4-S7). Overall the exon-skipping events are the best predictor of age compared to the other 6 types of events (all Wilcoxon test p-values <= 7.4e-3).

Next, we assessed whether a splicing-based model of age constructed using GTEx skin fibroblast samples can successfully predict relative ages of two independent longitudinal datasets of skin fibroblast (Methods). This analysis is limited by the data availability. The first dataset consists of cell passage (a standard proxy for cellular age) data, which includes young (11 passages), middle (16 passages) and old samples (21 or 20 Passages) for two healthy individual derived skin fibroblasts (6 samples). In addition, donor 1 is younger than donor 2. The second dataset 32 includes 10 pairs of longitudinal samples from 10 donors at two different ages separated by 15.7 years on average (20 samples). To specifically assess the contributions of age-associated splicing events, the model was constructed using only the significant age-associated splicing events detected in GTEx (Methods).

In the first validation dataset (Fig. 4-4B), our GTEx-trained model correctly predicts the lowest passage cells to be younger than the oldest passage cells in donor 1 (D1), but fails to correctly predict the age of middle passage cells. However, in donor 2 our model correctly predicts the relative ages of the three cell passages. Out of total 19 pairwise comparisons (based on donors' age and cellular age), we correctly order the samples in 16 (84%) of all the cases. A paired

Wilcoxon test of the 19 pairwise predicted ages showed significance with p-value is 0.0047. In the second longitudinal dataset (Fig. 4-4C), in 8 out of 10 cases, our model correctly predicts the relative ages of the two samples from the same individual.

### 4.2.5 Some of the age-associated splicing events may be driven by age-associated expression changes in the upstream splice factors

In exploring the mechanisms underlying age-associated splicing changes, we assessed whether certain motifs near the splicing event recognized by a splicing factor, along with age-associated changes in the expression level of the splicing factor, can together explain the changes in splicing. This analysis was restricted to exon skipping events. In each tissue independently, using the significant tissue-specific events, separately for up-regulated and down-regulated event, we identified the splicing regulators whose RNA-recognition motifs (obtained from (Ray et al., 2013) ) were significantly enriched in any of the 7 regions near the cassette exon (Fig. 4-5A), relative to the background cassette exons whose usage did not vary with age (Methods). An enrichment threshold (FDR <= 0.1) was applied to retained potential functional motifs.

Supplementary Table 4-S4 lists the 9 potential splicing factor drivers of age-associated splicing changes identified in skin fibroblast. Splice factor PTBP1 is known to inhibit exon retention by binding to exonic splicing enhancers (Spellman and Smith, 2006). We found that PTBP1 motifs are significantly enriched within the middle exon among up-regulated exon inclusion events and consistently, PTBP1 expression showed a significant decrease with age (standardized age covariate coefficient = -35.8). Illustrated in Fig. 4-5B, this example suggests a potential mechanism whereby an age-associated decrease in PTBP1 concentration lifts its inhibitory effect resulting in increased exon retention at multiple loci.

We sought for experimental support for splicing factor-mediated changes in splicing through age. We obtained 3546 potential PTBP1 targets in HeLa cell based on PTBP1 CLIP-seq data (Dror et al., 2016); such data is not available for skin. We then independently, using our approach, identified 46 genes whose age-associated splicing is potentially a downstream effect of PTBP1. We found that experimentally identified potential target genes of PTBP1 are highly enriched among the targets identified by our pipeline (Fisher test p-value = 1.3E-05; Odds-ratio = 2.4).

### 4.2.6 Age-associated splicing contributes to complex age-related diseases

Several complex diseases, many of which exhibit increased incidence with age, have been shown to be associated with distinctive tissue-specific gene expression profiles (Bruneau, 2008; Demichelis et al., 2012). Potential mechanisms linking alternative splicing to age-related diseases have been explored previously. Alternative splicing might change the transcript ratio leading to a greater fraction of impaired protein isoform, truncated wild type protein, or suboptimal isoform ratios, which might affect the cellular processes underlying age-related diseases

(Deschênes and Chabot, 2017; Li et al., 2017). Alternative splicing within genes EAAT2, SALL1 and TAU have been shown to contribute to age-related diseases (Li et al., 2017; Lin et al., 1998). Given our observed links between splicing and aging, we assessed the extent to which tissue-specific splicing profile potentially contributes to age-related diseases. We tested this for 4 diseases, including hypertension, for which there are a sufficient number of samples in GTEx in multiple tissues. For a given disease and tissue, Log Likelihood Ratio (LLR) test (Method) was used to assess the independent contribution of splicing profile to the disease by controlling for age, gender, and gene expression. As shown in Fig. 4-5C, relative to gene expression, age and gender, splicing can significantly (p-value <= 0.05) explain additional hypertension disease state variance in all of the 15 tissues tested. The three most significant tissues are Heart, Artery, and Adipose, which have well-established mechanistic links to hypertension. Results for three additional pathologies – Heart Attack, Chronic Respiratory Disease, and Diabetes mellitus type II, show consistent results (Supplementary Fig. 4-S8). Due to relatively small sample size, we analyzed fewer tissues for these three diseases. In 7, 4 and 3 tissues respectively for Diabetes mellitus type II, Chronic Respiratory Disease and Heart Attack, age-associated splicing events provide significant independent contribution in addition to age, gender and gene expression. In addition, we show that our results are robust to potential confounding by human ancestry (race) by additionally controlling for race in both the null and alternative models (concordance correlation coefficient of the two p-value distributions is

0.98). These results suggest links between splicing and complex age-related diseases independent of age and the genome-wide gene expression profile.

## 4.3 Materials and Methods

### 4.3.1 Splicing Level Quantification using GTEx data

The processed transcript expression data for ~8500 samples from 544 donors across 48 tissues were downloaded from Genotype-Tissue Expression (GTEx) database version 6 (GTEx Consortium, 2015). GENCODE genome annotation version 19 (Harrow et al., 2012) and SUPPA software package 48 was employed to extract 7 types of exon-centric splicing event annotations (exon skipping, alternative 5', alternative 3', mutually exclusive exons, alternative first exon, alternative last exon, intron retention). Then in each sample SUPPA was used to quantify the splicing level of each annotated event in terms of PSI values (Percent Splicing Index).

### 4.3.2 Model for detecting age-associated splicing events

To detect significant age-associated splicing events, we modeled the association between each event and age across multiple samples as follows :

$$PSI_{ij} = \alpha_i + \beta_i^1 AGE_j + \beta_i^2 GENDER_j + \sum_{k=1}^{n} \beta_i^{k+2} \mathrm{PEER}\left(\mathrm{CF}_\mathrm{j}^\mathrm{k}\right) + \varepsilon_{ij} \; (1)$$

Where $PSI_{ij}$ is the splicing level for event i in sample j, $AGE_j$ and $GENDER_j$ denote the age and gender of individual j respectively, $\mathrm{PEER}\left(\mathrm{CF}_\mathrm{j}^\mathrm{k}\right)$ denotes the kth confounding factor estimated using PEER packages 49 for individual j. $\alpha_i$ is the intercept for the model of event i, $\beta_i^1$ and $\beta_i^2$ are the coefficients respectively for age and gender covariate for event i, $\beta_i^{k+2}$ is the coefficient of the kth confounding factor for event i, $\varepsilon_{ij}$ is the error in the model for event i of individual j. In addition, in this model n is the number of hidden confounding factors we estimated (n= 20) compared to 15 hidden confounding factors used in Brinkmeyer-Langford et al's age-associated gene expression study 50.

Since some genes are not expressed in some of the samples, when modeling such splicing events corresponding to those genes, we excluded samples where the gene was not expressed (reported as -1 by SUPPA package). Further, to ensure statistical power, we only analyzed events having at least 50 samples where the corresponding gene had non-zero expression. For each event, we fitted the data to the model and examined the age covariate coefficient $\beta_i^1$, and assessed the significance for its deviation from zero, and applied FDR control across all tested events. In addition, we performed permutation test by shuffling the age distribution across all individuals. For each event, permutation test is performed for 1000 times and estimate the significance of the age covariate. Events with FDR <= 0.05 and fewer than 5% of the permuted data showing significance (p-value <= 0.05) were deemed significantly age-associated.

### 4.3.3 Correcting for confounding factors using PEER package

We ensured that our detected link between an event and age is not due to confounding factors, as follows. PEER software package is widely used in eQTL studies to correct for potential hidden confounding factors such as batch effects (Stegle et al., 2012). For each tissue, given the global PSI profiles (including all events of all types) for all individuals, we estimated 20 ‘PEER’ factors. Then we estimated Pearson correlation between each PEER factor and age across all individuals, and excluded the factors, in an event-specific way, that were significantly correlated with age (p-value < 0.05).

### 4.3.4 Functional Enrichment analysis

We map each significant age-associated splicing event to its corresponding gene, and identified significantly enriched (FDR <= 0.05) GO terms in a tissue-specific manner using NIH’s David online tool (Huang et al., 2008; Huang et al., 2009). We further retained only the terms that were enriched in at least three tissues.  Finally we performed hierarchical biclustering based on the enrichment level (fold change) of enriched functional terms across tissues. In addition, we used package “REVIGO” (Supek et al., 2011) to generate a TreeMap view of the enriched GO terms, and for each GO term the number of tissues in which it was significantly enriched (FDR <= 0.05) was used as the enrichment score for visualization.

### 4.3.5 Cross tissue similarity in age-associated splicing

Jaccard index is a metric to measure the similarity between two sets $\left(J(A,B) = \frac{(A\cap B)}{(A\cup B)}\ (0 \le J(A,B) \le 1)\right)$. We employed this metric to measure the similarity of age-associated splicing between two tissues. For each tissue, we identified the genes having at least one age-associated splicing event, and estimated the Jaccard index using the tissue-specific gene sets.

### 4.3.6 Age-associated gene, isoform and isoform ratio detection across tissues

In order to assess age-associated changes in gene expression, isoform expression, and isoform ratio levels, we developed three linear models.

Gene model:

$$G_{ij} = \alpha_{ij} + \beta_i^1 AGE_j + \beta_i^2 GENDER_j + \sum_{k=1}^{n} \beta_i^{k+2} \mathrm{PEER}\left(\mathrm{CFG}_\mathrm{j}^\mathrm{k}\right) + \varepsilon_{ij} \ (2)$$

Transcript model:

$$T_{ij} = \alpha_{ij} + \beta_i^1 AGE_j + \beta_i^2 GENDER_j + \sum_{k=1}^{n} \beta_i^{k+2} \mathrm{PEER}\left(\mathrm{CFT}_\mathrm{j}^\mathrm{k}\right) + \varepsilon_{ij} \ (3)$$

Transcript ratio model:

$$TR_{ij} = \alpha_{ij} + \beta_i^1 AGE_j + \beta_i^2 GENDER_j + \sum_{k=1}^{n} \beta_i^{k+2} \mathrm{PEER}\left(\mathrm{CFT}_\mathrm{j}^\mathrm{k}\right) + \varepsilon_{ij} \ (4)$$

i and j represent an event and an individual respectively, and n denotes the number of confounding PEER factors considerd. $\beta_i^1, \beta_i^2 and\ \beta_i^{k+2}$ denote the coefficient for age, gender and kth confounding factor for event i of individual j. G denotes gene expression, T represents for transcript expression and TR denotes transcript ratio ($\frac{T_x}{\sum T_x}$). CFG (equation 2) and CFT (equation 3 and 4) denote confounding factors derived from the genome-wide gene expression and transcript expression profiles respectively. The procedure is the same as that was used to detect age-associated splicing events.

### 4.3.7 Predicting Age using splicing level, isoform expression, and gene expression

To compare the power of splicing level, isoform expression and gene expression in predicting age, we built a LASSO regression model for each of them. For the splicing model, MDS analysis was performed over the population PSI values, and top 30 PCs were used as features in the linear model to predict age. In order to remove sampling bias, we performed randomized 10-fold cross validation 100 times and estimate the average cross-validation predicted age for each sample, and estimate the accuracy as the Pearson correlation between the predicted and the given age. For comparison, we implemented an identical procedure using gene-level expression as well as isoform-level expression.

### 4.3.8 Predicting relative ages in independent datasets using splicing-based model of age

To validate our splicing-based model of relative age, we build a lasso regression model based on our detected top age-associated splicing events in GTEx data to predict the relative ages in longitudinal data and cellular age data. Recall that in detecting significant age-associated splicing events, we perform a permutation test. Since the permutation is stochastic, we repeated it 10 times and selected 141 age-associated events detected in at least 8 permutations. These 141 events were then used to build a model of age in GTEx data, which is used to estimate the age of each sample in the independent validation set. The predicted age is transformed into a z-score using the predicted age distribution of GTEx data ro represent the relative ages. 100 rounds of model fitting were performed to remove sampling bias (the penalty parameter lambda was optimized based on randomized cross-validation) to generate a distribution of predicted z-scores for each sample in the independent dataset. Then Wilcoxon tests were performed to compare the relative ages of two samples.

### 4.3.9 Detecting potential upstream drivers of age-associated splicing changes

Here the goal was to test the hypothesis that the age-associated changes in expression of the upstream splicing factor gene contribute to the downstream age-associated splicing changes. We obtained 121 experimentally validated RNA motifs mediating splicing (Supplementary Table 4-S4), for which the corresponding splicing factors are also known. All significant age-associated exon skipping events from skin fibroblast tissue were divided to 3 classes based on the direction of their age-associate change ( class 1: increase with age, class 2: stable, class 3: decrease with age). We performed motif enrichment analysis between class 1 and 2, and also between 2 and 3. More specifically, the frequency of each motif between two classes was compared using Wilcoxon test, and FDR control was applied to select significantly enriched motifs.

Next, we performed differential gene expression (transcript level) analysis across aging using a model analogous to the model for splicing above, and detected the splicing factors (i) whose gene expression (transcript level) is significantly (p-value <= 0.05) associated with age, and (ii) whose motifs are enriched near the age-associated splicing events.

### 4.3.10 Estimating contribution of splicing profile to complex diseases

We build two nested linear models of age-related diseases in the GTEx population independetly in each of the 48 tissues. The first '*null'* model (Equation 5) relates the binary disease state to age (AGE), gender (GENDER), and gene information (GE), and the second '*splicing*' model (Equation 6) additionally uses splicing information; however, we only included significant age-associated splicing events.

$$Null\,model: \quad D_j = \alpha + \beta^1 AGE_j + \beta^2 GENDER_j + \beta^3 GE_j \ (5)$$

$$Splicing\,model: \quad D_j = \alpha + \beta^1 AGE_j + \beta^2 GENDER_j + \beta^3 GE_j + \beta^4 PSI_j \ (6)$$

$D_j$ denotes dieases status (0: normal, 1: disease). For both gene and splicing information, we reduced the dimensionality by performing MDS analysis using the top 100 PCs in both cases. Given the two model fits, we estimated the Log-Likelihood ratio and estimated the contribution of splicing information using Chi-sqaure test.

### 4.4 Summary and Discussion

Overall, exploiting ~8,500 tissue-specific transcriptomes in 544 individuals, we identified 49,869 age-related splicing events for 7 distinct types of splicing events across 48 tissues. In contrast to previous related works, our model stringently controls for potential hidden confounding factors. In addition to validating our splicing-based model of age in independent longitudinal and cell passage datasets, we show that splicing profiles are a better predictor of biological age than gene and transcript expression levels alone, and the splicing profile provides an independent contribution to age-related complex diseases. Finally, we propose a potential mechanism underlying age-associated splicing changes mediated by a concomitant change in the expression level of the upstream regulatory splice factor.

Mazin et al. identified 3,132 and 6,114 significant age-related splicing events in the two brain regions respectively, with 1,484 events in common, which represents ~5% of all events assessed. In contrast, we identified 1,066 events from the same regions. However, these represent ~0.06% of all events that we assessed, potentially reflecting the stringency of our approach. A direct comparison of events detected by their results and ours could not be made because of incompatibility of event definition. These differences could potentially be attributed to multiple factors related to sample sizes and controls.

Besides age-associated splicing studies mentioned above, recently, Yang et al. reported age-associated gene expression changes across 7 tissues from GTEx version 4 (Yang et al., 2015), and found that Blood has the most age-related gene expression changes, consistent with our splicing-based results. Lung, Muscle and Heart tissues were also shown to have significant age-associated changes in both our studies. Our study however uniquely identifies Skin to have a large number of age-associated splicing changes, which may suggest that age-associated effects in skin primarily affect splicing levels and are not reflected in gene expression levels. However, broadly, the genes revealed by both our and previous studies are related to common aging-related biological processes such as mitochondrial function, DNA repair, Cell Cycle, ATP-binding, etc. in Whole Blood, Muscle and Heart tissues.

Anomalous gene expression is often the first major factor considered when investigating aging and complex age-related diseases. However, our study suggests that tissue-specific splicing profiles may provide an additional contribution to aging and age-related diseases. Indeed previous studies have directly linked splicing dysregulation to diseases, independent of gene expression (Dror et al., 2016).

As the first multi-tissue study of age-associated splicing changes, we were able to compare such changes across tissues. Our observed lack of cross-tissue commonality is consistent with previous studies suggesting that the alternative splicing, as well as gene expression regulation, are highly tissue-specific (Ong and

Corces, 2011), and tissue-specific changes in the expression and splicing regulators can explain tissue-specificity of the age-related splicing changes.

Importantly, our analysis suggests one potential mechanism of age-associated splicing changes, namely, via age-associated expression changes of splicing regulators. Given the links between splicing and transcription (Naftelberg et al., 2015), it is conceivable that several other transcriptional mechanisms can contribute to age-related splicing changes. For instance, age-related changes in DNA methylation and histone modifications have been previously reported (Jung and Pfeifer, 2015; Kawakami et al., 2009). Specifically, DNA methylation has been shown to be an excellent biomarker of age (Horvath, 2013). Polymorphisms can also affect age-associated splicing changes, which may in turn manifest in variable vulnerability to age-related diseases. Our study provides a methodological framework and resource for future targeted investigation of links between splicing and aging.

# Chapter 5: Interactions of SNPs and age-related TFs are linked to gene regulation and complex diseases

## 5.1 Introduction

Normal aging is a critical “environmental” risk factor that is significantly associated with complex diseases such as hypertension, cardiovascular defects, macular degeneration, Parkinson's disease and cancers (Blasco et al., 2013; Niccoli and Partridge, 2012). Specifically, the prevalence of hypertension among people older than 60 years is around 65% compared to only 7.3% for people between 18 and 39 years old in 2011-2014 year based on the report of Centers for Disease Control and Prevention (CDC) in the USA (Sug et al., 2015). Numerous crucial biological functions like immune response, wound healing, DNA repair, metabolism and mitochondria function etc also significantly decline with aging (Blasco et al., 2013; Niccoli and Partridge, 2012; Wyss-coray, 2015), which further support potential pathological role of aging in complex diseases. However, the concrete mechanism are far from clear, although profound transcriptomic changes with both aging and complex diseases have been identified (Glass et al., 2013; Tollervey et al., 2011a; Wang et al., 2014b). In addition, genomic variations, associated with transcriptome variability shown by eQTL studies (Acharya et al., 2017; Das et al., 2015; Gilad et al., 2008; Westra and Franke, 2014), also potentially contribute to complex diseases revealed by GWAS studies (Adeyemo et al., 2009; Jiang et al., 2011; Visscher et al., 2017). ~120 SNPs have been found to be associated with elevated blood pressure and hypertension in GWAS studies (Franceschini et al., 2013; Ortega et al., 2015; Sofer et al., 2017; The UK Biobank Cardio-metabolic Traits Consortium Blood Pressure Working Group et al., 2018; Wain et al., 2017). Multiple independent studies detected common SNPs associated with colorectal cancer using GWAS (Dror et al., 2016; Lubbe et al., 2012; Wang et al., 2014a).

It is plausible, and indeed likely, that both aging and genome variations could contribute to complex diseases; however, potential interplay between the two factors has not been elucidated fully. It is highly likely that phenotypic effect of systemic molecular changes through aging may depend on the genotype of the individual. In other words, the genome variations might mediate age-dependent effect on phenotype or transcriptome variability. Such genotype-environment interactions have been previously investigated via incorporating SNP-Environment term in a regression model. Yao et al identified 10 age dependent eQTLs in Whole Blood (Yao et al., 2014). Using a meta-analysis approach, Simino et al detected 9 SNPs which have age dependent effect in blood pressure (Simino et al., 2014). Dongen et al also showed the methylome in whole blood could be affected by the interactions between SNPs and age (Dongen et al., 2016). However, insights about interaction and regulation mechanisms are limited in previous studies.

In our study, we incorporated an interaction mechanism in our model. More specifically, we approximate the environmental changes by age-associated changes in concentration of regulatory proteins (transcription factor), and with the hypothesis that age-associated TF might exhibit allelic binding imbalance, explore potential causal interaction effect between genotype and aging toward explaining age dependent transcriptional changes and ultimately, complex diseases.

To test our hypothesis, we proposed a novel pipeline, in which we first predicted allele-specific binding sites for putative age-associated TFs in accessible chromatin, then tested the association between interaction term and potential target gene expression and complex diseases. We performed the analysis in 25 tissues due to availability of chromatin accessibility data and detected ~591 SNP-TF-Gene triplets on average across 25 tissues. Interestingly, numerous detected interactions are related to hypertension. To assess the robustness, we estimated the replication rate in cross validation. In addition, significantly enriched epigenomic markers (chromatin accessibility, H3K27ac, H3K4me1, H3K4me3) related to regulatory elements in the interaction regions, suggest potential functionality of our detected interactions. Finally we showed target genes regulated by reported significant interactions could be linked to aging process and complex diseases by performing enrichment analysis.

In summary, we reported a novel framework to explore the potential causal interactions between SNPs and aging, in the context of transcriptional regulation, in which we incorporated a specific hypothesized interaction mechanism.

## 5.2 Results

### 5.2.1 Overall Pipeline to detect interactions between SNPs and age-associated TF

The overall pipeline is illustrated in Fig. 5-1 and described in Methods. We obtained the imputed SNPs and the corresponding transcriptome from ~570 individuals across 54 tissues from the GTEx consortium (GTEx Consortium, 2015).

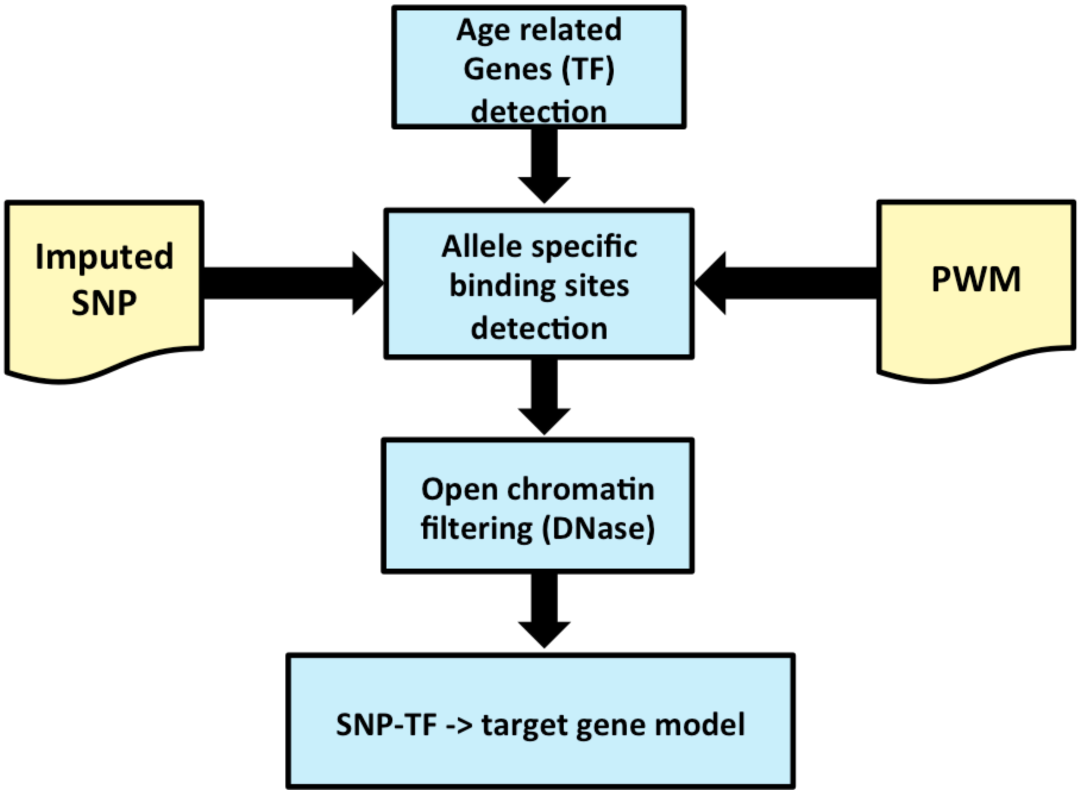

Age related Genes (TF) detection
Imputed SNP
Allele specific binding sites detection
PWM
Open chromatin filtering (DNase)
SNP-TF -> target gene model

**Figure 5-1: The overall Pipeline for detection of SNP-TF interactions.**

As a specific mechanism of SNP-Age interaction, we hypothesized that Age-associated TFs exhibiting allele-specific binding will lead to Age-associated expression of the target gene in an allele-specific manner (Fig. 5-2). We therefore retained only the SNPs within open chromatin regions (using tissue-specific DNase hypersensitivity (DHS) (Roadmap Epigenomics Consortium et al., 2015)) within 250Kb of genes (standard threshold for cis-eSNPs (Monlong et al., 2014; Rantalainen et al., 2015; Zhao et al., 2013)).

