ABSTRACT

Title of Dissertation : INVESTIGATION OF PROGERIN EXPRESSION IN NON-HUTCHINSON-GILFORD PROGERIA SYNDROME INDIVIDUALS

Reynold Yu, Doctor of Philosophy, 2023

Dissertation directed by: Dr. Kan Cao and Dr. Steve Mount
Department of Cell Biology and Molecular Genetics
University of Maryland

Hutchinson-Gilford Progerin Syndrome (HGPS) is a premature aging disease caused by a point mutation in the *LMNA* gene, which encodes A-type lamins. This mutation activates a cryptic splice donor in exon 11 and leads to the production of a toxic lamin variant called progerin. Interestingly, small amounts of progerin have also been found in cells and tissues of normal individuals. Here we examine the expression of progerin in publicly available RNA-seq data from normal individuals of the GTEx project. Among the 30 available tissues, progerin expression in normal individuals is highest in sun-exposed skin samples, and its expression in different tissues of the same donor is correlated. In addition, telomere shortening is significantly correlated with progerin expression. Transcriptome-wide correlation analyses suggest that the level of progerin expression is highly correlated with switches in gene isoform

expression patterns, perhaps reflecting widespread isoform shifts in these samples. Differential expression analyses show that progerin expression is correlated with significant changes in the level of transcripts from genes involved in splicing regulation and a significant reduction of mitochondrial transcripts. Interestingly, 5' splice sites whose use is correlated (either positively or negatively) with progerin expression have significantly altered frequencies of consensus trinucleotides within the core 5' splice site. Furthermore, introns whose alternative splicing is correlated with progerin have reduced GC content. Together, our study suggests that progerin expression in normal individuals is part of a global shift in splicing patterns and provides insight into the mechanism behind these changes.

INVESTIGATION OF PROGERIN EXPRESSION IN NON-HUTCHINSON-GILFORD PROGERIA SYNDROME INDIVIDUALS

by

REYNOLD YU

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
2023

Advisory Committee:

Dr. Kan Cao, Chair
Dr. Stephen Mount, Chair
Dr. Antony Jose
Dr. Adam Phillippy
Dr. Mihai Pop
Dr. Margaret Scull

Dedication

I dedicate my work and degree to my family for their unconditional love and support.

Acknowledgments

Pursuing a Ph.D. was a very challenging endeavor. There were many obstacles and hardships in my way. However, I was blessed with supporting and loving group of people that made the experience a fantastic and memorable one.

First, I would like to thank my advisor, Dr. Kan Cao, for her never-ending advice and support. The journey of Ph.D. is not only about academics and research, but also a lot of about how you grow as a human being. Dr. Cao has put so much time and effort into training me to be a better scientist, and also taught me about life and how to get over obstacles that may stand in my way. She also trained me well to be a good mentor and teacher. I have truly learned a lot from her, and I would like to express my endless gratitude towards Dr. Cao.

I would like to thank Dr. Steve Mount for co-mentoring me with Dr. Cao. Dr. Mount's enthusiasm, love and curiosity for science are unrivaled by anyone in this world. He has shown the qualities needed to be a great scientist, and that is to question everything that I do and everything I get as result. Dr. Mount pushed me every chance he got to be as meticulous and flawless as possible. I would like to express my gratitude to Dr. Mount for helping me grow as a bioinformatician.

I would like to thank my committee members Dr. Antony Jose, Dr. Adam Phillippy, Dr. Mihai Pop, and Dr. Margaret Scull, for taking the time and effort to give guidance and provide insights from multiple angles to bolster my research. Throughout my Ph.D. study, there were many times that I would be stuck on what I should do next or be pigeonholed into a single idea, leaving me tunnel-visioned. My committee members gave fresh ideas and views from angles that I had never thought about before. They have helped my research move forward with

great advice and constructive criticisms, and I would like to thank them for their attention and interest into my project.

I would like to show my appreciation to all Cao lab members, past and present, for being with me as friends and colleagues. Thank you, Dr. Zheng-Mei Xiong, for teaching a bioinformatician how to perform experimental techniques during my rotation and a good part of my first year in the Cao lab. I would like to thank Mason Trappio and Dr. Yantenew Gete for creating a lot of fun memories in the breakroom and also for having intellectual scientific discussions. I would like to thank Dr. Xiaojing Mao, Dr. Huijing Xue, Sahar Vakili, Catherine Harvey, Dr. Linlin Sun, Emma Gentry and Elizabeth Izydore for being great lab mates and collaborators throughout my Ph.D. study.

I would like to thank my family members for always being there for me when I am in need. I want to show my appreciation to my father and my mother for always giving inspirational speeches and also giving me comforting words when I needed them. There were definitely times when I felt I was at a roadblock, and their words and care helped me overcome all the difficulties. I would like to thank my brothers, Dennis and Hwangchan, for giving me laughs and advice when I needed them. They definitely made this endeavor much more bearable and helped me move along. I would like to thank my father-in-law and mother-in-law for giving me endless support and love. Thanks to many scientific discussions that I had with my father and mother-in-law, I have gained a lot of insight into viewing research from different angles.

Lastly, I would like to give my deepest and greatest gratitude and love to my newly formed family. Thank you, Ji Young Choi and Eleanor Yu for being part of my life. I have many things to thank Ji Young for, as she started as my lab mentor and developed our relationship

further to become partners for life. I want to thank Ji Young for teaching me basic laboratory techniques. I would like to thank Ji Young for teaching me basic human interaction skills I lacked when we first met. I would also like to thank Ji Young for teaching me basic husband skills. And I would also like to thank Ji Young for deciding to be my wife. I would like to thank my 6-month-old girl Eleanor for joining Ji Young and me on our journey called life. We are very fortunate to have such a blessing and gift in the form of you.

Again, I would like to show my gratitude to everyone who has been part of my Ph.D. journey. They helped me tremendously when I was in need, and I am very excited to begin a new chapter of my life afterward.

Table of Contents

List of Figures

List of Abbreviations

3'ss	3' splice site
5'ss	5' splice site
A3	Alternative 3' splice site
A5	Alternative 5' splice site
ABE	Adenine base editing
AF	Alternative first exon
AL	Alternative last exon
BP	Branch point
CAAX	C : Cystine, A: aliphatic amino acids, X: any amino acid
CPM	Counts per million
CS	Cryptic splicing
CSIM	Cystein-serine-isoleucine-methionine
DAVID	Database for annotation, visualization and integrated discovery
DEA	Differential expression analysis
DREME	Discriminative Regular Expression Motif Elicitation
EDMD	Emery Dreifuss Muscular Dystrophy
ESE	Exon splicing enhancer

ESS	Exon splicing silencer
FP	Fraction progerin
FPKM	Fragments per kilobase per million mapped fragments
FTase	Farnesyltransferase
FTI	Farnesyltransferase inhibitor
GC-content	Guanine-cytosine content
GO	Gene ontology
GTEx	Genotype tissue expression
GSEA	Gene set enrichment analysis
HGPS	Hutchinson-Gilford Progeria Syndrome
IF	Intermediate filament
INM	Inner nuclear membrane
ISE	Intron splicing enhancer
ISS	Intron splicing silencer
LA	Lamin A
LMNA	Lamin A/C gene
MB	Methylene blue
miRNA	Micro RNA
MEME	Multiple EM for Motif Elicitation

MX	Mutually exclusive exon
NES	Normalized enrichment score
NMD	Nonsense mediated decay
PG	Progerin
PSI	Percent spliced in
PTM	Post-translational modification
RBP	RNA binding protein
RI	Retained intron
ROS	Reactive oxygen species
RS	Recursive site
RPKM	Reads Per Kilobase per Million mapped reads
SE	Skipping-exon
siRNA	Small interfering ribonucleic acid
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein
STREAM	Sensitive, thorough, rapid, enriched motif elicitation
TF	Transcription factor
TMM	Trimmed mean of M-values
TPM	Transcript Per Million

UVA	Ultraviolet A
UVB	Ultraviolet B
UVC	Ultraviolet C
UVR	Ultraviolet radiation
Zmpste24	Zinc metallopeptidase STE 24 (CAAX prenyl protease 1)

Chapter I : Introduction

1.1 Hutchinson-Gilford Progeria Syndrome (HGPS)

1.1.1 Laminopathies

Laminopathies are genetic diseases caused by mutations in genes encoding for proteins in the nuclear lamina or nuclear envelope. Laminopathies may result in various phenotypes, from muscular dystrophy to neuropathy to premature aging syndromes (Schreiber & Kennedy, 2013). Nuclear lamins have been under intense study for everything that is related to the nucleus, including changes in the nuclear structure, organization, and functional changes. The nuclear lamina is composed of lamin A/C and B. Lamin A/C is encoded by the A-type lamin gene, *LMNA*, and lamin B is encoded by the B-type lamin genes, *LMNB1* and *LMNB2*. Intriguingly, it has been found that at least 15 unique diseases have been attributed to the *LMNA* gene. This number is greater than diseases attributed to any other single human gene (Worman, 2012). There are distinctions in the inherent properties between A-type lamins and B-type lamins. The difference in the isoelectric points makes it such that B-type lamins stay associated with the nuclear envelope throughout development and differentiation as opposed to A-type lamins, which become soluble during mitosis and are only expressed in terminally differentiated cells (T. Dechat et al., 2000; Thomas Dechat et al., 2008). The expression patterns also differ as well, where B-type lamins are expressed in almost all cell types, A-type lamins are most notably not expressed in neurons and glia due to a brain specific micro RNA (miRNA) Mir-9. (Jung et al., 2012). There are various mutations within the *LMNA* gene that contributes to different unique laminopathies (Figure 1-1). The most prevalent laminopathy, Emery Dreifuss Muscular Dystrophy (EDMD), has mutations all throughout the *LMNA* gene that may contribute to the disease. However, diseases such as HGPS, mandibuloacral dysplasia

familial lipodystrophy, and Dunnigan type familial lipodystrophy, have mutation in *LMNA* gene near the C-terminus, suggesting the importance of lamin A post-translational modification involving farnesylation (Navarro et al., 2004; Schreiber & Kennedy, 2013). There are other

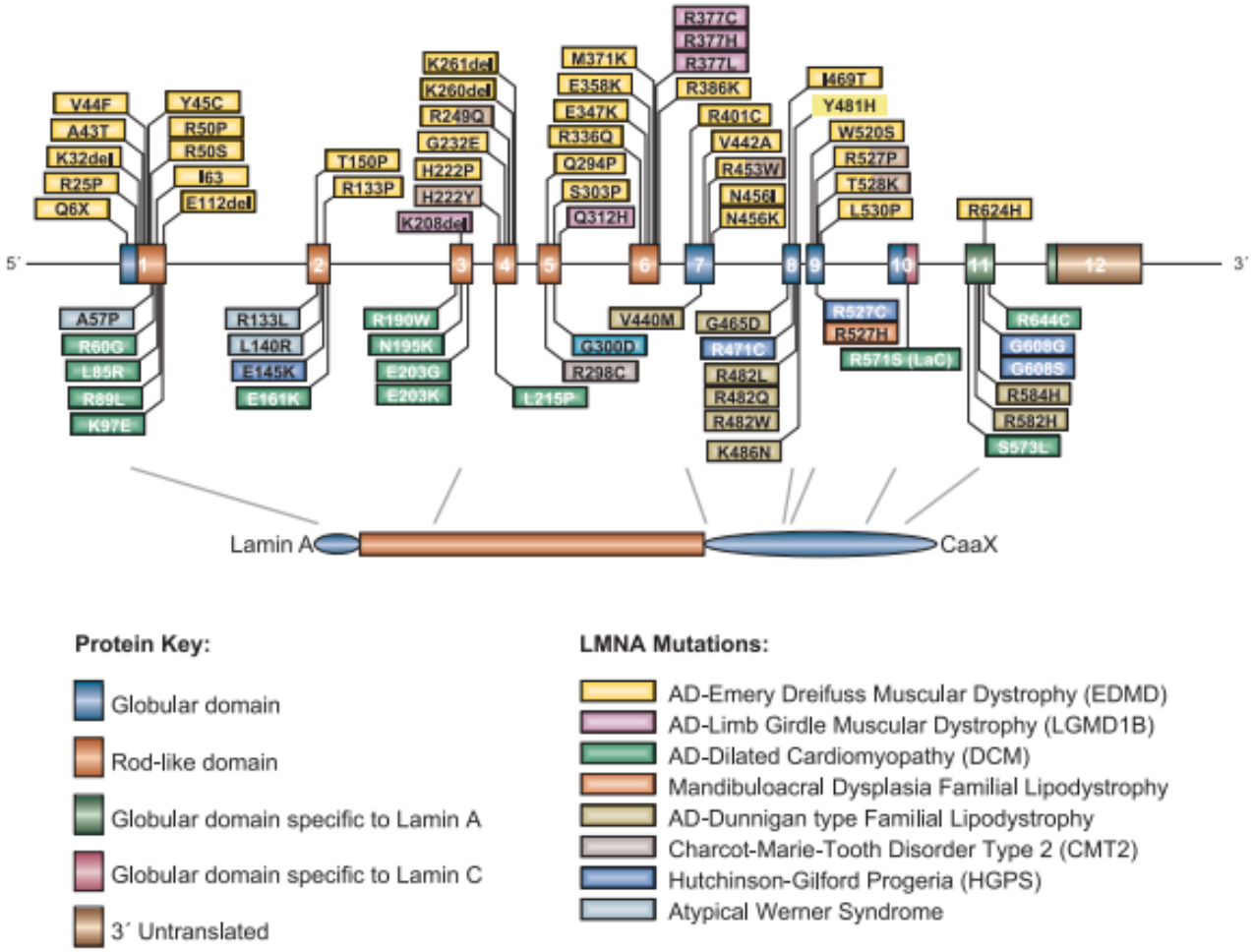


Figure 1-1 Location of mutation for various laminopathies in *LMNA*

Mutations resulting in diseases affecting striated muscle are evenly distributed throughout the *LMNA* gene. Other types of laminopathies, such as mandibuloacral dysplasia familial lipodystrophy, Dunnigan type familial lipodystrophy, and Hutchinson-Gilford Progeria Syndrome are clustered in the C-terminal globular domain.

Adapted from Burke & Stewart, 2006

lamin-associated genes that may contribute to various laminopathies, such as *SYNE1* being cause for EDMD, dilated cardiomyopathy and spinocerebellar ataxia type 8, *ZMPSTE24*

causing mandibuloacral dysplasia and restrictive dermopathy, and *LEMD3* being attributed for bone-related diseases. Some characteristics of laminopathies are abnormal cell signaling pathways, impaired nuclear function, and altered structural dynamics. However, the exact disease mechanisms and biological roles on dysfunctions of nuclear activity remain vague.

1.1.2 Mechanism of classical HGPS

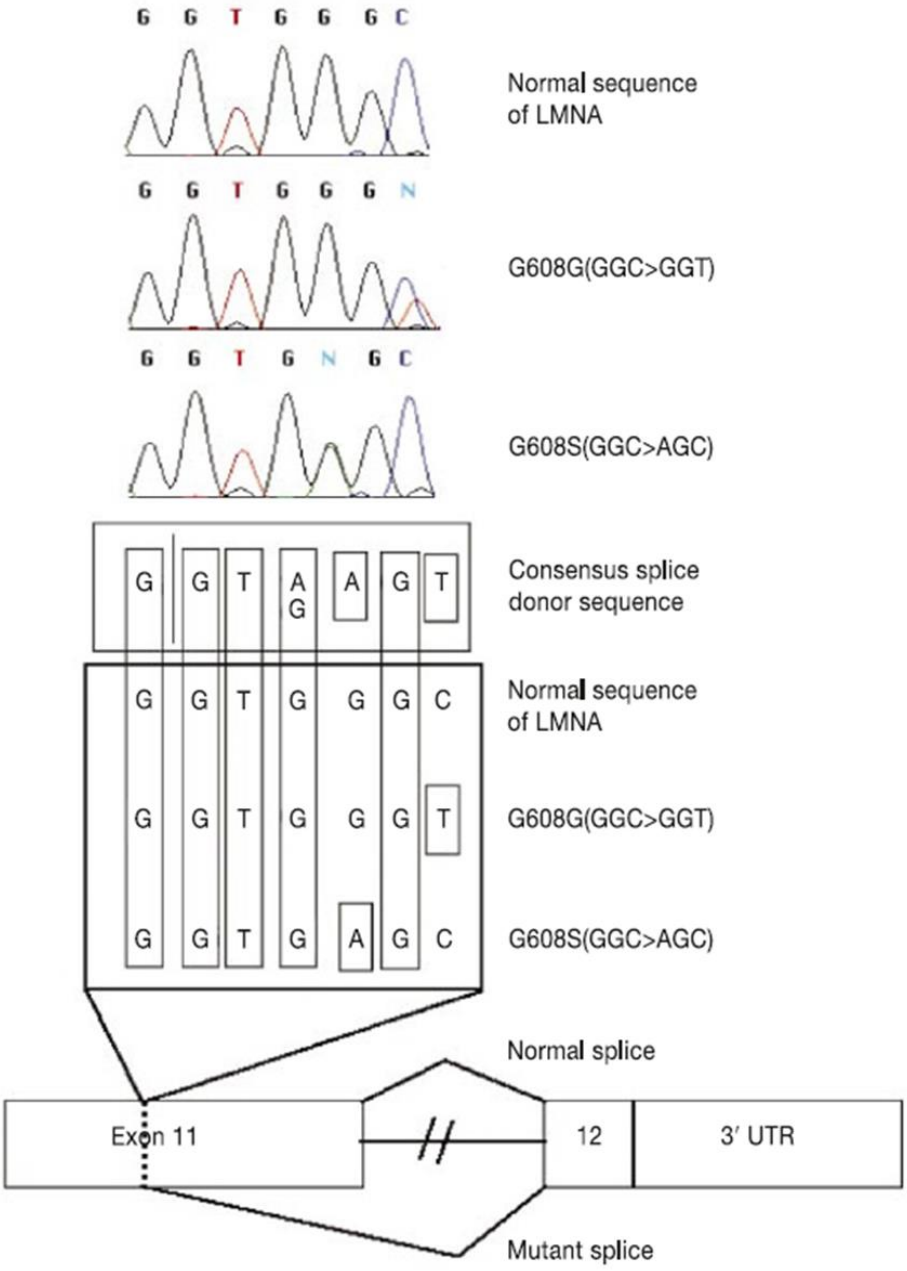


Figure 1-4 Classical HGPS single point mutation in *LMNA* exon 11

Single point mutation in exon 11 of *LMNA* at position 1824 from C to T is known to be the cause of classical HGPS. The point mutation activates cryptic splice site upstream, which leads to deletion of 150 nucleotides (50 amino acids). Usage of this cryptic splice site produces mutant protein called progerin.

Adapted from Eriksson et. al., 2003

HGPS is a rare, autosomal dominant lethal disorder characterized by premature aging. HGPS patients have clinical features mimicking those of normal aging, such as alopecia, loss of subcutaneous fat, short stature, prominent scalp vein, osteoporosis, and severe atherosclerosis (Merideth et al., 2008). Children with HGPS have normal intellectual development but will eventually develop severe atherosclerosis and die of strokes at an average age of 14.6 years (Gordon et al., 2018). Classic cases of HGPS are often attributable to a *de novo* mutation (C1824T, or, in rare cases, G1822A) in the exon 11 of the *LMNA* gene, which activates a cryptic splice site (Eriksson et al., 2003) (Figure 1-2). When this site is used, the subsequent 150 nucleotides of exon 11 are removed by splicing, resulting in an internal deletion of 50 amino acids from the encoded protein. As mentioned, *LMNA* encodes two A-type lamins: lamin A, which includes 12 exons, and lamin C, which uses an alternative 3' terminus after exon 10. After lamin A is translated, the precursor protein prelamin A undergoes several post-translational modification steps (Figure 1-3), including farnesylation of the CAAX tail at the C-terminus, where the isoprenyl group attaches to the cystine of the CAAX motif and removes the terminal AAX amino acids from the CAAX tail, methylation of the exposed cystine by ICMT, and upstream cleavage of the C-terminus by ZMPSTE24 at its recognition site (RSYLLG motif) (Young et al., 2006). ZMPSTE24 is a metalloprotease that plays important roles in post-translational modification of prelamin A by being involved in both deletion of the AAX amino acids and removal of the terminal 15 amino acids at the end of the post-

translational modification steps (Mallampalli et al., 2005). In the case of HGPS, the removal of the farnesyl tail at the end of the post-translational modification step is disturbed. The 50 amino acids missing from progerin include the recognition site for ZMPSTE24, progerin retains the C-terminus of lamin A, including a farnesyl tail that is highly hydrophobic and

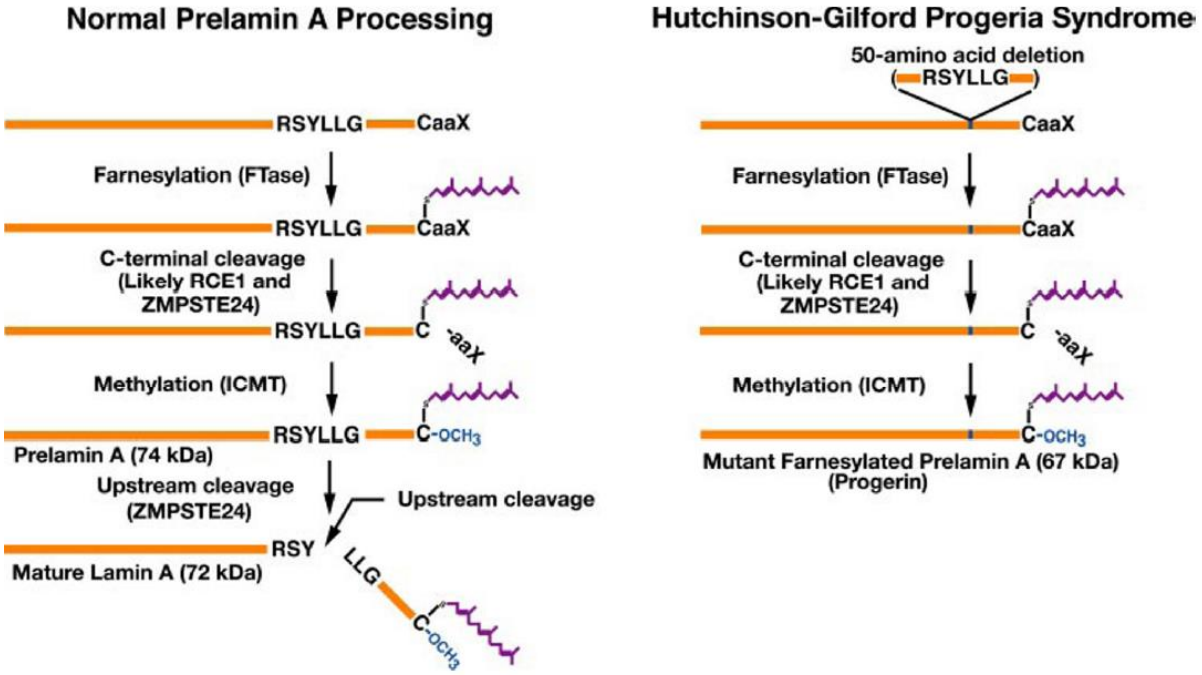


Figure 1-5 Prelamin A post-translational modification in normal and HGPS patients

Prelamin A must undergo post-translational modifications (PTMs) to become fully mature lamin A. The PTMs is initiated by farnesylation of the cystine of terminal CAAX motif. The terminal tripeptide is clipped off by RCE1 or ZMPSTE24. Isoprenylcysteine carboxyl methyltransferase (ICMT) methylates the exposed cystine. To complete the PTM, ZMPSTE24 removes the last 15 tail domain residues, which includes the farnesyl tail. The last step of the PTM, removal of the terminal 15 tail domain residue, is no longer possible in HGPS and thus permanently farnesylated.

Adapted from Young et. al., 2006

disrupts the protein's function (Gruenbaum et al., 2005). Thus, in HGPS, the prelamin A is permanently farnesylated and produces a mutant protein, progerin. Progerin interferes with wildtype lamin A protein function in a dominant negative manner, causing nuclear

abnormalities, genomic instability, mitochondrial dysfunction, and altered redox homeostasis (Gordon et al., 2007, 2012).

The inherent problem with the mutation associated with HGPS is that the C1824T (G608G) mutation causes alteration in the splicing pattern. The C1824T mutation activates a nearby upstream cryptic splicing donor site. The usage of this cryptic splice site then causes the aforementioned 150 nucleotide deletion in exon 11 and the 150 nucleotide deletion leads to production of progerin due to the lack of *ZMPSTE24* cleavage site. Understanding the mechanism of splicing may help us understand the mechanism behind progerin production in different settings.

1.2 Splicing

1.2.1 Function of splicing and alternative splicing

The central dogma of molecular biology states that the DNA undergoes transcription steps to produce RNA. This is the initial form of RNA, pre-mRNA, which is premature when it is first transcribed from the DNA. The pre-mRNA must be processed in order for it to become a mature version of RNA, which is the messenger RNA (mRNA). RNA splicing is a major step

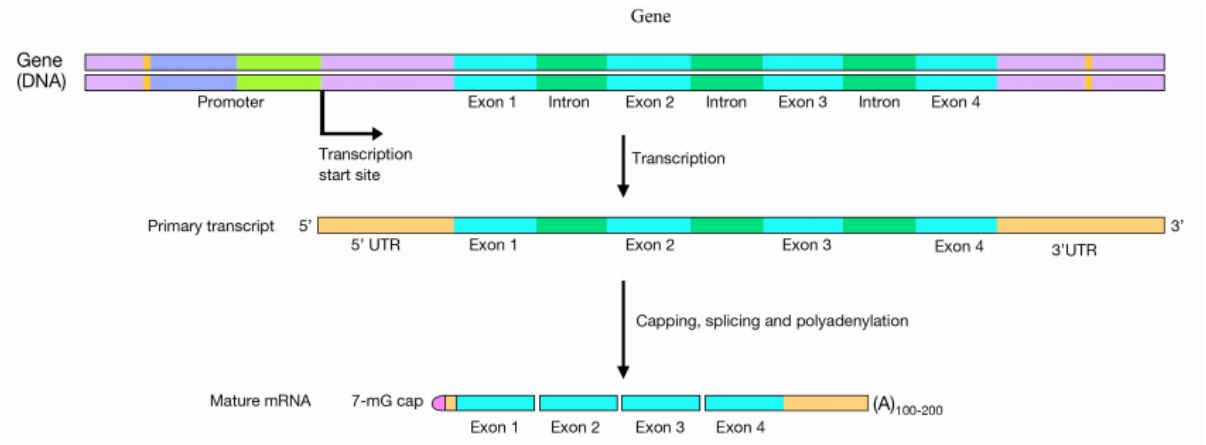


Figure 1-8 Transcription and splicing

From a gene's DNA template, the initial RNA (pre-mRNA) is made during a process called transcription. The pre-mRNA must be processed with a step called splicing to remove the non-coding sequences called introns (colored teal), followed by joining together the sequences to be included in the final RNA product (exons, colored green) to form the mature mRNA.

Adapted from Ahern & Rajagopal, 2023

in this process. To understand the mechanism behind splicing, we must understand the structure of a gene. A eukaryotic gene structure is largely subdivided into sequences known as exons, and introns (Figure 1-4). In the process of splicing, the exons are joined together and thus "splicing out" the noncoding sequences, introns, in order to create the mature mRNA. Alternative splicing is the process by which multiple different mature mRNA are produced from the same gene. This is possible because not every exon within a gene is utilized for splicing and creation of mature mRNA. The exons are "mix and matched" to have many different combinations. Up to 95% of human multi-exon genes go through alternative splicing to encode for multiple different proteins (Jiang & Chen, 2021). This, in turn, allows the alternative splicing to contribute to the majority of protein diversity.