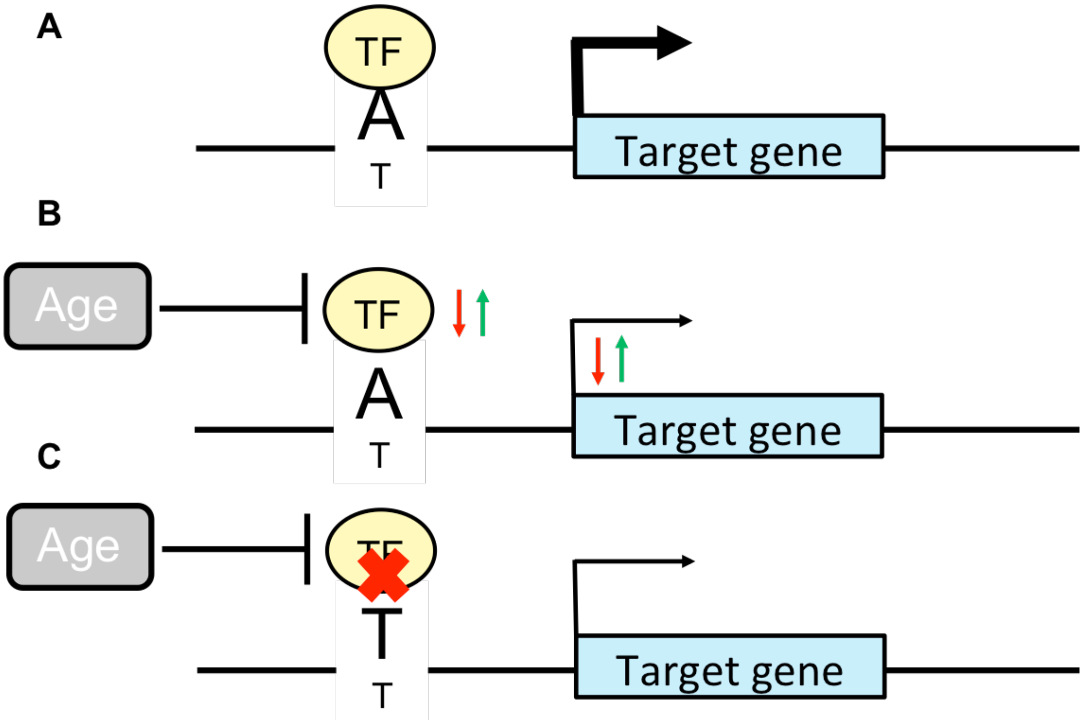
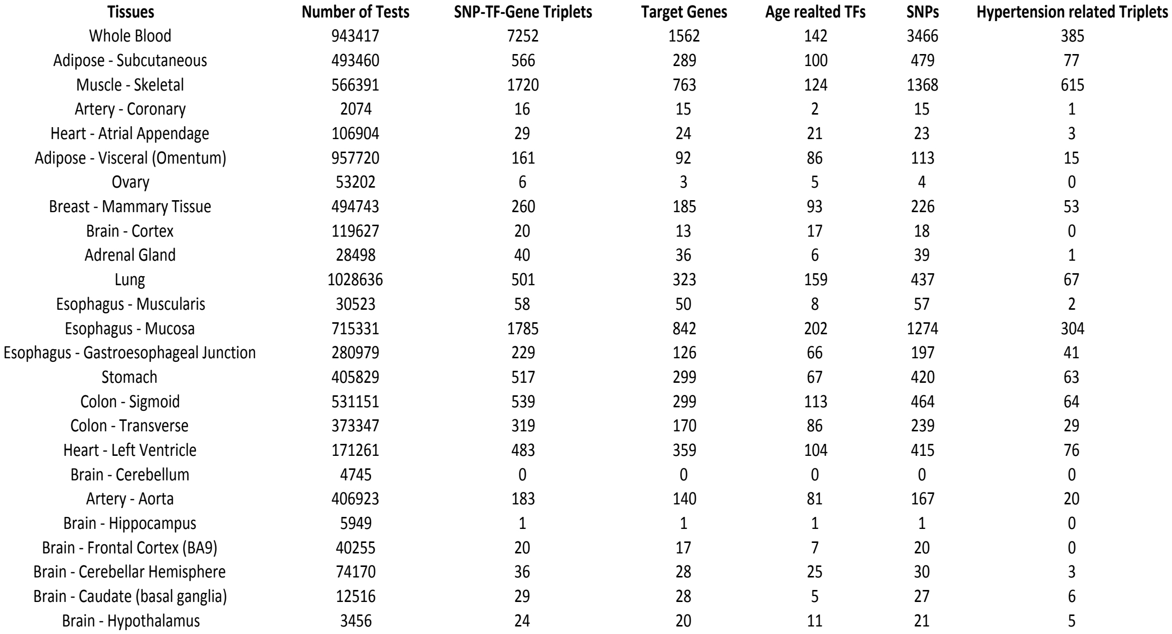


**Figure 5-2: The proposed mechanism for SNP-Age interactions.** (A): TF exhibits allelic binding preference (biased on binding to allele A here); (B) The concentration of the same TF could be affected by aging process, which induces the interaction between SNP and Age through that TF to regulate target genes; (C) The same TF could bind when individuals do not have allele A here and the interaction does not take place although aging still affects the same TF.

Our analysis is limited to 25 tissues for which DHS profiles are available. We detected Age-associated TFs based on their gene expression, controlling for genetic background and hidden variables. For Age-associated TFs, based on published DNA-binding motifs, we identified their putative allele-specific binding at all retained SNPs. For each SNP the genes within 250Kb are considered its potential targets. Finally, based on a linear model, we identified significant SNP-TF interactions associated with the target gene's expression.



| Tissues | Number of Tests | SNP-TF-Gene Triplets | Target Genes | Age realted TFs | SNPs | Hypertension related Triplets |
|---|---|---|---|---|---|---|
| Whole Blood | 943417 | 7252 | 1562 | 142 | 3466 | 385 |
| Adipose - Subcutaneous | 493460 | 566 | 289 | 100 | 479 | 77 |
| Muscle - Skeletal | 566391 | 1720 | 763 | 124 | 1368 | 615 |
| Artery - Coronary | 2074 | 16 | 15 | 2 | 15 | 1 |
| Heart - Atrial Appendage | 106904 | 29 | 24 | 21 | 23 | 3 |
| Adipose - Visceral (Omentum) | 957720 | 161 | 92 | 86 | 113 | 15 |
| Ovary | 53202 | 6 | 3 | 5 | 4 | 0 |
| Breast - Mammary Tissue | 494743 | 260 | 185 | 93 | 226 | 53 |
| Brain - Cortex | 119627 | 20 | 13 | 17 | 18 | 0 |
| Adrenal Gland | 28498 | 40 | 36 | 6 | 39 | 1 |
| Lung | 1028636 | 501 | 323 | 159 | 437 | 67 |
| Esophagus - Muscularis | 30523 | 58 | 50 | 8 | 57 | 2 |
| Esophagus - Mucosa | 715331 | 1785 | 842 | 202 | 1274 | 304 |
| Esophagus - Gastroesophageal Junction | 280979 | 229 | 126 | 66 | 197 | 41 |
| Stomach | 405829 | 517 | 299 | 67 | 420 | 63 |
| Colon - Sigmoid | 531151 | 539 | 299 | 113 | 464 | 64 |
| Colon - Transverse | 373347 | 319 | 170 | 86 | 239 | 29 |
| Heart - Left Ventricle | 171261 | 483 | 359 | 104 | 415 | 76 |
| Brain - Cerebellum | 4745 | 0 | 0 | 0 | 0 | 0 |
| Artery - Aorta | 406923 | 183 | 140 | 81 | 167 | 20 |
| Brain - Hippocampus | 5949 | 1 | 1 | 1 | 1 | 0 |
| Brain - Frontal Cortex (BA9) | 40255 | 20 | 17 | 7 | 20 | 0 |
| Brain - Cerebellar Hemisphere | 74170 | 36 | 28 | 25 | 30 | 3 |
| Brain - Caudate (basal ganglia) | 12516 | 29 | 28 | 5 | 27 | 6 |
| Brain - Hypothalamus | 3456 | 24 | 20 | 11 | 21 | 5 |

**Table 1: The number of detected interactions across 25 tissues.**

### 5.2.2 Numerous SNP-Age-associated TF interactions are detected across 25 tissues

The numbers of significant SNP-TF-Target genes triplets across 25 tissues are summarized in table 5-1. We tested a total of 7851107 SNP-TF-Gene triplets involving 447 TFs, 202978 SNPs, and 15315 Genes across the 25 tissues, and identified ~591 SNP-TF-Gene triplets on average in each of the 25 tissues. As a technical control, when we randomly shuffle gene expression across samples (while preserving gene-gene covariance among the target genes) the signals largely disappear. Interestingly, we detected the largest number of interactions in whole blood. We found that on average ~37% of the SNPs involved in an interaction were previously detected as eSNPs (GTEx Consortium, 2015), which is ~1.8 fold (Fisher test p-value<2.2e-16) greater than that for a control based on expression randomization, as above. What's more, on average ~44% of the target genes involved in an interactions are themselves significant associated with age (enriched over the background control; Fisher test p-value = 4e-3), consistent with a regulatory role of Age-associated TFs in driving Age-associated expression of the target gene.

A specific example of a detected SNP-TF-Gene interaction is illustrated in Fig. 5-3A, suggesting that TF GATA1 binds at SNP rs4857406 to regulate the expression of CPOX in whole Blood tissue. GATA1 exhibits a significant Age-associated expression (Fig. 5-4B). SNP rs4857406 was not captured by the GTEx eQTL study,

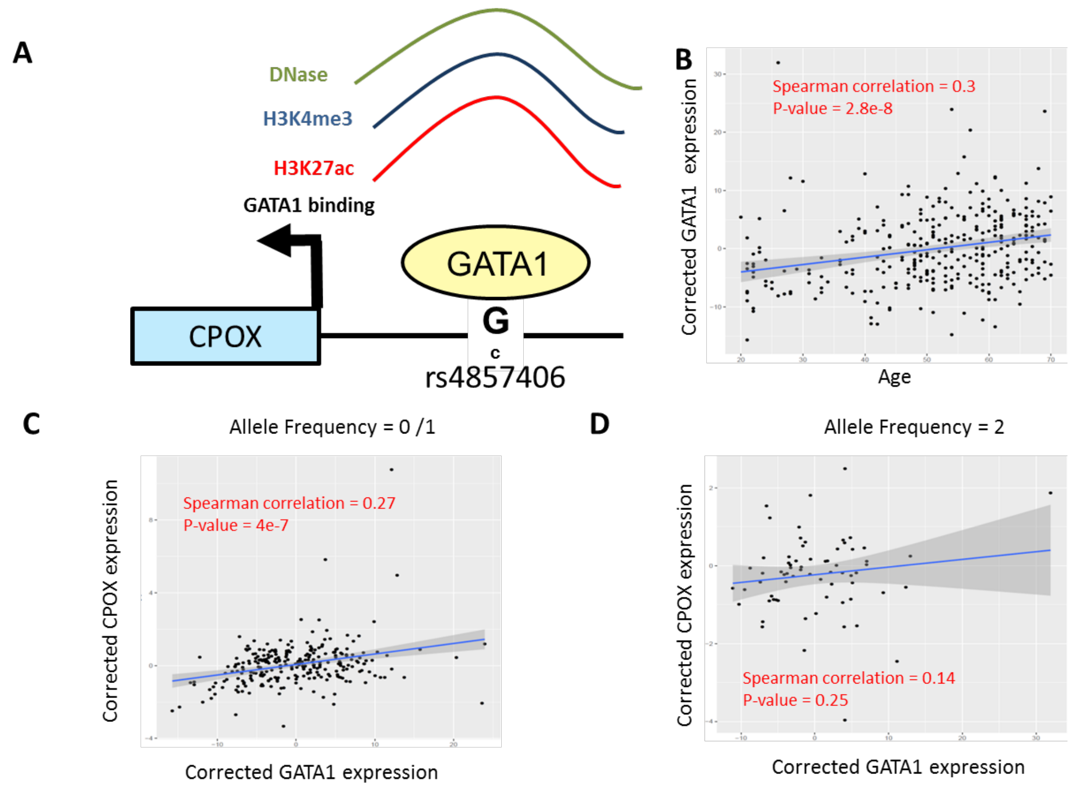


**Figure 5-3: One example for SNP-TF-Gene interaction triplet.** (A) GATA1 bind to the major allele (rs4857406) to regulate gene CPOX. The SNP region is enriched for DNase, H3K4me3, H3K27ac and GATA1 CHIP-seq binding signals; (B) GATA1 expression is correlated with age; (C) GATA1 expression is significantly correlated with CPOX expression among the population with at least one major allele (spearman correlation = 0.27 and p-value = 4e-7); (D) GATA1 expression is not correlated with CPOX expression among the population with only minor allele (spearman correlation = 0.14 and p-value = 0.25).

moreover its functional role is supported by DNase HS, H3K4me3, and H3K27ac peaks in K562 cell line (Epigenetic regulation of RNA processing : Nature ENCODE : Nature Publishing Group). In addition, the GATA1 ChIP-Seq peak also appears in the same region in K562 and PBDE cell lines, further supporting rs4857406's functional role. Based on GATA1 DNA-binding motifs, it is expected to differentially bind to the major allele at rs4857406. Accordingly, we observed genotype-specific correlation between GATA1 and CPOX; as shown in Fig. 5-3C-D, CPOX expression is significantly positively correlated (Spearman correlation = 0.27; p-value = 4e-7) with GATA1 expression when the individuals have at least one major allele, while for minor allele homozygotes, the correlation is not significant (p-value = 0.25).

### 5.2.3 Detected interactions are replicated in cross validation

To assess the robustness of the detected interactions, we estimated their replication rate following a standard cross-validation approach. We randomly divided two tissue samples (Whole Blood and Muscle –skeleton) into two equal subsets as discovery and validation datasets. FDR threshold of 0.1 was used for discovery, and p-value threshold of 0.05 was used as the validation threshold (Methods). Fig. 5-4 A-B show the replication rate for Top N (N = 20, 40, 60, 100) percent of detected interactions for standard cross validation respectively for whole blood and skeletal muscle tissues. The replication rate in Whole blood is ~85% (~797 interaction triplets) for top 20% of the discovered interactions and ~47% (~2206 interactions) for all discovered 4693 interactions. Muscle–skeleton exhibit relatively lower replication rate than whole blood tissue, the replication rates are still reasonable and significantly higher than the background (replication rate for random selected interactions). In addition, as expected both tissues exhibit a significant increasing trend with top selected interactions. These results support robustness of our detected interaction triplets.

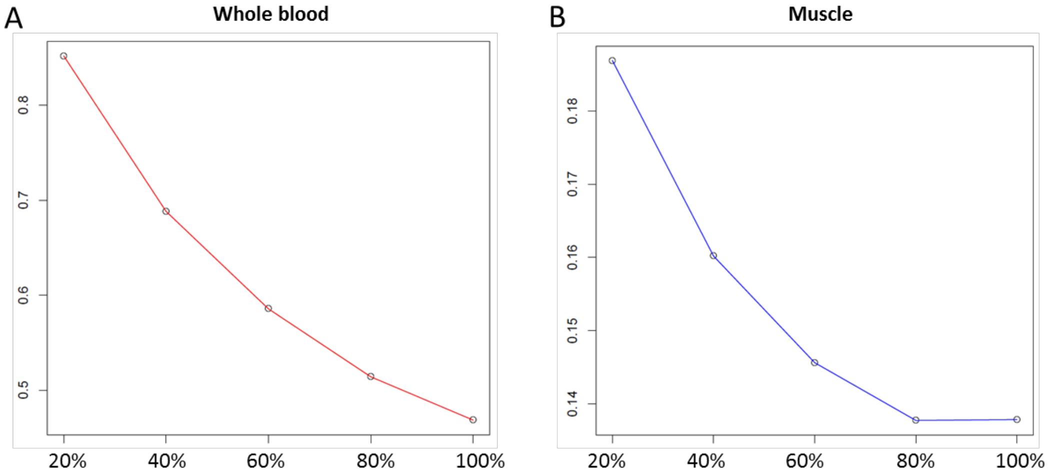


**Figure 5-4: Replication rate for interaction triplets in cross validation for top $\alpha$% ( $\alpha$ = 20, 40, 60, 80, 100) .** (A) Replication rates in cross validation for whole blood tissue. (B) Replication rates in cross validation for muscle tissue.

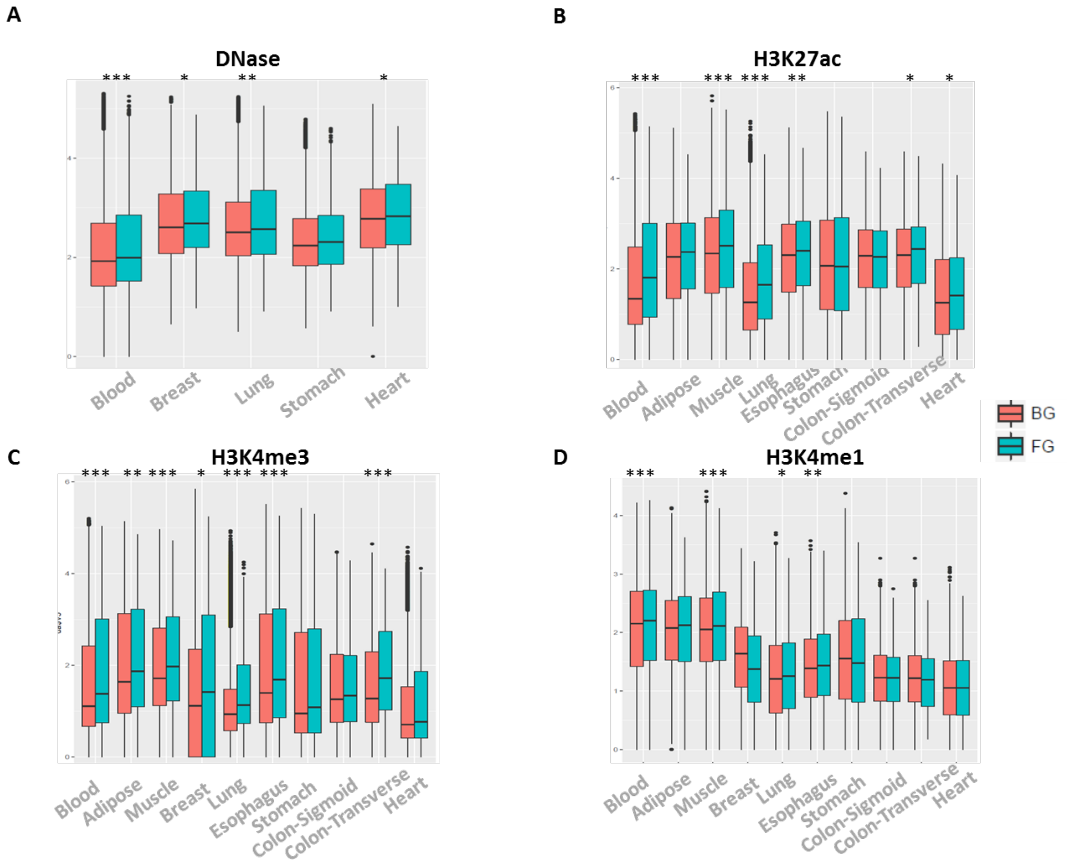


**Figure 5-5: Enrichment analysis for transcription related epigenetic signals.** Blue is for the foreground and red is for the background. (A): Enrichment for DNase signals across Blood, Breast, Lung, Stomach and Heart tissues; (B) Enrichment for H3K27ac across Blood, Adipose, Muscle, Lung, Esophagus, Stomach, Colon-Sigmoid, Colon-Transverse and Heart tissues; (C) Enrichment for H3K4me3 for Blood, Adipose, Muscle, Breast, Lung, Esophagus, Stomach, Colon-Sigmoid, Colon-Transverse and Heart tissues; (D) Enrichment for H3K4me1 across Blood, Adipose, Muscle, Breast, Lung, Esophagus, Stomach, Colon-Sigmoid, Colon-Transverse and Heart tissues.

### 5.2.4 SNPs mediating the detected interactions are likely to be functional

In view of our hypothesis that Age-associated TFs bind to SNP locus, the detected SNP loci are expected to be in potential cis-regulatory elements, such as enhancers and promoters. We therefore checked whether the detected SNP loci are enriched for open chromatin signals (DHS) and known active regulatory region related histone modifications (H3K27ac, H3K4me1 and H3K4me3). While H3K27ac and H3K4me1 are markers for active enhancers (Creyghton et al., 2010; Gibney and Nolan, 2010; Luco et al., 2010), H3K4me1 could be related to poised enhancers. H3K4me3 is also associated with gene promoters. Even though DHS peaks were used as an inclusion criterion for the SNPs, we tested whether the detected SNPs exhibit higher DHS intensity compared to the background SNPs, which also qualified the initial DHS filter. As the foreground for this analysis, we used 200 bps regions around the SNPs involved in a significant interaction, and as the background control we used 200 bps regions around all the SNPs that were tested but failed were not part of any detected interaction. Each of the 4 epigenomic marks was analyzed for the tissues in which the relevant data was available and at least 200 foreground SNP loci. As shown in Fig. 5-5A, the foreground DHS intensities are significant higher than the background in 4 tissues of 5. This was also broadly true for H3K27ac and H3K4me3 (Fig. 5-5B-C). However, we observed a significant difference between the foreground and the background SNPs for H3K4me1 in four of the ten tissues analyzed in Fig. 5-5D. Overall, these results strongly suggest that the SNPs involved in the detected interactions are likely to play a regulatory role.

### 5.2.5 Functional analysis of target genes suggests the link of detected interactions to age related processes and complex diseases

The genes involved in our detected interactions are expected to exhibit age-associated expression in an allele-specific manner, which may have implication on aging and age-associated complex diseases. To explore this further, first we performed functional enrichment analysis over target genes in a tissue-specific manner using GOstats Package in R (Falcon and Gentleman, 2007). Then a comprehensive TreeMap view of enriched GO terms across tissues (FDR <= 0.05; Methods) are generated in supplementary Fig. 5-S1. Among them, defense response and immune response, are well known to be linked to aging (Licastro et al., 2005) and several age-associated complex diseases including hypertension, diabetes, neurodegeneration, and even cancer (Blasco et al., 2013; Jin, 2010; Sinclair et al., 2012). Metabolic process is crucial for the maintenance of biological system homeostasis; especially for aging process which has to confront numerous

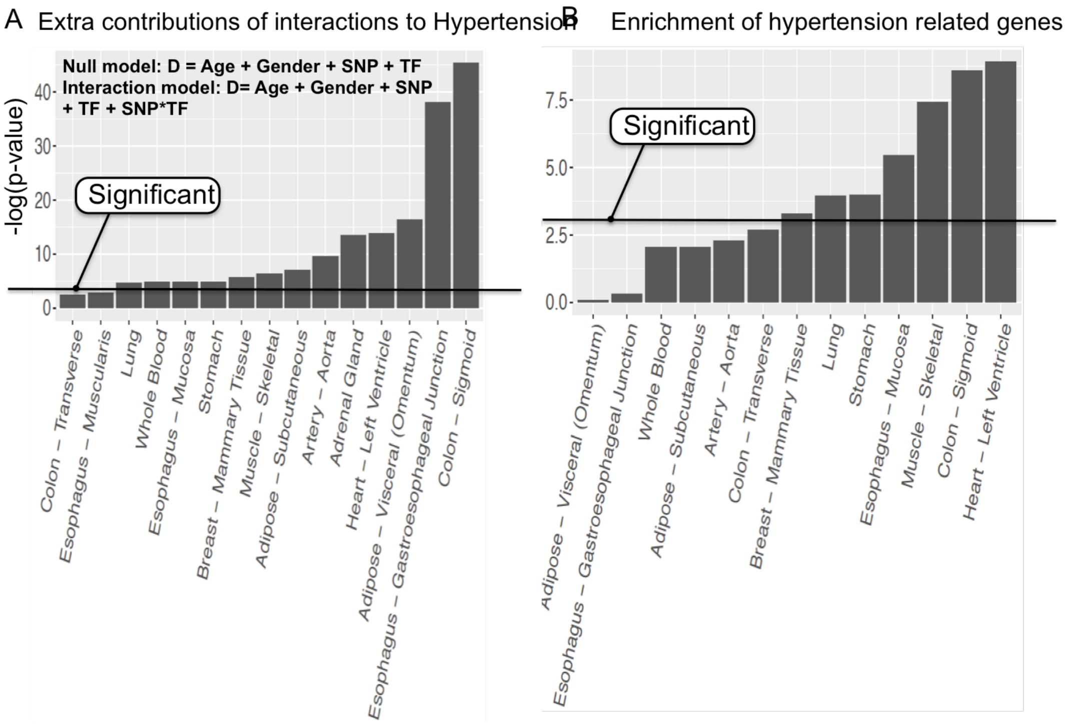


**Figure 5-6: Significant contributions of detected interactions to hypertension.** (A) Interactions are contributing to hypertension in 13 tissues showed by log likelihood ratio test; (B) in 7 tissues, target genes are enriched for hypertension related genes showed by Fisher's test.

harmful interventions (Barzilai et al., 2012). Actually extra cellular matrix component is related associated with aging process suggested by previous study (Ly et al., 2000) and it could contribute to stem cell maintenance and function (Kurtz and Oh, 2012). In addition, it is known that cell cycle regulation which drives reproduction is crucial for survival, and stressful stimulus to cell cycle could directly lead to cell senescence (Johnson et al., 1999; Molecular Biology of the Cell, 5th Edition: The Problems Book / Edition 5 by John Wilson | 9780815341109 | Paperback | Barnes & Noble).