1.2.2 Splicing mechanism

As described in the previously, RNA splicing is the process of removal of the noncoding sequence of the gene and joining together the protein coding sequence of the gene to create mature mRNA. The process of splicing is largely choreographed by the formation of a machinery called “spliceosome”. The spliceosome is composed of small nuclear ribonucleoproteins (snRNP)s U1, U2, U4, U5 and U6. These 5 snRNPs, along with other

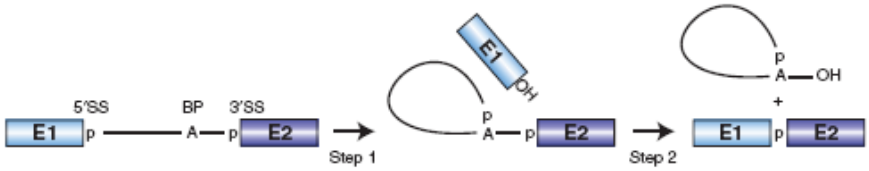


Figure 1-11 Pre-mRNA splicing by U2-type spliceosome

Schematic representation of the two-step mechanism of pre-mRNA splicing. Upstream and downstream exons are represented as E1 and E2, respectively. Adenosine for branch site is denoted by A.

Adapted from Will & Lührmann, 2011

numerous non-snRNP proteins, notably the RNA binding proteins (RBPs), make up the U2-dependent spliceosome.

Splicing is initiated by U1 snRNP identifying and binding to the 5’ss of the upstream exon (E1), and U2 snRNA identifying and binding to the branch site (Figure 1-5). This is called the prespliceosome or A complex. The prespliceosome then combines with the preassembled machinery consisting of U4/U6 and U5 snRNPs. Through this process, the fully assembled mature spliceosome complex is formed. The GU at the 5’ end of the intron is cleaved and then connected to the branch site adenine by transesterification reaction, in which a hydroxyl (OH) group of the adenine attacks the bond of the guanine at the splice site. This forms a loop and is

called the “lariat”. After the formation of the lariat, U1 and U4 snRNPs are released. Afterward, the 3’ss is cleaved as well, followed by the attaching of the E1 and E2 exons by transesterification reaction. The intron lariat, along with the U2, U6 and U5 snRNPs are released after this process.

It is important to know how and where the splicing machinery forms. How does the splicing machinery know where the *bona fide* splice site is? One of the most important tasks the spliceosome does is identifying the true splice site. The 50 bases into exonic and intronic regions around the splice site is known to contain many *cis*-acting elements that help recruit or deter splicing regulatory factors (X. H. F. Zhang et al., 2005), most notably splicing enhancers and splicing silencers. Depending on their location, they are subcategorized into exon splicing enhancers (ESE), intronic splicing enhancers (ISE), exonic splicing silencers (ESS) and intronic splicing silencers (ISS). Another important factor in determining the splice site is that there are a couple of ways that the spliceosome goes about identifying the splice site and be assembled at the proper location. Does the spliceosome identify the splice site from the exonic region? Or does it identify the splice site from the intronic region? The two modes of identifying the splice site is called “exon definition” and “intron definition” (De Conti et al., 2013). Whichever mode the spliceosome choses for splice site identification largely depends

on the architecture of the gene. Specifically, the exonic-intronic lengths between the flanking

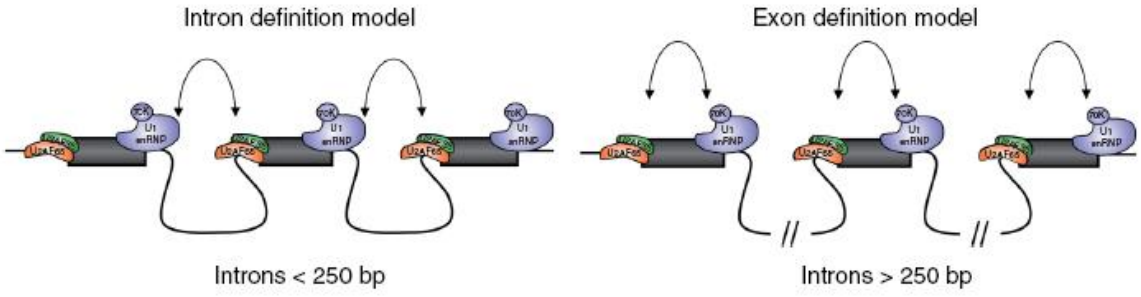


Figure 1-12 Two modes of splice site identification

There are two modes of splice site identification, which are intron definition model (depicted on the left side) and the exon definition model (depicted on the right side). Intron definition model takes place across an intron when the long exons are separated by short introns (< 250bp). Exon definition model takes place across exons when they are separated by long introns (> 250bp).

Adapted from Conti et. al., 2013.

exons are important determinants for the mode selection. In higher eukaryotes, short exons are separated by longer introns (> 250bp), whereas in lower eukaryotes, longer exons are separated by shorter introns (< 250bp). This architectural difference makes different modes of splice site identification more prevalent in one compared to other. When long exons are separated by short intron, the spliceosome is more likely to assemble over the intron, hence utilizing the intron definition mode. When exons are separated by long intron, the spliceosome is more likely to assemble over the exon, utilizing the exon definition mode.

1.2.3 Non-canonical splicing

Aside from the conventional splicing, there are other types of non-canonical splicing methods. Some of these include, cryptic splice sites, microexons, recursive splice sites, intron retentions, exitrons, circular RNAs, chimeric RNAs and atypical splice sites (Sibley et al.,

2016).

Cryptic splice sites are sites that are not normally used in the presence of consensus site. Cryptic splice sites can be activated directly or by mutation of a competing site, often times an authentic, consensus splice site. To assure splicing fidelity, there are exon definition mechanisms such that unauthentic splice sites, such as cryptic splice sites, are not rampantly utilized for splicing. The cause of HGPS is known to be a mutation within the gene *LMNA* at position 1824 of exon 11, which has a single base substitution from cytosine to thymine (Eriksson et al., 2003). This mutation activates a cryptic splice site upstream of the canonical site, resulting in production of mutant protein called progerin. Cryptic splicing is of heavy importance in studying HGPS and will be further discussed in future sections.

Microexons are defined as exons that are shorter than 30nt in length. Microexons are particularly interesting because despite how short they are in size, they are "compensated" by having additional enhancing signals for its utilization (Ustianenko et al., 2017). Interestingly, microexons are flanked by intronic motifs that attract RBPs which acts as enhancers, and in turn, promotes the inclusion of the exon. Notably, serine arginine repetitive matrix 4 (SRRM4), RBFOX, and polypyrimidine tract-binding protein 1 (PTBP1) are among RBPs that are found to be bound in the intronic regions flanking the micro exons. Utilization of microexons are significantly impactful in the resulting protein that is being produced in that, the microexons are often times central nodes in protein interaction networks.

Throughout the genome, there are many cases where there are extremely long introns between two consecutive exons. These pre-mRNA are processed in a manner that the introns are spliced in several differential, stepwise fashion. Within the long intron, there is a site that is composed of the non-canonical 3' splice site and 5' splice site that constitutes what's known

as the recursive site (RS). This site often follows the motif YAG|NN, where Y is C or T and N is any base. The RS divides the intron into two sections, the upstream intron, and the downstream intron. The upstream intron is first spliced out, initialized at the 3' splice site of the RS (YAG). The removal of the upstream intron then leaves the two flanking exons with the downstream intron in between. As the final step, the downstream intron is spliced out, and thus the two flanking exons are joined together to become mature mRNA. RS gives valuable insight into exon definition (Georgomanolis et al., 2016).

Intron retention is defined as when the intron between the two flanking exons are not spliced out and thus included in the final product for the mature mRNA. For a long time, intron retention was regarded as consequence of mis-splicing, most notable the inability to recognize

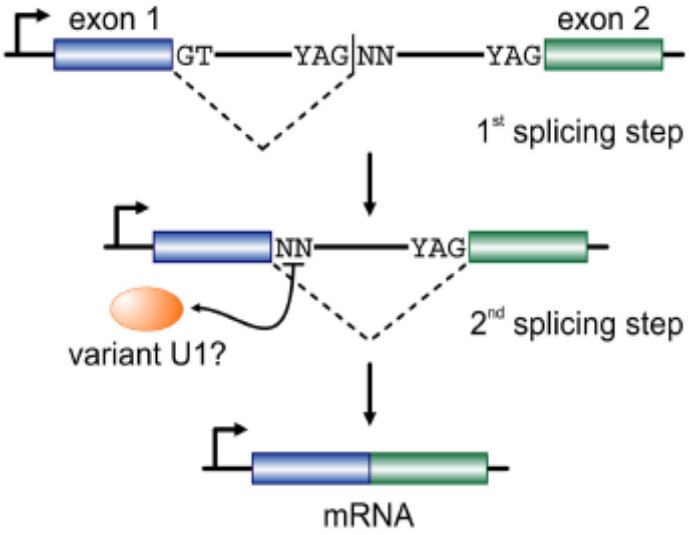


Figure 1-13 Schematic of recursive splicing processing

Two consecutive exons are separated by a long intron which has the recursive site with the motif YAG|NN. The GT of the 3' end of the upstream exon splices into the acceptor sequence, YAG, and thus the upstream half of the long intron is spliced out first. The NN sequence now becomes the new splice donor for the upstream exon. The downstream half of the long intron is spliced out, followed by joining of the two exons together.

Adapted from Georgomanolis et. al., 2016

a nearby canonical splice site (J-L Wong et al., 2015). It is one of many types of alternative splicing and in past studies, transcripts with intron retention were regarded as non-functional because of their susceptibility to non-sense mediated decay (NMD). This phenomenon is more often seen in transcripts with weaker 5' and 3' ss, combined with shorter intron lengths. The introns that are being retained are also found to have higher GC content that the normal counterparts. Recently, intron retention events are gaining attention to have integral role in disease mechanisms. Specifically, in intron retention events, rather than the intron being retained due to mis-splicing and triggering NMD pathway, many introns are being actively retained in mature mRNAs. Exitrons are considered a form of intron retention. Exitrons, by definition, are introns within exons that when retained, functions to have protein coding ability (Staiger & Simpson, 2015). Exitrons are also found to have higher GC content than normal introns, and thus may be less efficient to splice the exitrons, as it is with introns for intron retention events.

Circular RNAs (circRNAs) are a unique class of RNA. They are defined as single-stranded RNA molecules that are produced via a process called backsplicing. Backsplicing is a process where the downstream 5' ss is ligated with the upstream 3' ss. This process then encloses the RNA molecule to be in circular formation, leaving the 5' ss and 3' ss to be indistinguishable. circRNAs have been found to be involved in microRNA (miRNA) inhibition, epithelial-mesenchymal transition and tumorigenesis (Vo et al., 2019). circRNAs also have some unique features, such as rearranging the order of genomic information, protection from exonucleases and constraints on RNA folding (Lasda & Parker, 2014).

Chimeric RNAs are products of joining exons from two or more different gene loci (Frenkel-Morgenstern et al., 2012). There are several methods in which chimeric RNAs are produced in human, which are chromosomal rearrangement, *trans*-splicing of pre-mRNAs,

RNA transcription runoff, errors in other RNA transcription, and artifacts of RNA sequencing. Proteins that result from chimeric RNAs have been mostly associated with cancer in the past. However, recent findings suggest that chimeric RNAs are rather widely present in non-cancerous cells (Z. Li et al., 2018).

1.3 Progerin expression in non-HGPS individuals

Intriguingly, progerin, which is the protein responsible for HGPS, has been found in cells and tissues of non-HGPS normal individuals, although in very small amounts (Burtner & Kennedy, 2010; K. Cao et al., 2007; McClintock et al., 2007; Olive et al., 2010; Scaffidi & Misteli, 2006). Notably, it has been found that progerin can be and is produced by wild-type cells undergoing cellular senescence induced by telomere damage in the form of a telomere uncapping (Cao, Blair, Faddah, Kieckhaefer, et al., 2011). In another study, the impact of solar ultraviolet radiation (UVR) was assessed to mimic the effect of natural aging seen in the skin tissue (Lesiak et al., 2017). For this study, group of 30 healthy Caucasians was divided into two groups, where 15 were elderly (mean age : 64.1 ± 13.1 years), and the other 15 were young adults (mean age : 24.1 ± 4.8 years). The subjects participated in a 6-day sun holiday, getting approximately 5 hours of sun exposure daily. Skin samples before and after the sun-holiday were taken from the subjects. The results showed that the expression of progerin was the highest in the elderly group. Compared to that of young adults prior to and after the sun-holiday, the progerin expression in elderly was significantly higher ($p < 0.001$ and $p < 0.03$). Between young adults before and after sun-holiday, progerin was more abundant in samples post sun-holiday ($p = 0.001$). Takeuchi et. al conducted a study to observe progerin expression under conditions made by artificially induced UVR (Takeuchi & Rünger, 2013). Two types of UVR were used; UVA, which has wavelengths between 315nm and 400nm, and UVB, which has

wavelengths between 280nm and 315nm. Longwave UVA reaches deeper into the dermis than the shortwave UVB, which in part reaches to the upper part of the dermis but does not fully penetrate. They have excluded the use of UVC, which has the shortest range of wavelength, because they seek to mimic the environment that an individual may accumulate during their natural aging process and thus contributing to photoaging of the skin. From this study, they uncovered that UVA, but not UVB, induced progerin expression in both neonatal and aged cultured fibroblasts. The progerin induction was observed more in aged, cultured fibroblasts. Combining the results from these studies, DNA damage is a common theme contributing to progerin induction in non-HGPS normal individuals. It is known that DNA damage by UV irradiation and telomere attrition are both prime examples of phenotypes of chronological aging. It can be speculated that there is some trigger, such as DNA damage and telomere attrition, which causes aberrant splicing in normal individuals, including the progerin cryptic splicing through some unknown mechanism.

1.4 Genotype-Tissue Expression Data

The Genotype-Tissue Expression (GTEx) project was initiated by the GTEx Consortium (Lonsdale et al., 2013) in 2013 to build a database to combat limitations observed in genome-wide association studies (GWAS). From traditional GWAS, genomic loci contributing to thousands of diseases were identified. However, the mechanistic explanation for the disease susceptibility for the majority of the diseases remains unknown. The GTEx aimed to bolster these types of analyses by putting heavy emphasis on expression quantitative trait loci (eQTL). The proposed goals from the GTEx projects were:

1. To create a data resource to enable the systematic study of genetic variation and the

regulation of gene expression in multiple reference human tissues.

2. To provide the scientific community with a biospecimen resource including tissues, nucleic acids, and cell lines upon which to determine other molecular phenotypes.
3. To support and disseminate the results of a study of the ethical, legal, and social issues related to donor recruitment and consent.
4. To support the development of novel statistical methods for the analysis of human eQTLs, alone and in the context of other molecular phenotypes
5. To make data available to the research community as rapidly as possible.
6. To support the dissemination of knowledge, standards and protocols related to biospecimen collection and analysis methods developed during the project.

For the purposes of identifying genomic loci for common diseases and variants, GTEx provides robust and diverse dataset for researchers to investigate. The GTEx includes data from 8,555 samples collected from 544 donors across 30 different tissue types (GTEx version 6). Although the samples collected are from postmortem donors, the GTEx database serves as excellent resource to conduct the study of progerin expression in non-HGPS normal individuals. GTEx samples are collected and processed with a unified, standardized protocol developed by the Cancer Human Biobank (caHUB) of the National Cancer Institute. When studying a rare disease, often times it is difficult to obtain enough data to conduct research, especially for bioinformatics analyses. Using standardized large data, such as GTEx, will greatly increase the probability of finding useful results, especially in the setting of comparative study that I will conduct.

1.5 Significance of the Research

The significance of my research comes from the fact that the toxic protein progerin is found to be expressed in non-HGPS normal individuals. The amount of progerin in these samples is small compared to progerin expression levels in normal individuals. However, one may ask how these individuals have progerin expression, but are still unaffected by HGPS or phenotypes that are prevalent in HGPS patients. In previous studies (Cao, Blair, Faddah, & Al, 2011; Lesiak et al., 2017; Takeuchi & Rünger, 2013), it has been characterized that progerin appears to be expressed in conditions that create an environment causing DNA damage. Conditions such as cellular senescence, continuous sun exposure, and direct UV irradiation are shown to induce progerin in individuals that do not have HGPS. So far, there has not been any study trying to uncover the connection between the trigger that causes progerin expression and the effects of progerin expression in normal individuals. From this research, we seek to elucidate the relationship between the trigger, progerin expression, and the effects of progerin expression in normal individuals.

Chapter II : Exploration of progerin expression in normal individuals

2.0 Introduction

Hutchinson-Gilford Progeria Syndrome (HGPS) is known to be caused by a single point mutation in the *LMNA* gene, which leads to the production of the truncated prelamin A mRNA, which then would produce the mutant form of the lamina A protein called progerin. As described earlier, progerin is found to be expressed not only in HGPS patients, but also in non-HGPS normal individuals as well (Cao et al., 2007; Lesiak et al., 2017; Olive et al., 2010; Scaffidi & Misteli, 2006; Takeuchi & Rünger, 2013). The focus of this section is to explore and characterize progerin expression in Genotype-Tissue Expression (GTEx) data. There are total of 8,555 samples from over 30 different tissue types. Progerin expression will be measured in the GTEx samples across major tissues that are affected in HGPS patients, and also explore various biological processes that are associated with HGPS and also normal aging.

2.1 Progerin is expressed in normal individuals

GTEx version 6 has 8,555 post-mortem samples across 30 different tissues. As a prescreening measure to ask whether progerin is present in the GTEx samples, we took the

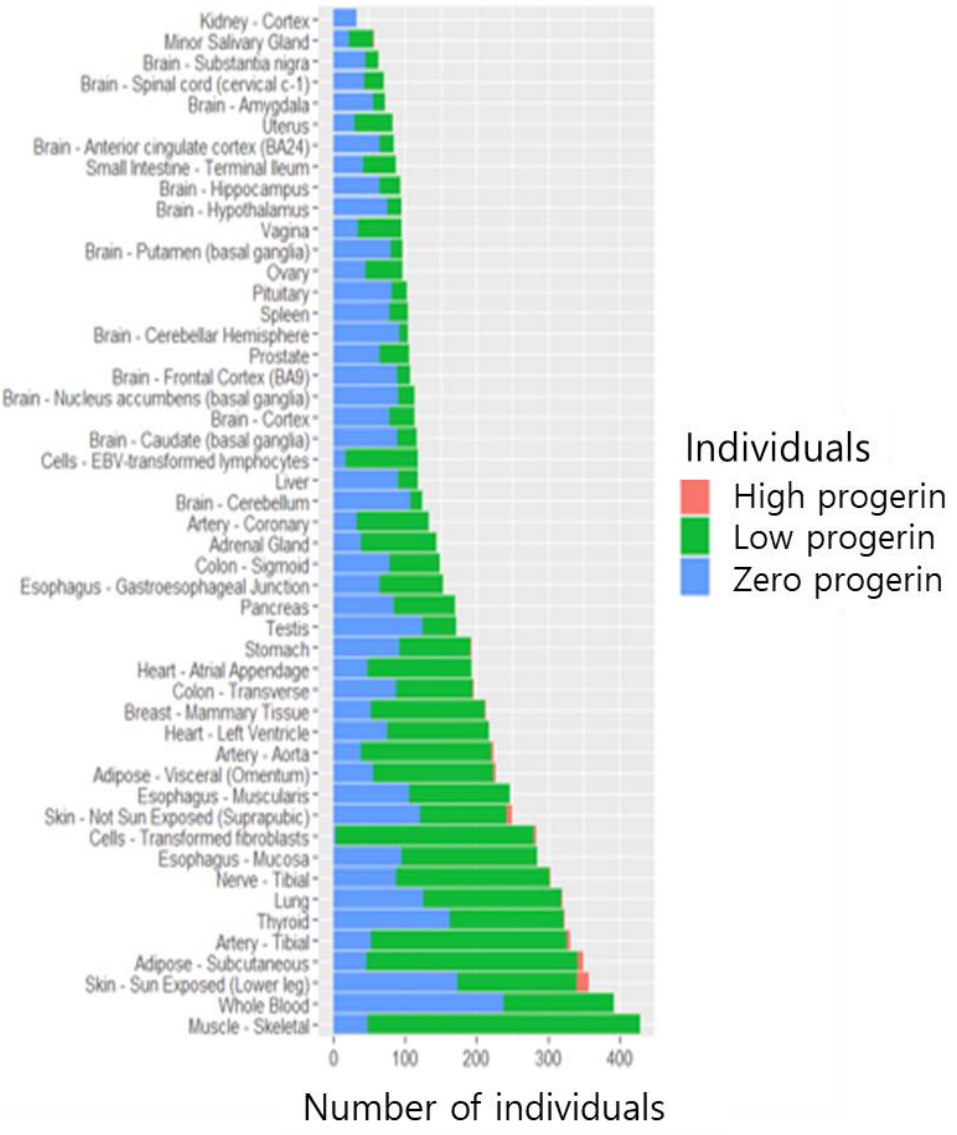


Figure 2-1 Progerin quantification in GTEx v6 count data

Progerin expression categorization was made based on read per kilobase per million mapped reads (RPKM) value on GTEx v6 count data. Zero progerin was defined as RPKM = 0, low progerin was defined as 0 ~ 33.5 RPKM and high progerin was defined as RPKM > 33.5. Zero, low and high progerin samples were indicated in each cell type as blue, green and red, respectively. The highest percentage of high progerin samples was found in sun-exposed skin tissue.

Data generated by Dr. Kun Wang

transcript count data to calculate the reads per kilobase per million mapped reads (RPKM) value of progerin. The samples were separated by tissue types to account for tissue specificity. The RPKM values showed that progerin transcript is present in some normal individuals. Initially, samples were categorized into 3 groups based on progerin RPKM, which consists of zero (0 RPKM), low (0 ~ 33.5 RPKM) and high (> 33.5 RPKM) (Figure 2-1). The "high" group was defined according to the abundance of progerin transcript in fibroblast cells from a progeria patient sample. Examining the abundance of high progerin expressing samples across different tissues, samples from sun-exposed skin tissue had the largest proportion of high progerin

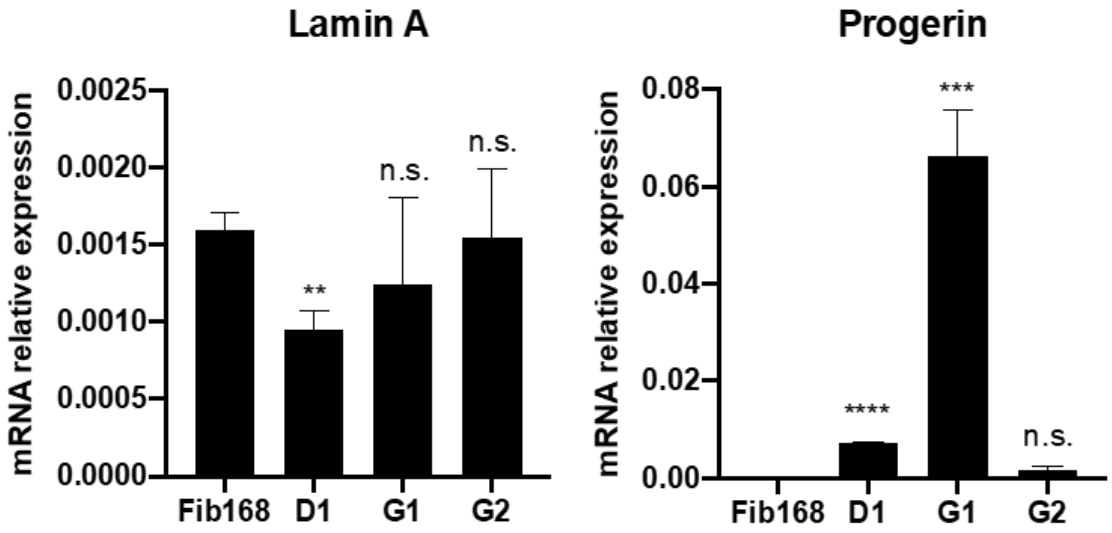


Figure 2-2 Quantitative RT-PCR analysis of lamin A and progerin mRNA levels in GTEx samples

Quantitative RT-PCR analysis of lamin A and progerin mRNA levels in the RNA from wild type fibroblasts (Fb168) and sun-exposed skin samples (D1, G1 and G2); n.s., no significant difference, *P<0.5, **P < 0.01, ***P < 0.001, ****P < 0.0001). Normalized against GAPDH.

Data generated by Dr. Huijing Xue

samples. To further verify the presence of progerin transcript in the GTEx, we chose three sun-exposed skin samples (Collaborator Participant ID 11OF3 (D1), Y5V5 (G1) and 13CF2 (G2)) that showed progerin expression and conducted quantitative RT-PCR analyses with progerin

specific primers (Kan Cao, Blair, Faddah, & Al, 2011). As a control, we used a well-characterized, normal human skin fibroblast line HGFDFN168 from the Progeria Research Foundation, which is known to be progerin negative. Quantitative RT-PCR analyses revealed that progerin is indeed expressed in sun-exposed skin samples, but at varying levels (Figure 2-2).

To measure the progerin RNA levels more precisely, we used fraction progerin (FP), derived from percent spliced in (PSI) quantification calculated with SUPPA2 (Trincado et al.,

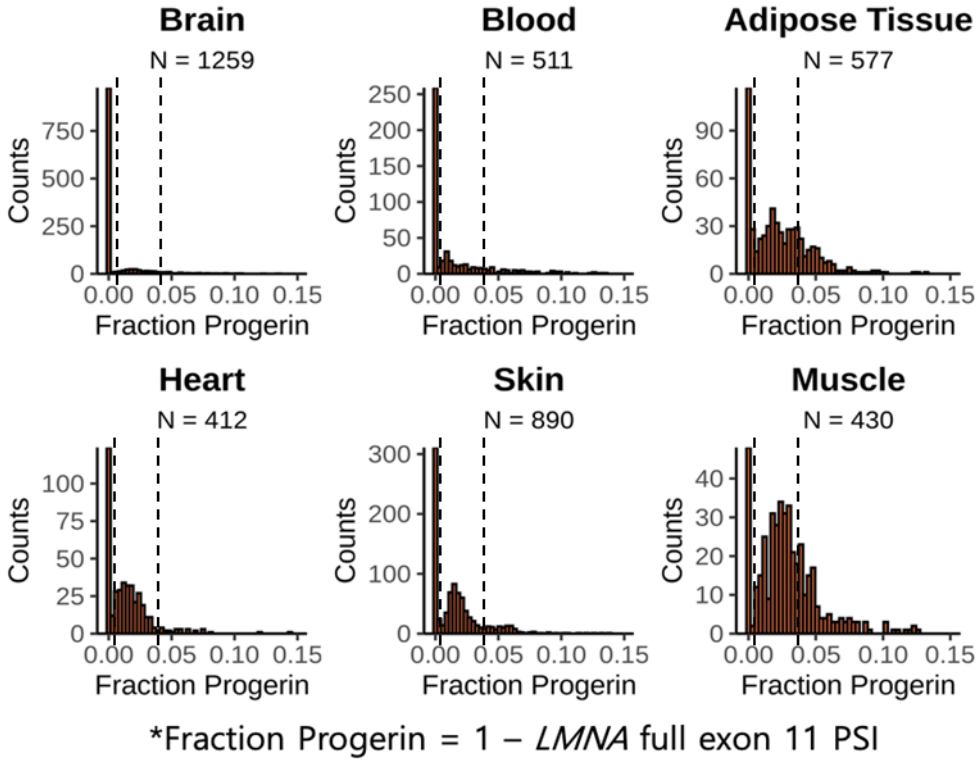


Figure 2-3 Quantifying progerin expression using fraction progerin (FP)

Progerin cryptic splice site usage was quantified using fraction progerin (FP) quantification on brain, blood, adipose, heart, skin and muscle tissue samples. Fraction progerin was predominantly zero in brain samples (negative control). Other tissues showed progerin expression followed a similar pattern. Using this distribution, progerin expression categorization was made, where zero progerin consisted of FP < 0.5%, low progerin consisted of FP value of 0.5% ~ 4% and high progerin consisted of FP > 4%. The dotted lines indicate the cutoff FP for categorization of zero, low and high FP samples.

2018). Tissues show a consistent pattern of a modal FP of about 0.02, but the fraction of samples with progerin expressing samples differs in brain tissue when compared to other tissue types (Figure 2-3). Brain samples showed only a few samples with high FP; this may be related to a brain-specific miRNA inhibiting both lamin A and progerin expression (Figure 2-3) (Jung et al., 2012). The FP distribution (Figure 2-3) was used to divide the samples into zero (FP < 0.005;), low (FP = 0.005 ~ 0.04) and high (FP > 0.04) categories for later analyses and is visually noted on Figure 2-3 with dotted lines. It is noteworthy that progerin levels in some samples approach 10% of all *LMNA* transcripts that include exon 11. To give a context, FP levels in fibroblast samples from HGPS patients ranged between 25% to 30%. While the average FP is relatively low in GTEx samples, there are very few samples that do express some moderate levels of progerin.

2.2 Progerin expression is correlated with shorter telomere lengths

To characterize the progerin-containing samples, we next explored the connection between progerin expression, biological aging (the molecular phenotypes of aging), and chronological aging (the actual age of the individual).

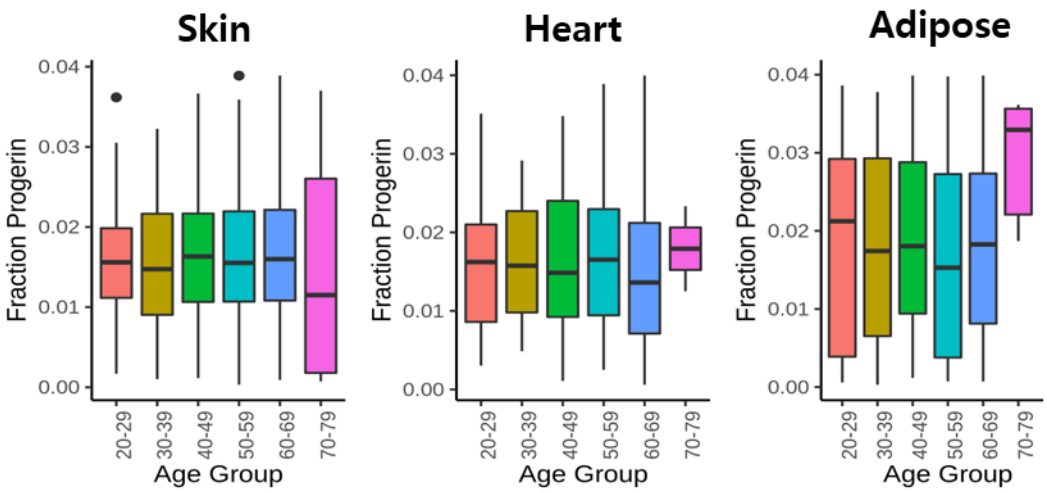


Figure 2-5 Fraction progerin of skin, heart and adipose tissues

Fraction progerin of skin, heart and adipose tissues were visualized in box plots based on their age. The age bins were divided into 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79. In all three tissue types, relationship between age and FP could not be determined.

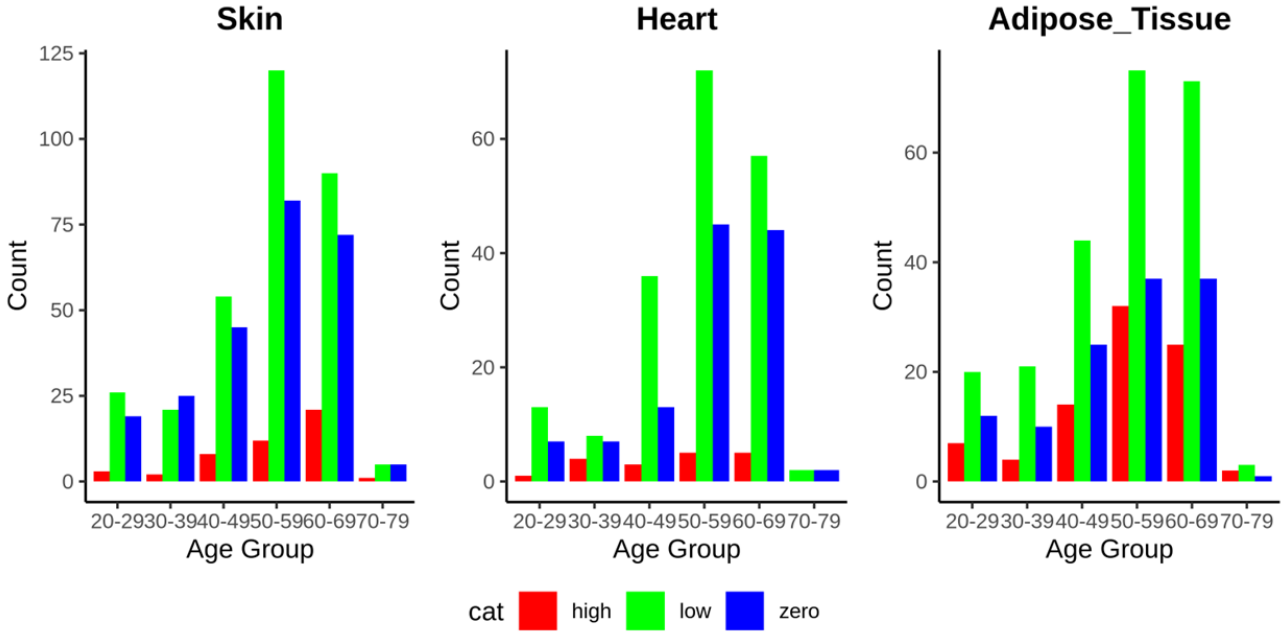


Figure 2-4 Count of zero, low and high progerin individuals in each age categories

The distribution of number of each progerin categorized GTEx samples are shown for age bins 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79. The samples were taken from skin, heart, and adipose tissues.

First, to investigate the age association in progerin-containing samples, we plotted the

distribution of FP grouped by age (Figure 2-5). The data was divided into age bins, in increments of 10 (bins: 20-29, 30-39, 40-49, 50-59, 60-69, 70+), and was examined in skin, heart, and adipose tissue samples. We found that the median values for FP were centered around 0.015, 0.016 and 0.017 across all age categories for skin, heart, and adipose tissue, respectively. The distribution also shows that progerin expression has no substantial correlation with chronological age in these samples (Figure 2-4 and Figure 2-5). However, interestingly, there appears to be a significant increase in FP (to about 0.33) in the 13 samples from adipose tissue from the age 70+ cohort ($p = 0.0344$ for groups 20-69 vs. 70-79, and $p = 0.0384$ for groups 60-69 vs. 70-79).

Next, we asked whether progerin expression associates with some molecular signatures of aging. One such biomarker of aging is the shortening of telomeres (Kan Cao,

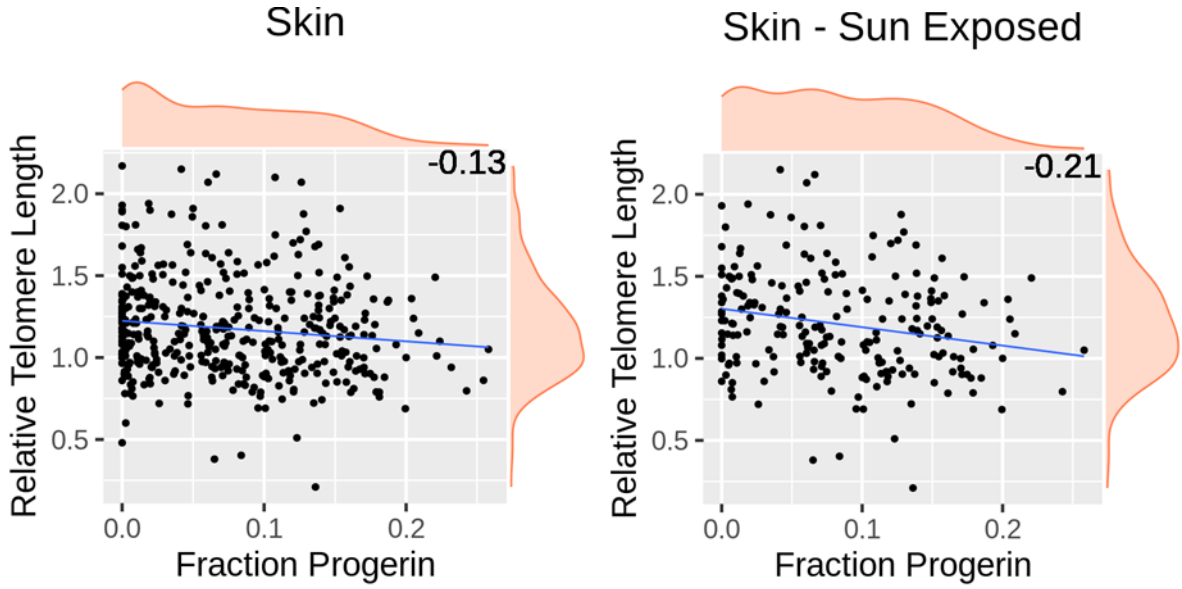


Figure 2-6 Fraction progerin against relative telomere length in skin and skin – sun exposed samples

Relative Telomere Length of skin and sun exposed skin samples from GTEx was plotted. Both skin data show negative correlation between fraction progerin and relative telomere lengths, but the relationship is more profound in sun exposed skin samples.

Blair, Faddah, & Al, 2011; Grammatikakis et al., 2014). To test the correlation between progerin expression and telomere lengths, we utilized the GTEx Relative Telomere Length

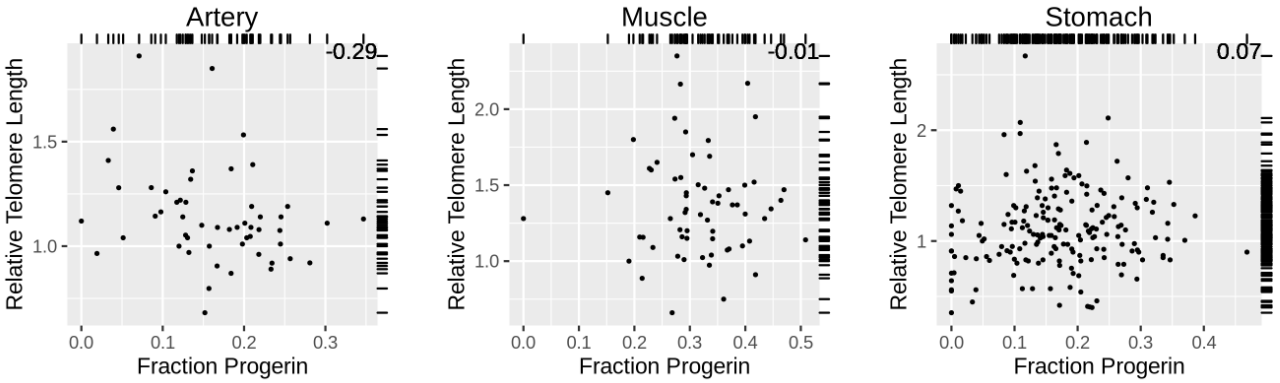


Figure 2-7 Fraction progerin against relative telomere length in artery, muscle and stomach samples

Relative Telomere Length of artery, muscle, and stomach samples from GTEx was plotted. The relationship between FP and RTL in muscle and stomach was close to none. However, there was a visible negative correlation in arterial samples.

(RTL) data from Demanelis et al (Demanelis et al., 2020), in which RTLs in various GTEx tissue samples were calculated by determining the telomeric repeats relative to a standard DNA sample. The correlations between FP and RTL in artery, muscle, skin, and stomach samples were examined (Figure 2-6, 2-7). A negative correlation (r = -0.13) between FP and RTL was observed in skin tissue (Figure 2-6, left). Interestingly, the strength of correlation was strongest in sun-exposed skin samples, with a correlation coefficient of r = -0.21 (Figure 2-6, right). These results support that in these normal GTEx samples, progerin abundance is negatively correlated with telomere lengths. When FP and RTL were examined in muscle and stomach tissues, no relationships were observed for muscle and stomach (r = -0.01, r = 0.07, respectively). However, negative correlation was observed for artery (r = -0.29). This may be related to progerin expression being observed in cardiovascular aging, specifically in coronary artery (Olive et al., 2010).

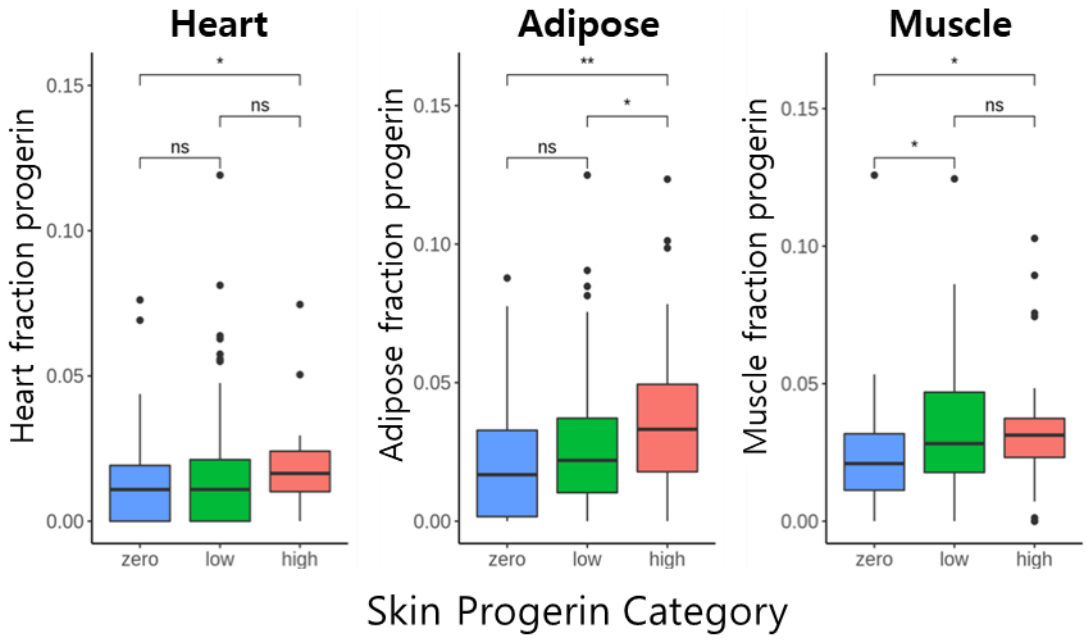


Figure 2-8 Fraction progerin of zero, low, and high progerin for skin samples in other tissues

Quantification of fraction progerin levels of samples with zero, low and high progerin expression in skin tissue compared to heart, adipose and muscle tissue (t test, *: $p \leq 0.05$, **: $p \leq 0.01$).

We also examined the progerin expression patterns in different tissues from the donors who have progerin expression in skin tissue. Fraction progerin of samples categorized based on skin tissue levels was measured in heart, muscle, adipose tissue (Figure 2-8). Fraction progerin is weakly but significantly correlated across different tissues ($p < 0.05$). This result suggests that if progerin is expressed in one tissue, it is likely that progerin is also expressed in other tissues from the same donor.

2.3 Summary

Our initial characterization and investigation of progerin expression on GTEx data was conducted. Progerin quantification was made using a more splicing-centric method called fraction progerin, which measures the fraction of all *LMNA* isoforms that does not utilize the full size exon 11. In this case, that turns out to be only progerin. Using FP, progerin quantification was made *in silico* in brain, blood, adipose, heart, skin and muscle samples. The FP distribution was used to create categorization to do downstream RNA-seq analyses (zero, low and high FP samples). Progerin expression's association with aging could not be determined, possibly due to the low resolution of the data being in bins of 10 years. However, negative correlation between FP and telomere length was observed in skin samples. The negative correlation became stronger if we take the subset of the skin sample so that only sun-exposed skin samples were examined. There was a negative correlation between FP and telomere length in arterial samples, although less samples were available to do the correlation analysis compared to skin. There were no correlation between FP and telomere length in stomach and muscle. Lastly, we have uncovered that if progerin is expressed in one tissue, it is likely that progerin is also expressed in other tissues when examining the same donor.

Chapter III : RNA-seq analyses of GTEx data

3.0 Introduction

In the previous section, categorizations of GTEx samples were made based on the distribution of the FP observed in brain, blood, adipose, heart, skin, and muscle tissues. The three categories were zero (FP < 0.005;), low (FP = 0.005 ~ 0.04) and high (FP > 0.04). In this section, comparative study between the zero FP samples and high FP samples have been conducted. The comparative study included differential expression analysis to identify candidate transcripts that are differentially expressed between zero and high FP samples, gene ontology analysis to identify commonly associated biological processes that the differentially expressed transcripts are involved in, and gene set enrichment analysis to validate and strengthen the outcome from the differential expression analysis and gene ontology analysis. Based on the outcome of the aforementioned analyses, we have conducted transcriptome-wide correlation analysis to assess the changed in splicing landscape.

3.1 Differential expression analysis (DEA)

With the previously-made categorization of the FP of the samples based on FP distribution (Figure 2-3), a comparative study needed to be conducted to identify any transcripts or biological functions that are associated with progerin expression. RNA-seq analyses were performed to identify potential candidate transcripts. Specifically, we conducted differential expression analyses contrasting zero-progerin samples with high-progerin samples from skin and heart tissues. These tissues were selected because of the relatively large number

of high-progerin samples (Figure 2-1, 2-3) and the prominence of phenotypes in those tissues in HGPS patients (Eriksson et al., 2003; Hennekam, 2006). There are a total of 890 and 412 samples from skin and heart, respectively. As shown in the volcano plots (Figure 3-1), 10,698

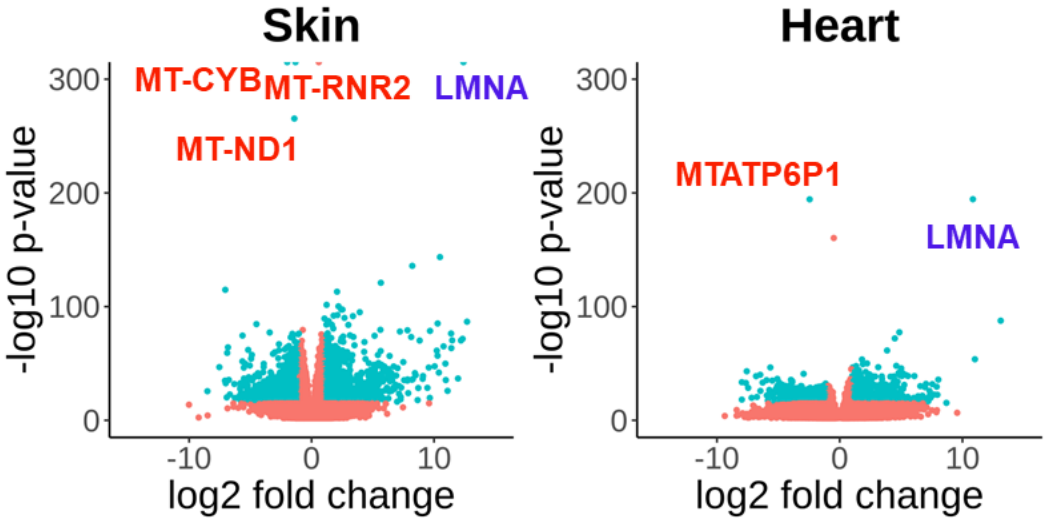


Figure 3-1 Volcano plot of differentially expressed transcripts in skin and heart samples

Volcano plot of differentially expressed transcripts in skin and heart tissues is being shown. The red label corresponds to down-expressed transcripts and blue corresponds to up-expressed transcripts. The cutoff value for log2 fold change was 1.5, and the cutoff value for -log10 p-value was 5.

and 8,393 transcripts were up-expressed in progerin-expressing samples from skin and heart, respectively; while 9,807 and 8,313 transcripts were down-expressed in skin and heart, respectively.

3.2 Gene ontology (GO) and Gene set enrichment analysis (GSEA)

Next, the sets of differentially expressed transcripts in skin and heart samples were compared (Figure 3-2). There were 2,764 commonly positively correlated transcripts and 2,778 commonly negatively correlated transcripts between skin and heart samples. Gene ontology (GO) analyses of these common transcripts revealed that the top 1,000 commonly down-expressed transcripts obtained from the ranking of p-values concentrated in GO terms highly related to splicing ($p < 10^{-8}$) (Figure 3-3). Splicing GO terms were also associated with 55 and 50 genes encoding transcripts differentially expressed in skin and heart samples, respectively. The GO analyses in each individual tissue also showed splicing GO terms ranked high with significant p-values.

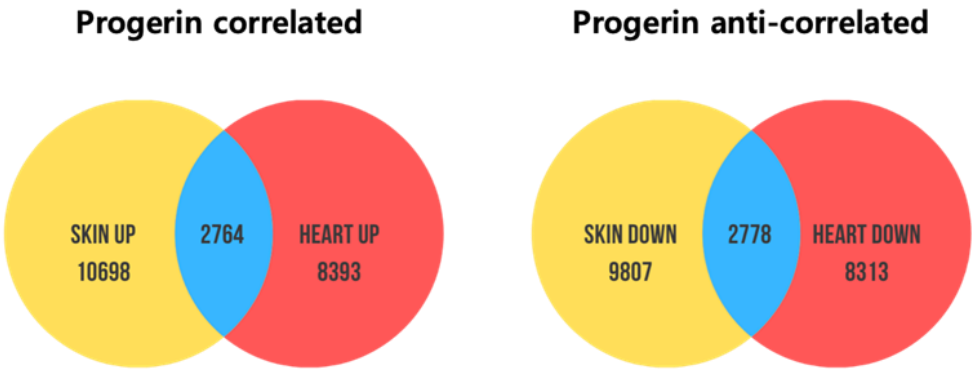


Figure 3-2 Venn diagram of correlated and anti-correlated differentially expressed transcripts in skin and heart tissues

Overlap of differentially expressed transcripts in skin and heart tissues are being shown for both up and down-expressed transcripts. 2,764 transcripts were up-expressed, and 2,778 transcripts were down-expressed in both skin and heart tissues.

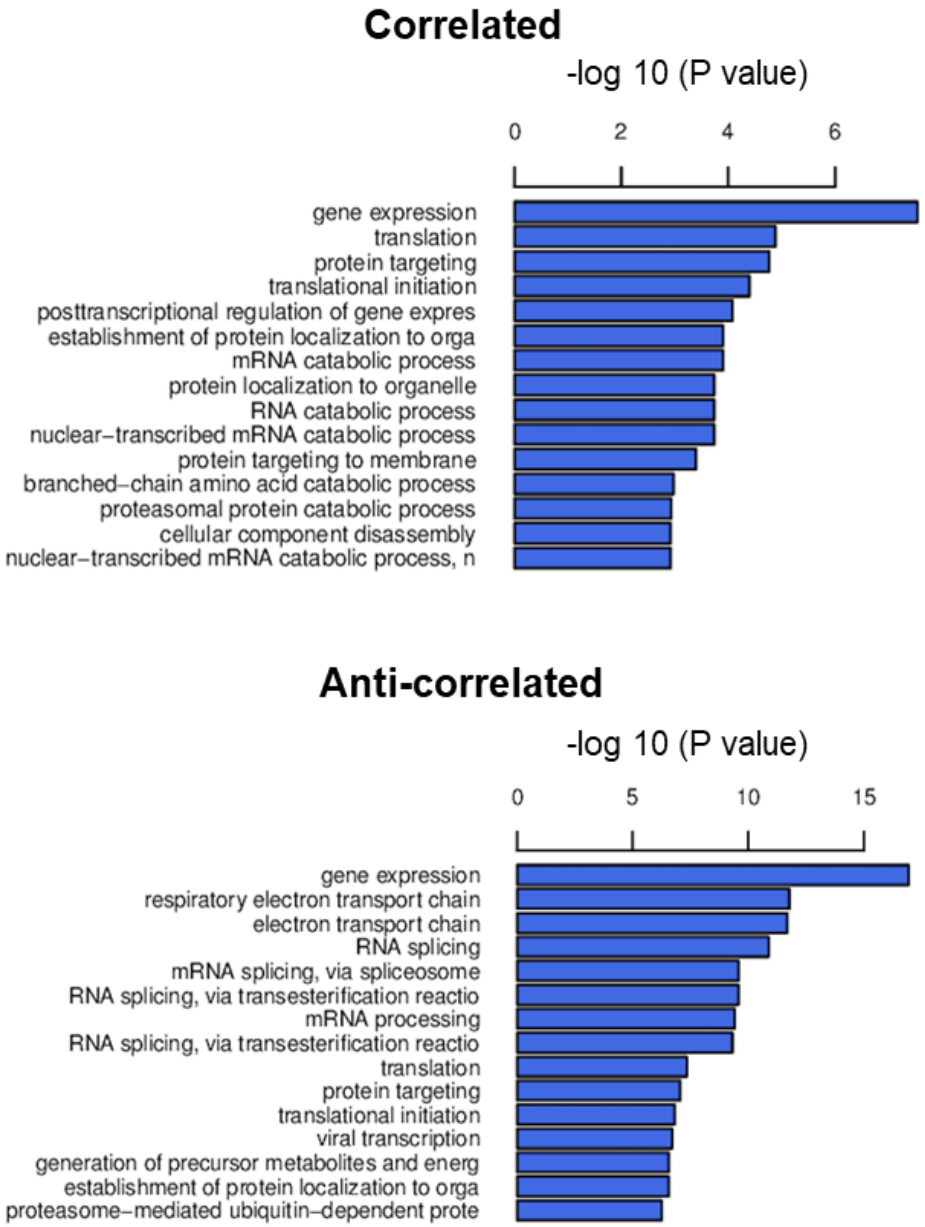


Figure 3-3 Gene ontology analysis of the overlapped correlated and anti-correlated transcripts in skin and heart

Gene ontology analysis was made on overlapping down-expressed transcripts in skin and heart tissues. Biological processes related to splicing GO terms are being shown. Terms such as RNA splicing, RNA splicing via transesterification reaction and mRNA splicing via spliceosome are shown to be significant (p value < 10^{-8}).