Next, using Hypertension as an exemplar age-associated complex disease, we assessed the association of each detected SNP-TF-Gene triplet with Hypertension by detecting hypertension related target genes in the same model (Method). The numbers of hypertension-associated interactions across 25 tissues are summarized in Table 5-1. What's more, we performed enrichment test on our detected hypertension related interaction targets genes by comparing with previous reported hypertension related genes (Rouillard et al., 2016). More specifically, we performed fisher's test to compare our detected foreground and the background (target genes which failed in the significance test) in both tissue-specific and comprehensive manner. Interestingly we showed hypertension genes are enriched in 7 out of 13 tissues as Figure 5-6B shows. And also comprehensively across all tissues our detected targets are enriched for hypertension related genes with fold change = ~1.8 (fisher's test p-value = 3.9e-18). Moreover by performing log likelihood ratio test (interactions are included as the extra covariate in the alternative model), we showed in 13 out of 15 tested tissues, detected interactions significantly contribute to hypertension (Figure 5-6A).

In addition, we obtained GWAS signals for 16 age related diseases and directly performed enrichment for our detected SNPs in both tissue-specific and comprehensive manner. The same as above, fisher's test is used for the enrichment test. We specifically found our detected SNPs in Stomach tissue are enriched for association with aging (p-value = 0.049 and fold change = 23). What's more, detected SNPs in both Skeletal Muscle and Heart tissues are significantly associated with cardiovascular diseases (p-values are 0.043 and 0.048 respectively; fold changes are 2.4 and 3.9 respectively). Detected SNPs in whole blood tissue are related to immune response.  In addition, our comprehensive detected SNPs across 25 tissues are enriched for ageing, cardiovascular diseases and immune response associated SNPs.

Overall, these results suggest that some of the detected SNP-Age-associated TF interactions could affect aging linked processes and age-associated complex diseases (including hypertension) by mediating the target gene expression.

## 5.3 Materials and Methods

### 5.3.1 Transcriptome and Epigenomic data

We obtained the processed transcriptome data across 25 tissues from Genotype-Tissue Expression (GTEx) database version 6 (GTEx Consortium, 2015). Corresponding tissue-specific DNase-seq, broad peaks data for H3K27ac, H3K4me1 and H3K4me3 are downloaded from Roadmap consortium (Roadmap Epigenomics Consortium et al., 2015). In addition, top three principal components generated over SNP profile and PEER factors generated over gene expression profile are also obtained from GTEx database. GENCODE genome annotation version 19 (hg19) (Harrow et al., 2012) is used consistently in this study.

### 5.3.2 Age-associated TF detection

We derived the linear regression model from previous studies (Glass et al., 2013; Yang et al., 2015) to detect significant age-associated genes as follows:

$$g_{ij} = \alpha_i + \beta_i^1 AGE_j + \beta_i^2 GENDER_j +$$

$$\sum\nolimits_{l=1}^{3} \beta_i^{l+2} S_j^l + \sum\nolimits_{k=1}^{n} \beta_i^{k+5} \mathrm{PEER}\left(\mathrm{CF}_\mathrm{j}^\mathrm{k}\right) + \varepsilon_{ij} \; (1)$$

Where $g_{ij}$ is the expression of target gene i for jth sample and $\alpha_i$ is the basal expression and the intercept on y axis for gene i. $\beta_i$ is the coefficient for covariates in the ith gene model. $AGE_j$ and $GENDER_j$ are the age and gender for jth sample respectively. $S_j$ is the covariate derived from SNPs for jth sample. $\mathrm{PEER}\left(\mathrm{CF}_\mathrm{j}^\mathrm{k}\right)$ is the kth PEER factor (hidden variable) derived over gene expression profile for jth sample. $\varepsilon_{ij}$ denotes the error term.

We removed PEER factors, which significantly correlate with age (Pearson correlation coefficient < 0.05). Then we checked whether the coefficient of age is significantly deviated from zero (FDR <= 0.1) and selected the significant age-associated TF genes.

### 5.3.3 Putative allele specific binding site prediction

We generated allele specific sequence (~100 bps) around all the SNP loci and scan the sequences for age-associated TFs binding sites using PWM-scan, which is an approach for putative TF binding sites prediction based on positional weighted matrix (PWM). Tissue-specific DNase-seq broad peaks data is applied to remove false positive prediction regarding of chromatin accessibility. Allele specific binding events are selected as interaction candidates (SNP-TF).

### 5.3.4 Interaction detection

We modeled the association between the interaction (SNP-TF) and potential target gene expression (within 250kb bps) as follows:

$$g_{ij} = \alpha_i + \beta_i^1 AGE_j + \beta_i^2 GENDER_j + \beta_i^3 SNP_j + \beta_i^4 \widetilde{TF}_j + \beta_i^5 D_j + \beta_i^6 SNP_j * \widetilde{TF}_j + \beta_i^7 D_j * \widetilde{TF}_j + \beta_i^8 SNP_j * D_j + \beta_i^9 SNP_j * D_j * \widetilde{TF}_j + \sum_{l=1}^{3} \beta_i^{l+9} S_j^l + \sum_{k=1}^{n} \beta_i^{k+13} \mathrm{PEER}\left(\mathrm{CF}_\mathrm{j}^\mathrm{k}\right) + \varepsilon_{ij} \; (2)$$

Where $g_{ij}$ is the expression of gene i for ith sample, $\alpha_i$ is the basal expression and the intercept on y axis for gene i. $\beta_i$ is the coefficient for covariates in the ith gene model. $AGE_j$ and $GENDER_j$ are the age and gender respectively for jth sample. $SNP_j$is allele frequency for ith sample. $\widetilde{TF}_j$ is adjusted concentration of TF which only includes age component (residuals after controlled for gender, covariates derived from SNP profile and PEER factors that are not correlated with age) for jth sample. $D_j$ is the hypertension disease state for jth sample (0/1). All the four types of interactions among $SNP_j$, $\widetilde{TF}_j$ and $D_j$ are respectively denoted as

$SNP_j * \widetilde{TF}_j, D_j * \widetilde{TF}_j, SNP_j * D_j$ and $SNP_j * D_j * \widetilde{TF}_j$. $S_j$ is the covariate derived from SNPs for jth sample and $\mathrm{PEER}(\mathrm{CF}_j^k)$ is the kth PEER factor (hidden variable) derived over gene expression profile for jth sample. $\varepsilon_{ij}$ denotes the error term.

Whether $\beta_i^6$ is significantly deviated from zero (FDR <= 0.1) implies potential interaction mechanism for general gene regulation, moreover coefficients $\beta_i^5$, $\beta_i^7$, $\beta_i^8$ and $\beta_i^9$ together could suggest the ith target gene is related to hypertension.

### 5.3.5 Functional analysis

We performed functional analysis on target genes involved in significant interaction across 25 tissues using GOstats (Falcon and Gentleman, 2007). The frequency of each GO term is observed across 25 tissues is assigned as the weight for GO terms. Then a TreeMap view is generated using Revigo package (Supek et al., 2011). For the background control, we picked up random target genes, which also fall within 250kb from the significant SNPs, to repeat the same functional analysis.

### 5.3.6 Enrichment test for Epigenomic signals

We extracted 100 bps windows centering around significant SNP loci and map raw DNase-seq, ChIP-seq signals for H3K27ac, H3K4me1 and H3K4me3 to those window regions using “bedtools”. The average score across all the base pairs within the window is referred as the signal score for each significant SNP region. For the background control, the same procedure is repeated over tested nonsignificant SNPs. One tailed Wilcoxon test is performed to compare the two distributions (foreground and background).

### 5.3.7 Estimation of the replication rate

To estimate the replication rate, we detect age-associated TF in both training and testing datasets respectively first, then extract overlapped interaction triplet candidates following the pipeline. We identified significant interactions in training dataset using a stringent threshold (FDR <= 0.1), following which we validate whether those reported significant interactions still retain in testing dataset with a relaxed threshold (p-value <= 0.05). Moreover, for visualization we measure the replication rate for top $\alpha$% ($\alpha$ = 20, 40, 60, 80, 100) of detected significant interaction triplets respectively. For cross validation approach, we equally divided blood tissue as training and testing dataset randomly.

## 5.4 Summary and Discussion

In summary, we reported a novel pipeline and framework, which incorporates a specific hypothesized interaction mechanism, to test the association between SNP-age interactions and phenotypes (gene expression/diseases). We hypothesized that age-related TF might preferentially bind to one of the alleles, which actually forms the basis for the interplay between SNPs and age. Numerous SNP-TF-Gene interaction triplets (average ~591) are detected across 25 tissues, some of which are also associated with hypertension. The detected significant SNP regions are enriched for epigenetic signals related to regulatory elements, which further validates their functionality. In addition, we showed the robustness by estimating the replication rates of detected interactions in cross validation. Enrichment tests for both biological terms and previous reported diseases associated genes also suggest that our detected interactions are linked to aging process and age related diseases.

As we have mentioned, although multiple previous studies incorporating intuitive SNP-Age terms reported some potential interactions, both the detection power and insights about explicit interaction mechanisms are extremely limited, which could be addressed by our approach. The approximation of age factor by age-associated TF changes does not only increase the detection power but also provide evidences for a specific potential mechanism of the interactions. What's more, tissue-specific studies are missed previously; relatively in our study analysis across 25 tissues are carried out, which will definitely provide more insight into tissue-specific gene regulation and complex diseases.

As an "environmental" risk factor, aging has to play a role by interacting with gene regulation, which could be complex and diverse. Age-associated TF may be just one specific type of aging agent, and some other factors could be also involved. For instance, it could be also age-associated splicing factor or epigenetic markers, both of which play essential roles in gene regulation and potentially could be linked to complex diseases. We explicitly observed age-associated splicing factor expression changes across aging in our study. Age-associated splicing factor need to bind to RNA motif regions, which may contain polymorphisms, which thus forms the basis for interaction between SNPs and aging. As for DNA methylation, Horvath et al.

have shown methylation to be a robust age biomarker (Horvath, 2013) and moreover it is well known that methylation usually blocks TF binding (Maurano et al., 2015; Moore et al., 2012; Yin et al., 2017). Thus, it is reasonable to expect that age-associated methylation changes could affect TF binding landscape across aging, in which age factor would interact with allele specific binding sites to regulate gene expression. It is especially exciting since DNA methylation has also been identified as a robust cancer biomarker. Combination of genome variations, aging factor, methylation and gene regulation might be able provide novel insights into cancer.

Tissue-specificity complicates our understanding of gene regulation and complex diseases. We expect our analysis across 25 tissues could shed light on molecular basis underlying tissue-specific genetic regulation. In fact, we specifically observed more interactions in tissues like whole blood and adipose, which might imply they are affected more severely by aging process. What's more, tissue-specific genetic investigations might be able to suggest tissues causality involved in complex diseases. In our studies, multiple intuitively known hypertension related tissues have relatively more interactions detected associated with hypertension disease.

In addition, our study might promote the annotation and interpretation of GWAS and eQTL signals. In recent years, large scales of GWAS and eQTL signals are reported; however, it is still difficult to decode the observed associated into explicit pathways and mechanisms. Alignment of our detected interaction with reported associations by GWAS and eQTL could suggest potential explicit links between genetics and phenotypes.

# Chapter 6:  Conclusion and future directions

### 6.1 Summary of the dissertation

Overall, my dissertation work focuses on gene regulation (transcription and alternative splicing) in normal aging and HGPS.

First, in order to understand the common and distinct pathways underlying HGPS and normal aging, we developed a novel regression based clustering approach to identify gene clusters whose expression co-vary with aging and HGPS. Our approach could specifically address problems of limited sample size, which might be advantageous especially when used for rare disease studies. In this study, we also provided the first novel set of matched normal aging and HGPS RNA-seq data. We applied our approach to our RNA-seq profile and suggested numerous unique insightful biological processes underlying aging and HGPS.

Second, we performed an investigation on alternative splicing regulation, which is related to both aging and HGPS. Besides re-assessment the predictive ability of genomic features for alternative splicing, we also predicted alternative splicing using numerous epigenomic features based on our novel benchmark datasets using a deep neural network model (DNN). Most importantly, we provided the first comparative investigation of the predictability of genomic and epigenomic features for alternative splicing across three tissues. Consistently across three tissues we concluded that genomic features play the primary role in alternative splicing compared to epigenomic features and it seems that epigenomic features are not contributing independent regulatory information relative to genomic features. We also provided potential location specific interaction map between genomic and epigenomic features, which might imply context specific regulatory roles of epigenomic features.

Third, we performed investigation on age-associated splicing changes across numerous human primary tissues. Alternative splicing regulation could be linked to diverse phenotype and cross-tissue variability complicates these links. In this study, we provided the first systematic comprehensive map of age-associated splicing changes across multiple tissues. We detected 49869 tissue-specific age-associated splicing events of 7 types across 48 tissues in 544 individuals by analyzing ~8500 processed RNA-seq samples employing a stringent linear model controlling for multiple confounding factors. Functional analysis showed those age-associated splicing changes potentially are linked to aging process, in addition, we specifically showed they could contribute to age related complex diseases. In addition, we performed the first comparisons on predictability of age for splicing level, gene expression and transcription expression, which showed splicing information could be more correlated with age. We suggested that age-associated splicing factor concentration change could be the potential driver for age-associated splicing changes.

Lastly, we performed investigations on interactions between genome variations and normal aging, both of which have been linked to gene regulation and complex diseases. Instead of simply incorporating the SNP-age interaction term in the eQTL model, we hypothesized an interaction mechanism and proposed a novel pipeline to detect SNP-age interaction. More specifically, we approximate age factor by age-associated transcription factor concentration and expect interplay might happen while age-associated TFs specially bind to one allele instead of the other one. We detected ~591 SNP-TF-Gene interaction triplets in average across 25 tissues and we also showed some of them are associated with hypertension diseases. Our detected significant SNPs are more likely to be functional since those regions are enriched for epigenomic signals related to regulatory activities. In the meantime, high replication rates in cross validation and similar tissues suggest robustness for our detections. In this study, we provided a new framework for investigation on interactions between genome polymorphisms and environment factors.

In summary, my dissertation work specifically focus on fundamental gene regulation mechanism (transcription and alternative splicing) in aging and HGPS, by measuring the association between molecular levels (gene expression, splicing level and transcript expression) and phenotype (age, complex diseases), we provided more insight about the underpinning of aging and HGPS.

### 6.2 Future directions

As mentioned in the introduction, although the causal de novo mutation in HGPS has already been identified, the mechanism underlying the syndrome is still not clear. Importantly, the common symptoms in normal aging suggest HGPS as a reasonable aging model, which hypothesizes mutated protein--progerin would play an essential role in aging process. Whether progerin is one key driver of aging process and how much phenotypic variability it affects has been hotly debated.

Intriguingly, we observed significant progerin transcript expressed in non-HGPS individual in GTEx consortium. It would be extremely interesting to identify specific mechanism that generate progerin in normal individuals from GTEx data and assess the potential phenotypic effect of progerin in that population, which might provide novel insights into both HGPS and aging process.

In our second study, we predicted alternative splicing using k-mer motif and epigenomic features around splicing events; actually polymorphisms could also affect splicing regulation. Moreover splicing factor proteins play essential roles in splicing regulation and could explain large tissue-specific splicing variations (Black, 2003; Chen et al., 1999; Linares et al., 2015). It would be exciting to develop a mixture model to impute splicing level across tissues by appropriately combining all those information. It would enhance our understanding of the fundamental question about tissue-specific splicing; additionally it will provide imputed splicing data for related studies.

We have detected numerous age-associated splicing events across 48 tissues in our third study, however it is still difficult to explicitly show those age-associated splicing events are actually functional and playing important roles in aging and complex diseases. It would be interesting and helpful to develop an approach to estimate the likelihood that a specific splicing event is functional by combining all existed annotation datasets.

Age-associated TF probably is not the unique aging “agent”, splicing factor and methylation, could be another two factors involved in interaction mechanisms.

Splicing factors plays essential roles in splicing regulation and could be affected by aging and bind to splicing regulatory cis-elements (splicing enhancers, silencers, branch sites, splicing sites) containing polymorphisms in an allele-specific fashion, which forms the basis for interaction between SNPs and aging. As for methylation, it is well known that as a robust age biomarker, DNA methylation could block TF binding, thus age-associated DNA methylation changes could significantly affect the landscape of TF binding tissue-specifically. What's more, there are many evidence that methylation is an important biomarker for cancers. Colorectal cancer is significantly related to genetic background, ~29 related SNPs have been identified (Lubbe et al., 2012; Spring). A novel and exciting project could be to explore the potential interactions between SNPs and TF binding affected by age-associated methylation profile, which might provide insights for cancer from a new perspective.