We further confirmed our finding from the GO analysis by conducting gene set

enrichment analysis (GSEA) (Korotkevich et al., 2021). GSEA was done on skin data, considering all the transcripts from both up and down-regulated sets. GSEA of the biological process (BP) of the skin data revealed that mRNA processing is the top enriched term (NES = 3.70, p.adj = 3.3e-69) (Figure 3-4). Zooming in specifically the splicing-related terms within the mRNA processing, the result indicates that RNA splicing (NES = 3.67, p.adj = 5.1e-62),

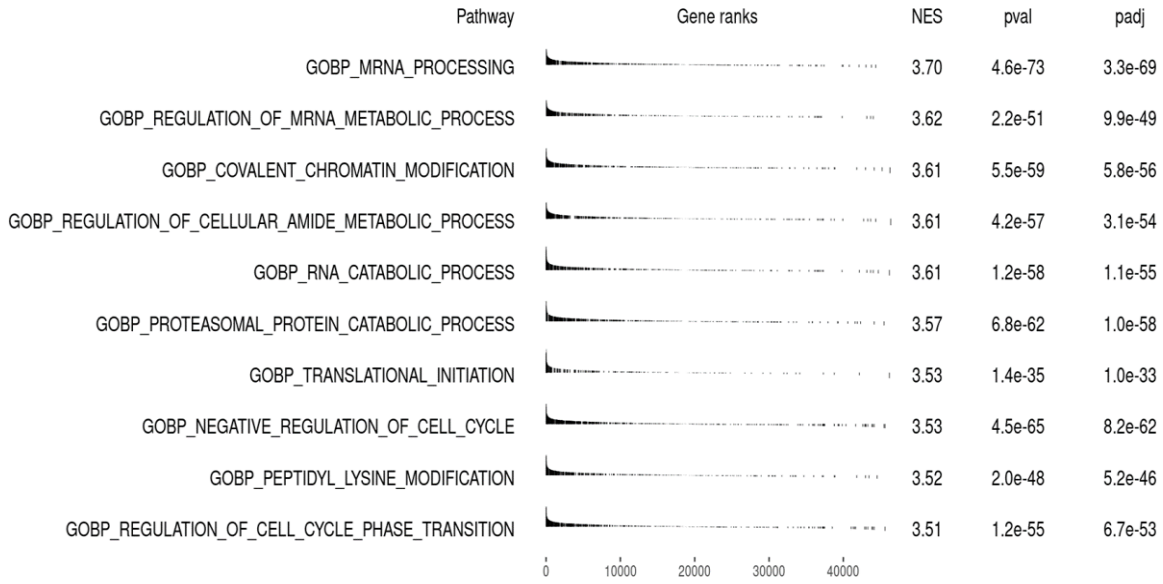


Figure 3-4 Gene set enrichment analysis of the altered expressed transcripts in skin sample

Gene set enrichment analysis (GSEA) was done on differentially expressed transcripts in skin samples. Top ranked biological processes were shown in the figure. mRNA processing term is ranked the highest (NES = 3.70, p.adj = 3.3e-69).

RNA splicing via transesterification reactions (NES = 3.57, p.adj = 6.3e-48), and regulation of RNA splicing (NES = 3.41, p.adj = 9.5e-26) are the top 3 terms with a high significance level (Figure 3-5). The results from GSEA align with the GO analysis, suggesting that transcripts from multiple splicing-related genes are differentially expressed.

Some of the high contributing genes (leading edge) were PRMT5, RBM5, SRSF2 and SRSF7, all of which are known to be involved in major splicing pathways. We found that a transcript from protein arginine methyltransferase 5 (*PRMT5*), which is a well-known splicing regulator and transcriptional cofactor (Bezzi et al., 2013; Hamard et al., 2018; Sanchez et al., 2010; Tan et al., 2019), is significantly down-expressed with progerin expression (FDR < 0.05)

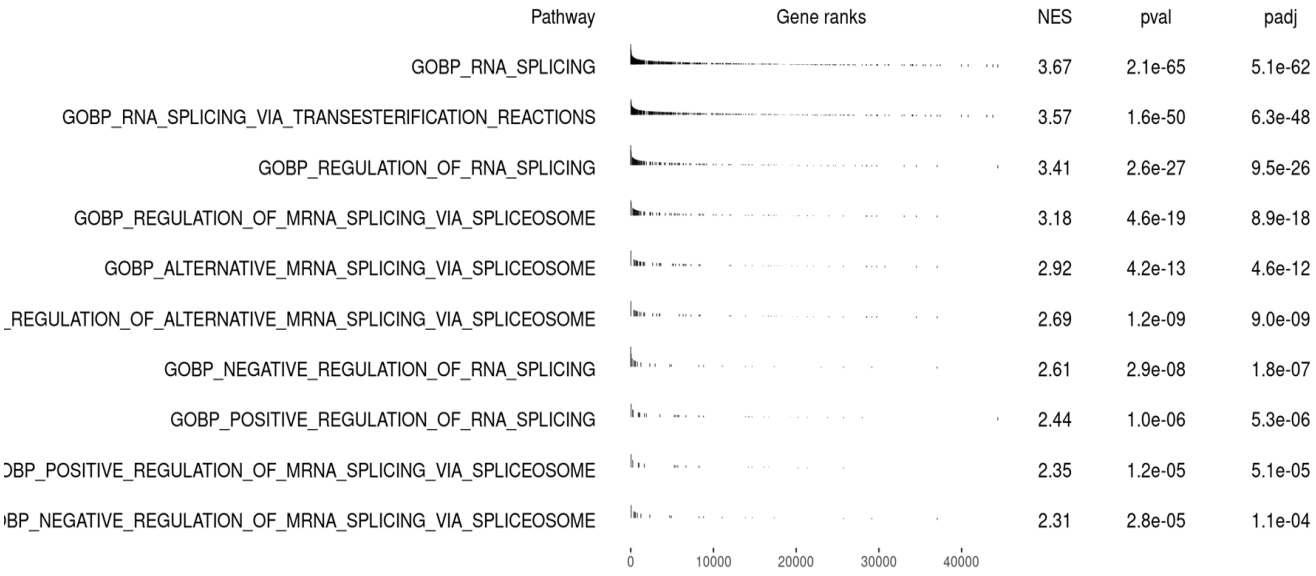


Figure 3-5 Subset of gene set enrichment analysis, consisting of splicing related terms

Gene set enrichment analysis revealed mRNA processing term to be highest ranked with NES = 3.70, p.adj = 3.3e-69. Within the mRNA processing term, RNA splicing related terms were ranked significantly high, with RNA splicing to be at top with NES = 3.67, p.adj = 5.1e-62 and RNA splicing via transesterification reactions with NES = 3.57, p.adj = 6.3e-48.

in both skin and heart samples. Preliminary correlation analysis shows that in all GTEx samples, gene-level PRMT5 expression is negatively correlated with progerin expression (Figure 3-6).

Transcript	Gene	log FC	P-Value
ENST00000361789.2	MT-CYB	-1.985	~ 0
ENST00000387347.2	MT-RNR2	-1.315	~ 0
ENST00000361390.2	MT-ND1	-1.406	4.323E-266
ENST00000557443.1	PRMT5	-5.638	3.184E-75
ENST00000535326.1	PRPF19	-7.070	1.377E-35

Figure 3-6 Notable differentially expressed transcript.

Along with mitochondrial transcripts, ENST00000557443.1, a transcript of key splicing regulator PRMT5, is significantly down expressed. ENST00000535326.1, a transcript of pre-mRNA processing factor PRPF19, is also significantly down expressed. Significance is noted by the log fold change (log FC) and the p-value.

The results combined hint towards a potential alteration in splicing patterns when progerin is being expressed.

In addition, multiple transcripts from mitochondrial genes, including cytochrome b (MT-CYB), 16S RNA (MT-RNR2) and NADH-ubiquinone oxidoreductase chain 1 (MT-ND1) are significantly down-regulated in high-progerin samples (FDR < 0.01, Figure 3-7). This result implies that samples high in progerin typically have low levels of mitochondrial activities,

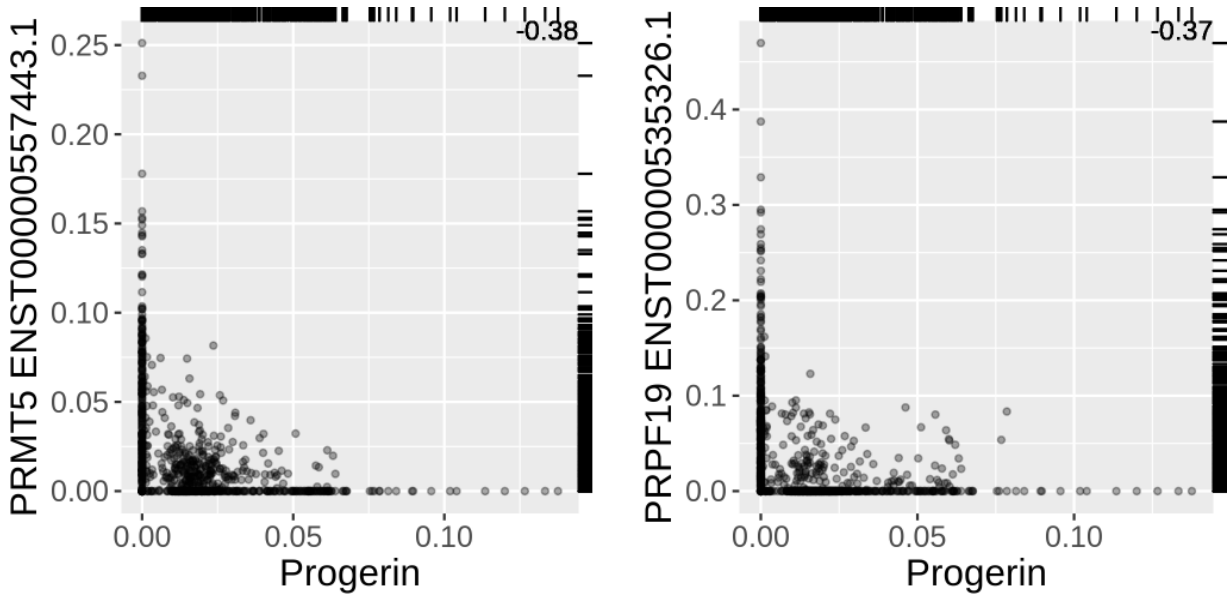


Figure 3-7 Preliminary correlation analysis of progerin expression against FP in PRMT5 and PRPF19

Pearson correlation plot of PRMT5 transcript ENST00000557443.1 and PRPF19 transcript ENST00000535326.1 against fraction progerin levels on all samples show a clear downward trend of PRMT5 as FP level increases.

which is consistent with previous studies (Xiong et al., 2016).

3.3 Transcriptome-wide correlation analysis

We observed both positive and negative correlations between progerin expression and isoforms from some genes, suggesting a widespread mis-regulation of splicing in progerin-containing samples. To characterize how progerin expression is correlated with the expression pattern of all known transcripts, we conducted transcriptome-wide correlation analyses in skin samples. Here, all samples and all transcripts were used for the analyses. Many transcripts that were highly correlated and anticorrelated with progerin expression (Pearson's $r > 0.6$ and $r < -0.6$). Remarkably, among the 200 transcripts with the greatest absolute value of r, we found 84 genes with both correlated and anticorrelated transcripts, suggesting widespread switches in alternative splicing correlated with progerin expression (Figure 3-8, 3-9). For example, RPL4 transcripts RPL4-006 and RPL4-018 ($r = 0.63$, $r= -0.53$) and SLC25A6 transcripts SLC25A6-

Transcript	Gene	r
ENST00000369299.3	LMNA	1
ENST00000561775.1	RPL4	0.63
ENST00000484026.1	SLC25A6	0.62
ENST00000270792.5	SH3BGRL3	0.61
ENST00000393564.2	G6PD	0.59
ENST00000489915.1	RPL6P27	0.58
ENST00000571732.1	YWHAE	0.57
ENST00000473620.1	PPDPF	0.57
ENST00000568562.1	ATP6V0C	0.56
ENST00000547387.1	RP11-386G11.10	0.55
ENST00000217652.3	MYL12A	0.54
ENST00000407537.1	COMT	0.52
ENST00000392168.2	TCP1	0.52
ENST00000261443.5	CSDE1	0.52
ENST00000536647.1	GNAI2	0.51
ENST00000398603.1	GSTP1	0.51
ENST00000340687.6	SERPING1	0.51
ENST00000331457.4	DRG1	0.51
ENST00000394214.3	RPL18AP3	0.5
ENST00000357865.2	STMN1	0.5

Transcript	Gene	r
ENST00000546120.1	UBC	-0.5
ENST00000540828.2	PTMS	-0.5
ENST00000476216.1	NDUFA10	-0.5
ENST00000469673.1	DRG1	-0.5
ENST00000462885.1	RPL18AP3	-0.5
ENST00000396861.1	GAPDH	-0.51
ENST00000250498.4	DAD1	-0.51
ENST00000578038.1	MYL12A	-0.52
ENST00000576965.1	RNF167	-0.52
ENST00000552893.1	RP11-386G11.10	-0.52
ENST00000533308.1	COMMD9	-0.52
ENST00000565723.1	RPL4	-0.53
ENST00000393562.2	G6PD	-0.53
ENST00000370179.3	PPDPF	-0.53
ENST00000524204.1	NPM1	-0.56
ENST00000583065.1	RPL6P27	-0.58
ENST00000466227.2	YWHAE	-0.58
ENST00000493597.1	RNF187	-0.59
ENST00000381401.5	SLC25A6	-0.61
ENST00000319041.6	SH3BGRL3	-0.61

Figure 3-8 Pearson transcriptome-wide correlation analysis with correlation coefficients

Table of the Pearson correlation analysis of GTEx skin samples. The correlation analyzed between fraction progerin and fraction of all factors in the transcriptome.

003 and SLC25A6-001 (r = 0.62, r = -0.61, respectively), illustrate this pattern of reciprocal correlations (Figure 3-10). In many cases, these are simple differences an alternative splicing

event; in the case of SLC25A6, exon 2 is retained in the isoform SLC25A6-003 but skipped in SLC25A6-001 (Figure 3-10). This was the case for many of the top-ranked genes found from

Transcript name	Gene name	r	Transcript type
ENST00000561775.1	RPL4	0.63	Retained intron
ENST00000565723.1		-0.53	lncRNA
ENST00000484026.1	SLC25A6	0.62	lncRNA
ENST00000381401.5		-0.61	Protein coding (primary)
ENST00000270792.5	SH3BGRL3	0.61	Protein coding (primary)
ENST00000319041.6		-0.61	Protein coding
ENST00000393564.2	G6PD	0.59	Protein coding
ENST00000393562.2		-0.53	Protein coding (primary)
ENST00000489915.1	RPL6P27	0.58	Transcribed processed pseudogene
ENST00000583065.1		-0.58	lncRNA
ENST00000571732.1	YWHAE	0.57	Protein coding
ENST00000466227.2		-0.58	Nonsense mediated decay
ENST00000473620.1	PPDPF	0.57	lncRNA
ENST00000370179.3		-0.53	Protein coding (primary)
ENST00000547387.1	RP11-386G11.10	0.55	Antisense
ENST00000552893.1		-0.52	Antisense
ENST00000217652.3	MYL12A	0.54	Protein coding (primary)
ENST00000578038.1		-0.52	lncRNA
ENST00000261443.5	CSDE1	0.52	Protein coding
ENST00000339438.6		-0.45	Protein coding

Figure 3-9 Genes with multiple top ranked transcripts from Pearson's correlation analysis that are correlated and anti-correlated to progerin expression

Table of top ranked genes with isoforms that are positively and negatively correlated with respect to fraction progerin levels. It is interesting to note that there are many genes with one isoform being strongly correlated to progerin expression and another isoform that is strongly negatively correlated to progerin expression.

the correlation analysis. The examples PRMT5, and PRPF19, two genes whose transcripts

ranked highly in our differential expression analysis, are illustrated in Figure 3-11. These widespread isoform switches suggest a program of alternative splicing for which progerin may serve as a marker.

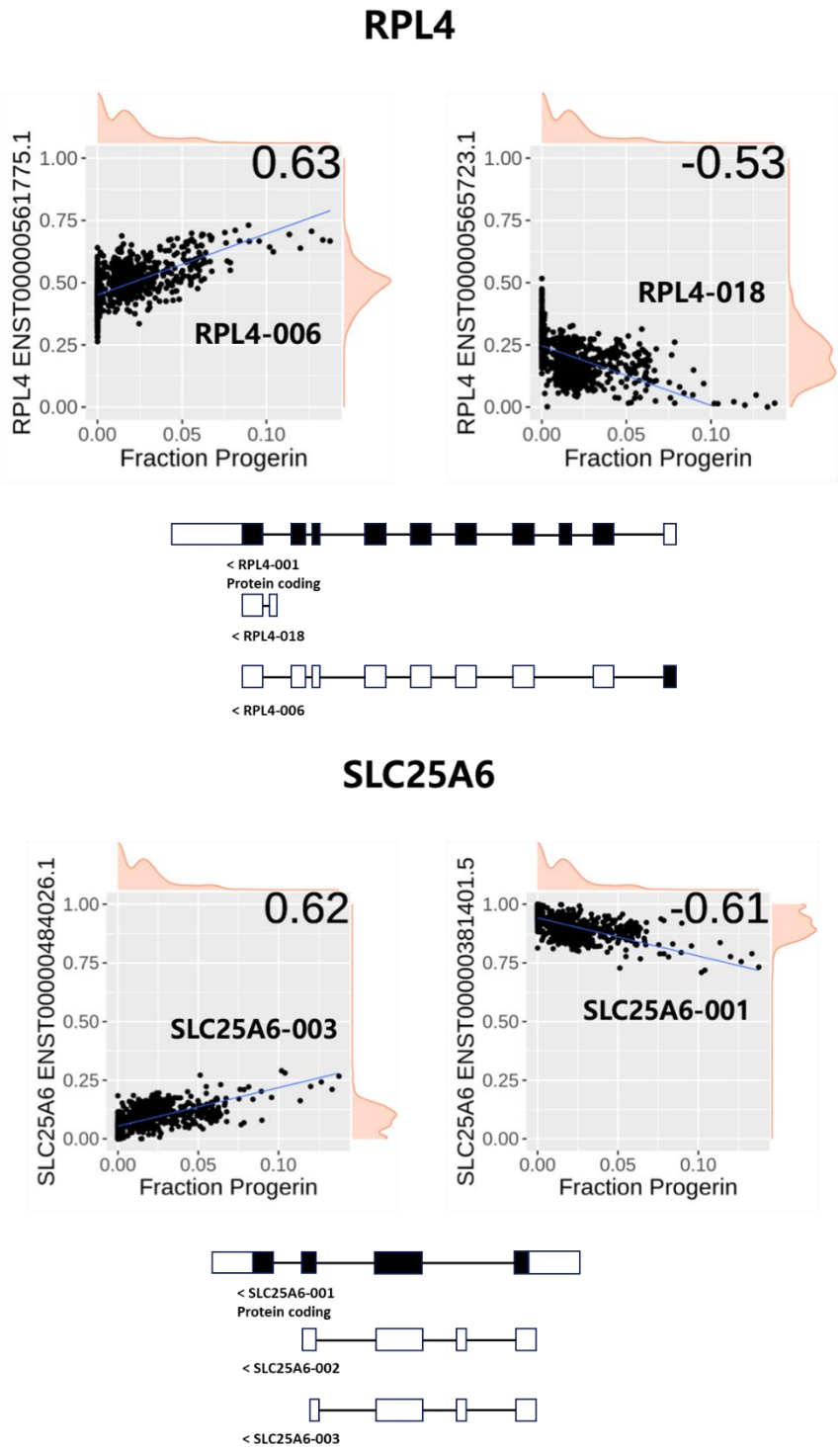


Figure 3-10 Correlation plot of FP vs expression levels of RPL4 and SLS24A6

Pearson correlation plot of RPL4 and SLC25A6, which were the genes with top ranked isoforms that are correlated and anticorrelated with increasing progerin expression. The plot shows clear upward and downward trends with respect to progerin at significant levels (r = 0.63, r = -0.53 for RPL4 and r = 0.62, r = -0.61 for SLC25A6).

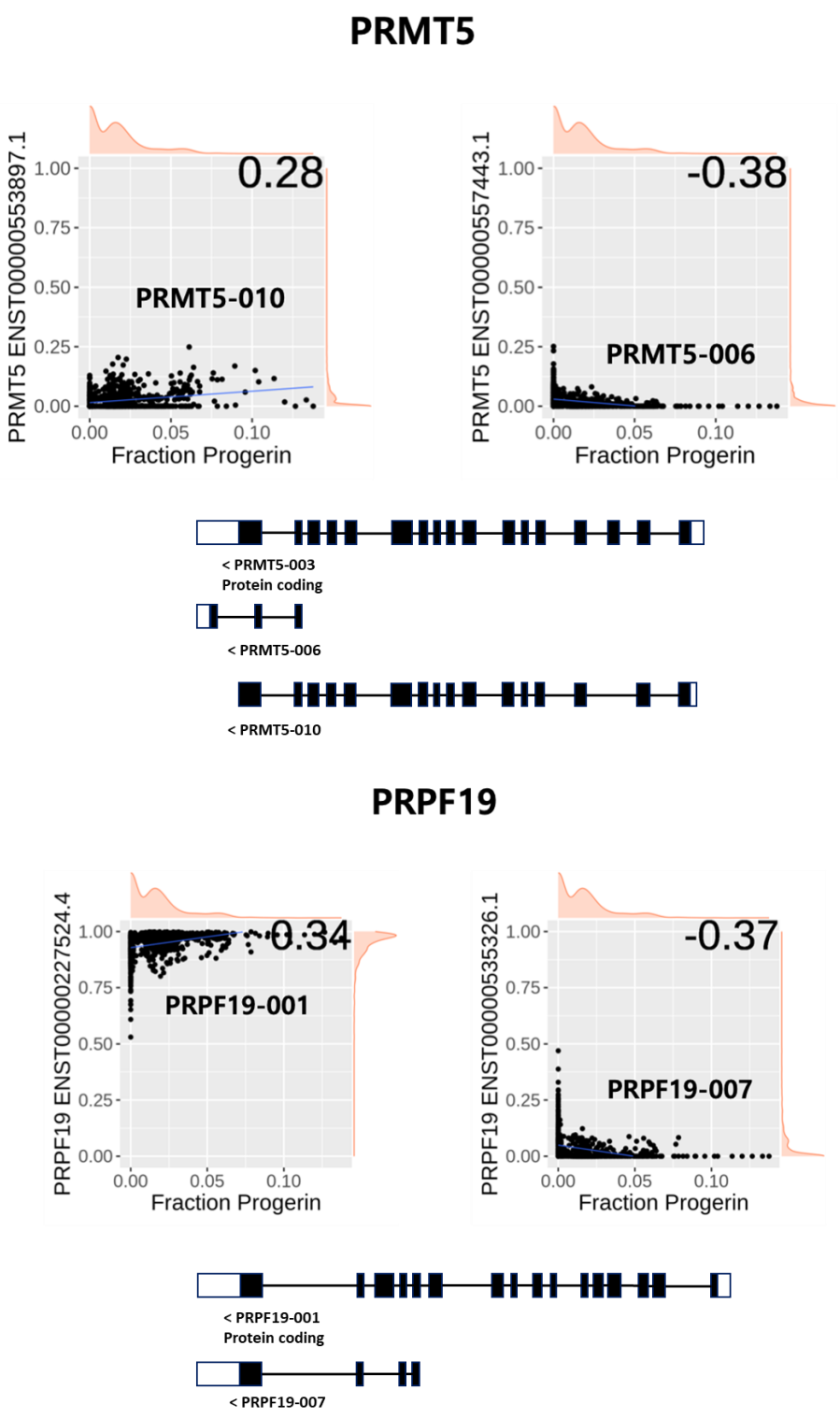


Figure 3-11 Correlation plot of FP vs expression levels of PRMT5 and PRPF19

Pearson correlation plot for PRMT5 and PRPF19 are also shown (r = 0.28, r = -0.38 for PRMT5 and r = 0.34, r = -0.37 for PRPF19). Isoform exon-intron structure for RPL4 (cor:RPL4-006 vs anti-cor:RPL4-018), SLC25A6 (cor:SLC25A6-003 vs anti-cor:SLC25A6-001), PRMT5-010 (cor:PRMT5-010 vs PRMT5-006), and PRPF19 (cor:PRPF19-001 vs anti-cor:PRPF19-007) are illustrated.

3.4 Summary

In this section, we have examined the differences of zero FP samples and the high FP samples, in the context that the samples were derived from non-HGPS normal individuals. The differential expression analyses were conducted on skin and heart samples. Skin and heart tissues were selected because of HGPS's most prominent phenotypes are observed in skin and heart. The differential expression analyses revealed that many mitochondrial transcripts were differentially expressed. Also, many transcripts from splicing related genes were differentially expressed as well. To see what the commonly affected transcripts and the biological implications of it are, the overlaps of differentially expressed transcripts between skin and heart (both correlated and anti-correlated to FP) were taken to conduct gene ontology analysis. Many splicing related terms were significantly enriched, especially in the anti-correlated category. Additional analysis was done in the form of gene set enrichment analysis. The gene set enrichment analysis revealed that mRNA processing is the top ranked term which is enriched for the differentially expressed transcripts. We took the splicing subset of the mRNA processing term and observed that RNA splicing and RNA splicing via transesterification reaction were top tanked terms that would also be among the top terms if they were transposed to the superset of the gene set enrichment analysis. To see the if there are any changes in the splicing landscape, transcriptome-wide correlation analysis was done, where we observed the correlation between FP and all available transcripts in the transcriptome. We found that there were numerous genes that had one of its isoforms to be high correlated with FP, while another isoform being highly negatively correlated, simultaneously. This showed that there are switches in the splicing pattern as FP levels increase.

Chapter IV : Mechanistic correlations to splicing events altered by progerin expression

4.0 Introduction

We have obtained a list of genes, and transcript isoforms from those genes that have demonstrated the switch in expression levels as FP levels increased. In this section, the mechanistic correlates to the progerin expression will be examined. There are various factors that affect splicing, ranging from the 5' splice site sequence to *cis* and *trans* factors that recruit splicing factors. To search for the mechanism, initial screening for differential motif enrichment was made around the 5'ss. It is crucial what type of splicing events and distinguish what type of splice sites were examined. Two types of alternative splicing events were of interests: Skipping exon and alternative 5' splice site events. Skipping exon was chosen because it is the most widely observed alternative splicing events. Alternative 5' splice site event was chosen because the mechanism in which the progerin is produced (cryptic site activation) is inherently an alternative 5' splice site event. We also were looking to identify other factors that could contribute to the changes in the splicing activity, such as GC content and exonic/intronic length. Overall, we made efforts to shed light on the mechanistic correlation seen in the switches in splicing pattern observed when progerin levels increased.

4.1 Motif analyses

To explore the mechanism behind the shift in isoforms, we used a splicing-centric quantification, percent spliced in (PSI), which is the fraction of transcripts with the longer of two products of alternative splicing. The PSI value for splicing events in GTEx samples was calculated from transcript abundances using SUPPA2, and events that are correlated or anti-correlated with progerin were identified and ranked by their Pearson's correlation coefficient (r). We focused specifically on skipped exon (SE) and alternative 5' (A5) events. Skipped exon events allow us to observe differential exon usage of one specific exon and are the most common form of alternative splicing. Alternative 5' events were of interest because progerin cryptic splice site usage falls into this category, and we focused on 5' splice sites for the same reason. We separately investigated eight types of 5' splice site sequences, according to whether

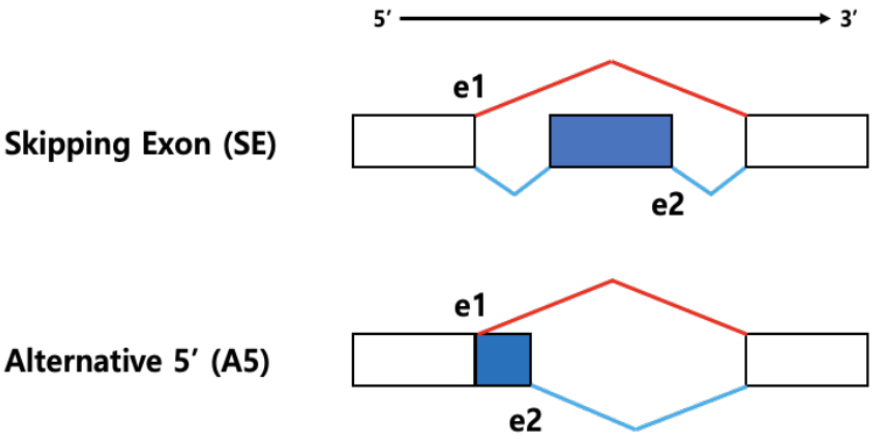


Figure 4-1 Schematic and nomenclature of the sites analyzed for skipping exon (SE) and alternative 5′ (A5) splicing events

The schema and nomenclature of which sites were examined is illustrated. e1 and e2 sites for skipping exon and alternative 5′ splicing events were examined. e1 refers to the alternatively spliced site, and e2 refers to the site that e1 is alternative to.

the event was of the SE or A5 type; whether the 5' splice site is upstream, denoted as e1, or downstream (e2); and whether the PSI values were correlated or anti-correlated.