# Supplementary Information

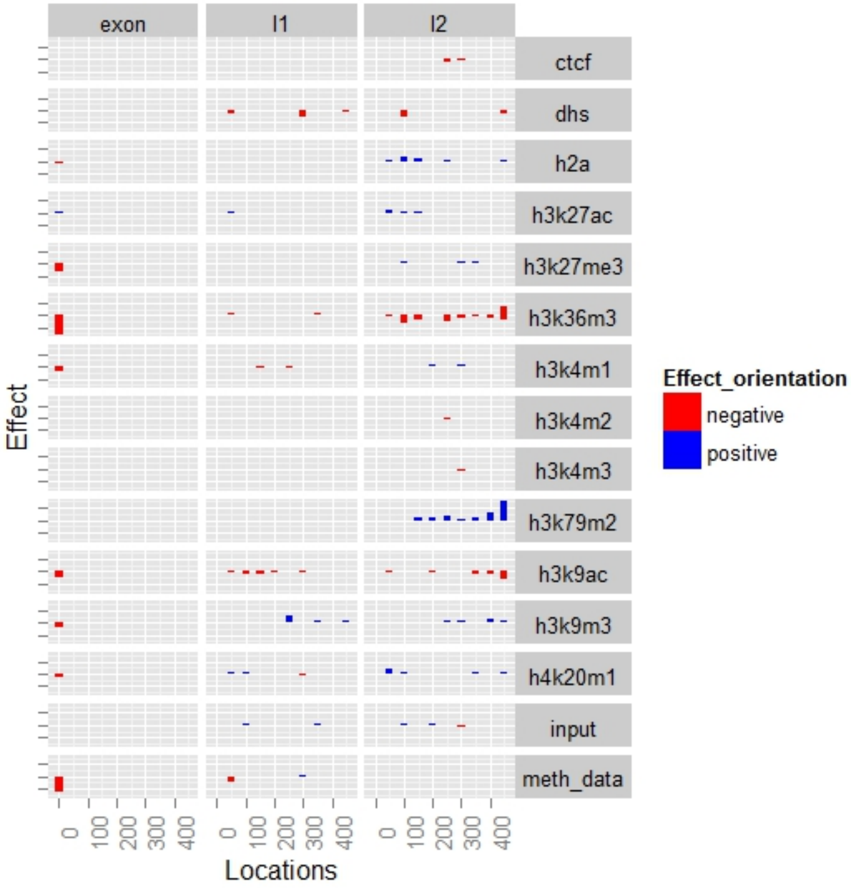


**Figure 3-S1: The chromatin feature map for tissue GM12878.**

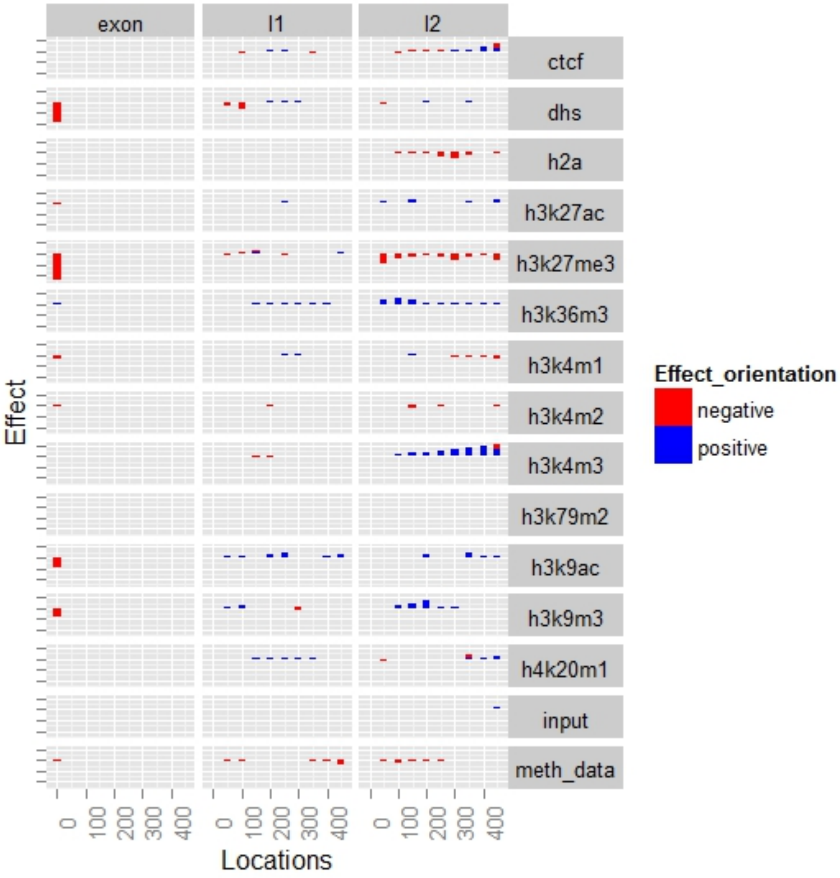


**Figure 3-S2: The chromatin feature map for tissue K562.**

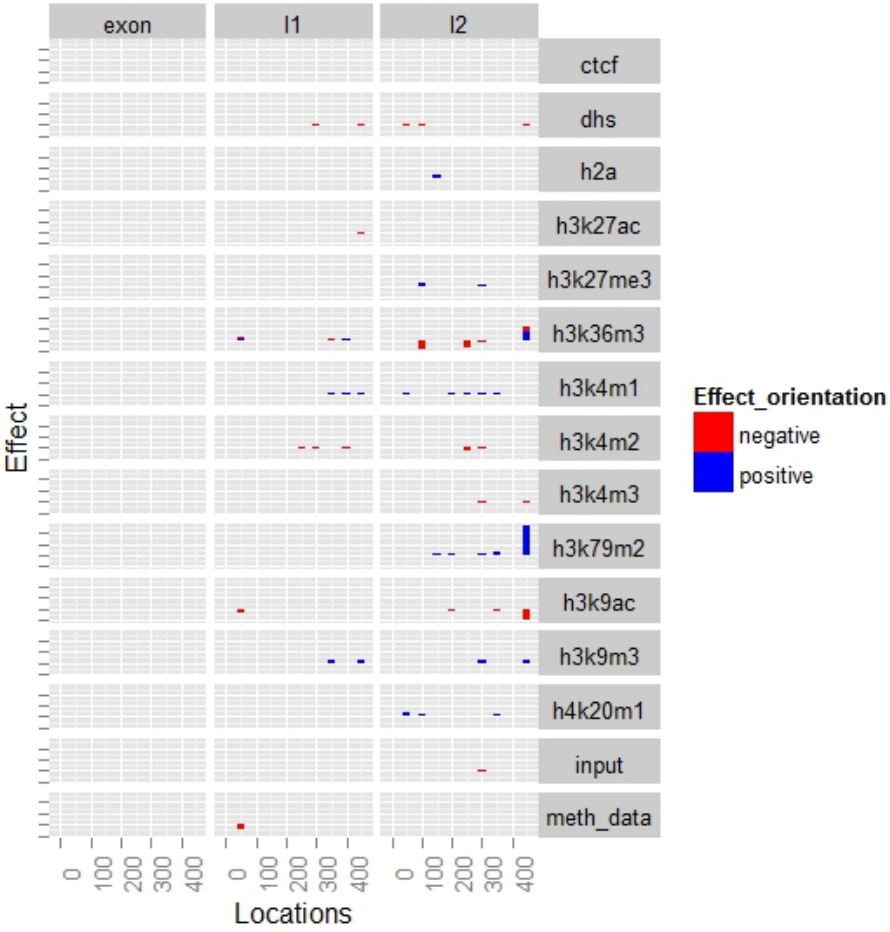


**Figure 3-S3: Adjusted feature enrichment map for GM12878.**

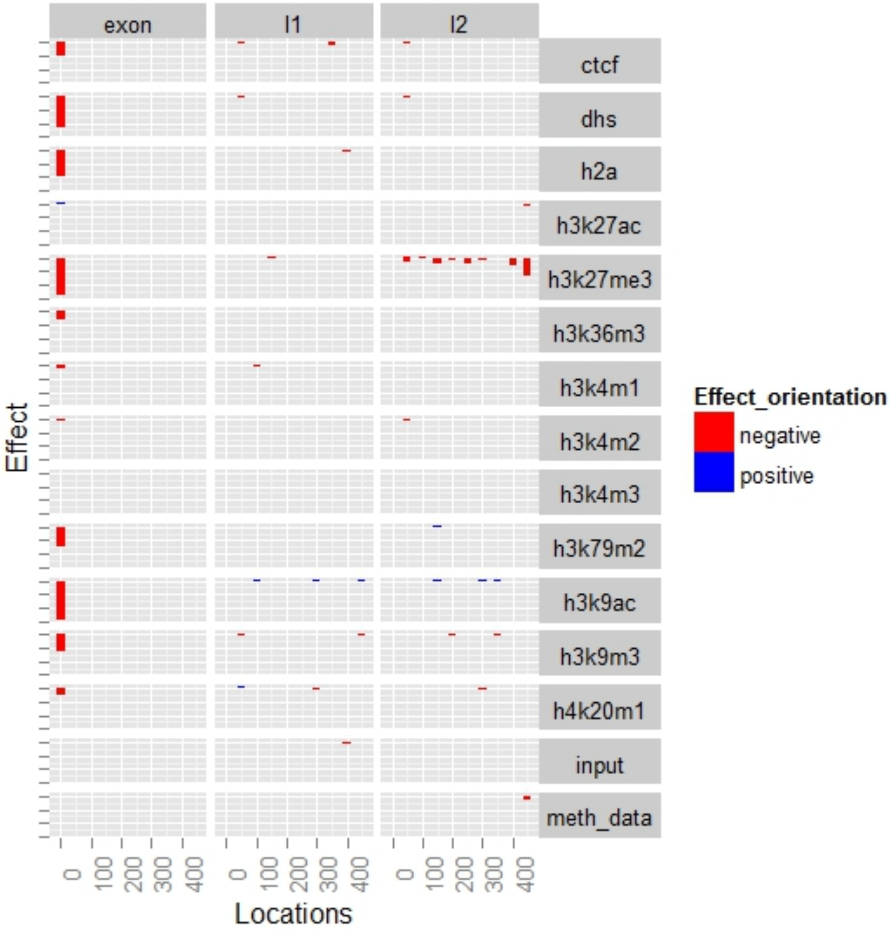


**Figure 3-S4: Adjusted feature enrichment map for h1-hESC.**

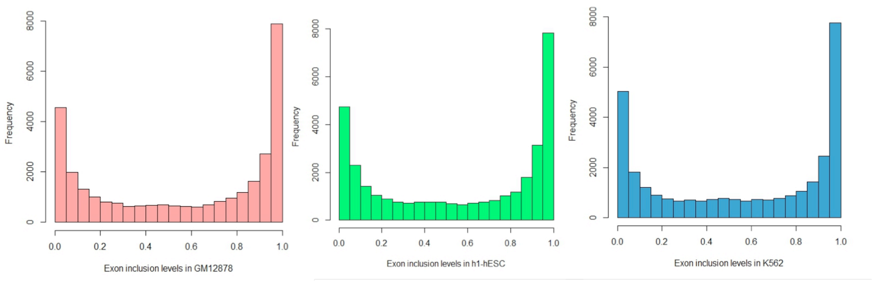


**Figure 3-S5: Exon inclusion level distributions for GM12878, h1-hESC and K562**

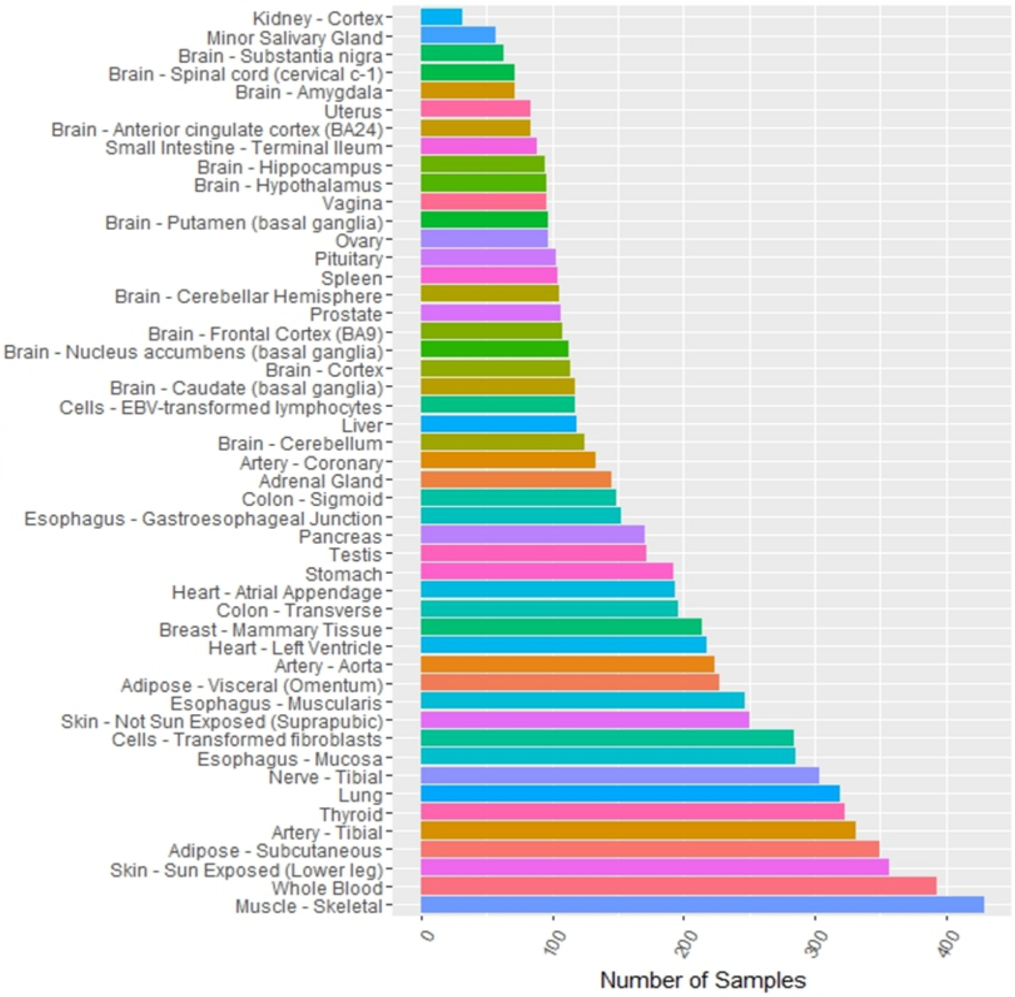


**Figure 4-S1A: Number of samples across 54 tissues.**

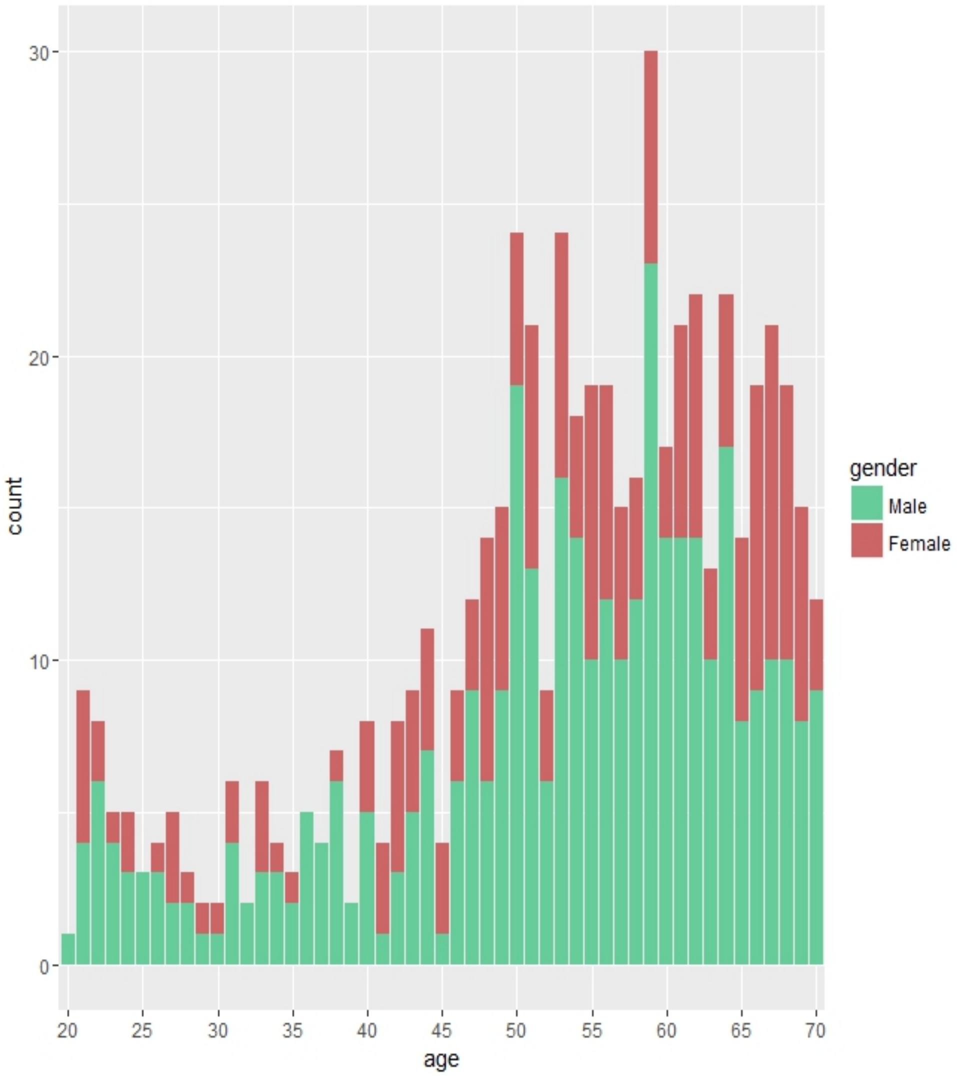


**Figure 4-S1B: Distribution of sample age and gender.**

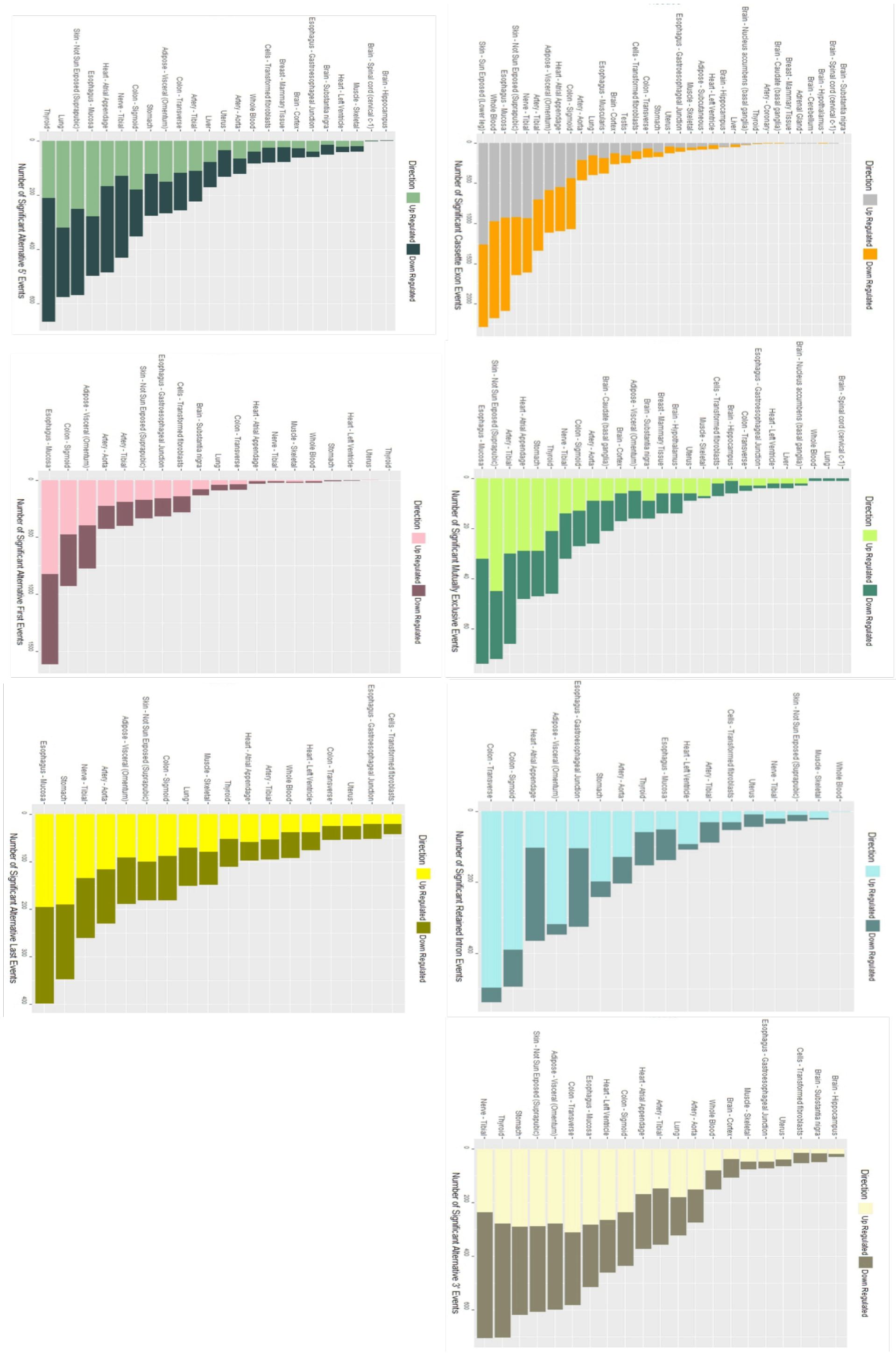
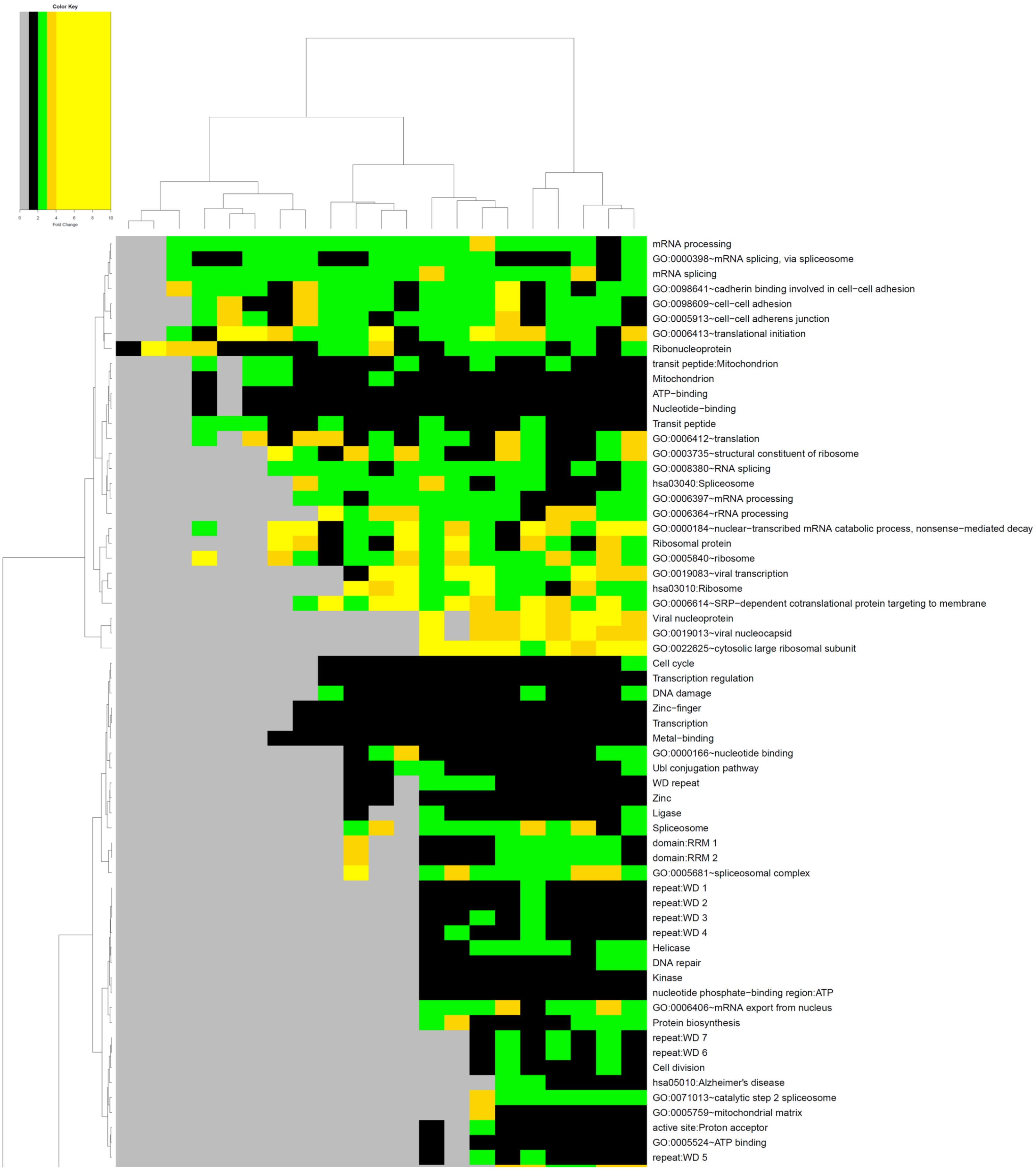


**Figure 4-S2: Number of significant up and down-regulated splicing events across tissues for the 7 types of splicing events.**


Color Key
0
2
4
6
8
10
Fold Change
mRNA processing
GO:0000398~mRNA splicing, via spliceosome
mRNA splicing
GO:0098641~cadherin binding involved in cell−cell adhesion
GO:0098609~cell−cell adhesion
GO:0005913~cell−cell adherens junction
GO:0006413~translational initiation
Ribonucleoprotein
transit peptide:Mitochondrion
Mitochondrion
ATP−binding
Nucleotide−binding
Transit peptide
GO:0006412~translation
GO:0003735~structural constituent of ribosome
GO:0008380~RNA splicing
hsa03040:Spliceosome
GO:0006397~mRNA processing
GO:0006364~rRNA processing
GO:0000184~nuclear−transcribed mRNA catabolic process, nonsense−mediated decay
Ribosomal protein
GO:0005840~ribosome
GO:0019083~viral transcription
hsa03010:Ribosome
GO:0006614~SRP−dependent cotranslational protein targeting to membrane
Viral nucleoprotein
GO:0019013~viral nucleocapsid
GO:0022625~cytosolic large ribosomal subunit
Cell cycle
Transcription regulation
DNA damage
Zinc−finger
Transcription
Metal−binding
GO:0000166~nucleotide binding
Ubl conjugation pathway
WD repeat
Zinc
Ligase
Spliceosome
domain:RRM 1
domain:RRM 2
GO:0005681~spliceosomal complex
repeat:WD 1
repeat:WD 2
repeat:WD 3
repeat:WD 4
Helicase
DNA repair
Kinase
nucleotide phosphate−binding region:ATP
GO:0006406~mRNA export from nucleus
Protein biosynthesis
repeat:WD 7
repeat:WD 6
Cell division
hsa05010:Alzheimer's disease
GO:0071013~catalytic step 2 spliceosome
GO:0005759~mitochondrial matrix
active site:Proton acceptor
GO:0005524~ATP binding
repeat:WD 5

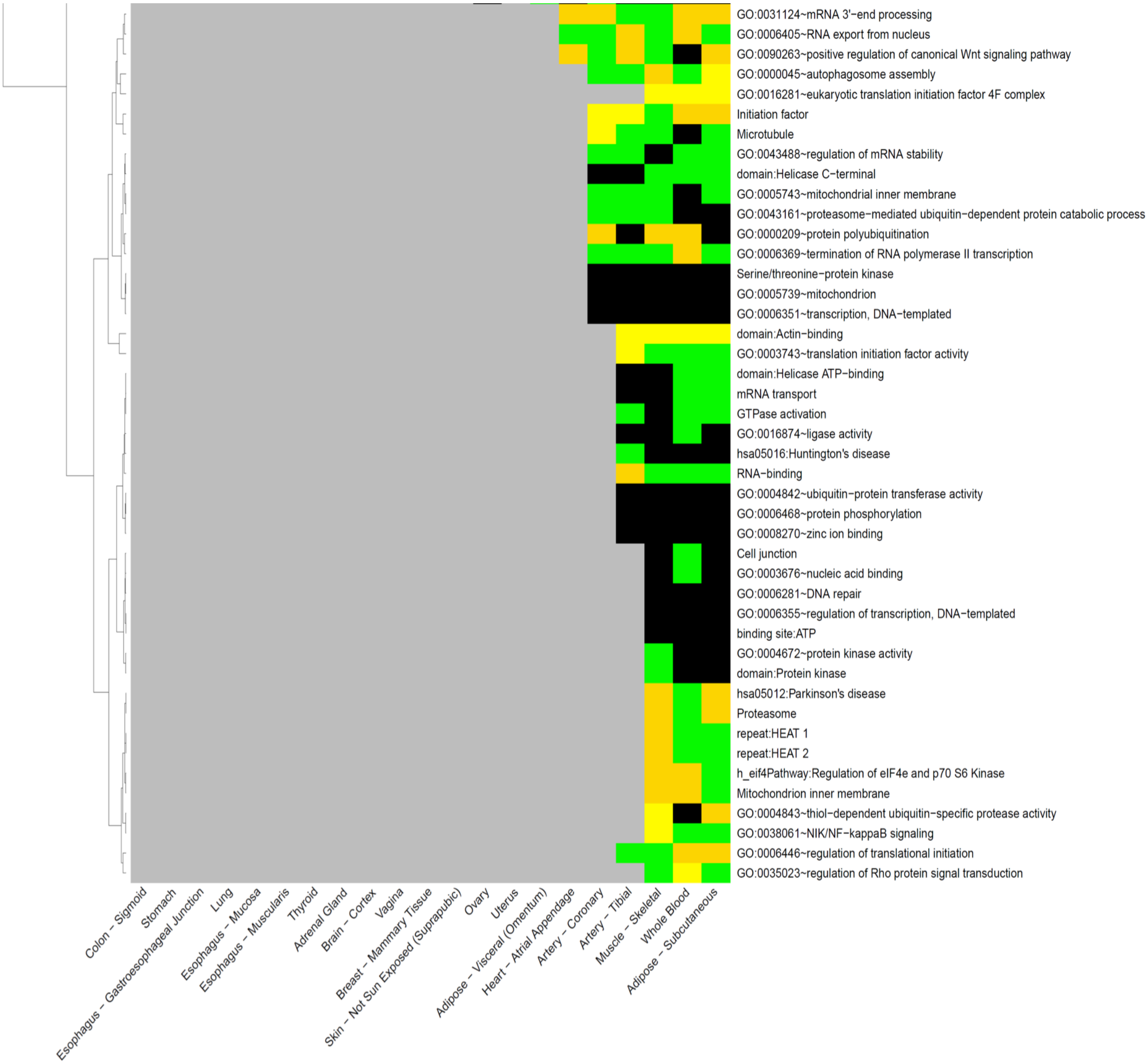
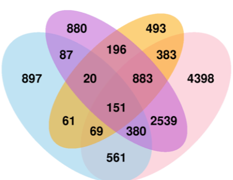
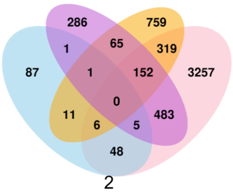


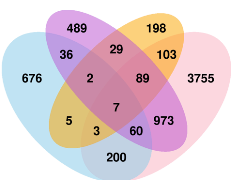
**Figure 4-S3: Functional enrichment among genes affected by age-associated splicing across tissues.** Columns correspond to tissues, rows to biological functions, colors indicate fold-change.



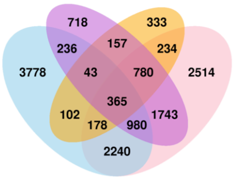
1



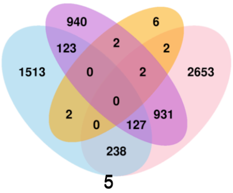
2



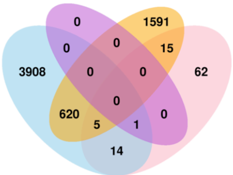
3



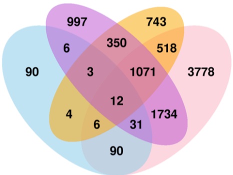
4



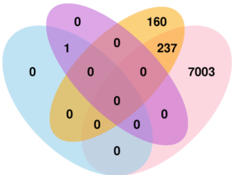
5



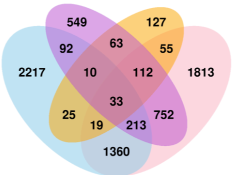
6



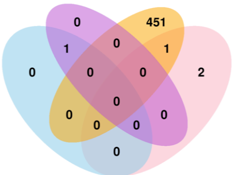
7



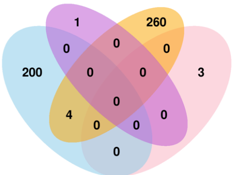
8



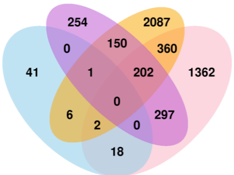
9



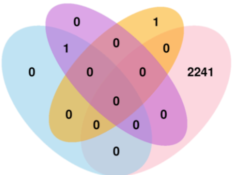
10



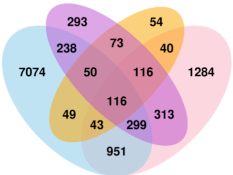
11



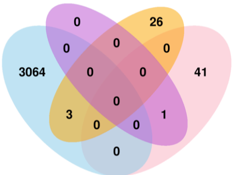
12



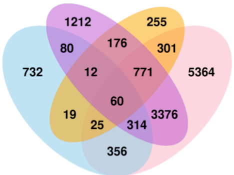
13



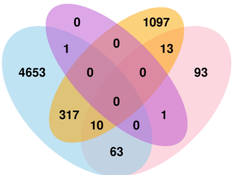
14



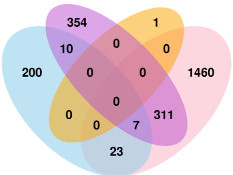
15



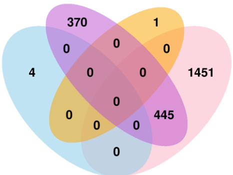
16



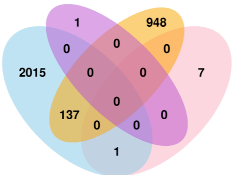
17



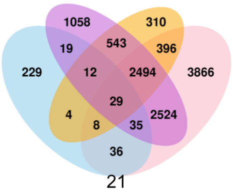
18



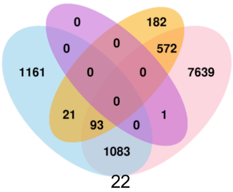
19



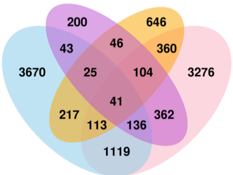
20



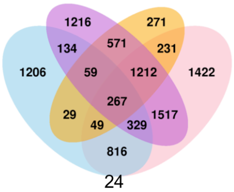
21



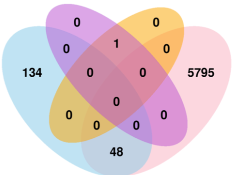
22



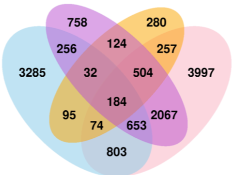
23



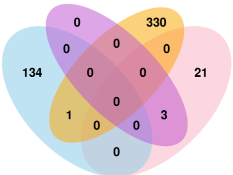
24



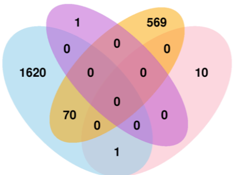
25



26



27



28

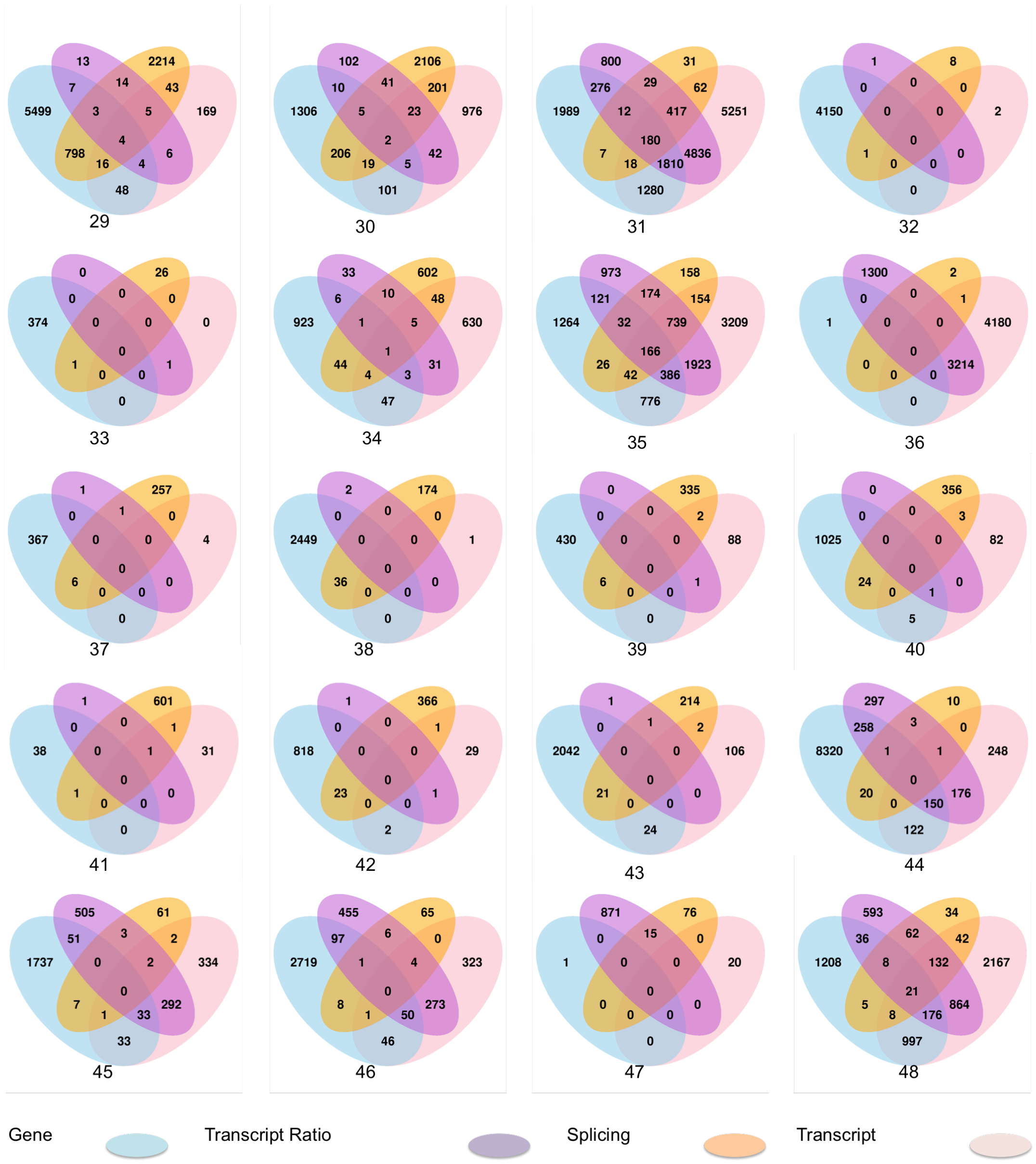


**Figure 4-S4: Overlap of gene sets affected by age-associated changes detected at the level of gene expression (sky blue), transcript ratio (medium orchid), splicing events (orange) or transcript expression (pink).** (1) – (48) show the data for  Whole Blood,Adipose - Subcutaneous,Muscle - Skeletal,Artery - Tibial,Artery - Coronary,Heart - Atrial Appendage,Adipose - Visceral (Omentum),Ovary,Uterus,Vagina,Breast - Mammary Tissue,Skin - Not Sun Exposed (Suprapubic), Minor Salivary Gland,Brain - Cortex,Adrenal Gland,Thyroid,Lung,Spleen,Pancreas,Esophagus - Muscularis,Esophagus - Mucosa,Esophagus - Gastroesophageal Junction,Stomach,Colon - Sigmoid,Small Intestine - Terminal Ileum,Colon - Transverse,Prostate,Testis,Skin - Sun Exposed (Lower leg),Nerve - Tibial,Heart - Left Ventricle,Pituitary,Brain - Cerebellum,Cells - Transformed fibroblasts,Artery - Aorta,Cells - EBV-transformed lymphocytes,Liver,Brain - Hippocampus,Brain - Substantia nigra,Brain - Anterior cingulate cortex (BA24),Brain - Frontal Cortex (BA9),Brain - Cerebellar Hemisphere,Brain - Caudate (basal ganglia),Brain - Nucleus accumbens (basal ganglia),Brain - Putamen (basal ganglia),Brain - Hypothalamus,Brain - Spinal cord (cervical c-1) and Brain – Amygdala respectively.

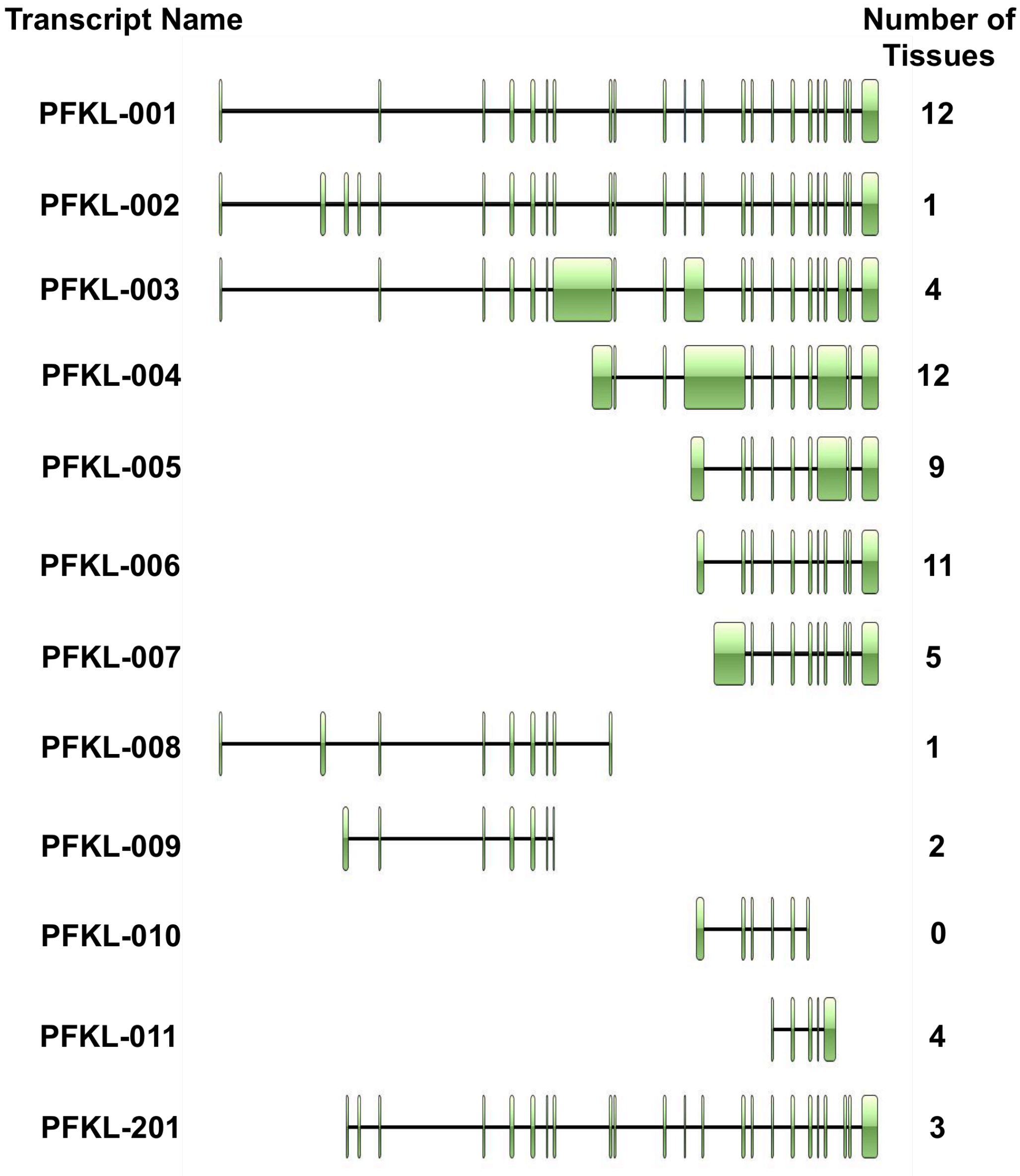


**Figure 4-S5: The annotated 12 transcripts of PFKL gene and the number of tissues in which each transcript's expression changes significantly with age.**

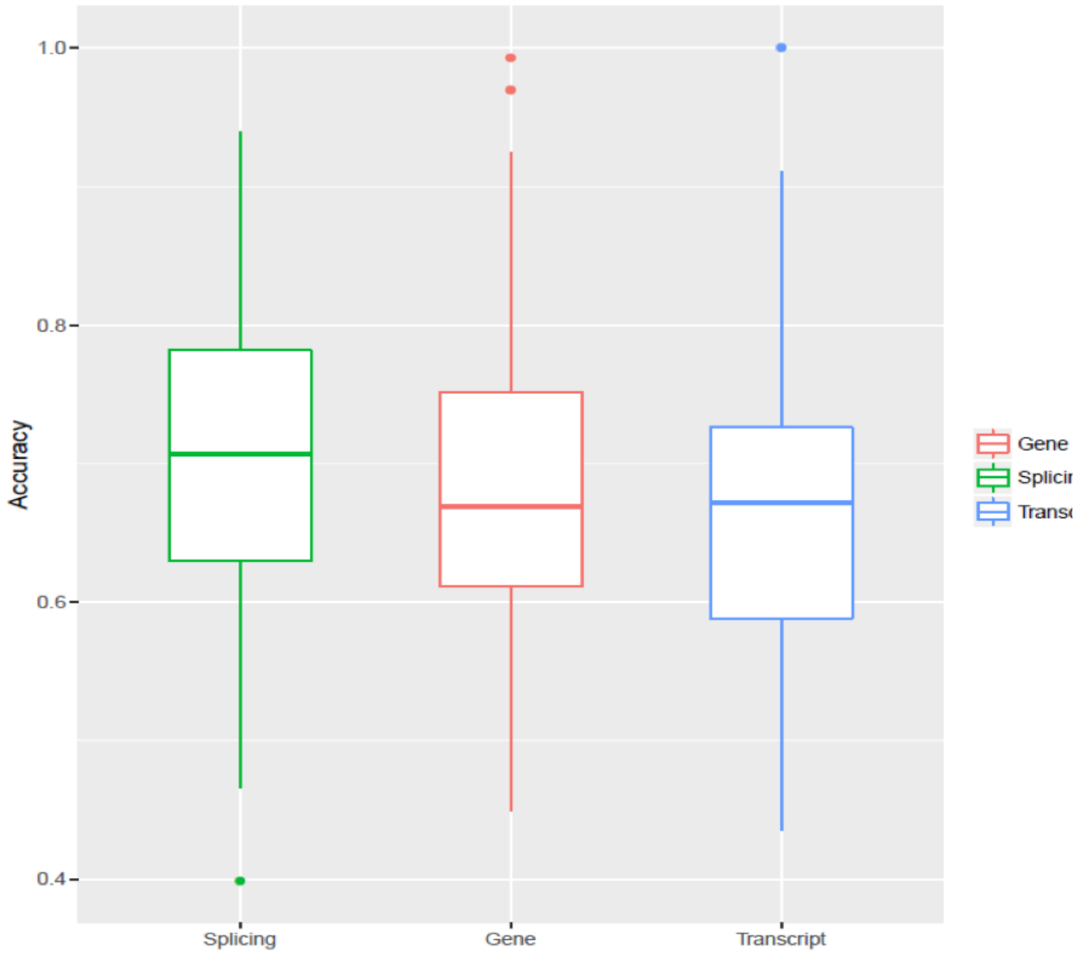


**Figure 4-S6: Age classification accuracy for three models. The green one is for splicing model, the red is for gene model, the blue is for transcript model.**

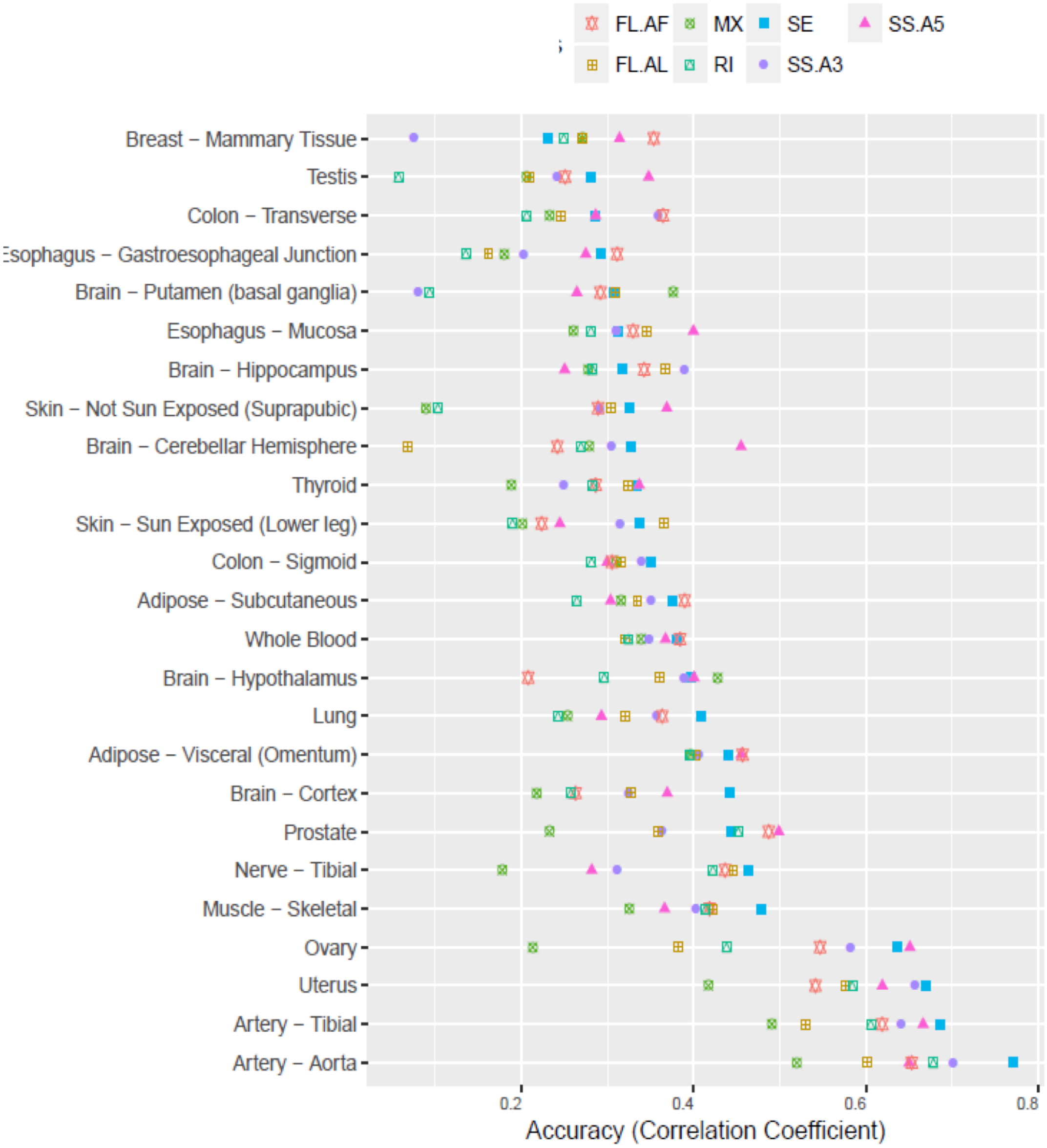


**Figure 4-S7: Accuracies of prediction of age using models based on 7 different types of splicing events across tissues.** The symbols represent the 7 types of splicing events (as defined in Fig. 5). The accuracy is measured by Pearson correlation between predicted and true ages based on cross-validation.

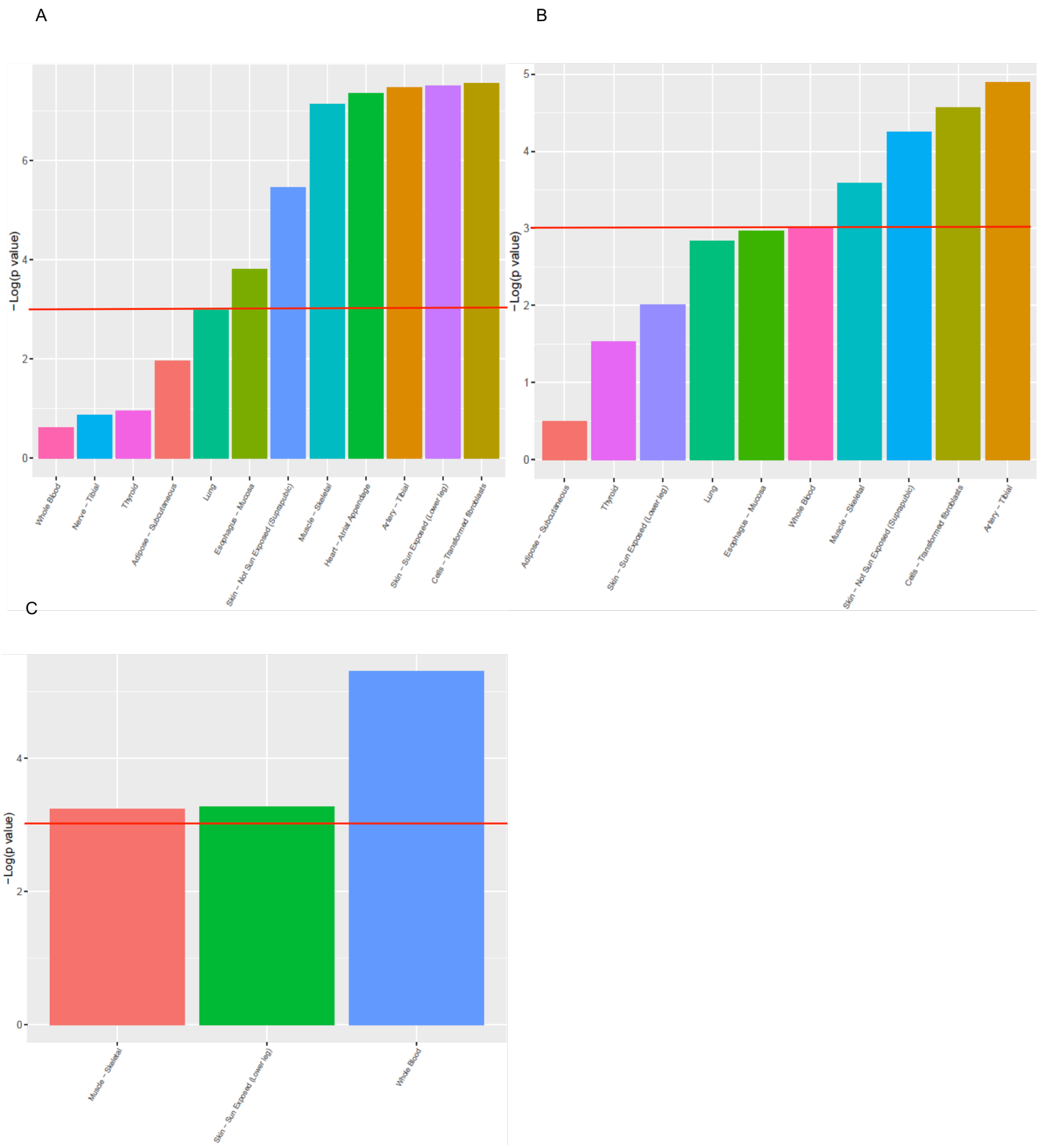


**Figure 4-S8: Additional contribution of splicing events to the explained variance of complex diseases:** (A) Diabetes mellitus type II, (B) Chronic Respiratory Disease, (C) Heart Attack, in multiple tissues shown on x-axis. The y-axis denotes the significance (-log(p-value)) based on a log-likelihood ratio test. The red line indicates p-value = 0.05.