As a preliminary screening, For each of the eight types of the splicing sites, we looked for enriched motifs using the Multiple EM for Motif Elicitation (MEME) suite (Bailey et al., 2009). MEME suite includes 4 different types of motif analyses: motif discovery, motif-motif database searching, motif-sequence database searching and assignment of function. For the purposes of preliminary motif searching, Discriminative Regular Expression Motif Elicitation (DREME) was used to identify short, ungapped motifs that are relatively enriched in the identified GTEx sequences compared to the background reference sequences, and thus obtaining the differentially enriched motifs in 5'ss of splicing events that are altered by progerin cryptic splicing (Bailey, 2011). To prepare for motif searching, the rank of the splicing events that are associated with progerin expression was sorted based on Pearson's correlation coefficient calculated from PSI values. With the new ranked list, the sequence around the 5'ss were obtained (-20bp to +20bp). Sequence around all splicing events labeled A5 or SE were obtained to be used as background sequences. Using DREME, differentially enriched motifs around 5'ss in events associated with progerin expression were searched.

For SE events, we identified 259 and 356 events whose PSI levels are correlated and anti-correlated with progerin cryptic splicing event (threshold $r \geq 0.3$), respectively. For A5 events, we identified 94 and 82 events whose PSI levels are correlated and anti-correlated with progerin cryptic splicing event (threshold $r \geq 0.3$), respectively. The background sequences were selected by identifying all the alternative splicing events that are labeled as SE or A5 (37,395 and 14,109 sequences, respectively) and were also separated by e1 and e2 sites. For the A5 and SE events identified to have $r \geq 0.3$, sequences 20 base pairs around the 5' splice site were obtained and further analyzed for differentially enriched motifs compared to the

background sequences using the MEME suite. Top-ranked differentially enriched motifs were obtained at e1 and e2 sites for SE and A5 events (Figure 4-2, 4-3, 4-4, 4-5). We observed that

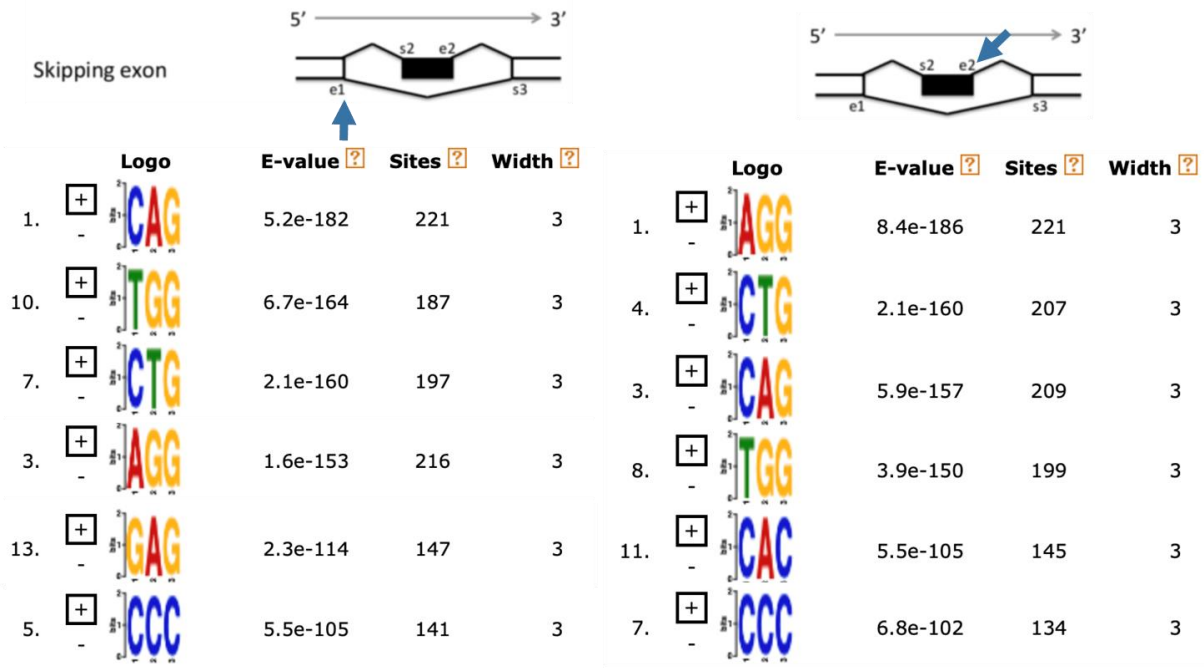


Figure 4-2 DREME differential motif enrichment analysis (Skipping exon, correlated)

Differential motif enrichment analysis was done for correlated skipping exon events at e1 splice site (left) and e2 splice site (right).

the CAG motif was consistently ranked as one of the top differentially enriched motifs, regardless of the splicing type (SE/A5) or the splice site (e1/e2). This was an interesting outcome, as we expected a motif to be differentially enriched specific to either e1 or e2. To confirm that the result is not due to the disparity in the number of sequences used for our identified events against the background, we conducted the bootstrap analysis to repeat the same differential motif enrichment analyses on a randomly selected smaller subset of sequences 100 times. The results confirmed that the CAG motif is consistently ranked high to be differentially enriched. The distribution of the CAG motif across -40bp to +40bp around the 5'ss was then analyzed. For events that were identified to be correlated or anti-correlated to

progerin splicing event and the background sequences, we observed that the CAG motif was the most abundant at the -3 position from the consensus GT motif of the 5'ss (Figure 4-7). The distribution of CAG motif occurrence across the bootstrapped sequences at the -3 position for

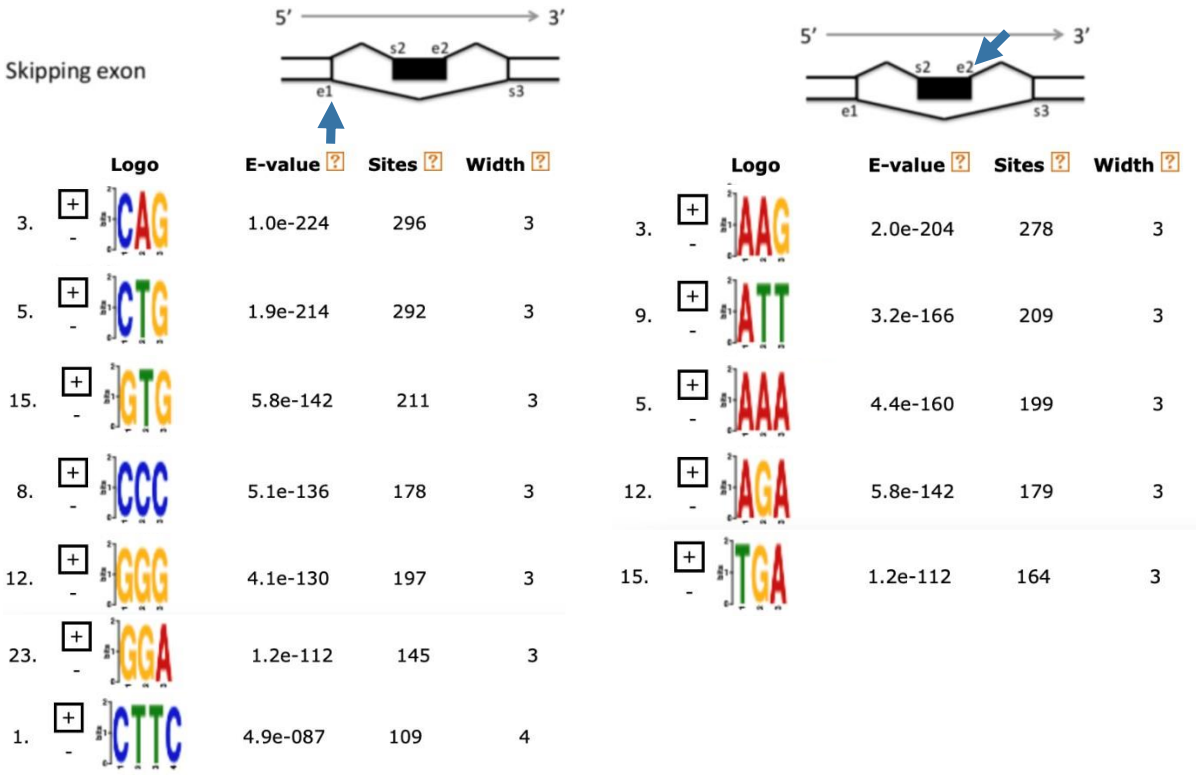


Figure 4-3 DREME differential motif enrichment analysis (Skipping exon, anti-correlated)

Differential motif enrichment analysis was done for anti-correlated skipping exon events at e1 splice site (left) and e2 splice site (right).

both background and our identified sequences were examined (Figure 4-7). Remarkably, the CAG motif distributions were significantly altered between the background and the sequences at the e2 site. Especially in the case of A5 sequences, there is a sharp decrease of the CAG motif occurrence, where we observed around 3% and 7% decrease of CAG motif for A5 correlated and anti-correlated events. For SE events, there is around 1% decrease in CAG motif occurrence in the anti-correlated events. It was interesting to see that the CAG motif occurrence for the background sequences to be roughly identical between the e1 site and e2 site, suggesting

that progerin cryptic splicing is correlated with the reduction of representation CAG motif at the e2 site in the events whose PSI levels are correlated or anti-correlated to progerin splicing event. The correlation was far stronger for A5 events.

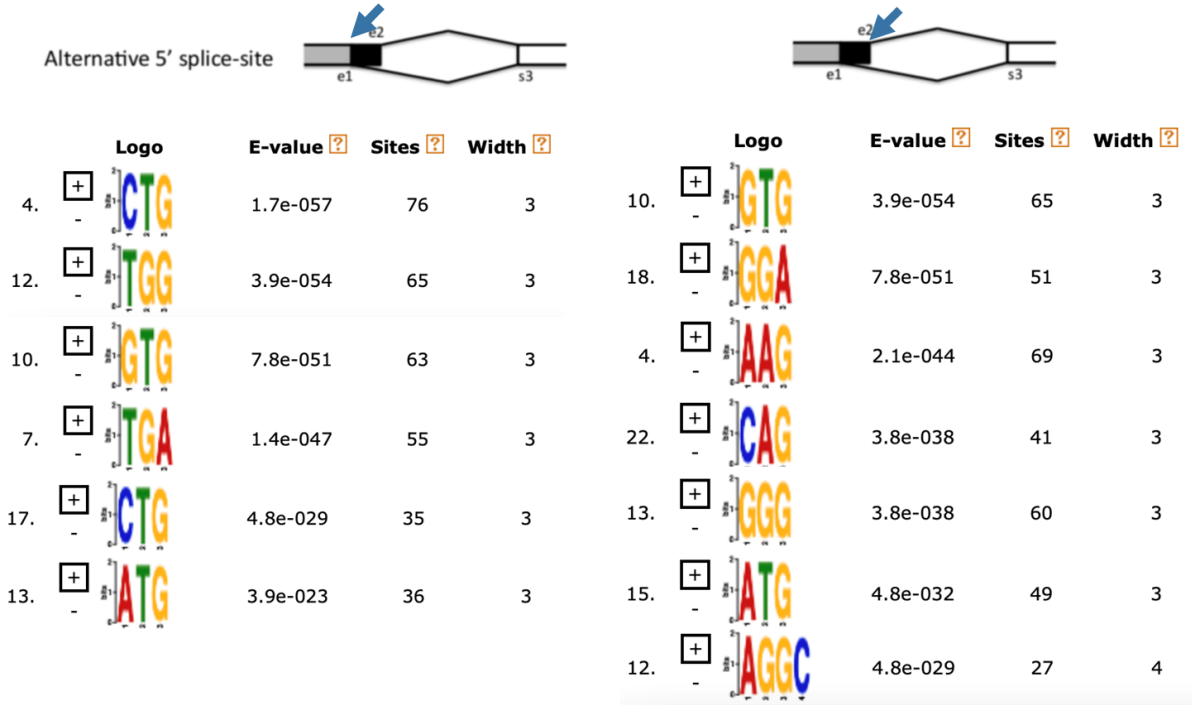


Figure 4-4 DREME differential motif enrichment analysis (Alternative 5′, correlated)

Differential motif enrichment analysis was done for correlated alternative 5′ splicing events at e1 splice site (left) and e2 splice site (right).

	Logo	E-value	Sites	Width
6.		1.2e-063	71	3
7.		1.3e-052	65	3
5.		3.7e-049	61	3
12.		3.7e-049	62	3
1.		8.7e-046	58	3
4.		8.6e-030	42	4

	Logo	E-value	Sites	Width
8.		8.2e-068	63	3
11.		1.3e-052	48	3
3.		8.7e-046	63	3
6.		8.7e-046	60	3
4.		3.1e-039	55	3

Figure 4-6 DREME differential motif enrichment analysis (Alternative 5′, anti-correlated)

Differential motif enrichment analysis was done for anti-correlated alternative 5′ splicing events at e1 splice site (left) and e2 splice site (right).

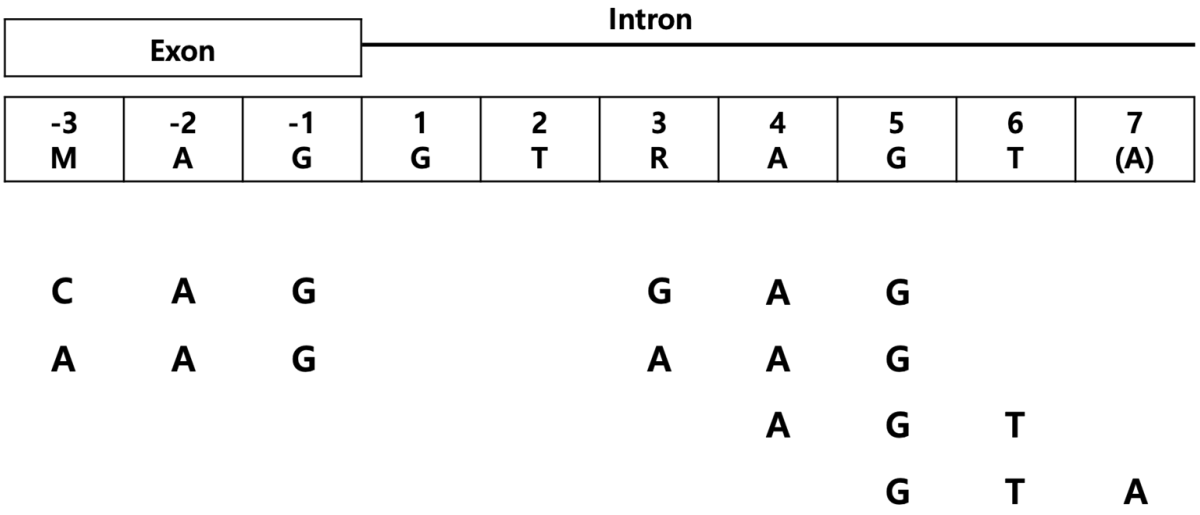


Figure 4-5 Highly conserved tri-nucleotide sequences around 5′ss

The nucleotide sequence (from -3 to +7 position) MAG|GTRAGTA is the consensus 5′ss sequence for *homo sapiens*. Tri-nucleotides abundance for CAG and AAG at -3, AAG and GAG at +3, AGT at +4, and GTA for +5 were examined.

To examine whether the effect seen in the CAG motif at the -3 position is part of a

larger phenomenon, we have expanded the search to other known motifs around the 5' splice site. We examined the -3, +3, +4 and +5 positions around the 5'ss. Motif percentage distribution

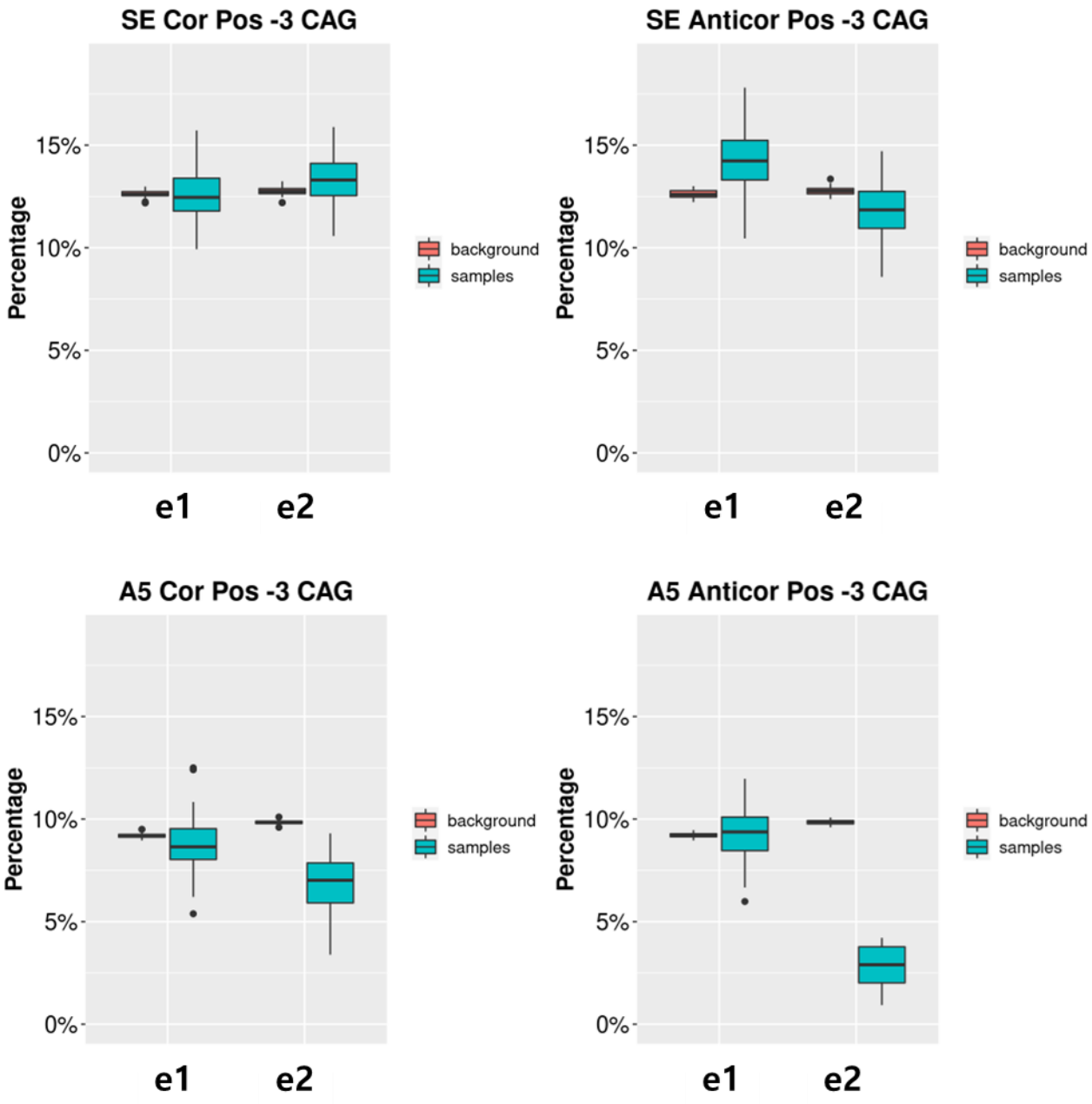


Figure 4-7 CAG tri-nucleotide representation at -3 position from 5'ss

CAG motif distribution at the -3 position from the bootstrapped runs are shown. The distribution was compared between the background and the events that were identified as either correlated or anti-correlated with progerin cryptic splicing event. Significant reduction in CAG motif representation is observed at the e2 site for SE progerin anti-correlated events, A5 progerin correlated events, and A5 progerin anti-correlated events.

for CAG and AAG at -3, AAG and GAG at +3, AGT at +4, and GTA for +5 was measured for both e1 and e2 sites. Several shifts in motif abundance were observed. The most significant of these were underrepresentation of CAG at position -3 at the e2 site, overrepresentation of AAG at -3 at the e1 sites with correlated PSI values, and underrepresentation of GAG at +3 at e1 sites with correlated PSI values.

4.2 GC content and exon/intron length analysis

To further investigate signatures that are associated with the alterations in the splicing landscape, we examined the GC content in the exons and the introns that are involved in alternative splicing events correlated or anti-correlated with progerin. In previous studies, it has been found that different splicing patterns are associated with differences in GC content (Amit et al., 2012; J. Zhang et al., 2011). For each correlated or anti-correlated A5 event identified in the transcriptome-wide correlation analysis, we examined GC content in four regions (Figure 4-8 : the upstream exon to e1, the region between the two alternative splice sites, the intron between e2 and the downstream exon and the intron e1 and the downstream exon) for the top 10, top 20 and top 50 events most highly correlated events. With these

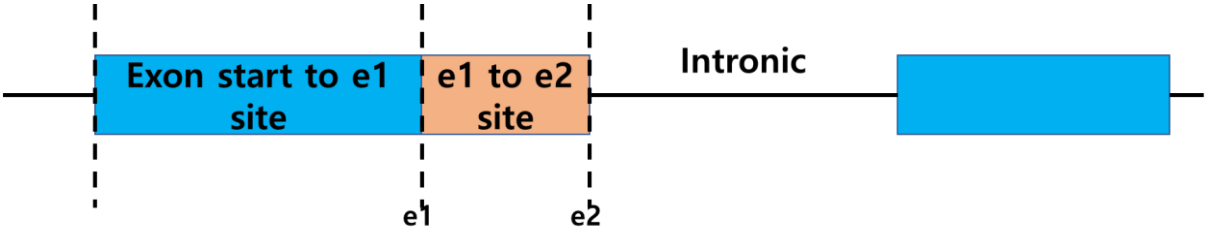


Figure 4-8 Schematic of the division of the exon and intron

The exon and intron were divided such that there are 4 different areas for investigations : First area marks the start of the upstream exon up to the alternative splice site (e1 site), second area marks the alternative splice site (e1 site) to the end up the upstream exon (e2 site), and third area marks the intron between the upstream and downstream exon. We have analyzed an additional area for the resulting intron that would be made if the e1 site used for splicing.

categorizations, also including background A5 events, we have quantified GC content for each regions, at 3 tiers of sample divisions (Figure 4-9). A5 events that are anti-correlated with respect to progerin expression have significantly reduced GC content within the introns.

We have also investigated the lengths of the respective areas. The length distribution

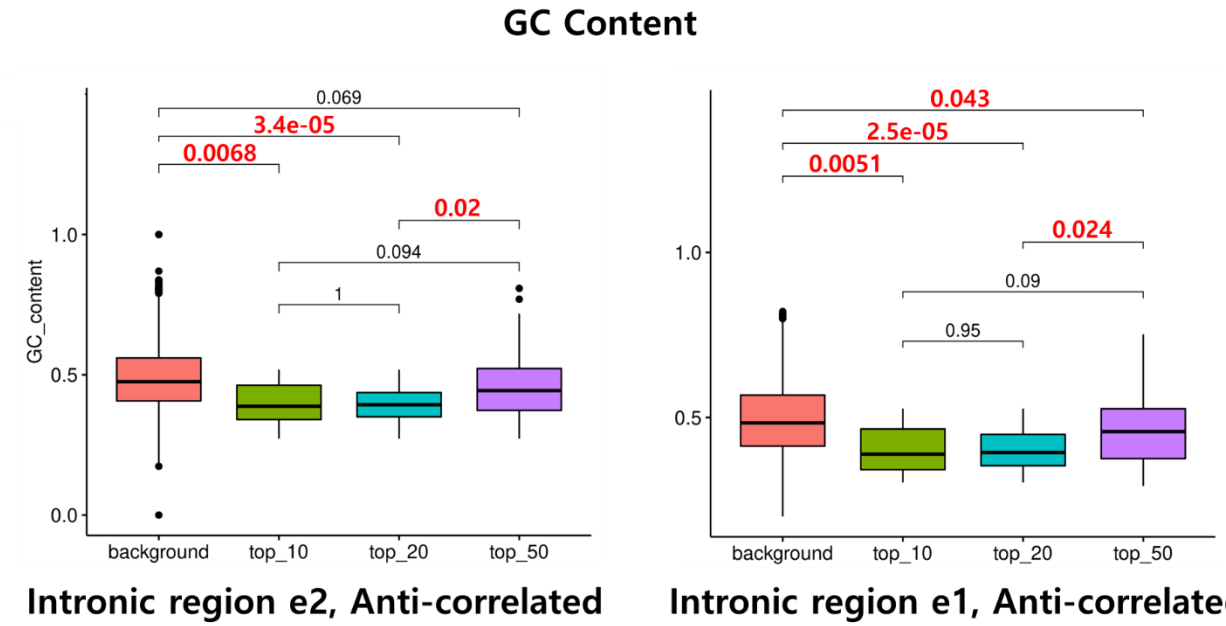


Figure 4-9 GC content comparison between background and the top 10, 20 and 50 ranked (by Pearson's correlation coefficient) A5 events

The distribution of GC content for intronic regions for anti-correlated A5 events show that GC content is reduced in the identified events compared to background (e2 background vs top 10 : p.val = 0.0068, e2 background vs top 20 : p.val = 3.4e-05, e1 background vs top 10 : p.val = 0.0051, e1 background vs top 20 : p.val = 2.5e-05, e1 background vs top 50 : p.val = 0.043).

in the tiered events were compared against the length distribution for the background set (Figure 4-10). The distance between alternative sites in anti-correlated events was significantly shorter than that of the background events (85 bp vs. 52 bp, p.val = 4.17e-04; vs. 42 bp, 3.72e-06; and vs. 44 bp, 7.76e-18 for background against top 10, top 20, and top 50, respectively).