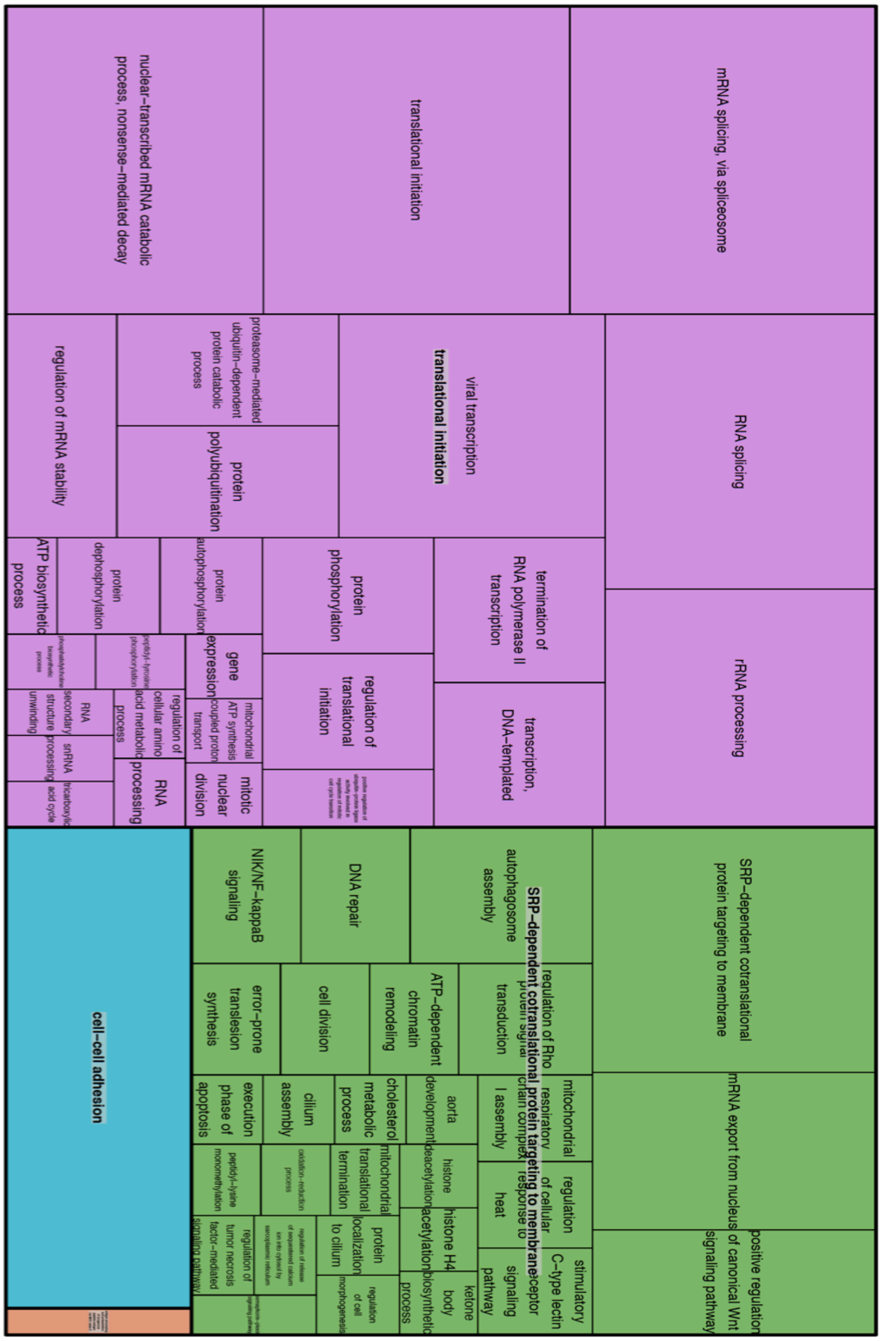
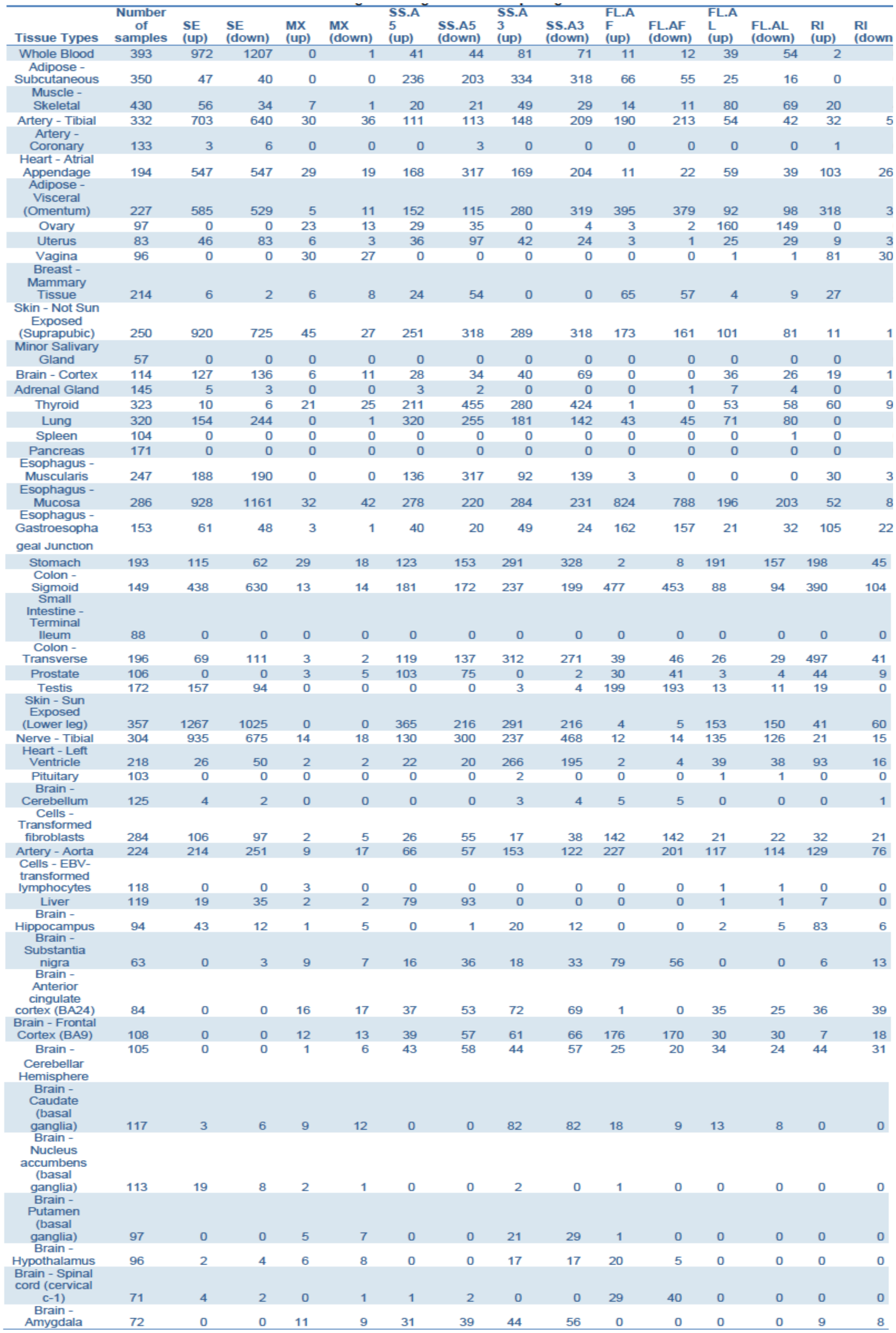


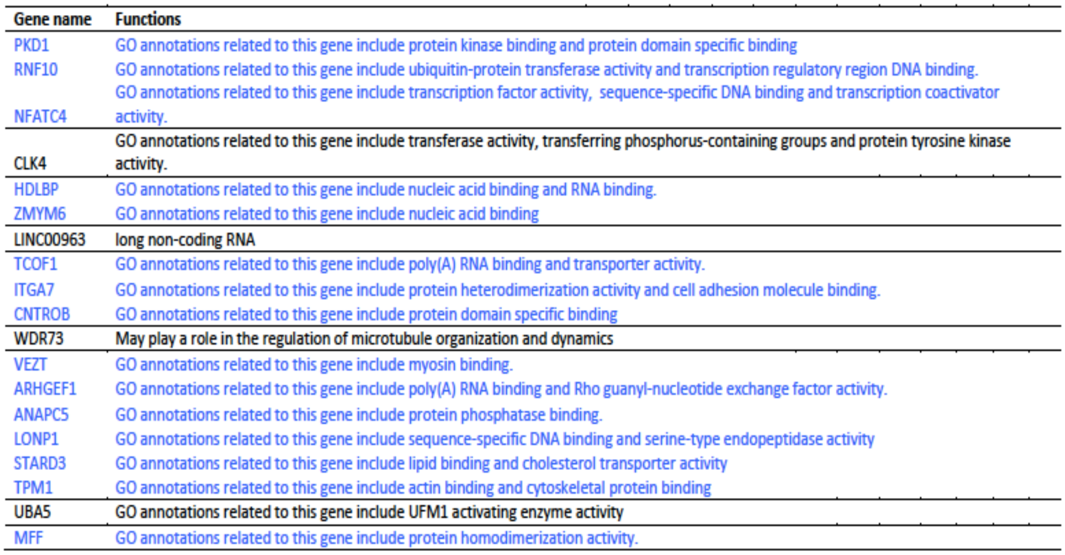
**Figure 4-S9: TreeMap view of enriched functional terms.**



| Tissue Types | Number of samples | SE (up) | SE (down) | MX (up) | MX (down) | SS.A5 (up) | SS.A5 (down) | SS.A3 (up) | SS.A3 (down) | FL.AF (up) | FL.AF (down) | FL.AL (up) | FL.AL (down) | RI (up) | RI (down |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Whole Blood | 393 | 972 | 1207 | 0 | 1 | 41 | 44 | 81 | 71 | 11 | 12 | 39 | 54 | 2 | |
| Adipose - Subcutaneous | 350 | 47 | 40 | 0 | 0 | 236 | 203 | 334 | 318 | 66 | 55 | 25 | 16 | 0 | |
| Muscle - Skeletal | 430 | 56 | 34 | 7 | 1 | 20 | 21 | 49 | 29 | 14 | 11 | 80 | 69 | 20 | |
| Artery - Tibial | 332 | 703 | 640 | 30 | 36 | 111 | 113 | 148 | 209 | 190 | 213 | 54 | 42 | 32 | 5 |
| Artery - Coronary | 133 | 3 | 6 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| Heart - Atrial Appendage | 194 | 547 | 547 | 29 | 19 | 168 | 317 | 169 | 204 | 11 | 22 | 59 | 39 | 103 | 26 |
| Adipose - Visceral (Omentum) | 227 | 585 | 529 | 5 | 11 | 152 | 115 | 280 | 319 | 395 | 379 | 92 | 98 | 318 | 3 |
| Ovary | 97 | 0 | 0 | 23 | 13 | 29 | 35 | 0 | 4 | 3 | 2 | 160 | 149 | 0 | |
| Uterus | 83 | 46 | 83 | 6 | 3 | 36 | 97 | 42 | 24 | 3 | 1 | 25 | 29 | 9 | 3 |
| Vagina | 96 | 0 | 0 | 30 | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 81 | 30 |
| Breast - Mammary Tissue | 214 | 6 | 2 | 6 | 8 | 24 | 54 | 0 | 0 | 65 | 57 | 4 | 9 | 27 | |
| Skin - Not Sun Exposed (Suprapubic) | 250 | 920 | 725 | 45 | 27 | 251 | 318 | 289 | 318 | 173 | 161 | 101 | 81 | 11 | 1 |
| Minor Salivary Gland | 57 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Brain - Cortex | 114 | 127 | 136 | 6 | 11 | 28 | 34 | 40 | 69 | 0 | 0 | 36 | 26 | 19 | 1 |
| Adrenal Gland | 145 | 5 | 3 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 1 | 7 | 4 | 0 | |
| Thyroid | 323 | 10 | 6 | 21 | 25 | 211 | 455 | 280 | 424 | 1 | 0 | 53 | 58 | 60 | 9 |
| Lung | 320 | 154 | 244 | 0 | 1 | 320 | 255 | 181 | 142 | 43 | 45 | 71 | 80 | 0 | |
| Spleen | 104 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | |
| Pancreas | 171 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| Esophagus - Muscularis | 247 | 188 | 190 | 0 | 0 | 136 | 317 | 92 | 139 | 3 | 0 | 0 | 0 | 30 | 3 |
| Esophagus - Mucosa | 286 | 928 | 1161 | 32 | 42 | 278 | 220 | 284 | 231 | 824 | 788 | 196 | 203 | 52 | 8 |
| Esophagus - Gastroesophageal Junction | 153 | 61 | 48 | 3 | 1 | 40 | 20 | 49 | 24 | 162 | 157 | 21 | 32 | 105 | 22 |
| Stomach | 193 | 115 | 62 | 29 | 18 | 123 | 153 | 291 | 328 | 2 | 8 | 191 | 157 | 198 | 45 |
| Colon - Sigmoid | 149 | 438 | 630 | 13 | 14 | 181 | 172 | 237 | 199 | 477 | 453 | 88 | 94 | 390 | 104 |
| Small Intestine - Terminal Ileum | 88 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Colon - Transverse | 196 | 69 | 111 | 3 | 2 | 119 | 137 | 312 | 271 | 39 | 46 | 26 | 29 | 497 | 41 |
| Prostate | 106 | 0 | 0 | 3 | 5 | 103 | 75 | 0 | 2 | 30 | 41 | 3 | 4 | 44 | 9 |
| Testis | 172 | 157 | 94 | 0 | 0 | 0 | 0 | 3 | 4 | 199 | 193 | 13 | 11 | 19 | 0 |
| Skin - Sun Exposed (Lower leg) | 357 | 1267 | 1025 | 0 | 0 | 365 | 216 | 291 | 216 | 4 | 5 | 153 | 150 | 41 | 60 |
| Nerve - Tibial | 304 | 935 | 675 | 14 | 18 | 130 | 300 | 237 | 468 | 12 | 14 | 135 | 126 | 21 | 15 |
| Heart - Left Ventricle | 218 | 26 | 50 | 2 | 2 | 22 | 20 | 266 | 195 | 2 | 4 | 39 | 38 | 93 | 16 |
| Pituitary | 103 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| Brain - Cerebellum | 125 | 4 | 2 | 0 | 0 | 0 | 0 | 3 | 4 | 5 | 5 | 0 | 0 | 0 | 1 |
| Cells - Transformed fibroblasts | 284 | 106 | 97 | 2 | 5 | 26 | 55 | 17 | 38 | 142 | 142 | 21 | 22 | 32 | 21 |
| Artery - Aorta | 224 | 214 | 251 | 9 | 17 | 66 | 57 | 153 | 122 | 227 | 201 | 117 | 114 | 129 | 76 |
| Cells - EBV-transformed lymphocytes | 118 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| Liver | 119 | 19 | 35 | 2 | 2 | 79 | 93 | 0 | 0 | 0 | 0 | 1 | 1 | 7 | 0 |
| Brain - Hippocampus | 94 | 43 | 12 | 1 | 5 | 0 | 1 | 20 | 12 | 0 | 0 | 2 | 5 | 83 | 6 |
| Brain - Substantia nigra | 63 | 0 | 3 | 9 | 7 | 16 | 36 | 18 | 33 | 79 | 56 | 0 | 0 | 6 | 13 |
| Brain - Anterior cingulate cortex (BA24) | 84 | 0 | 0 | 16 | 17 | 37 | 53 | 72 | 69 | 1 | 0 | 35 | 25 | 36 | 39 |
| Brain - Frontal Cortex (BA9) | 108 | 0 | 0 | 12 | 13 | 39 | 57 | 61 | 66 | 176 | 170 | 30 | 30 | 7 | 18 |
| Brain - Cerebellar Hemisphere | 105 | 0 | 0 | 1 | 6 | 43 | 58 | 44 | 57 | 25 | 20 | 34 | 24 | 44 | 31 |
| Brain - Caudate (basal ganglia) | 117 | 3 | 6 | 9 | 12 | 0 | 0 | 82 | 82 | 18 | 9 | 13 | 8 | 0 | 0 |
| Brain - Nucleus accumbens (basal ganglia) | 113 | 19 | 8 | 2 | 1 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Brain - Putamen (basal ganglia) | 97 | 0 | 0 | 5 | 7 | 0 | 0 | 21 | 29 | 1 | 0 | 0 | 0 | 0 | 0 |
| Brain - Hypothalamus | 96 | 2 | 4 | 6 | 8 | 0 | 0 | 17 | 17 | 20 | 5 | 0 | 0 | 0 | 0 |
| Brain - Spinal cord (cervical c-1) | 71 | 4 | 2 | 0 | 1 | 1 | 2 | 0 | 0 | 29 | 40 | 0 | 0 | 0 | 0 |
| Brain - Amygdala | 72 | 0 | 0 | 11 | 9 | 31 | 39 | 44 | 56 | 0 | 0 | 0 | 0 | 9 | 8 |

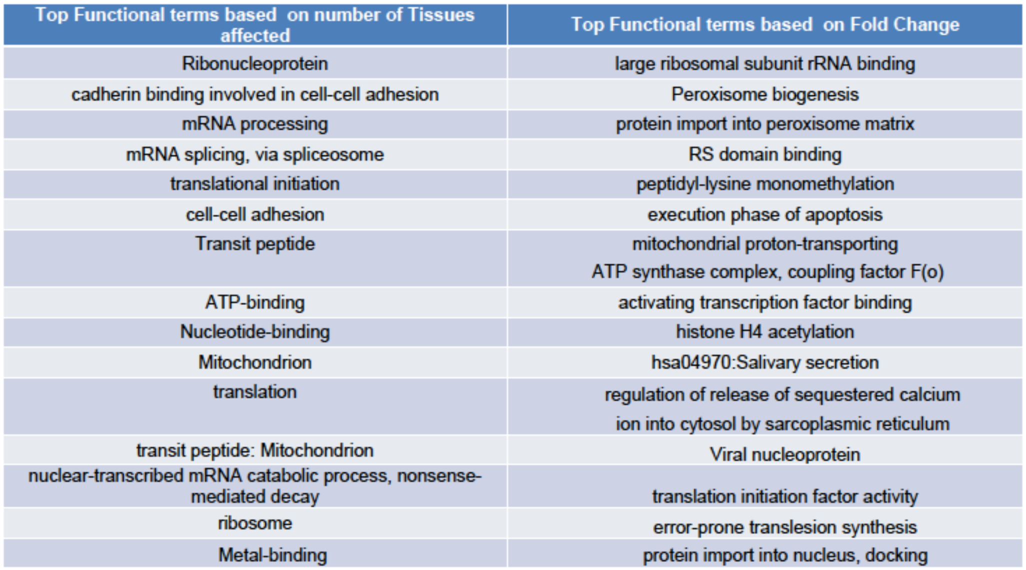
**Table 4-S1: Number of age-associated splicing events across tissues.**

Table 4-S2 Nineteen alternatively spliced genes shared by 9 tissues.



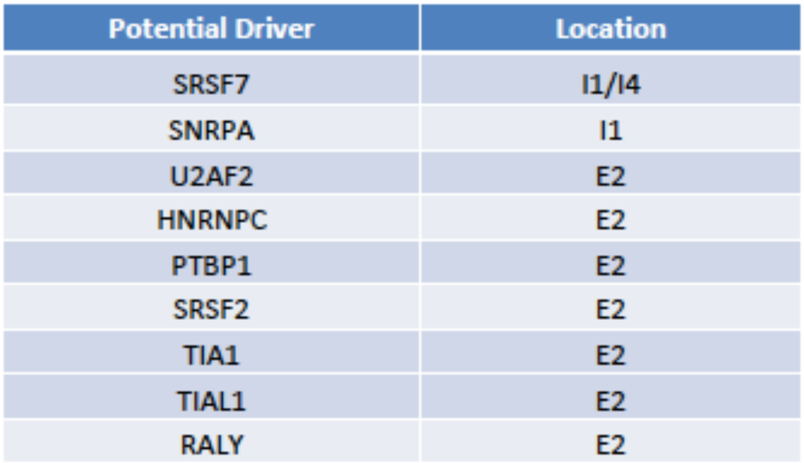
| Gene name | Functions |
|---|---|
| PKD1 | GO annotations related to this gene include protein kinase binding and protein domain specific binding |
| RNF10 | GO annotations related to this gene include ubiquitin-protein transferase activity and transcription regulatory region DNA binding. |
| NFATC4 | GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and transcription coactivator activity. |
| CLK4 | GO annotations related to this gene include transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity. |
| HDLBP | GO annotations related to this gene include nucleic acid binding and RNA binding. |
| ZMYM6 | GO annotations related to this gene include nucleic acid binding |
| LINC00963 | long non-coding RNA |
| TCOF1 | GO annotations related to this gene include poly(A) RNA binding and transporter activity. |
| ITGA7 | GO annotations related to this gene include protein heterodimerization activity and cell adhesion molecule binding. |
| CNTROB | GO annotations related to this gene include protein domain specific binding |
| WDR73 | May play a role in the regulation of microtubule organization and dynamics |
| VEZT | GO annotations related to this gene include myosin binding. |
| ARHGEF1 | GO annotations related to this gene include poly(A) RNA binding and Rho guanyl-nucleotide exchange factor activity. |
| ANAPC5 | GO annotations related to this gene include protein phosphatase binding. |
| LONP1 | GO annotations related to this gene include sequence-specific DNA binding and serine-type endopeptidase activity |
| STARD3 | GO annotations related to this gene include lipid binding and cholesterol transporter activity |
| TPM1 | GO annotations related to this gene include actin binding and cytoskeletal protein binding |
| UBA5 | GO annotations related to this gene include UFM1 activating enzyme activity |
| MFF | GO annotations related to this gene include protein homodimerization activity. |

Table 4-S3: Top 15 functional terms based on number of tissues affected and enrichment (fold change).



| Top Functional terms based on number of Tissues affected | Top Functional terms based on Fold Change |
|---|---|
| Ribonucleoprotein | large ribosomal subunit rRNA binding |
| cadherin binding involved in cell-cell adhesion | Peroxisome biogenesis |
| mRNA processing | protein import into peroxisome matrix |
| mRNA splicing, via spliceosome | RS domain binding |
| translational initiation | peptidyl-lysine monomethylation |
| cell-cell adhesion | execution phase of apoptosis |
| Transit peptide | mitochondrial proton-transporting ATP synthase complex, coupling factor F(o) |
| ATP-binding | activating transcription factor binding |
| Nucleotide-binding | histone H4 acetylation |
| Mitochondrion | hsa04970:Salivary secretion |
| translation | regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum |
| transit peptide: Mitochondrion | Viral nucleoprotein |
| nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | translation initiation factor activity |
| ribosome | error-prone translesion synthesis |
| Metal-binding | protein import into nucleus, docking |

**Table 4-S4: Potential splicing factor drivers for skin fibroblast tissues.**



| Potential Driver | Location |
|---|---|
| SRSF7 | I1/I4 |
| SNRPA | I1 |
| U2AF2 | E2 |
| HNRNPC | E2 |
| PTBP1 | E2 |
| SRSF2 | E2 |
| TIA1 | E2 |
| TIAL1 | E2 |
| RALY | E2 |

## Supplementary Note 4-1

We ascertained that the number of significant events detected was not correlated with the number of samples across tissues; Pearson correlation between sample size and fraction of events found to be 0.06 (p-value = 0.8). We also assessed, for each event type, the potential bias between up-regulated and down-regulated events using paired Wilcoxon test, and found most types of splicing events to be relatively balanced. There was a modest bias in alternative 5’ usage (toward down-regulated events; p-value = 0.047) and alternative first exons events (toward longer isoforms; p-value = 0.03).