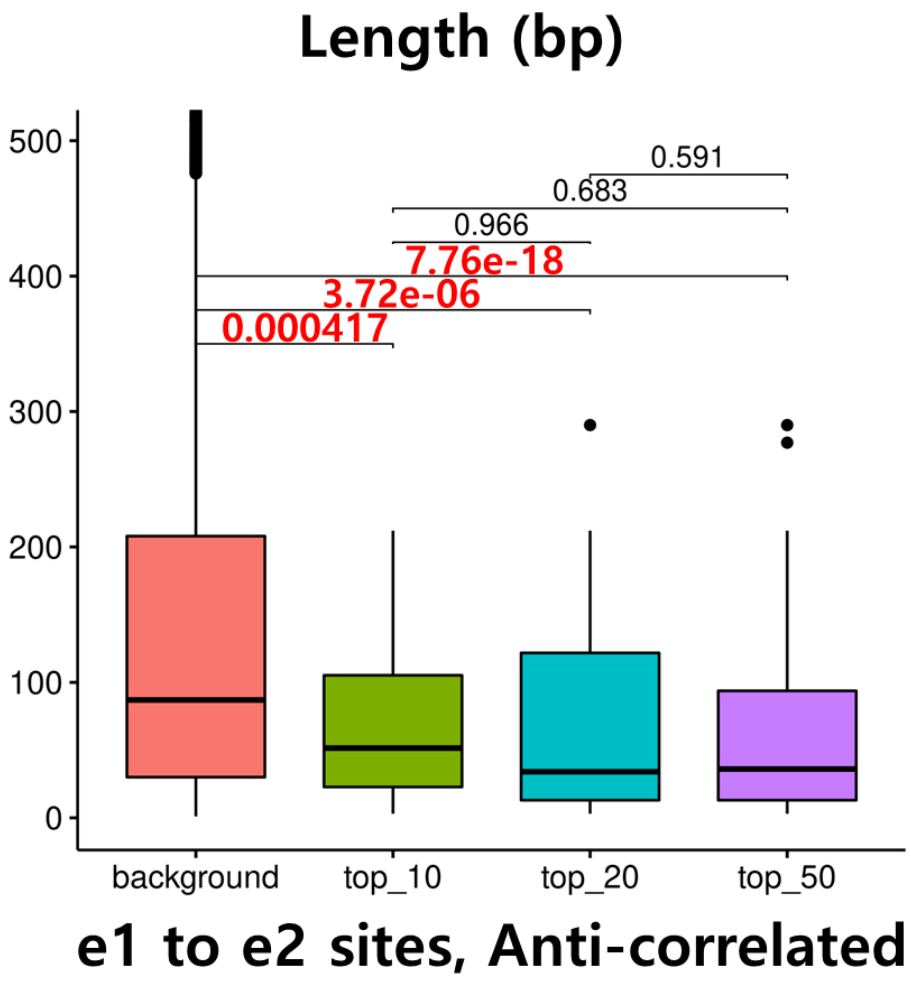


Figure 4-10 Distribution of length between e1 site and e2 site for background and the top 10, 20 and 50 ranked (by Pearson's correlation coefficient) A5 events

The distribution of length (in bp) for area between e1 and e2 for anti-correlated events reveal that area for the identified events are significantly shorter than that of the background.

4.3 Summary

Results from the previous section identified list of genes that had one of its isoforms being highly correlated, while another isoform to be highly negatively correlated. We have

made efforts to uncover the mechanistic correlation to this phenomenon. Initial differential motif enrichment analyses were made to identify candidate tri-nucleotide motif around the 5' splice site that is differentially enriched in the events identified compared to the background reference sequences. The differential motif enrichment analysis revealed that the CAG motif around the 5' splice site is significantly differentially enriched. Examining the consensus 5' splice site sequence for human, the CAG motif at the -3 position is crucial for the splice site to be competitive. We examined the CAG motif abundance at the -3 positions for e1 and e2 sites for skipping exon event and alternative 5' splice site events and uncovered that the CAG motif at -3 position is significantly differentially enriched. Especially for the alternative 5' splicing events, the CAG motif was significantly reduced compared to the background. The GC content and exonic/intronic length analyses revealed that the GC content is significantly lower in the intronic region, and that the e1 to e2 site length is significantly shorter in the events associated with progerin expression compared to the background reference sequences.

Chapter V : Summary, discussion, and future directions

5.1 Research summary

This research endeavor was first inspired by previous research work investigating phenotypes of chronological aging, specifically DNA damage induced either by natural means (sun exposure), cellular senescence, or artificial means (UV irradiation) (Kan Cao, Blair, Faddah, Kieckhaefer, et al., 2011; Lesiak et al., 2017; Takeuchi & Rünger, 2013). Also, a preliminary screening of the GTEx data revealed that GTEx samples, which are from non-HGPS normal individuals, had small amounts of progerin expression. The highest levels of progerin expression were seen in sun-exposed skin samples (Figure 2-1). The preliminary GTEx screening was done using reads per kilobase of transcript, per million mapped reads (RPKM) values. RPKM quantification is not the preferred method for comparative analyses cross samples (sample-to-sample comparisons). Using a splice-site based quantification method, Percent Spliced In (PSI), the fraction of *LMNA* isoforms that does not utilized the full size exon 11 was calculated. This, in our case, is left with only progerin and is named fraction progerin (FP). Using the FP quantification method, it was uncovered that progerin is indeed expressed in samples from 5 different tissue types (excluding brain) (Figure 2-3). After examining the distribution of progerin expression in 5 different tissues, categorization was made so that there would be labeled samples zero, low and high progerin. Progerin expression was compared with other metrics to observe any correlations between the two factors. One interesting observation was that progerin had weak to no apparent correlation with age (Figure 2-4). We observed that FP had negative correlation with telomere length in skin samples. The effect was observed at a greater strength when we subset sun-exposed samples from total skin samples (Figure 2-6).

With the categorization previously made from the FP distribution, differential expression analyses were conducted to identify transcripts differentially expressed between

high-progerin and zero-progerin samples. DEA revealed many mitochondrial transcripts to be differentially expressed in skin and heart tissue. Mitochondrial dysfunction is one of many well-known phenotypes of HGPS (Rivera-Torres et al., 2013; Xiong et al., 2016). Progerin expression in non-HGPS normal individuals also seems to mimic the mitochondrial phenotype. Aside from mitochondrial transcripts, many transcripts from splicing related genes were differentially expressed as well, such as PRMT5 and PRPF19. Deeper investigation on the overlapping correlated and anti-correlated transcripts between skin and heart tissue samples was done in the form of gene ontology (GO) and gene set enrichment analysis (GSEA). GO and GSEA revealed that the transcripts that are commonly associated with progerin expression in skin and heart tissues are involved in various mRNA splicing regulations (Figure 3-3, 3-4, 3-5). This gave us a strong idea that splicing landscape may be altered when progerin is expressed in non-HGPS normal individuals. As a preliminary screening, FP against PRMT5 and PRPF19 transcripts were examined. It was shown that with respect to FP levels, the PRMT5 and PRPF19 transcripts were negatively correlated, with correlation coefficients of $r = -0.38$ and $r = -0.37$, respectively.

A transcriptome-wide correlation analysis was conducted, with FP against all available annotated transcripts in the transcriptome in GENCODE version 19 (Harrow et al., 2012). The initial finding was very promising, seeing multiple transcripts that are either positively or negatively correlated, with high Pearson's correlation coefficients. Closer investigation of the results revealed that there were many genes that had multiple isoforms being highly correlated or anti-correlated with respect to FP, and also strikingly that while one isoform is strongly positively correlated, another isoform was strongly negatively correlated to FP. Examples from RPL4 and SLC25A6 show clear positive expression trend for one isoform when FP levels increase and clear negative expression trend for the other isoform when FP levels increase

(Figure 3-10). The two splicing related genes identified from DEA, PRMT5 and PRPF19, also showed clear positive and negative trend by two different isoforms (Figure 3-11). This confirms that splicing changes are correlated with FP levels in non-HGPS normal individuals.

I sought to elucidate the mechanistic explanation for such a widespread change in the splicing landscape. Motifs around the splice sites affect the splicing efficiency the most. To search for potential candidate motifs in top ranked splicing events, DREME was used to search for differentially enriched motif around the 5'ss of top ranked splicing events based on the correlation analyses done against FP. It was found that CAG motif at -3 position of the 5'ss seemed to be the most consistent and significantly differentially enriched motif (Figure 4-2, 4-3, 4-4, 4-5). Next, we observed the distribution of CAG motif representation at -3 position for both e1 splice site and e2 splice site. CAG motif representation for A5 events at e2 sites specifically were shown to be significantly less represented compared to that of in the background sequences (Figure 4-7). Lastly, other factors that may contribute to splice pattern changes were examined. GC content and exonic/intronic length divided at 4 different points were examined. The results revealed that GC content for intronic regions, both e1 to downstream exon and e2 to downstream exon, showed significantly reduced GC content compared to background. Exonic area between the e1 and e2 sites were shown to be significantly reduced for A5 evens that are anti-correlated with respect to FP.

We have characterized progerin expression in non-HGPS normal individuals by using GTEx database, establish differential expression of transcripts from splicing related genes in samples expressing high FP, observe widespread splicing pattern changes in splicing events when progerin is being expressed and explore the mechanistic explanations that could contribute to this phenomenon.

Combining together the findings from my research, a working model has been devised to explain what may be occurring when progerin is being expressed in non-HGPS normal individuals (Figure 5-1). We believe that there is some sort of a trigger, most likely factors derived from DNA damaging agent such as telomere attrition, DNA damaging reagent or UV irradiation, which leads to alteration of the splicing landscape. When this alteration occurs, it affects skipping exon and alternative 5' splicing events the most. Included within this alteration

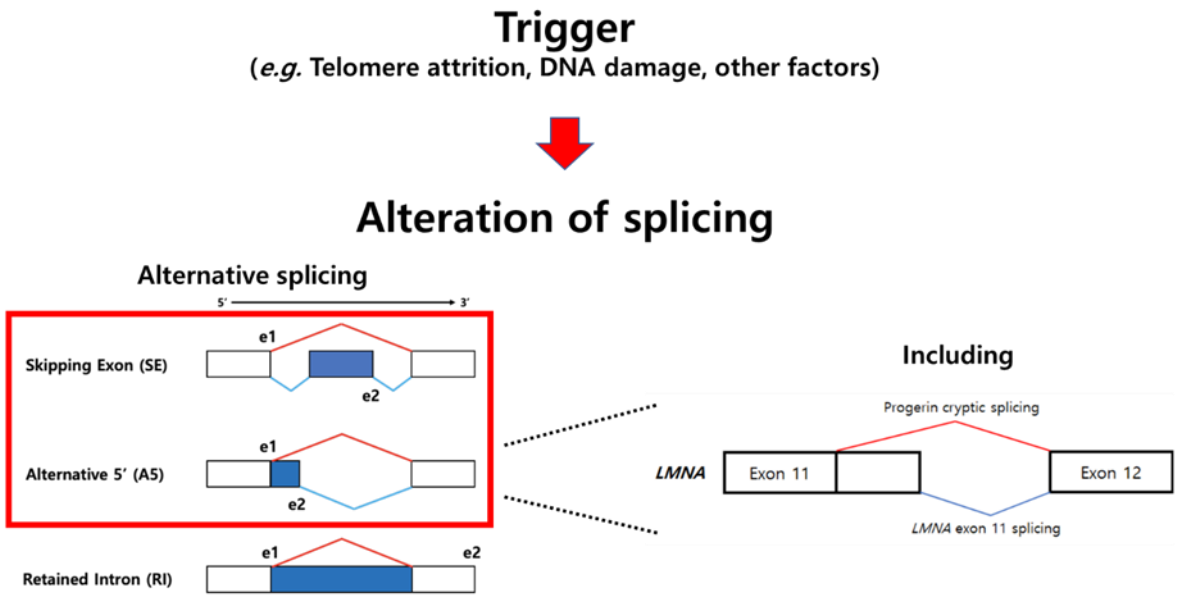


Figure 5-1 Working model

The figure illustrates the working model for the research done on progerin expression and its association with changes in the splicing landscape.

of A5 events is what we observe in *LMNA* cryptic splicing event. The cryptic splice site it utilized and produces progerin in small amounts.

5.2 Discussion

Progerin expression in normal individuals has been characterized in several previous studies. We observed that small amounts of RNA that could encode progerin are found in many GTEx samples, from all tissues (Figure 2-2, 2-3). Interestingly, use of the progerin cryptic splice site appears to activate some of the rarer isoforms, implying the correlation with the activation of other cryptic splicing events (Figure 5-2). Since HGPS is a premature aging

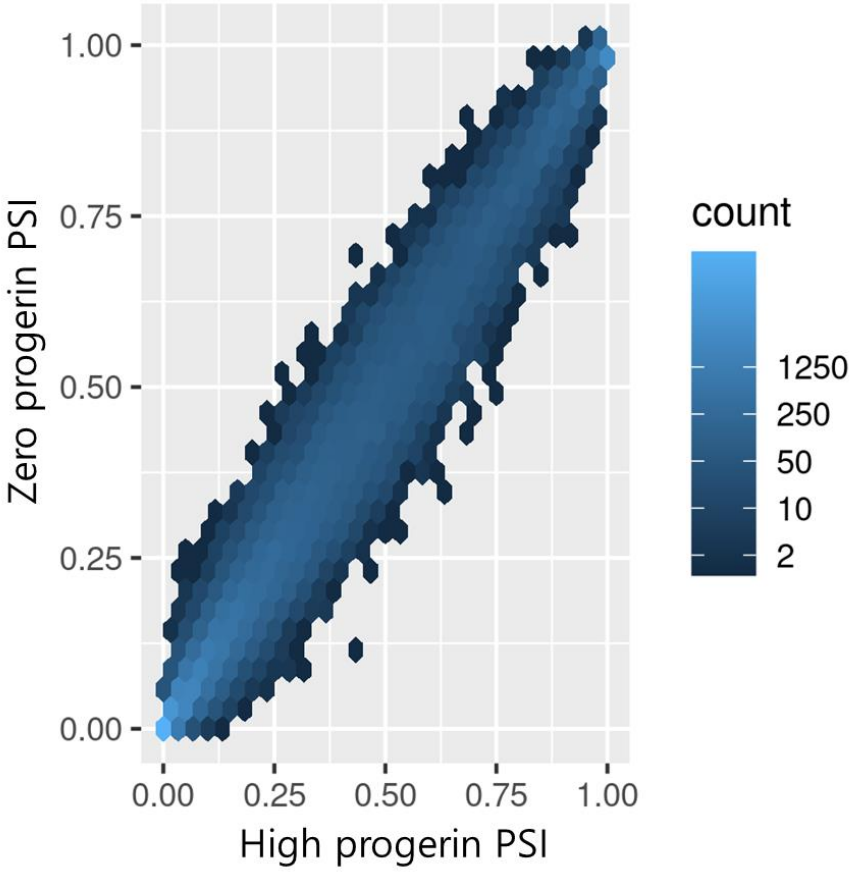


Figure 5-2 Plot of individual splicing events based on their PSI observed in zero progerin samples vs high progerin expressing samples

disease, we expected chronological age to play a role in progerin expression in normal individuals. However, we did not identify a significant association between the use of the progerin cryptic site and chronological age. However, we did find that progerin expression is

correlated with telomere shortening, a major molecular phenotype of aging (Grammatikakis et al., 2014), especially in sun-exposed skin samples. Among the top-ranked differentially expressed transcripts, we observed a number of mitochondrial transcripts that are down-expressed in the zero to high progerin sample contrast (Figure 3-1), which suggests progerin is correlated with depletion of mitochondria. Mitochondrial defects are staple phenotypes of HGPS (Mei Xiong et al., 2017; Rivera-Torres et al., 2013; Xiong et al., 2016), and remarkably this effect is also seen in non-HGPS individuals with progerin expression.

Analysis of alternative 5' splice sites (the type of splicing observed in HGPS) reveals that changes correlated with the use of the progerin site tend to involve close-spaced alternative sites. While the mean distance between alternative sites in the A5 set described by is about 85 nucleotides., the 50 most strongly correlated switches are to a downstream site on average about 30 nucleotides away (Figure 4-10). Furthermore, the 5' splice sites involved tend to have a slightly different core consensus. The majority of consensus nucleotides at positions -3, +3, +4 and +5 have distinct abundance compared to the background. Specifically, in the AG motif at -2 and G motif at +5 positions, we have observed altered abundance in our identified events compared to the background, and we saw a trend in their abundance as we loosened the r value cutoff. In general, the core 5’ splice site of both correlated and anti-correlated events, shows reduced conservation. We suspect that the altering of the representation of consensus motifs at the -3, +3, +4 and +5 positions of both e1 and e2 sites is highly correlated to the unknown mechanism which is involved in altering splicing patterns in multiple genes (Figure 5-1). And, previous studies may hint that progerin expression and its correlation to altered consensus motif representation at the 5’ss may be correlated to weaker splice site strength, and thus search for a nearby splice site candidate, leading to alternative splicing.

In accordance with previous research (Amit et al., 2012; Tammer et al., 2022; J. Zhang

et al., 2011), GC content also plays a role. We uncovered that while the GC content and the length distributions were not significantly affected in events that are correlated with respect to progerin expression, the anti-correlated events told a different story (Figure 4-9, 4-10). Here, intron GC content is significantly reduced, especially between the two alternative sites, and the distance between these sites is significantly shorter than that of the background (Figure 4-10). We speculate that the switch to the downstream site (e2) is facilitated if the two sites are close together and are not separated by GC-rich RNA.

While the mechanistic basis for these changes remain unknown, but there are some suggestive associations. First, the most highly correlated changes in isoform abundance include switches in factors known to be involved in alternative splicing, including PRMT5, SRSF9 and PRPF19 (Figure 3-1). Furthermore, there is a correlation with telomeres, and telomere shortening is known to lead to increased use of the progerin cryptic site (Kan Cao, Blair, Faddah, Kieckhaefer, et al., 2011). It is also possible that these changes are associated with alterations in nuclear architecture, which are known to result in altered splicing (Tammer et al., 2022).

On a macroscopic scale, telomeres attrition, as well as general DNA damage, have both been connected to alternative splicing (Kan Cao, Blair, Faddah, Kieckhaefer, et al., 2011; Paronetto et al., 2009; Shkreta & Chabot, 2015; Sprung et al., 2011). It is worth noting that sunlight is accompanied by ultraviolet (UV) radiation and past studies have revealed that UV radiation is highly correlated with DNA damage and aberrant splicing (Liu et al., 2005; Muñoz et al., 2009; Nicholls et al., 2004; Takeuchi & Rünger, 2013; H. Zhang et al., 2014). Telomere attrition is also found to have a causative relationship with aberrant splicing, as well as the production of progerin during senescence in human fibroblasts (Kan Cao, Blair, Faddah, Kieckhaefer, et al., 2011). In our study, progerin levels in sun-exposed skin samples had the highest amount of high progerin expressing samples. DNA damage by UV irradiation and

telomere attrition are both prime examples of phenotypes of chronological aging. Combining the past research with our own results, we suspect that there is some trigger, such as DNA damage and telomere attrition, which causes aberrant splicing in normal individuals, including the progerin cryptic splicing through some unknown mechanism.

We also found that if progerin is expressed in one tissue, it is more likely to be co-expressed in other tissues. Together with the result from the motif analyses, it is suggestive that cis-regulatory elements that affect the changes in splicing may affect equally in different tissues. Although we observed widespread change in splicing paradigm, we have yet to determine any kind of physical phenotype by having progerin expression in normal individuals. It could be possible that there is a safe threshold for progerin expression, since the level of progerin expression in normal individuals is nowhere close to the amount of expression seen in HGPS patients. Other speculations can be made, such as potential progerin phenotype repressors acting upon individuals with progerin expression, which can make the difference between normal individual versus HGPS patients. However, differential expression analysis revealed that many transcripts from mitochondrial genes are significantly anti-correlated in GTEx samples. In HGPS patients, it is well-documented that multitude of mitochondrial dysfunctions have been observed (Mei Xiong et al., 2017; Rivera-Torres et al., 2013; Xiong et al., 2016). It is evident from previous studies that progerin induces DNA-damage signaling and repair pathways, alongside with activation of p53 pathway that would trigger premature senescence (Benson et al., 2010; Burtner & Kennedy, 2010; Kan Cao, Blair, Faddah, & Al, 2011; McClintock et al., 2007). There is also a possibility that progerin expression feeds back into the DNA-damage loop by creating more DNA-damage, making it a vicious cycle, in which the effect must be explored further. It is possible that the progerin expression we see occurs in a small number of cells that produce high levels of progerin and are on a path towards cell death.

In any case we do not know what clinical phenotypes, if any, are associated with progerin expression in normal individuals.

The function of progerin in normal individuals, if any, remains unclear, but from this study, we have established a correlation between progerin, and molecular phenotype related to chronological aging and alternative splicing. Progerin expression in normal individuals may indeed be associated with aspects of the molecular phenotypes of HGPS patients at the cellular level. However, in the absence of mutation, we speculate that there is some trigger that induce progerin expression in normal individuals. We have uncovered basic differences between the background versus the events that are altered in splicing patterns. However, further studies must be done to establish a causative relationship between progerin expression, DNA-damage, changes in splicing and the regulatory element that are responsible for the changes, and the subsequent phenotypes that may not be readily visible in normal individuals.

5.3 Future directions

Throughout my research, I have used GTEx database as the sole source of expression data for non-HGPS normal individuals. One of the ways that can greatly bolster the result of my research and to validate my findings is to incorporate another large database that is on a similar scale to the GTEx. The Cancer Genome Atlas (TCGA) (McLendon et al., 2008; Weinstein et al., 2013) database contains clinical, biospecimen, molecular characterization, and imaging data for samples from 11,000 patients over 33 different cancer types (phs000178.v11.p8). While the condition may be different from that of non-HGPS normal individuals, usage and incorporation of comparative study done on TCGA data may provide us valuable insight. We may explore the progerin expression in environments that induces DNA damage in a different setting compared to that of the previously explored methods, such as sun

exposure, direct UV irradiation, and telomere attrition. The systemic differences between cancer patients and non-cancer patients may give us unique understanding of potential progerin expression in different settings of non-HGPS conditions.

Another path that can be explored more would be the upstream part of the working model. The connection between the trigger and the splicing landscape change can be made stronger. The study done on cellular senescence being the trigger for progerin induction is robust (Kan Cao, Blair, Faddah, Kieckhaefer, et al., 2011). However, the study on natural source of UV irradiation from sun exposure (Lesiak et al., 2017), and thus contribute to phenotype for chronological aging, could be bolstered. New studies, consisting of both short term and longitudinal, can be developed and conducted. Short term studies can be useful in determining the acute effect of longwave UV irradiation in the form of UVA and partially UVB. There has been only 1 study of such type, and as such, acquiring more subjects and providing a stricter, more controlled method of delivering the DNA damage can help making connection between UV irradiation via sunlight and progerin expression stronger. Longitudinal study will provide insights on the much needed long term effects of environmental agents that may induce DNA damage to individuals. The degree of exposure to DNA damaging agent, along with the duration per instance and the duration of the study will provide valuable results to help make causal claims on relationship between DNA damage and progerin expression in non-HGPS normal individuals.

Lastly, it has been uncovered in previous study that progerin is expressed in small amount of cells, rather than equally spread across a population of cells (Kan Cao, Blair, Faddah, Kieckhaefer, et al., 2011). In this study, they have designed a double-colored progerin splicing reporter construct, where cells expression progerin are indicated by both green (GFP) and red signals (DsRed). This reporter construct combined with fluorescence-activated cell sorting

(FACS) will allow us to identify the subpopulation of cells that has the cryptic splice site activated in *LMNA*. Another study that can be conducts is the use of single-cell RNA-seq (scRNA-seq) technique for identifying subpopulation of cells that express progerin. To assess the effect of progerin expression more precisely, technique that can provide us higher resolution results are crucial. Advantage of using scRNA-seq is getting high resolution cell-to-cell analysis, capture the cellular heterogeneity within a tissue, as well as assessing expression patterns based on cell type, subtypes and states. Studies including progerin induction proves to be very difficult because of lack of proper control samples. Latest advances in technology have shown that adenine base editing (ABE) can be extremely useful to creating isogenic control for progerin expressing samples without the concern of having unintended off-target effects (Koblan et al., 2021). Combining the method to generate isogenic controls, as well as identifying subpopulations of cells that expresses progerin, a deeper understanding of progerin's role maybe elucidated.

Chapter VI : Materials and methods

6.1 Genotype Tissue Expression data

Genotype Tissue Expression (GTEx) data was obtained from the GTEx consortium. GTEx version 6 raw transcript and gene count data downloaded in February 2018 were used in this study (GTEx_Analysis_v6_RNA-seq_Flux1.6_transcript_reads.txt.gz and GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_reads.gct.gz). GTEx version 6 contains a total of 8,555 samples across 30 different tissues. Per GTEx consortium, the RNA-seq data were collected using Illumina TruSeq library construction protocol. Alignment was done using STAR v2.5.3a (Dobin et al., 2013) against human reference genome GRCh38/hg38 based on GENCODE version 19 annotation. Transcript level quantifications were done using RSEM v1.3.0 (B. Li & Dewey, 2011).

6.2 Reads Per Kilobase per Million mapped reads (RPKM) calculation

RPKM is a normalization method to account for the total library size. RPKM values were calculated using the following formula:

$$RPKM_i = \frac{X_i}{(\frac{l_i}{10^3})(\frac{N}{10^6})}$$

where X_i is the read count of gene i, l_i is the length of gene i, and N is the total number of mapped reads.

6.3 Sample categorization following percent spliced in quantification

Transcript count data from GTEx version 6 (Lonsdale et al., 2013) was used to calculate percent spliced in (PSI) using SUPPA2 (Trincado et al., 2018). Subsequently, fraction progerin (FP) was then calculated by subtracting progerin PSI from 1 to reflect progerin

abundance relative to the abundance all *LMNA* isoforms that utilize exon 11. Human GENCODE version 19 (hg38) (Harrow et al., 2012) was used as a reference transcriptome. SUPPA2's option -e SE SS MX RI FL was used to calculate skipping exon, alternative 5'/3' splice sites, mutually exclusive exons, retained intron, and alternative first/last exons, respectively. IOE file was generated, as a part of the pipeline for SUPPA2, to quantify isoform inclusion levels of all transcripts from GTEx samples. The fraction progerin distribution was examined and used to divide the samples into zero (FP < 0.5%; progerin PSI > 99.5%), low (FP 0.5% ~ 4%) and high (FP > 4%; PSI < 96%) categories. NAs were treated as zeroes.

6.4 Bioinformatics analyses

Differential expression analyses were done using R (version 3.5.1 "Feather Spray") package limma (Smyth, 2005) (version 3.38.3). Limma is designed to analyze expression data using linear models and it is specialized to find differentially expressed genes in the context of multifactor designed experiments. The progerin expression categories were treated as different conditions. The expression data were normalized using limma's voom normalization. Voom normalization transforms count data to log2-counts per million (logCPM). lmFit function was used to calculate the linear fit for every transcript and determine the significance. Significant transcripts were defined as having a p-value cutoff of $p < 0.05$. Gene ontology analysis of transcripts that were identified as significantly differentially expressed was done using Enrichr (Chen et al., 2013). Enrichr analysis provided GO terms for 4 different biological domains: Biological processes (BP), cellular components (CC), molecular functions (MF) and KEGG pathway. Gene set enrichment analysis (GSEA) was done using the FGSEA package from Bioconductor (Korotkevich et al., 2021). The annotated gene set used was provided by molecular signatures database (MSigDB) (Liberzon et al., 2011). C5 ontology gene set was

used, where we obtained the biological process (BP) subset to do the GSEA analysis. Empirical false discovery rate (FDR) was calculated with the bootstrap method. 40 samples from the selected tissue samples were picked at random and differential expression analysis was done on 20 samples versus 20 samples contrast. This process was repeated 100 times to calculate the empirical FDR. Pearson's pairwise correlation coefficient was used to measure the strength of the correlation between samples. Motif analyses were done using Multiple Expectation maximizations for Motif Elicitation (MEME) suite (Bailey et al., 2009) with the following options: -objfun de (used for setting objective function as differential expression), -nmotifs 30 (stops after finding 30 motifs), -minw 3 (minimum length of the motif to be 3) and -maxw 5 (maximum length of the motif to be 5). Bootstrapping for the motif analysis was done by subsetting 150 sequences in SE events, and 60 sequences in A5 events. As for background, we subset 10,000 sequences for both SE and A5 events. With the sequence subsets, we conducted the differential motif enrichment analyses 100 times. For GC content and length analysis, the top 10, 20 and 50 events were selected based on the r-value obtained from transcriptome-wide correlation analysis (Pearson's correlation coefficient). Statistical tests were done using t-test.