## Supplementary Note 4-2

PEER factors are expected to capture a variety of known hidden confounders such as batch effects and a global population variance. We explicitly remove the PEER factors that are correlated with age to minimize false negatives, because such PEER factors will occlude the signal for specific splicing events. Since some of the removed PEER factors are also correlated with race or ethnicity, we may falsely detect splicing events associated with race/ethnicity instead of age. To test whether our results are robust to human ancestry (ethnicity/race here), we assessed our false discovery rate by comparing our results with those without controlling ethnicity/race for skipped exons events using two approaches as following:

First, naively we controlled the ethnicity/race by including all the confounding factors which are correlated with both age and ethnicity/race (Pearson correlation p-value <= 0.05). In Skin (sun exposed) tissue, we found that on average across all 23192 Skipped exon events, (1) Of the removed factors, almost none (8e-3%) are correlated with ethnicity and 2.4% are correlated with race. (2) If we include PEER factors correlated with ethnicity, none of detected events would be deemed insignificant. (3) If we include PEER factors correlated with race, only 0.05% of detected events would be deemed insignificant. We also have done this assessment for all tissues and found that (1) Of the removed factors, on average 4.8% are correlated with ethnicity and on average 8.2% are correlated with race. (2) If we include PEER factors correlated with ethnicity, on average only 0.3% of detected events would be deemed insignificant. (3) If we include PEER factors correlated with race, around 0.35% of detected events would be deemed insignificant in 45 tissues, even though significant difference (~95%) were observed in three tissues (Blood, Lung and  Colon – Sigmoid).

However, since the correlation between age and race can be true due to population sampling bias, the “95% significant difference” in three tissues by directly including PEER factors correlated with both age and race is more likely to be false negatives. In addition, the ethnicity information is extremely limited compared to race information in GTEx. Thus instead of keeping PEER factors correlated with both age and race, we have evaluated the false positive rate by including race information directly in our model as a confounding covariate, which is more reasonable and fair. In this way, on average only 3% of detected events would be deemed insignificant. The three significant affected tissues do not show significant differences (only ~5%) either.

Thus, overall, the possible false positive due to exclusion of PEER factors that might be associated with race/ethnicity is not substantial and our model should be robust to human ancestry.

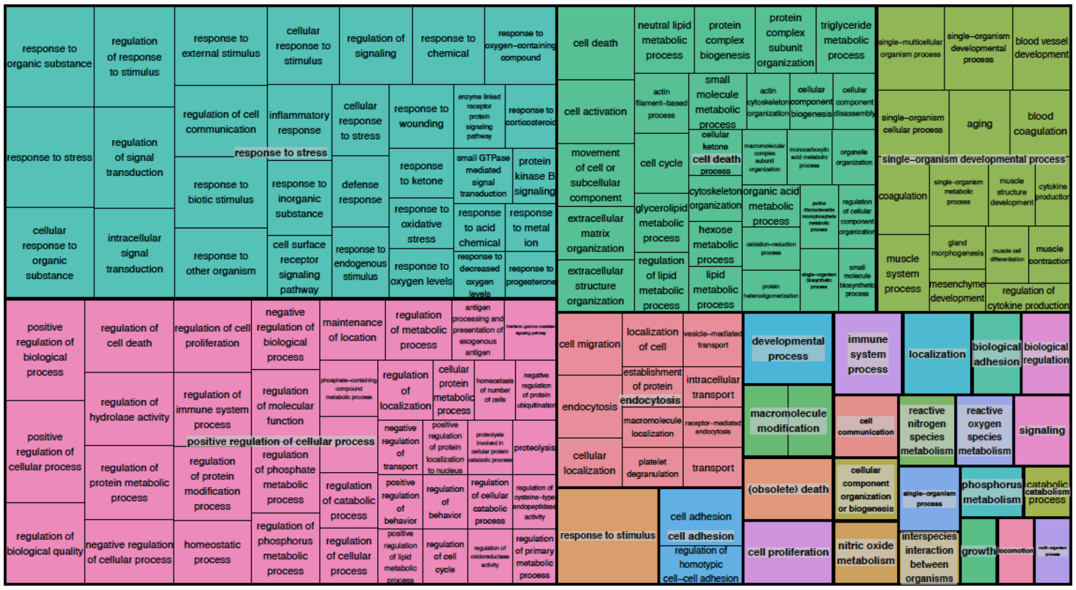


**Figure 5-S1: TreeMap view of enriched functional terms related to the targets genes in the interactions.**

# Bibliography

**Acharya, C. R., Owzar, K., Allen, A. S. and Carlo, M.** (2017). Mapping eQTL by leveraging multiple tissues and DNA methylation. *BMC Bioinformatics* 1–11.

**Adeyemo, A., Gerry, N., Chen, G., Herbert, A., Doumatey, A., Huang, H., Zhou, J., Lashley, K., Chen, Y., Christman, M., et al.** (2009). A Genome-Wide Association Study of Hypertension and Blood Pressure in African Americans. **5**, 1–11.

**Akerman, M., David-eden, H., Pinter, R. Y. and Mandel-, Y.** (2009). A computational approach for genome-wide mapping of splicing factor binding sites. *Genome Biol.* **10**,.

**Alamancos, G. P., Pages, A., Trincado, J. L., Bellora, N. and Eyras, E.** (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21**, 1521–1531.

**Andrés, V. and González, J. M.** (2009). Role of A-type lamins in signaling , transcription , and chromatin organization. **187**, 945–957.

**Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. and Noble, W. S.** (2009). MEME SUITE : tools for motif discovery and searching. **37**, 202–208.

**Barash, Y., Calarco, J. A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B. J. and Frey, B. J.** (2010). Deciphering the splicing code. *Nature* **465**, 53–9.

**Barzilai, N., Huffman, D. M., Muzumdar, R. H. and Bartke, A.** (2012). The Critical Role of Metabolic Pathways in Aging. *Diabetes* **61**, 1315–1322.

**Basu, M., Sharmin, M., Das, A., Nair, N. U. and Wang, K.** (2017). Prediction and subtyping of Hypertension from pan-tissue transcriptomic and genetic analyses.

**Bell, J. T. and Spector, T. D.** (2011). A twin approach to unraveling epigenetics. *Trends Genet.* **27**, 116–125.

**Berdyyeva, T. K., Woodworth, C. D. and Sokolov, I.** (2005). Human epithelial cells increase their rigidity with ageing in vitro: direct measurements. *Phys. Med. Biol.* **50**, 81–92.

**Berget, S. M., Moore, C. and Sharp, P. A.** (1977). Spliced segments at the 5 ' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci.* **74**, 3171–3175.

**Black, D. L.** (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.* **72**, 291–336.

**Blasco, M. A., Partridge, L., Serrano, M., Kroemer, G. and Lo, C.** (2013). Review The Hallmarks of Aging.

**Bosu, W. K., Aheto, J. M. K., Zucchelli, E. and Reilly, S.** (2017). Prevalence , awareness , and associated risk factors of hypertension in older adults in Africa : a systematic review and meta- analysis protocol. 4–11.

**Broderick, P., Carvajal-carmona, L., Pittman, A. M., Webb, E., Howarth, K., Rowan, A., Lubbe, S., Sullivan, K., Fielding, S., Jaeger, E., et al.** (2007). A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.* **39**, 1315–1317.

**Brodsky, A. S., Meyer, C. A., Swinburne, I. A., Hall, G., Keenan, B. J., Liu, X. S., Fox, E. A. and Silver, P. A.** (2005). Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol.* **6**, 1–9.

**Bruneau, B. G.** (2008). The developmental genetics of congenital heart disease. *Nature* **451**, 943–948.

**Bush, W. S. and Moore, J. H.** (2012). Chapter 11 : Genome-Wide Association Studies. *Plos Comput. Biol.* **8**,.

**Cartegni, L.** (2003). ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* **31**, 3568–3571.

**Cassano, P., Lezza, A. M. S., Leeuwenburgh, C., Cantatore, P. and Gadaleta, M. N.** (2004). Measurement of the 4,834-bp mitochondrial DNA deletion level in aging rat liver and brain subjected or not to caloric restriction diet. *Ann. N. Y. Acad. Sci.* **1019**, 269–273.

**Chaves, D. F. S., Carvalho, P. C., Lima, D. B., Nicastro, H., Lorenzeti, F. M., Siqueira-Filho, M., Hirabara, S. M., Alves, P. H. M., Moresco, J. J., Yates, J. R., et al.** (2013). Comparative proteomic analysis of the aging soleus and extensor digitorum longus rat muscles using TMT labeling and mass spectrometry. *J. Proteome Res.* **12**, 4532–4546.

**Chen, M. and Manley, J. L.** (2010). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. **10**, 741–754.

**Chen, C. D., Kobayashi, R. and Helfman, D. M.** (1999). Binding of hnRNP H to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat ??-tropomyosin gene. *Genes Dev.* **13**, 593–606.

**Chen, X., Yu, B., Carriero, N., Silva, C. and Bonneau, R.** (2017a). Mocap :

large-scale inference of transcription factor binding sites from chromatin accessibility. **45**, 4315–4329.

**Chen, L., Luo, C., Shen, L., Liu, Y., Wang, Q., Zhang, C., Guo, R. and Zhang, Y.** (2017b). SRSF1 Prevents DNA Damage and Promotes Tumorigenesis through Regulation of DBF4B Pre- mRNA Splicing. *CellReports* **21**, 3406–3413.

**Cheng, C., Alexander, R., Min, R., Leng, J., Yip, K. Y., Rozowsky, J., Yan, K., Dong, X., Djebali, S., Ruan, Y., et al.** (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. 1658–1667.

**Chiang, K., Yang, H. and Pan, W.** (2018). A Two-Stage Whole-Genome Gene Expression Association Study of Young-Onset Hypertension in Han Chinese Population of Taiwan. *Sci. Rep.* 1–11.

**Cogent Study** (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–1435.

**Cook, K. B., Hughes, T. R. and Morris, Q. D.** (2014). High-throughput characterization of protein-RNA interactions. *Br. Funct. genomics* **14**, 74–89.

**Create linear regression model using stepwise regression - MATLAB stepwiselm**.

**Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al.** (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–6.

**Crick, F.** (1970). Central dogma of molecular biology. 561–563.

**Csoka, A. B., English, S. B., Simkevich, C. P., Ginzinger, D. G., Butte, A. J., Schatten, G. P., Rothman, F. G. and Sedivy, J. M.** (2004). Genome-scale expression profiling of Hutchinson – Gilford progeria syndrome reveals widespread transcriptional misregulation leading to mesodermal / mesenchymal defects and accelerated atherosclerosis. *Aging Cell* 235–243.

**Curradi, M., Izzo, A., Badaracco, G. and Landsberger, N.** (2002). Molecular Mechanisms of Gene Silencing Mediated by DNA Methylation. **22**, 3157–3173.

**Cusack, B. P., Arndt, P. F., Duret, L. and Roest Crollius, H.** (2011). Preventing dangerous nonsense: selection for robustness to transcriptional error in human genes. *PLoS Genet.* **7**, e1002276.

**Das, S. and Krainer, A.** (2014). Emerging functions of SRSF1, splicing factor and oncoprotein, in RNA metabolism and cancer. *Mol. Cancer Res.* **12**, 1195–1204.

**Das, A., Morley, M., Moravec, C. S., Tang, W. H. W., Hakonarson, H., Consortium, M., Margulies, K. B., Cappola, T. P., Jensen, S. and Hannenhalli, S.** (2015). Bayesian integration of genetics and epigenetics detects causal regulatory SNPs underlying expression variability. *Nat. Commun.* **6**, 1–11.

**de Magalhães, J. P., Curado, J. and Church, G. M.** (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875–81.

**Deep Neural Networks for Acoustic Modeling in Speech Recognition - Microsoft Research**.

**Demichelis, F., Setlur, S. R., Banerjee, S., Chakravarty, D., Chen, J. Y. H., Chen, C. X., Huang, J., Beltran, H., Oldridge, D. A., Kitabayashi, N., et al.** (2012). Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *Proc. Natl. Acad. Sci.* **109**, 6686–6691.

**Dermitzakis, E. T. and Clark, A. G.** (2002). Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions : Conservation and Turnover. *Mol. Biol. Evol.* 1114–1121.

**Deschênes, M. and Chabot, B.** (2017). The emerging role of alternative splicing in senescence and aging. *Aging Cell* **16**, 918–933.

**Diao, H., Aplin, J. D., Xiao, S., Chun, J., Li, Z., Chen, S. and Ye, X.** (2011). Altered Spatiotemporal Expression of Collagen Types I , III , IV , and VI in Lpar3 -Deficient Peri-Implantation Mouse Uterus 1. **265**, 255–265.

**Dogan, R. I., Getoor, L., Wilbur, W. J. and Mount, S. M.** (2007). SplicePort — An interactive splice-site analysis tool. *Nucleic Acids Res.* **35**, 285–291.

**Dongen, J. Van, Nivard, M. G., Willemsen, G., Hottenga, J., Helmer, Q., Dolan, C. V, Ehli, E. A., Davies, G. E., Iterson, M. Van, Breeze, C. E., et al.** (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7**, 1–13.

**Drazner, M. H.** (2011). The Progression of Hypertensive Heart Disease. 327–334.

**Dror, H., Donyo, M., Atias, N., Mekahel, K., Melamed, Z., Yannai, S., Lev-**

**Maor, G., Shilo, A., Schwartz, S., Barshack, I., et al.** (2016). A network-based analysis of colon cancer Splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from ELK1. *Genome Res.* **26**, 541–553.

**Eccleston, A., Cesari, F. and Skipper, M.** (2013). Transcription and epigenetics. **502**, 7472.

**Ehret, G. B. and Teresa Ferreira et al** (2016). The genetics of blood pressure regulation and its target organs from association studies in 342 , 415 individuals. *Nat. Genet.* **48**, 1171–1184.

**Eisenberg, E. and Levanon, E. Y.** (2013). Human housekeeping genes , revisited. *Trends Genet.* **29**, 569–574.

**Enroth, S., Bornelöv, S., Wadelius, C. and Komorowski, J.** (2012). Combinations of histone modifications mark exon inclusion levels. *PLoS One* **7**, e29911.

**Epigenetic regulation of RNA processing : Nature ENCODE : Nature Publishing Group**.

**Epigenetic similarities between Wilms tumor cells and normal kidney stem cells found -- ScienceDaily**.

**Eriksson, M., Brown, W. T., Gordon, L. B., Glynn, M. W., Singer, J., Scott, L., Erdos, M. R., Robbins, C. M., Moses, T. Y., Berglund, P., et al.** (2003). Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature* **423**, 293–8.

**Ewald, C. Y., Landis, J. N., Abate, J. P., Murphy, C. T., Keith, T. and Road,**

**W.** (2015). remodelling in longevity. **519**, 97–101.

**Fairbrother, W. G., Yeh, R.-F., Sharp, P. A. and Burge, C. B.** (2002). Predictive identification of exonic splicing enhancers in human genes. *Science* **297**, 1007–13.

**Falcon, S. and Gentleman, R.** (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**, 257–258.

**Finkel, T., Serrano, M. and Blasco, M. A.** (2007). The common biology of cancer and ageing. **448**,.

**Flores, K., Wolschin, F., Corneveaux, J. J., Allen, A. N., Huentelman, M. J. and Amdam, G. V** (2012). Genome-wide association between DNA methylation and alternative splicing in an invertebrate. *BMC Genomics* **13**, 480.

**Fong, C., Ko, D. C., Wasnick, M., Radey, M., Miller, S. I. and Brittnacher, M.** (2018). GWAS Analyzer : integrating genotype , phenotype and public annotation data for genome-wide association study analysis. *Bioinformatics* **26**, 560–564.

**Franceschini, N., Fox, E., Zhang, Z., Edwards, T. L., Nalls, M. A., Sung, Y. J., Tayo, B. O., Sun, Y. V, Gottesman, O., Adeyemo, A., et al.** (2013). Genome-wide Association Analysis of Blood-Pressure Traits in African-Ancestry Individuals Reveals Common Associated Genes in African and Non-African Populations. 545–554.

**Franklin, S. S. and Wong, N. D.** (2013). Hypertension and Cardiovascular Disease: Contributions of the Framingham Heart Study. *Glob. Heart* **8**, 49–57.

**Frith, M. C., Li, M. C. and Weng, Z.** (2003). Cluster-Buster : finding dense clusters of motifs in DNA sequences. **31**, 3666–3668.

**Fuchs, F. D.** (2006). Why Do Black Americans Have Higher Prevalence of Hypertension? 379–381.

**Geiger, H., Koehler, A. and Gunzer, M.** (2007). Stem cells, aging, niche, adhesion and Cdc42: A model for changes in cell-cell interactions and hematopoietic stem cell aging. *Cell Cycle* **6**, 884–887.

**Gibney, E. R. and Nolan, C. M.** (2010). Epigenetics and gene expression. *Heredity (Edinb).* **105**, 4–13.

**Gilad, Y., Rifkin, S. A. and Pritchard, J. K.** (2008). Revealing the architecture of gene regulation : the promise of eQTL studies. *Trends Genet.* **24**, 408–415.

**Glass, D., Viñuela, A., Davies, M. N., Ramasamy, A., Parts, L., Knowles, D., Brown, A. A., Hedman, A. K., Small, K. S., Buil, A., et al.** (2013). Gene expression changes with age in skin, adipose tissue, blood and brain. *Genome Biol.* **14**, R75.

**Goldman, R. D., Shumaker, D. K., Erdos, M. R., Eriksson, M., Goldman, A. E., Gordon, L. B., Collins, F. S., Gruenbaum, Y., Khuon, S. and Mendez, M.** (2004). Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson – Gilford progeria syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 8963–8968.

**Gotzmann, J. and Foisner, R.** (2006). A-type lamin complexes and regenerative potential : a step towards understanding laminopathic diseases ? *Histochem. Cell Biol.* 33–41.

**Grant, C. E., Bailey, T. L. and Noble, W. S.** (2011). FIMO : scanning for occurrences of a given motif. **27**, 1017–1018.

**GTEx Consortium, Gte.** (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–60.

**Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., et al.** (2005). Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration. *Science (80-. ).* **308**, 419–422.

**Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al.** (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774.

**Heinzen, E. L., Yoon, W., Tate, S. K., Sen, A., Wood, N. W., Sisodiya, S. M. and Goldstein, D. B.** (2007). Nova2 Interacts with a Cis -Acting Polymorphism to Influence the Proportions of Drug-Responsive Splice Variants of SCN1A. *Am. J. Hum. Genet.* **80**, 876–883.

**Hinton, G. E., Osindero, S. and Teh, Y.-W.** (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–54.

**Horvath, S.** (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115.

**Hu, M., Yu, J., Taylor, J. M. G., Chinnaiyan, A. M. and Qin, Z. S.** (2010). On the detection and refinement of transcription factor binding sites using ChIP-

Seq data. *Nucleic Acids Res* **38**, 2154–2167.

**Huan, T., Esko, T., Peters, M. J. and Pilling, L. C.** (2015). A Meta-analysis of Gene Expression Signatures of Blood Pressure and Hypertension. 1–29.

**Huang, D. W., Sherman, B. T. and Lempicki, R. A.** (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57.

**Huang, D. W., Sherman, B. T. and Lempicki, R. A.** (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13.

**Human, T., Beadchip, E., Chip, E., Chip, T. E., Chip, E., Chip, E., Chip, E., New, R., Fig, S., Table, S., et al.** (2016). Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet.* **48**,.

**Ismene Karakasilioti, G. A. G.** (2014). Tissue-specific aging: a tale of functional asymmetry. **6**, 7–8.

**Jaenisch, R. and Bird, A.** (2003). Epigenetic regulation of gene expression : how the genome integrates intrinsic and environmental signals. **33**, 245–254.

**Jensen, K. B. and Darnell, R. B.** (2015). CLIP: Crosslinking and ImmunoPrecipitation of In Vivo RNA Targets of RNA-Binding Proteins. *Methods Mol. Biol.* 85–98.

**Jiang, Y., Shen, H., Liu, X., Dai, J., Jin, G., Qin, Z., Chen, J., Wang, S., Wang, X., Hu, Z., et al.** (2011). Genetic variants at 1p11.2 and breast cancer risk: A two-stage study in Chinese women. *PLoS One* **6**,.

**Jin, K.** (2010). Modern Biological Theories of Aging. **1**, 72–74.

**Johnson, F. B., Sinclair, D. A. and Guarente, L.** (1999). Molecular biology of aging. *Cell* **96**, 291–302.

**Jung, M. and Pfeifer, G. P.** (2015). Aging and DNA methylation. *BMC Biol.* **13**, 7.

**Ka, J. and La, H.** (2015). BinDNase : a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. **31**, 2852–2859.

**Kato, N., Loh, M. and Takeuchi, F. et al** (2015). Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* 1282–1293.

**Katz, Y., Wang, E. T., Airoldi, E. M. and Burge, C. B.** (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–15.

**Kaufmann, R. G., Ahrens, K., Koop, R. and Smale, S. T.** (1998). CIF150 , a Human Cofactor for Transcription Factor IID-Dependent Initiator Function. **18**, 233–239.

**Kawakami, K., Nakamura, A., Ishigami, A., Goto, S. and Takahashi, R.** (2009). Age-related difference of site-specific histone modifications in rat liver. *Biogerontology* **10**, 415–421.

**Keren, H., Lev-Maor, G. and Ast, G.** (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.* **11**, 345–55.

**Kokubo, Y.** (2014). Prevention of Hypertension and Cardiovascular Diseases. 655–660.

**Kouadjo, K. E., Nishida, Y., Cadrin-girard, J. F., Yoshioka, M. and St-amand, J.** (2007). Housekeeping and tissue-specific genes in mouse tissues. **16**,.

**Krivega, I. and Dean, A.** (2013). Enhancer and promoter interactions — long distance calls. **22**, 79–85.

**Kumar, A., Gibbs, J. R., Beilina, A., Dillman, A., Kumaran, R., Trabzuni, D., Ryten, M., Smith, C., Traynor, B. J., Hardy, J., et al.** (2014). Age associated changes in gene expression in human brain and isolated neurons. **34**, 1199–1209.

**Kuneš, J. and Zicha, J.** (2009). The Interaction of Genetic and Environmental Factors in the Etiology of Hypertension. **58**,.

**Kurtz, A. and Oh, S. J.** (2012). Age related changes of the extracellular matrix and stem cell maintenance. *Prev. Med. (Baltim).* **54**, S50–S56.

**Lee, T. I. and Young, R. A.** (2013). Review Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251.

**Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., Mccallion, A. S. and Beer, M. A.** (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nat. Publ. Gr.* **47**, 955–961.

**Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. and Storey, J. D.** (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883.

**Levy, S. and Hannenhalli, S.** (2002). Identification of transcription factor binding sites in the human genome sequence. **514**, 510–514.

**Li, L. and Neaves, W. B.** (2006). Normal stem cells and cancer stem cells: the niche matters. *Cancer Res.* **66**, 4553–7.

**Li, Y., Willer, C., Sanna, S. and Abecasis, G.** (2009). Genotype Imputation. *Annu. Rev. Genomics Hum. Genet.* 387–406.

**Li, J. J., Bickel, P. J. and Biggin, M. D.** (2014). System wide analyses have underestimated protein abundances and the importance of transcription in mammals.

**Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y. and Pritchard, J. K.** (2016). RNA splicing is a primary link between genetic variation and disease. *Science (80-. ).* **352**, 600–604.

**Li, H., Wang, Z., Ma, T., Wei, G. and Ni, T.** (2017). Alternative Splicing in Aging and Age-related Diseases. *Transl. Med. Aging* 1–9.

**Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K. and Pritchard, J. K.** (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158.

**Licastro, F., Candore, G., Lio, D., Porcellini, E., Colonna-romano, G., Franceschi, C. and Caruso, C.** (2005). Innate immunity and inflammation in ageing : a key for understanding age-related diseases. *Immun. Aging* **14**, 1–14.

**Lin, C. L. G., Bristol, L. A., Jin, L., Dykes-Hoberg, M., Crawford, T., Clawson, L. and Rothstein, J. D.** (1998). Aberrant RNA processing in a neurodegenerative disease: The cause for absent EAAT2, a glutamate transporter, in amyotrophic lateral sclerosis. *Neuron* **20**, 589–602.

**Linares, A. J., Lin, C., Damianov, A., Adams, K. L., Novitch, B. G. and Black, D. L.** (2015). The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation. *Elife* 1–25.

**Liu, W., Xie, Y., Ma, J., Luo, X., Nie, P., Zuo, Z., Lahrmann, U., Zhao, Q., Zheng, Y., Zhao, Y., et al.** (2015). IBS: An illustrator for the presentation and visualization of biological sequences. *Bioinformatics* **31**, 3359–3361.

**Lubbe, S. J., Pittman, a M., Olver, B., Lloyd, a, Vijayakrishnan, J., Naranjo, S., Dobbins, S., Broderick, P., Gómez-Skarmeta, J. L. and Houlston, R. S.** (2012). The 14q22.2 colorectal cancer variant rs4444235 shows cis-acting regulation of BMP4. *Oncogene* **31**, 3777–84.

**Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M. and Misteli, T.** (2010). Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000.

**Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. and Misteli, T.** (2011). Epigenetics in alternative pre-mRNA splicing. *Cell* **144**, 16–26.

**Luo, R. X. and Dean, D. C.** (1999). Chromatin Remodeling and Transcriptional Regulation. **91**, 1288–1294.

**Ly, D. H., Lockhart, D. J. and Lerner, R. A.** (2000). Mitotic Misregulation and Human Aging. **287**, 2486–2493.

**Macarthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., Mcmahon, A., Milano, A., Morales, J., et al.** (2017). The new NHGRI-EBI Catalog of published genome-wide association studies ( GWAS Catalog ). *Nucleic Acids Res* **45**, 896–901.

**Maquat, L. E.** (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* **5**,.

**Marji, J., Donoghue, I. O., Mcclintock, D., Satagopam, V. P., Schneider, R., Ratner, D., Worman, H. J., Gordon, L. B. and Djabali, K.** (2010). Defective Lamin A-Rb Signaling in Hutchinson-Gilford Progeria Syndrome and Reversal by Farnesyltransferase Inhibition. *PLoS One* **5**,.

**Marsman, J. and Hors, J. A.** (2012). Long distance relationships : Enhancer – promoter communication and dynamic gene transcription. **1819**, 1217–1227.

**Matthews, K. S.** (1992). DNA looping. *Microbiol. Rev.* **56**, 123–36.

**Maurano, M. T., Wang, H., John, S., Canfield, T., Lee, K. and Stamatoyannopoulos, J. A.** (2015). Role of DNA Methylation in Modulating Transcription Article Role of DNA Methylation in Modulating Transcription Factor Occupancy. *CellReports* **12**, 1184–1195.

**Mazin, P., Xiong, J., Liu, X., Yan, Z., Zhang, X., Li, M., He, L., Somel, M., Yuan, Y., Phoebe Chen, Y.-P., et al.** (2013a). Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.* **9**, 633.

**Mazin, P., Xiong, J., Liu, X., Yan, Z., Zhang, X., Li, M., He, L., Somel, M., Yuan, Y., Phoebe Chen, Y.-P., et al.** (2013b). Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.* **9**, 633.

**Mccord, R. P., Nazario-toole, A., Zhang, H., Chines, P. S., Zhan, Y., Erdos, M. R., Collins, F. S., Dekker, J. and Cao, K.** (2013). Correlated alterations in genome organization , histone methylation , and DNA – lamin A / C interactions in Hutchinson-Gilford progeria syndrome. *Genome Res.* 260–269.

**Menopause** (2015). *Med. Clin. North Am.* **99**, 521–534.

**Merideth, M. A., Gordon, L. B., Clauss, S., Sachdev, V., Smith, A. C. M.,**

**Perry, M. B., Brewer, C. C., Zalewski, C., Kim, H. J., Solomon, B., et al.** (2008). Phenotype and course of Hutchinson-Gilford progeria syndrome. *N. Engl. J. Med.* **358**, 592–604.

**Milne, L.** Feature Selection Using Neural Networks with Contribution Measures.

**Milne, L.** (1995). Feature Selection Using Neural Networks with Contribution Measures. 1–8.

**Molecular Biology of the Cell, 5th Edition: The Problems Book / Edition 5 by John Wilson | 9780815341109 | Paperback | Barnes & Noble**.

**Monlong, J., Calvo, M., Ferreira, P. G. and Guigo, R.** (2014). Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat. Commun.*

**Moore, L. D., Le, T. and Fan, G.** (2012). DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38**, 23–38.

**Mounkes, L., Kozlov, S., Burke, B. and Stewart, C. L.** (2003). The laminopathies : nuclear structure meets disease. *Curr. Opin. Genet. Dev.* 223–230.

**Mount, S. M.** (2000). Genomic Sequence , Splicing , and Gene Annotation. *Am. J. Hum. Genet.* **6**, 788–792.

**Myatt, S. S. and Lam, E. W. F.** (2007). The emerging roles of forkhead box ( Fox ) proteins in cancer. *Nat. Rev. cancer* **7**, 847–859.

**Naftelberg, S., Schor, I. E., Ast, G. and Kornblihtt, A. R.** (2015). Regulation of Alternative Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.* **84**, 165–198.

**Niccoli, T. and Partridge, L.** (2012). Ageing as a Risk Factor for Disease. *Curr. Biol.* **22**, R741–R752.

**Nilsen, T. W. and Graveley, B. R.** (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–63.

**Olden, K., Freudenberg, N. and Dowd, J. B.** (2014). Discovering how environmental exposures alter genes and could lead to new treatments for chronic illnesses. **30**, 1–9.

**Ong, C. and Corces, V.** (2011). Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–93.

**Ortega, L. M., Sedki, E. and Nayer, A.** (2015). Hypertension in the African American population : A succinct look at its epidemiology , pathogenesis , and therapy. *Nefrol. (English Ed.* **35**, 139–145.

**Paek, K. Y., Hong, K. Y., Ryu, I., Park, S. M., Keum, S. J., Kwon, O. S. and Jang, S. K.** (2015). Translation initiation mediated by RNA looping. *Proc. Natl. Acad. Sci.* **112**, 201416883.

**Park, W., Hwang, C., Kang, M., Seo, J. Y., Chung, J. H., Kim, Y. S., Lee, J., Kim, H., Kim, K., Yoo, H., et al.** (2001). Gene Profile of Replicative Senescence Is Different from Progeria or Elderly Donor. **939**, 934–939.

**Patikoglou, G. A., Kim, J. L., Sun, L., Yang, S., Kodadek, T. and Burley, S. K.** (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. 3217–3230.

**Paz, I., Akerman, M., Dror, I., Kosti, I. and Mandel-gutfreund, Y.** (2010). SFmap : a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res.* **38**, 281–285.

**Pennacchio, L. A., Bickmore, W., Dean, A. and Nobrega, M. A.** (2013). Enhancers : five essential questions. *Nat. Publ. Gr.* **14**, 288–295.

**Petretto, E., Mangion, J., Dickens, N. J., Cook, S. A., Kumaran, M. K., Lu, H., Fischer, J., Maatz, H., Kren, V., Pravenec, M., et al.** (2006). Heritability and Tissue Specificity of Expression Quantitative Trait Loci. *Plos* **2**,.

**Pickrell, J. K., Pai, A. A., Gilad, Y. and Pritchard, J. K.** (2010). Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genet.* **6**,.

**Pinto, E.** (2007). Blood pressure and ageing.

**Piva, F., Giulietti, M., Nocchi, L. and Principato, G.** (2018). SpliceAid : a database of experimental RNA target motifs bound by splicing proteins in humans. *Bioinformatics* **25**, 1211–1213.

**Portales-casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W. and Sandelin, A.** (2010). JASPAR 2010 : the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* **38**, 105–110.

**Pradeepa, M. M., Sutherland, H. G., Ule, J., Grimes, G. R. and Bickmore, W. A.** (2012). Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genet.* **8**, e1002717.

**Prediction as a candidate for learning deep hierarchical models of data**.

**Prince, P. R., Emond, M. J. and Monnat, R. J.** (2001). Loss of Werner syndrome protein function promotes aberrant mitotic recombination. 933–938.

**Qin, Q. and Feng, J.** (2017). Imputation for transcription factor binding predictions based on deep learning. 1–20.

**Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A. and Wysocka, J.** (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–83.

**Rantalainen, M., Lindgren, C. M. and Holmes, C. C.** (2015). Robust Linear Models for Cis-eQTL Analysis. *PLoS One* 1–16.

**Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al.** (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177.

**Reiter, F., Wienerroither, S. and Stark, A.** (2017). Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* **43**, 73–81.

**Renz, H., Autenrieth, I. B., Brandtzæg, P. and Cookson, W. O.** Gene-environment interaction in chronic disease : A European Science Foundation Forward Look. *J. Allergy Clin. Immunol.* **128**, S27–S49.

**Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al.** (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 317–330.

**Rodriguez, S., Coppedè, F., Sagelius, H. and Eriksson, M.** (2009). Increased expression of the Hutchinson-Gilford progeria syndrome truncated lamin A transcript during cell aging. *Eur. J. Hum. Genet.* **17**, 928–937.

**Rouillard, A. D., Gundersen, G. W., Fernandez, N. F., Wang, Z., Monteiro, C. D., McDermott, M. G. and Ma'ayan, A.** (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford).* **2016**, 1–16.

**Rueda, D., Lamichhane, R., Auweter, S. D., Manatchal, C., Austin, K. S., Valniuk, O. and Allain, F.** (2009). Evidence of RNA looping by PTB using Fluorescence Resonance Energy Transfer and NMR spectroscopy. *&lt;I&gt;The FASEB Journal&lt;/I&gt;* **23**,.

**Salekin, S., Zhang, J. M. and Huang, Y.** (2017). A deep learning model for predicting transcription factor binding location at Single Nucleotide Resolution. 57–60.

**Sarda, S. and Hannenhalli, S.** (2014). Next-Generation Sequencing and Epigenomics Research: A Hammer in Search of Nails. **12**, 2–11.

**Schor, I. E., Gómez Acuña, L. I. and Kornblihtt, A. R.** (2013). Coupling between transcription and alternative splicing. *Cancer Treat. Res.* **158**, 1–24.

**Schwartz, S., Meshorer, E. and Ast, G.** (2009). Chromatin organization marks exon-intron structure. *Nat. Struct. Mol. Biol.* **16**, 990–995.

**Shabalin, A. A.** (2012). Matrix eQTL : ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358.

**Shen, S., Wang, Y., Wang, C., Wu, Y. N. and Xing, Y.** (2016). SURVIV for survival analysis of mRNA isoform variation. *Nat. Commun.* **7**, 11548.

**Shi, W., Fornes, O., Mathelier, A. and Wasserman, W. W.** (2016). Evaluating the impact of single nucleotide variants on transcription factor binding.

*Nucleic Acids Res.* **44**, 10106–10116.

**Shindo, Y., Nozaki, T., Saito, R. and Tomita, M.** (2013). Computational analysis of associations between alternative splicing and histone modifications. *FEBS Lett.* **587**, 516–21.

**Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B. and Kashlev, M.** (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79.

**Shumaker, D. K., Dechat, T., Kohlmaier, A., Adam, S. A., Bozovsky, M. R., Erdos, M. R., Eriksson, M., Goldman, A. E., Khuon, S., Collins, F. S., et al.** (2006). Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. *Proc. Natl. Acad. Sci. U. S. A.* **21**, 21–26.

**Simino, J., Shi, G., Bis, J. C., Chasman, D. I., Ehret, G. B., Lyytika, L., Nolte, I. M., Sim, X., Dehghan, A., Eiriksdottir, G., et al.** (2014). Gene-Age Interactions in Blood Pressure Regulation : A Large-Scale Investigation with the CHARGE , Global BPgen , and ICBP Consortia. *Am. J. Hum. Genet.* 24–38.

**Sinclair, D., North, B., Editors, G., North, B. J. and Sinclair, D. A.** (2012). The Intersection Between Aging and Cardiovascular Disease. **2115**, 1097–1108.

**Sofer, T., Wong, Q., Hartwig, F. P., Taylor, K., Warren, H. R., Evangelou, E., Cabrera, C. P., Levy, D., Kramer, H., Lange, L. A., et al.** (2017). Genome-Wide Association Study of Blood Pressure Traits by Hispanic / Latino Background : the Hispanic Community Health Study / Study of Latinos. 1–12.

**Somech, R., Shaklai, S., Geller, O., Amariglio, N., Simon, A. J., Rechavi, G. and Gal-yam, E. N.** (2005). The nuclear-envelope protein and transcriptional repressor LAP2 NL interacts with HDAC3 at the nuclear periphery , and induces histone H4 deacetylation. *J. Cell Sci.* **2**, 4017–4025.

**Spellman, R. and Smith, C. W. J.** (2006). Novel modes of splicing repression by PTB. *Trends Biochem. Sci.* **31**, 73–76.

**Spike, B. T. and Wahl, G. M.** (2011). p53, Stem Cells, and Reprogramming: Tumor Suppression beyond Guarding the Genome. *Genes Cancer* **2**, 404–19.

**Spring, C.** A network-based analysis of colon cancer splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from ELK1.

**Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.** (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958.

**Steger, D. J. and Workman, J. L.** (1996). Remodeling chromatin structures for transcription: What happens to the histones? *BioEssays* 875–84.

**Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R.** (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–7.

**Stork, C. and Zheng, S.** (2017). Genome-Wide Profiling of RNA–Protein Interactions Using CLIP-Seq. *Methods Mol. Biol.* 137–151.

**Stormo, D. and Schneider, T. D.** (1982). Use of the “Perceptron” algorithm to distinguish translational initiation sites in E. coli. **10**, 2997–3011.

**Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C.,**

**Ingle, C. E., Dunning, M., Flicek, P., Koller, D., et al.** (2007). Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224.

**Streuli, M. and Saito, H.** (1989). Regulation of tissue-specific alternative splicing: exon-specific cis-elements govern the splicing of leukocyte common antigen pre-mRNA. **8**, 787–796.

**Sug, S., Yoon, S., Ph, D., Fryar, C. D. and Carroll, M. D.** (2015). Hypertension Prevalence and Control Among Adults : United States, 2011-2014. 2011–2014.

**Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T.** (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**,.

**Sveen, A., Kilpinen, S., Ruusulehto, A., Lothe, R. A. and Skotheim, R. I.** (2015). Aberrant RNA splicing in cancer ; expression changes and driver mutations of splicing factor genes. *Oncogene* **35**, 2413–2427.

**The ENCODE Project Consortium** (2012). An Integrated Encyclopedia of DNA Elements in the Human Genome. **489**, 57–74.

**The UK Biobank Cardio-metabolic Traits Consortium Blood Pressure Working Group, Warren, H. R. and Al, E. E. et** (2018). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. **49**, 403–415.

**Thomas, B. J., Rubio, E. D., Krumm, N., Broin, P. O., Bomsztyk, K., Welcsh, P., Greally, J. M., Golden, A. A. and Krumm, A.** (2011). Allele-specific transcriptional elongation regulates monoallelic expression of the IGF2BP1 gene. *Epigenetics Chromatin* **4**, 14.

**Todeschini, A., Georges, A. and Veitia, R. A.** (2014). Transcription factors : specific DNA binding and specific gene regulation. *Trends Genet.* **30**, 211–219.

**Tollervey, J. R., Wang, Z., Hortobágyi, T., Witten, J. T., Zarnack, K., Kayikci, M., Clark, T. A., Schweitzer, A. C., Rot, G., Curk, T., et al.** (2011a). Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. *Genome Res.* **21**, 1572–1582.

**Tollervey, J. R., Wang, Z., Hortobágyi, T., Witten, J. T., Zarnack, K., Kayikci, M., Clark, T. A., Schweitzer, A. C., Rot, G., Curk, T., et al.** (2011b). Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. *Genome Res.* **21**, 1572–82.

**Tomlinson, I. P. M., Carvajal-carmona, L. G., Dobbins, S. E., Tenesa, A., Jones, A. M., Howarth, K., Palles, C., Broderick, P., Jaeger, E. E. M., Farrington, S., et al.** Multiple Common Susceptibility Variants near BMP Pathway Loci GREM1 , BMP4 , and BMP2 Explain Part of the Missing Heritability of Colorectal Cancer. *PLoS Genet.* **7**, 2–12.

**Tomovic, A. and Ã, E. J. O.** (2007). Sequence analysis Position dependencies in transcription factor binding sites. **23**, 933–941.

**Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. and Pachter, L.** (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–78.

**Tu, Y. and Ara, D.** (2016). LMNA missense mutations causing familial partial lipodystrophy do not lead to an accumulation of prelamin A. *Nucleus* **7**, 512–521.

**Tullet, J. M. A., Hertweck, M., An, J. H., Baker, J., Yun, J., Liu, S., Oliveira, R. P., Baumeister, R., Blackwell, T. K. and Hwang, J. Y.** (2008). Direct inhibition of the longevity promoting factor SKN-1 by insulin-like signaling in C. elegans. *Cell* **132**, 1025–1038.

**Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., Mccarthy, M. I., Brown, M. A. and Yang, J.** (2017). 10 Years of GWAS Discovery : Biology , Function , and Translation. *Am. J. Hum. Genet.* **101**, 5–22.

**Wahl, M. C., Will, C. L. and Lührmann, R.** (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718.

**Wain, L. V, Vaez, A., Jansen, R., Joehanes, R., Most, P. J. Van Der, Erzurumluoglu, A. M., Reilly, P. O., Cabrera, C. P., Warren, H. R., Rose, L. M., et al.** (2017). Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets From Blood and the Kidney.

**Wallace, D. C.** (2005). A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.* **39**, 359–407.

**Wang, Z. and Burge, C. B.** (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**, 802–13.

**Wang, G.-S. and Cooper, T. A.** (2007a). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **8**, 749–61.

**Wang, G.-S. and Cooper, T. A.** (2007b). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet.* **8**, 749–61.

**Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B.** (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–6.

**Wang, H., Burnett, T., Kono, S., Haiman, C. A., Iwasaki, M., Wilkens, L. R., Loo, L. W. M., Berg, D. Van Den, Kolonel, L. N., Henderson, B. E., et al.** (2014a). Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat. Commun.* 1–7.

**Wang, K., Das, A., Xiong, Z., Cao, K. and Hannenhalli, S.** (2014b). Phenotype-dependent coexpression gene clusters: application to normal and premature ageing. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 1–1.

**Wang, K., Cao, K. and Hannenhalli, S.** (2015). Chromatin and Genomic determinants of alternative splicing. In *ACM-BCB*, pp. 345–354.

**Wang, K., Cao, K. and Hannenhalli, S.** (2017). Chromatin and Genomic determinants of alternative splicing. 345–354.

**Wang, K., Wu, D., Zhang, H., Das, A., Basu, M., Malin, J., Cao, K. and Hannenhalli, S.** (2018). Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases. *Sci. Rep.* **8**, 10929.

**Watson, I. R., Takahashi, K., Futreal, P. A. and Chin, L.** (2013). Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14**, 703–18.

**Westra, H. and Franke, L.** (2014). From genome to function by studying eQTLs.

*BBA - Mol. Basis Dis.* **1842**, 1896–1902.

**Whitaker, J. W., Chen, Z. and Wang, W.** (2014). Predicting the human epigenome from DNA motifs. *Nat. Methods* **12**, 265–272.

**Will, C. L. and Lührmann, R.** (2011). Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3**,.

**Wilson, K. L. and Foisner, R.** (2010). Lamin-binding Proteins. *Cold Spring Harb. Perspect. Biol.* 1–17.

**Wingender, E., Dietze, P., Karas, H. and Knüppel, R.** (1996). TRANSFAC : a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**, 238–241.

**Wyss-coray, T.** (2015). Ageing, neurodegeneration and brain rejuvenation.

**Xiong, H. Y., Barash, Y. and Frey, B. J.** (2011). Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* **27**, 2554–62.

**Xu, Q., Modrek, B. and Lee, C.** (2002). Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Res.* **30**, 3754–66.

**Yang, C., Bolotin, E., Jiang, T., Sladek, F. M. and Martinez, E.** (2007). Prevalence of the Initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. **389**, 52–65.

**Yang, J., Huang, T., Petralia, F., Long, Q., Zhang, B., Argmann, C., Zhao, Y., Mobbs, C. V, Schadt, E. E., Zhu, J., et al.** (2015). Synchronized age-related gene expression changes across multiple tissues in human and the link to complex diseases. *Sci. Rep.* **5**, 15145.

**Yao, C., Joehanes, R., Johnson, A. D., Huan, T., Ying, S., Freedman, J. E., Murabito, J., Lunetta, K. L., Metspalu, A., Munson, P. J., et al.** (2014). Sex- and age-interacting eQTLs in human complex diseases. *Hum. Mol. Genet.* **23**, 1947–1956.

**Yeo, G. and Burge, C. B.** (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *J. Comput. Biol.* **11**, 377–394.

**Yeo, G., Hoon, S., Venkatesh, B. and Burge, C. B.** (2004). Variation in sequence and organization of splicing regulatory elements in vertebrate genes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15700–5.

**Yi, W., Clark, P. M., Mason, D. E., Keenan, M. C., Hill, C., Goddard, W. A., Peters, E. C., Driggers, E. M. and Hsieh-Wilson, L. C.** (2012). PFK1 Glycosylation Is a Key Regulator of Cancer Cell Growth and Central Metabolic Pathways. *Science* **337**, 975–980.

**Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., et al.** (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. **2239**,.

**Yu, X., Lin, J., Zack, D. J. and Qian, J.** (2006). Computational analysis of tissue-specific combinatorial gene regulation: Predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.* **34**, 4925–4936.

**Zentner, G. E. and Scacheri, P. C.** (2012). The Chromatin Fingerprint of Gene Enhancer Elements. **287**, 30888–30896.

**Zhang, C., Zhang, Z., Castle, J., Sun, S., Johnson, J., Krainer, A. R. and Zhang, M. Q.** (2008a). Defining the regulatory network of the tissue-specific splicing factors. 2550–2563.

**Zhang, W., Duan, S., Kistner, E. O., Bleibel, W. K., Huang, R. S., Clark, T. A., Chen, T. X., Schweitzer, A. C., Blume, J. E., Cox, N. J., et al.** (2008b). Evaluation of Genetic Variation Contributing to Differences in Gene Expression between Populations. *Am. J. Hum. Genet.* 631–640.

**Zhao, K., Lu, Z., Park, J. W., Zhou, Q. and Xing, Y.** (2013). GLiMMPS: robust statistical model for regulatory variation of alternative splicing using RNA-seq data. *Genome Biol.* **14**, R74.

**Zhou, Y., Lu, Y. and Tian, W.** (2012). Epigenetic features are significantly associated with alternative splicing. *BMC Genomics* **13**, 123.

**Zhou, H.-L., Luo, G., Wise, J. A. and Lou, H.** (2014). Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic Acids Res.* **42**, 701–13.

**Zhu, S., Wang, G., Liu, B. and Wang, Y.** (2013). Modeling exon expression using histone modifications. *PLoS One* **8**, e67448.

**Zhu, H., Wang, G., Qian, J., Sciences, M., Miller, E., Kimmel, S., Cancer, C. and Building, T. S.** (2017). Transcription factors as readers and effectors of DNA methylation. **17**, 551–565.

UNIVERSITY OF MARYLAND

BIOLOGICAL SCIENCES GRADUATE PROGRAM

1247 Biology-Psychology Building
College Park, Maryland 20742-4415
301.405.6905 TEL, bisi@umd.edu

Dr. Steve Fetter
Dean of the Graduate School
The Graduate School
2123 Lee Building
University of Maryland
College Park, MD 20742

Dear Dean Fetter,

This letter is written to signify that the dissertation committee, committee chair, and the graduate director have all approved the use of previously published co-authored work in the final dissertation of Kun Wang, Biological Sciences, UID 112530525.

Citations for the published work(s):

1. Kun Wang, Di Wu, Haoyue Zhang, Avinash Das, Kan Cao, Sridhar Hannenhalli. *Comprehensive map of age-associated splicing changes across human tissues and their contributions to age-associated diseases*. **Scientific Reports 2018**

2. Kun Wang, Kan Cao, and Sridhar Hannenhalli. *Chromatin and Genomic determinants of alternative splicing*. **Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM BCB), Atlanta, 2015**

3. Kun Wang, Avinas Das, Zheng-Mei Xiong, Kan Cao and Sridhar Hannenhalli. *Phenotype-dependent coexpression gene clusters: application to normal and premature ageing*. **IEEE Transactions on Computational Biology and Bioinformatics(TCBB) (doi: TCBB.2014.2359446), 2014**

4. Kun Wang, A. Das, Z. Xiong, K. Cao, S. Hannenhalli. *Identification of gene clusters with phenotype-dependent expression with application to normal and premature ageing*. **Proceedings of ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM BCB),Washington DC, 2013**

In accordance with the Graduate School's policy the dissertation committee has determined that they made substantial contributions to the included work.

Per Graduate School policy the dissertation foreword will identify the scope and nature of the student's contributions to the jointly authored work included in the dissertation and a copy of this letter will be submitted with the dissertation.

Sincerely,

Hssridhar

Dr. Sridhar Hannenhalli, Dissertation Committee Chair,
Professor, CBMG/BISI

Kan Cao
Digitally signed by Kan Cao
DN: cn=Kan Cao, o=University of Maryland, ou, email=kcao@umd.edu, c=US
Date: 2018.07.26 15:54:20 -04'00'

Dr. Kan Cao, Co-Advisor,
Associate Professor, CBMG/BISI

Dr. Charles F. Delwiche
Director of Graduate Studies, Biological Sciences
Professor, CBMG/BISI