6.5 Cell line used

The HGFDFN168 (Fib 168) and HGADFN167 (Fib 167) fibroblast lines were obtained from the Progeria Research Foundation. HGFDFN168 is from the father of HGADFN167, an HGPS patient.

6.6 qPCR analyses

Total RNA from fibroblasts was extracted with Trizol (Life Technologies, Carlsbad,

California, United States) and purified using the RNeasy Mini kit (Qiagen, USA) according to the manufacturer's instructions. The RNA yield was determined using the NanoDrop 2000 spectrophotometer. RNA samples from sun-exposed skin were requested from GTEx. RNA was converted to cDNA using iScript Select cDNA Synthesis kit (Bio-Rad, USA). Quantitative RT-PCR was performed in triplicate using SYBR Green Supermix (Bio-Rad) on CFX96 real-time system (C1000 Thermal Cycler; Bio-Rad). Primer sequences are:

lamin A forward: 5'-GCAACAAGTCCAATGAGGACCA-3',

lamin A reverse: 5'-CATGATG

CTGCAGTTCTGGGGGCTCTGGAT-3',

progerin forward: 5'-GCAACAAGTCCAATGAG

GACCA-3',

progerin reverse: 5'-CATGATGCTGCAGTTCTGGGGGCTCTGGAC-3',

β-actin forward: 5'-CTGGAACGGTGAAGGTGACA-3',

β-actin reverse: 5'-AAGGGACTTCCTGTAACAATGCA-3'.

Chapter VII : Appendix

7.1 Purpose and results

Over the course of conducting research on progerin expression in non-HGPS normal individuals, many different approaches have been taken, as well as looking at the problem at different angles. The appendix section is dedicated to shed light on some of the approaches that were taken prior to or after some of the methods used in the previous sections. Different approaches may or may not have yielded meaningful results, but nevertheless were important for me to examine to find potential mechanisms or association with various factors that were found to be associated with progerin expression. In this section, I will also describe my contribution to a collaborative work with Dr. Xiaojing Mao on "Impaired LEF1 activation accelerates iPSC-derived keratinocytes differentiation in Hutchinson-Gilford Progeria Syndrome" (Mao et al., 2022).

Transcript	Gene	log FC	P-Value
ENST00000556647.1	SRSF5	-4.264	8.921E-24
ENST00000554929.1	SRSF5	-1.275	7.125E-08
ENST00000556436.1	SRSF5	-1.003	3.097E-04
ENST00000576585.1	PRPF8	-1.283	1.032E-34
ENST00000572723.1	PRPF8	-1.402	1.901E-27
ENST00000573681.1	PRPF8	-0.597	5.509E-12
ENST00000575116.1	PRPF8	-0.733	5.113E-11

Figure 7-1 Transcripts from splicing related genes that are differentially expressed.

Many transcripts from SRSF5 and PRPF8 were found to be significantly differentially expressed in skin samples. SRSF5 promotes splice site usage, and PRPF8 encodes for protein that is a core part of the U2 and the U12-dependent spliceosomes.

Figure 7-1 shows the table of differentially expressed transcripts from two prominent splicing related genes, SRSF5 and PRPF8, gathered from initial screening of differential expression

analysis. This is the results of differential expression analysis prior to correcting for multiple testing.

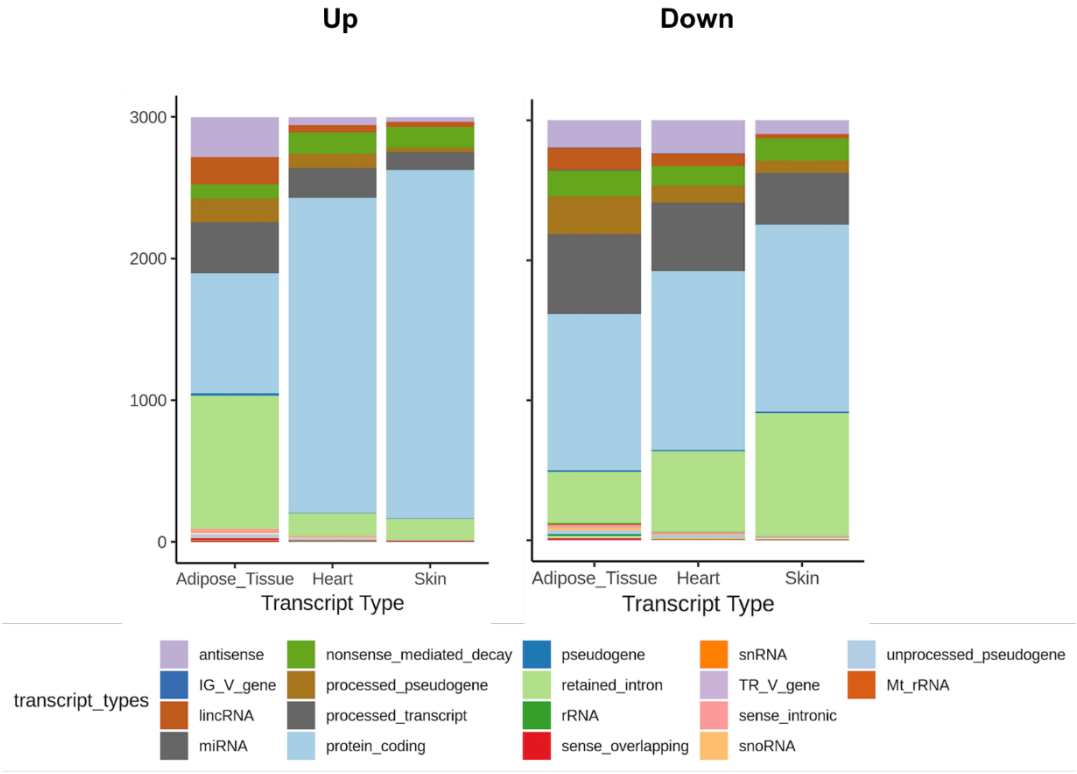
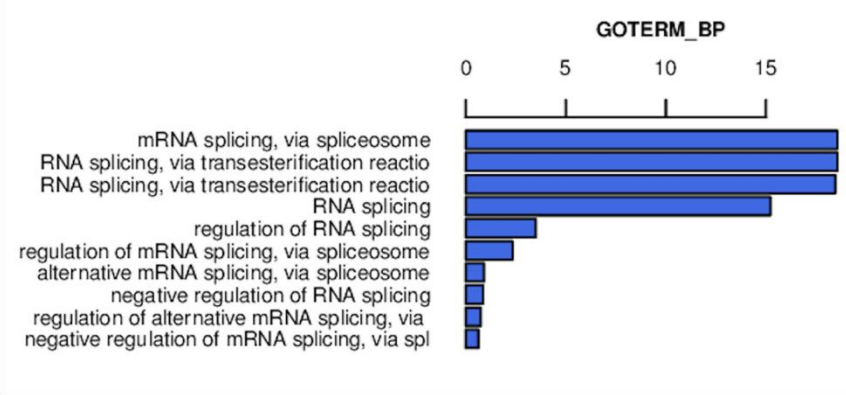


Figure 7-2 Transcript types of differentially expressed transcripts.

Transcript types of the identified differentially expressed transcripts in adipose, heart and skin were observed.

Figure 7-2 shows the proportions of different transcript types of the differentially expressed transcripts in adipose, heart and skin tissues. Correlated and anti-correlated transcripts were both examined. We expected to see anomalies in the composition of various transcript types, however none were observed.



GO Terms	Genes
mRNA splicing, via spliceosome (GO:0000398)	ISY1 PPWD1 HNRNPR DDX41 PRPF19 EFTUD2 MAGOH SNRPD3 RBM5 SRRM2 RBM17 AQR MBNL1 TFIP11 PRPF4B GTF2F2 SFPQ DDX39A HNRNPUL1 HNRNPH1 PRPF3 SRSF2 PPIE SNRPA1 SRSF4 HNRNPH3 SNRPF PABPC1 SRSF5 SNRPA SF1 SRSF9 DDX5 RBM8A SRSF1 U2AF1 PRPF8 PABPN1 POLR2B PCBP2 DHX38 SF3B1 POLR2J SRSF10 SRSF11 HNRNPA3 CPSF1 FUS ALYREF CWC15 HNRNPL PHF5A HNRNPD HNRNPC HNRNPA1L2
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile (GO:0000377)	ISY1 PPWD1 HNRNPR DDX41 PRPF19 EFTUD2 MAGOH SNRPD3 RBM5 SRRM2 RBM17 AQR MBNL1 TFIP11 PRPF4B GTF2F2 SFPQ DDX39A HNRNPUL1 HNRNPH1 PRPF3 SRSF2 PPIE SNRPA1 SRSF4 HNRNPH3 SNRPF PABPC1 SRSF5 SNRPA SF1 SRSF9 DDX5 RBM8A SRSF1 U2AF1 PRPF8 PABPN1 POLR2B PCBP2 DHX38 SF3B1 POLR2J SRSF10 SRSF11 HNRNPA3 CPSF1 FUS ALYREF CWC15 HNRNPL PHF5A HNRNPD HNRNPC HNRNPA1L2

Figure 7-3 Subset of gene ontology analysis (splicing related terms).

Collection of the splicing related GO terms are shown. For mRNA splicing, via spliceosome, and RNA splicing, via transesterification reactions, the genes associated with the respective terms are shown.

Subset (splicing related terms) of the gene ontology analysis was examined and shown in figure 7-3. The gene compositions for mRNA splicing via spliceosome and RNA splicing via transesterification reactions are shown. Genes from this list were taken for further investigation for alterations in expression pattern in their isoforms.

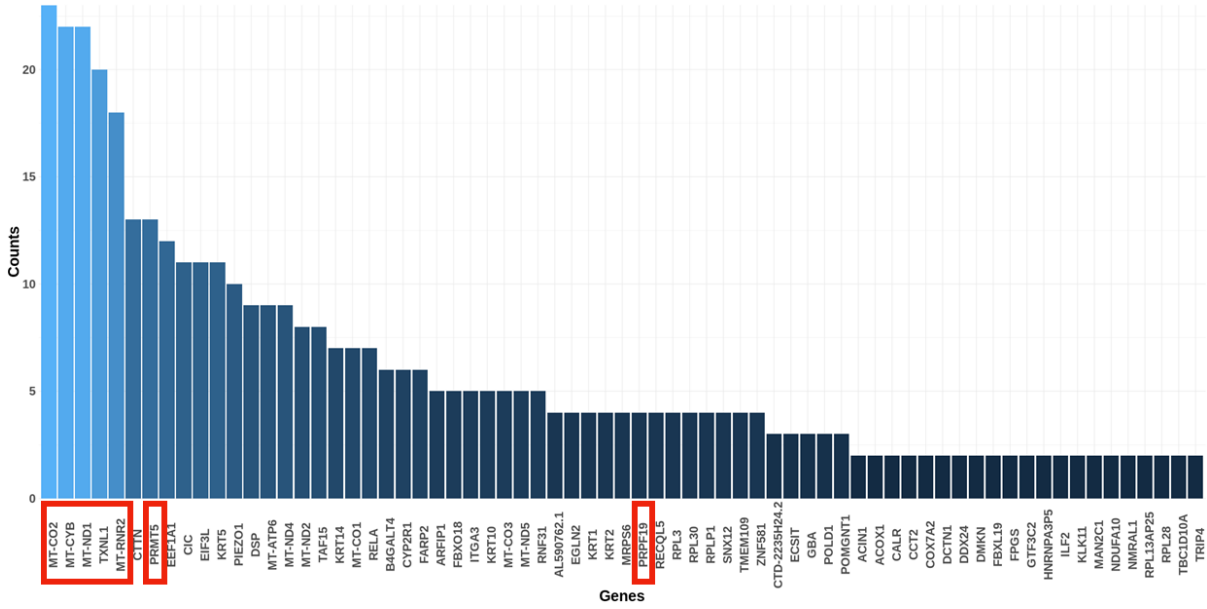


Figure 7-4 Top ranked differentially expressed transcripts, after correcting for multiple testing.

Many mitochondrial transcripts and transcripts from splicing related genes were found to be significantly differentially expressed, after correcting for multiple testing.

Figure 7-4 shows results of differential expression analysis after correcting for multiple testing. The transcripts denoted by the red box (MT-CO2, MT-CYB, MT-ND1, TXNL1, MT-RNR2, PRMT5 and PRPF19) were of particular interest, because of their relation to the mitochondria or splicing function.

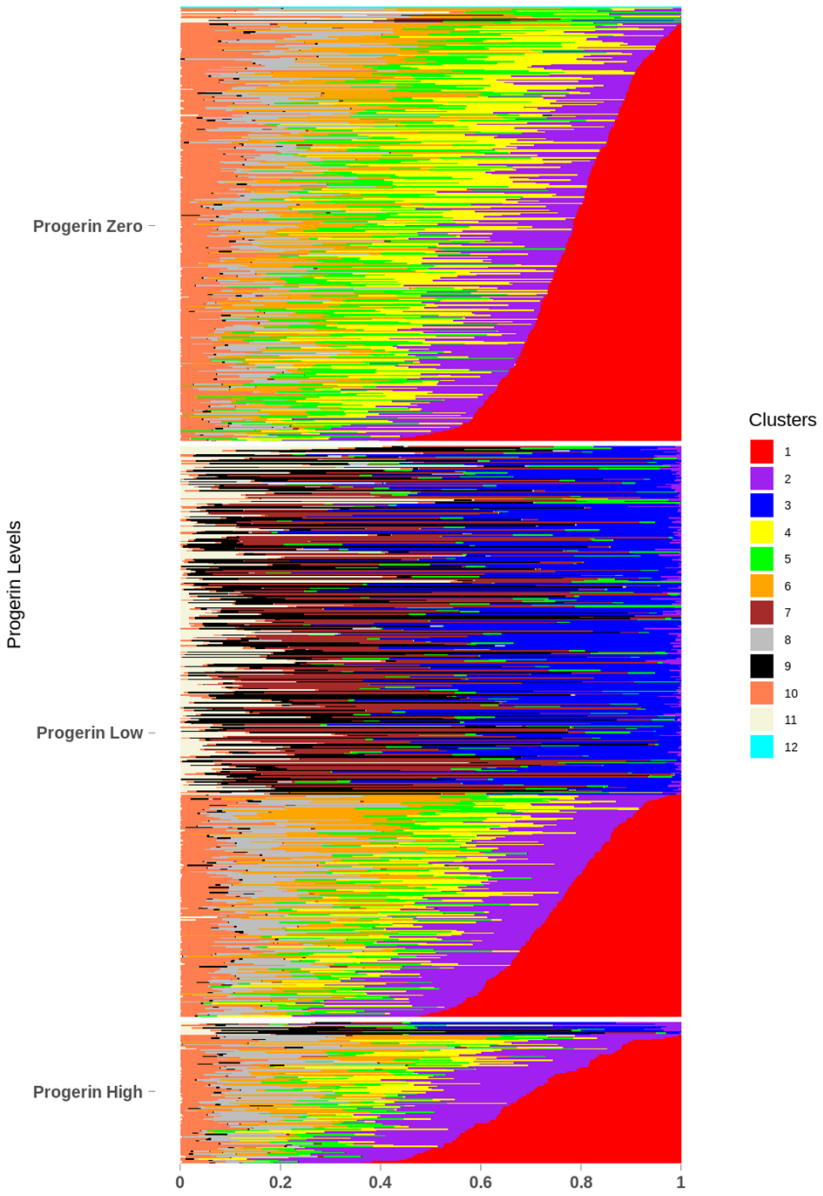


Figure 7-5 Latent Dirichlet Allocation of GTEx skin samples

Latent Dirichlet Allocation (LDA) model was used to assess the differences in gene clusters for zero, low, and high fraction progerin samples, as well as finding composition of driving genes for each of the clusters.

Latent Dirichlet Allocation model was used to further investigate the differences between the 3 categories of fraction progerin levels in the GTEx samples. From the LDA analysis, we were not able to find significant compositional differences between zero, low and high fraction

progerin samples.

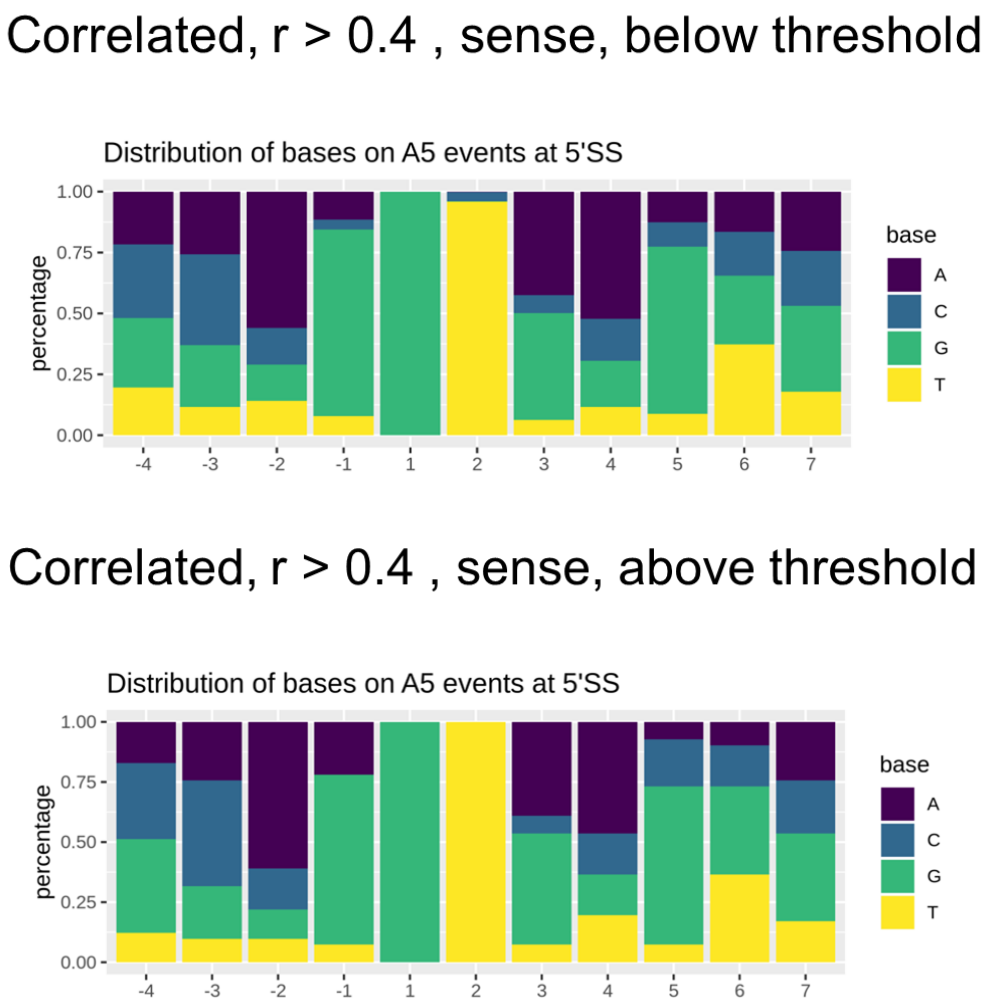


Figure 7-6 Representation of bases around the 5′ splice site for A5 events.

Representation of the 4 bases around the 5′ splice site for A5 events were observed. This was done for correlated, cutoff r-value of 0.4. Those above the threshold were compared to those that were below the threshold.

Base composition around the 5' splice site (from -4 to +7 positions) were examined. the comparison was made between r > 0.4 vs r < 0.4. No visible differences were observed between the two groups.

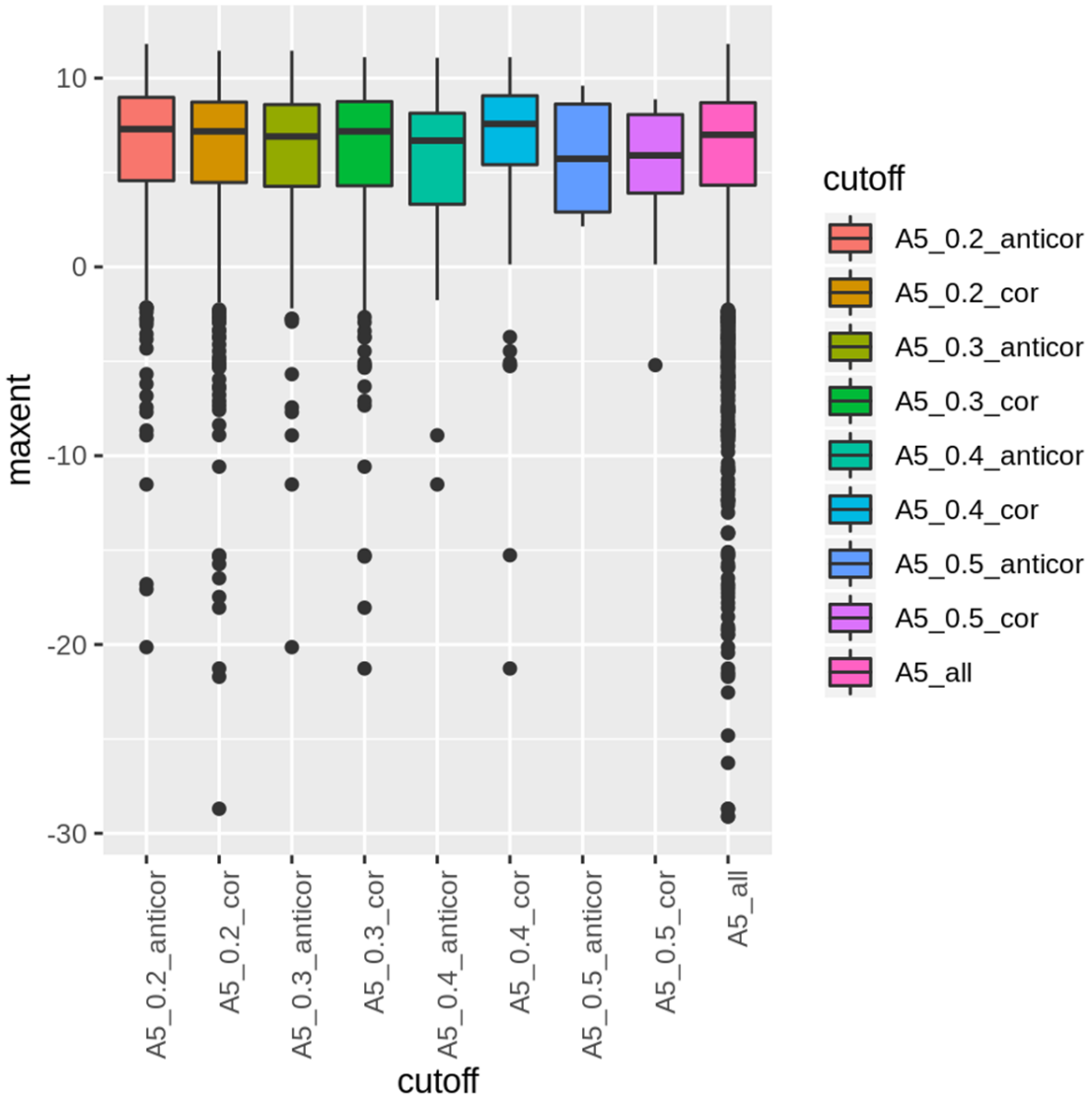


Figure 7-7 Splice site strength analysis using MaxEnt (A5).

Splice site strength of the identified splicing events were observed, based on the various Pearson's correlation coefficient cutoffs for A5 events.

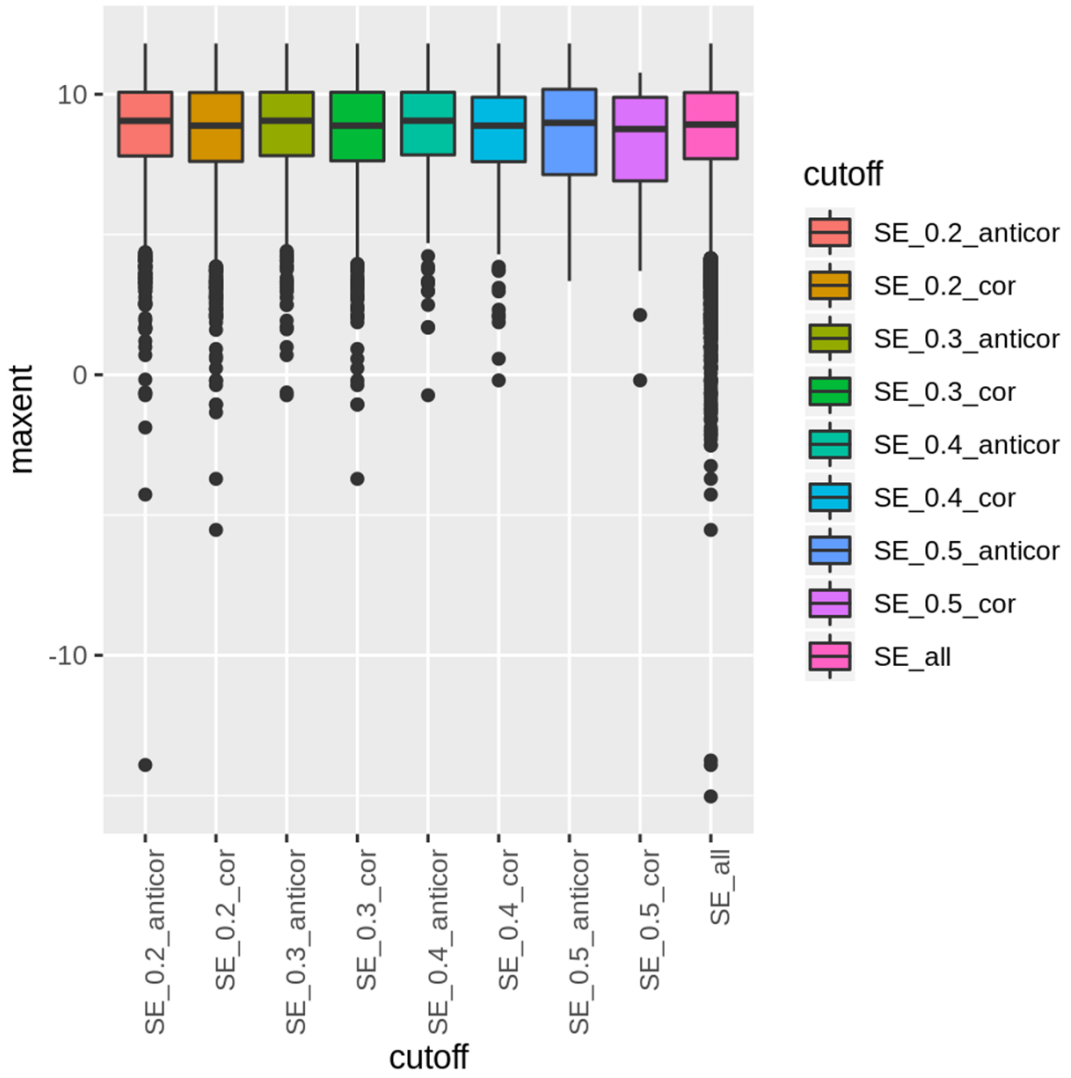


Figure 7-8 Splice site strength analysis using MaxEnt (SE).

Splice site strength of the identified splicing events were observed, based on the various Pearson's correlation coefficient cutoffs for SE events.

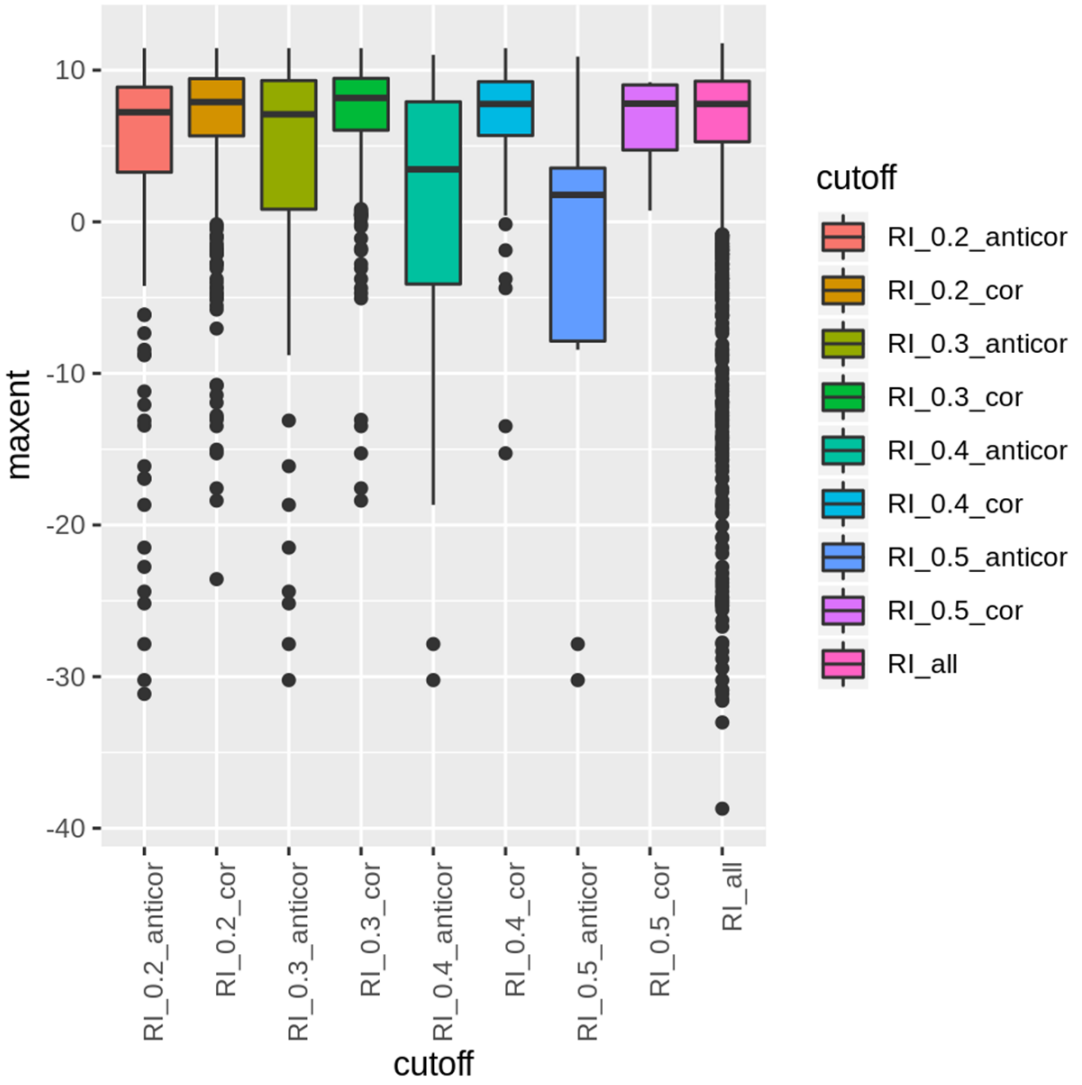


Figure 7-9 Splice site strength analysis using MaxEnt (RI).

Splice site strength of the identified splicing events were observed, based on the various Pearson's correlation coefficient cutoffs for RI events.

Splice site strength analysis was done using MaxEnt methodology for A5, SE, and RI events (Figure 7-6, 7-7, 7-8). For A5 and SE events, there were little to no noticeable difference between the background (all) and the various cutoffs. However, there seems to be a preference for progerin expression to be associated with splice sites of intron retention events at $r > 0.5$

and r > 0.4.

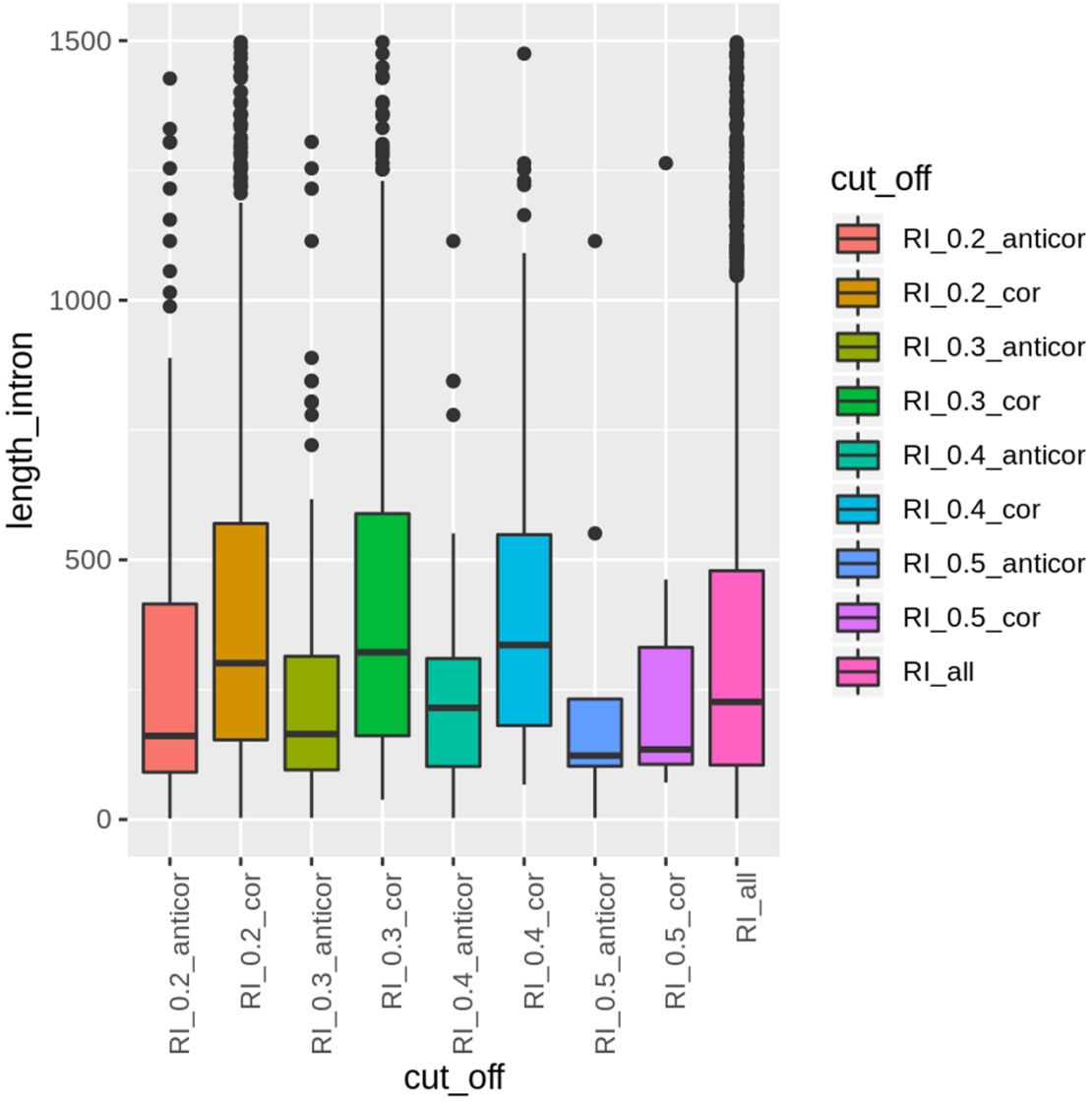


Figure 7-10 Intron length distribution for retained intron events.

Intron lengths of the identified RI splicing events were observed, based on the various Pearson's correlation coefficient cutoffs.

Figure 7-9 shows the intron length distribution for intron retention events. It appears that RI events that are correlated to the progerin expression tend to have longer introns than the anti-

correlated counterparts.

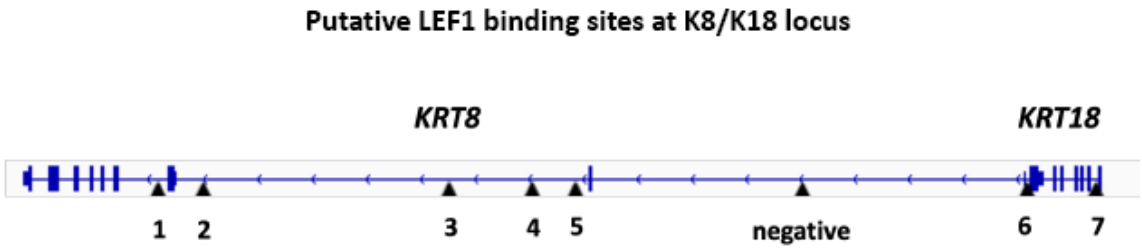


Figure 7-11 Identification of LEF1 binding sites around KRT8 and KRT18 loci

In a collaborative work with Dr. Xiaojing Mao, LEF1 binding sites around KRT8 and KRT18 loci were identified. The binding sites were assessed with various cell types.

I have collaborated with Dr. Xiaojing Mao on the study of "Impaired LEF1 activation accelerates iPSC-derived keratinocytes differentiation in Hutchinson-Gilford Progeria Syndrome", published in 2022. My contribution to the work is identifying the LEF1 binding site around KRT8 and KRT18 loci. The loci were assessed using HEK293T and HUES64 cell lines, and identified 7 different LEF1 binding sites.

Bibliography

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., Lev-Maor, G., Burstein, D., Schwartz, S., Postolsky, B., Pupko, T., & Ast, G. (2012). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Reports*, *1*(5), 543–556. https://doi.org/10.1016/J.CELREP.2012.03.013

Bailey, T. L. (2011). *DREME: motif discovery in transcription factor ChIP-seq data*. *27*(12), 1653–1659. https://doi.org/10.1093/bioinformatics/btr261

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, *37*(SUPPL. 2). https://doi.org/10.1093/nar/gkp335

Benson, E. K., Lee, S. W., & Aaronson, S. A. (2010). Role of progerin-induced telomere dysfunction in HGPS premature cellular senescence. *Journal of Cell Science*, *123*(15), 2605–2612. https://doi.org/10.1242/jcs.067306

Bezzi, M., Teo, S. X., Muller, J., Mok, W. C., Sahu, S. K., Vardy, L. A., Bonday, Z. Q., & Guccione, E. (2013). Regulation of constitutive and alternative splicing by PRMT5 reveals a role for Mdm4 pre-mRNA in sensing defects in the spliceosomal machinery. *Genes and Development*. https://doi.org/10.1101/gad.219899.113

Burtner, C. R., & Kennedy, B. K. (2010). Progeria syndromes and ageing: What is the connection? In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm2944

Cao, K., Capell, B. C., Erdos, M. R., Djabali, K., & Collins, F. S. (2007). A lamin A protein isoform overexpressed in Hutchinson-Gilford progeria syndrome interferes with mitosis in progeria and normal cells. *Proceedings of the National Academy of Sciences*, *104*(12), 4949–4954. https://doi.org/10.1073/pnas.0611640104

Cao, Kan, Blair, C. D., Faddah, D. A., & Al, E. (2011). Progerin and Telomer Dysfunction Collaborate to Trigger Celluler Senescence in Normal Human Fibroblast. *The Journal of Clinical Investigation*, *121*(7), 2833–2844. https://doi.org/10.1172/JCI43578.associated

Cao, Kan, Blair, C. D., Faddah, D. A., Kieckhaefer, J. E., Olive, M., Erdos, M. R., Nabel, E. G., & Collins, F. S. (2011). Progerin and telomere dysfunction collaborate to trigger cellular senescence in normal human fibroblasts. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI43578

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., & Ma'ayan, A. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-128

De Conti, L., Baralle, M., & Buratti, E. (2013). Exon and intron definition in pre-mRNA splicing. *Wiley Interdisciplinary Reviews: RNA*, *4*(1), 49–60. https://doi.org/10.1002/WRNA.1140

Dechat, T., Korbei, B., Vaughan, O. A., Vlcek, S., Hutchison, C. J., & Foister, R. (2000). Lamina-associated polypeptide 2alpha binds intranuclear A-type lamins. *Journal of Cell Science*, *113*(19), 3473–3484. https://doi.org/10.1242/JCS.113.19.3473

Dechat, Thomas, Pfleghaar, K., Sengupta, K., Shimi, T., Shumaker, D. K., Solimando, L., & Goldman, R. D. (2008). Nuclear lamins: Major factors in the structural organization and

function of the nucleus and chromatin. In *Genes and Development*. https://doi.org/10.1101/gad.1652708

Demanelis, K., Jasmine, F., Chen, L. S., Chernoff, M., Tong, L., Delgado, D., Zhang, C., Shinkle, J., Sabarinathan, M., Lin, H., Ramirez, E., Oliva, M., Kim-Hellmuth, S., Stranger, B. E., Lai, T. P., Aviv, A., Ardlie, K. G., Aguet, F., Ahsan, H., … Pierce, B. L. (2020). Determinants of telomere length across human tissues. *Science*, *369*(6509). https://doi.org/10.1126/SCIENCE.AAZ6876

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts635

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., Dutra, A., Pak, E., Durkin, S., Csoka, A. B., Boehnke, M., Glover, T. W., & Collins, F. S. (2003). Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature*, *423*(6937), 293–298. https://doi.org/10.1038/nature01629

Frenkel-Morgenstern, M., Lacroix, V., Ezkurdia, I., Levin, Y., Gabashvili, A., Prilusky, J., Del Pozo, A., Tress, M., Johnson, R., Guigo, R., & Valencia, A. (n.d.). *Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts*. https://doi.org/10.1101/gr.130062.111

Georgomanolis, T., Sofiadis, K., & Papantonis, A. (2016). Cutting a long intron short: Recursive splicing and its implications. *Frontiers in Physiology*, *7*(NOV), 234102. https://doi.org/10.3389/FPHYS.2016.00598/BIBTEX

Gordon, L. B., Kleinman, M. E., Miller, D. T., Neuberg, D. S., Giobbie-Hurder, A., Gerhard-Herman, M., Smoot, L. B., Gordon, C. M., Cleveland, R., Snyder, B. D., Fligor, B., Bishop, W. R., Statkevich, P., Regen, A., Sonis, A., Riley, S., Ploski, C., Correia, A., Quinn, N., … Kieran, M. (2012). Clinical trial of a farnesyltransferase inhibitor in children with Hutchinson-Gilford progeria syndrome. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1202529109

Gordon, L. B., McCarten, K. M., Giobbie-Hurder, A., Machan, J. T., Campbell, S. E., Berns, S. D., & Kieran, M. W. (2007). Disease progression in Hutchinson-Gilford progeria syndrome: Impact on growth and development. *Pediatrics*. https://doi.org/10.1542/peds.2007-1357

Gordon, L. B., Shappell, H., Massaro, J., D'Agostino, R. B., Brazier, J., Campbell, S. E., Kleinman, M. E., & Kieran, M. W. (2018). Association of lonafarnib treatment vs no treatment with mortality rate in patients with Hutchinson-Gilford progeria syndrome. *JAMA - Journal of the American Medical Association*. https://doi.org/10.1001/jama.2018.3264

Grammatikakis, I., Panda, A. C., Abdelmohsen, K., & Gorospe, M. (2014). Long noncoding RNAs (lncRNAs) and the molecular hallmarks of aging. *Aging*, *6*(12), 992–1009. https://doi.org/10.18632/aging.100710

Gruenbaum, Y., Margalit, A., Goldman, R. D., Shumaker, D. K., & Wilson, K. L. (2005). The nuclear lamina comes of age. In *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm1550

Hamard, P. J., Santiago, G. E., Liu, F., Karl, D. L., Martinez, C., Man, N., Mookhtiar, A. K.,

Duffort, S., Greenblatt, S., Verdun, R. E., & Nimer, S. D. (2018). PRMT5 Regulates DNA Repair by Controlling the Alternative Splicing of Histone-Modifying Enzymes. *Cell Reports*. https://doi.org/10.1016/j.celrep.2018.08.002

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., … Hubbard, T. J. (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*. https://doi.org/10.1101/gr.135350.111

Hennekam, R. C. M. (2006). Hutchinson–Gilford progeria syndrome: Review of the phenotype. *American Journal of Medical Genetics Part A*, *140A*(23), 2603–2624. https://doi.org/10.1002/ajmg.a.31346

J-L Wong, J., M Au, A. Y., Ritchie, W., & J Rasko, J. E. (n.d.). *Prospects & Overviews Intron retention in mRNA: No longer nonsense Known and putative roles of intron retention in normal and disease biology*. https://doi.org/10.1002/bies.201500117

Jiang, W., & Chen, L. (2021). *Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing*. https://doi.org/10.1016/j.csbj.2020.12.009

Jung, H.-J., Coffinier, C., Choe, Y., Beigneux, A. P., Davies, B. S. J., Yang, S. H., Barnes, R. H., Hong, J., Sun, T., Pleasure, S. J., Young, S. G., & Fong, L. G. (2012). Regulation of prelamin A but not lamin C by miR-9, a brain-specific microRNA. *Proceedings of the National Academy of Sciences*, *109*(7), E423–E431. https://doi.org/10.1073/pnas.1111780109

Koblan, L. W., Erdos, M. R., Wilson, C., Cabral, W. A., Levy, J. M., Xiong, Z. M., Tavarez,

U. L., Davison, L. M., Gete, Y. G., Mao, X., Newby, G. A., Doherty, S. P., Narisu, N., Sheng, Q., Krilow, C., Lin, C. Y., Gordon, L. B., Cao, K., Collins, F. S., … Liu, D. R. (2021). In vivo base editing rescues Hutchinson–Gilford progeria syndrome in mice. *Nature 2020 589:7843*, *589*(7843), 608–614. https://doi.org/10.1038/s41586-020-03086-7

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2021). Fast gene set enrichment analysis. *BioRxiv*, 060012. https://doi.org/10.1101/060012

Lasda, E., & Parker, R. (2014). Circular RNAs: diversity of form and function. *RNA*, *20*(12), 1829–1842. https://doi.org/10.1261/RNA.047126.114

Lesiak, A., Bednarski, I., Rogowski-Tylman, M., Sobjanek, M., Woźniacka, A., Danilewicz, M., Young, A., & Narbutt, J. (2017). One week of exposure to sunlight induces progerin expression in human skin. In *Postepy Dermatologii i Alergologii*. https://doi.org/10.5114/pdia.2016.62416

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-12-323

Li, Z., Qin, F., & Li, H. (2018). Chimeric RNAs and their implications in cancer. *Current Opinion in Genetics & Development*, *48*, 36–43. https://doi.org/10.1016/j.gde.2017.10.002

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, *27*(12),

1739–1740. https://doi.org/10.1093/BIOINFORMATICS/BTR260

Liu, B., Wang, J., Chan, K. M., Tjia, W. M., Deng, W., Guan, X., Huang, J. D., Li, K. M., Chau, P. Y., Chen, D. J., Pei, D., Pendas, A. M., Cadiñanos, J., López-Otín, C., Tse, H. F., Hutchison, C., Chen, J., Cao, Y., Cheah, K. S. E., … Zhou, Z. (2005). Genomic instability in laminopathy-based premature aging. *Nature Medicine*. https://doi.org/10.1038/nm1266

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, *45*(6), 580–585. https://doi.org/10.1038/ng.2653

Mallampalli, M. P., Huyer, G., Bendale, P., Gelb, M. H., & Michaelis, S. (2005). *Inhibiting farnesylation reverses the nuclear morphology defect in a HeLa cell model for Hutchinson-Gilford progeria syndrome*. www.pnas.orgcgidoi10.1073pnas.0503712102

Mao, X., Xiong, Z. M., Xue, H., Brown, M. A., Gete, Y. G., Yu, R., Sun, L., & Cao, K. (2022). Impaired LEF1 Activation Accelerates iPSC-Derived Keratinocytes Differentiation in Hutchinson-Gilford Progeria Syndrome. *International Journal of Molecular Sciences*, *23*(10). https://doi.org/10.3390/IJMS23105499

McClintock, D., Ratner, D., Lokuge, M., Owens, D. M., Gordon, L. B., Collins, F. S., & Djabali, K. (2007). The mutant form of Lamin A that causes Hutchinson-Gilford progeria is a biomarker of cellular aging in human skin. *PLoS ONE*, *2*(12). https://doi.org/10.1371/journal.pone.0001269

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., Olson, J. J., Mikkelsen, T., Lehman, N., Aldape, K., Yung, W. K. A., Bogler, O., Weinstein, J. N., VandenBerg, S., Berger, M., Prados, M., Muzny, D., Morgan, M., Scherer, S., … Thomson, E. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, *455*(7216), 1061–1068. https://doi.org/10.1038/NATURE07385

Mei Xiong, Z., O'Donovan, M., Sun, L., Young Choi, J., Ren, M., & Cao, K. (2017). Anti-Aging Potentials of Methylene Blue for Human Skin Longevity. *Scientific Reports*, *7*(1), 1–12. https://doi.org/10.1038/s41598-017-02419-3

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., Brooks, B. P., Gerber, L. H., Turner, M. L., Domingo, D. L., Hart, T. C., Graf, J., Reynolds, J. C., Gropman, A., Yanovski, J. A., … Introne, W. J. (2008). Phenotype and course of Hutchinson-Gilford progeria syndrome. *New England Journal of Medicine*. https://doi.org/10.1056/NEJMoa0706898

Muñoz, M. J., Santangelo, M. S. P., Paronetto, M. P., de la Mata, M., Pelisch, F., Boireau, S., Glover-Cutter, K., Ben-Dov, C., Blaustein, M., Lozano, J. J., Bird, G., Bentley, D., Bertrand, E., & Kornblihtt, A. R. (2009). DNA Damage Regulates Alternative Splicing through Inhibition of RNA Polymerase II Elongation. *Cell*, *137*(4), 708–720. https://doi.org/10.1016/j.cell.2009.03.010

Navarro, C. L., De Sandre-Giovannoli, A., Lle Bernard, R., Ne Boccaccio, I., Boyer, A., Geneviè Ve, D., Hadj-Rabia, S., Gaudy-Marqueste, C., Smitt, H. S., Vabres, P., Faivre, L., Verloes, A., Flori, E., Hennekam, R., Beemer, F. A., Laurent, N., Le Merrer, M.,

Cau, P., & Vy, N. L. (2004). Lamin A and ZMPSTE24 (FACE-1) defects cause nuclear disorganization and identify restrictive dermopathy as a lethal neonatal laminopathy. *Human Molecular Genetics*, *13*(20), 2493–2503. https://doi.org/10.1093/hmg/ddh265

Nicholls, C. D., Shields, M. A., Lee, P. W. K., Robbins, S. M., & Beattie, T. L. (2004). UV-dependent alternative splicing uncouples p53 activity and PIG3 gene function through rapid proteolytic degradation. *Journal of Biological Chemistry*, *279*(23), 24171–24178. https://doi.org/10.1074/jbc.M401049200

Olive, M., Harten, I., Mitchell, R., Beers, J. K., Djabali, K., Cao, K., Erdos, M. R., Blair, C., Funke, B., Smoot, L., Gerhard-Herman, M., MacHan, J. T., Kutys, R., Virmani, R., Collins, F. S., Wight, T. N., Nabel, E. G., & Gordon, L. B. (2010). Cardiovascular pathology in Hutchinson-Gilford progeria: Correlation with the vascular pathology of aging. *Arteriosclerosis, Thrombosis, and Vascular Biology*. https://doi.org/10.1161/ATVBAHA.110.209460

Paronetto, M. P., Mata, M. De, Pelisch, F., Glover-cutter, K., Ben-dov, C., Lozano, J. J., Bird, G., Bentley, D., Bertrand, E., & Kornblihtt, A. R. (2009). *DNA Damage Regulates Alternative Splicing through Inhibition of RNA Polymerase II Elongation*. 708–720. https://doi.org/10.1016/j.cell.2009.03.010

Rivera-Torres, J., Acín-Perez, R., Cabezas-Sánchez, P., Osorio, F. G., Gonzalez-Gómez, C., Megias, D., Cámara, C., López-Otín, C., Enríquez, J. A., Luque-García, J. L., & Andrés, V. (2013). Identification of mitochondrial dysfunction in Hutchinson-Gilford progeria syndrome through use of stable isotope labeling with amino acids in cell culture. *Journal of Proteomics*. https://doi.org/10.1016/j.jprot.2013.08.008

Sanchez, S. E., Petrillo, E., Beckwith, E. J., Zhang, X., Rugnone, M. L., Hernando, C. E.,

Cuevas, J. C., Godoy Herz, M. A., Depetris-Chauvin, A., Simpson, C. G., Brown, J. W. S., Cerdán, P. D., Borevitz, J. O., Mas, P., Ceriani, M. F., Kornblihtt, A. R., & Yanovsky, M. J. (2010). A methyl transferase links the circadian clock to the regulation of alternative splicing. *Nature*. https://doi.org/10.1038/nature09470

Scaffidi, P., & Misteli, T. (2006). Lamin A-dependent nuclear defects in human aging. *Science*. https://doi.org/10.1126/science.1127168

Schreiber, K. H., & Kennedy, B. K. (2013). When Lamins Go Bad: Nuclear Structure and Disease. *Cell*, *152*(6), 1365–1375. https://doi.org/10.1016/J.CELL.2013.02.015

Shkreta, L., & Chabot, B. (2015). The RNA Splicing Response to DNA Damage. *Biomolecules 2015, Vol. 5, Pages 2935-2977*, *5*(4), 2935–2977. https://doi.org/10.3390/BIOM5042935

Sibley, C. R., Blazquez, L., & Ule, J. (2016). Lessons from non-canonical splicing. *Nature Reviews Genetics 2016 17:7*, *17*(7), 407–421. https://doi.org/10.1038/nrg.2016.46

Smyth, G. K. (2005). limma: Linear Models for Microarray Data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. https://doi.org/10.1007/0-387-29362-0_23

Sprung, C. N., Li, J., Hovan, D., McKay, M. J., & Forrester, H. B. (2011). Alternative Transcript Initiation and Splicing as a Response to DNA Damage. *PLOS ONE*, *6*(10), e25758. https://doi.org/10.1371/JOURNAL.PONE.0025758

Staiger, D., & Simpson, G. G. (2015). Enter exitrons. *Genome Biology*, *16*(1), 1–3. https://doi.org/10.1186/S13059-015-0704-3/METRICS

Takeuchi, H., & Rünger, T. M. (2013). Longwave UV light induces the aging-associated

progerin. *Journal of Investigative Dermatology*. https://doi.org/10.1038/jid.2013.71

Tammer, L., Hameiri, O., Keydar, I., Roy, V. R., Ashkenazy-Titelman, A., Custódio, N., Sason, I., Shayevitch, R., Rodríguez-Vaello, V., Rino, J., Lev Maor, G., Leader, Y., Khair, D., Aiden, E. L., Elkon, R., Irimia, M., Sharan, R., Shav-Tal, Y., Carmo-Fonseca, M., & Ast, G. (2022). Gene architecture directs splicing outcome in separate nuclear spatial regions. *Molecular Cell*, *82*(5), 1021-1034.e8. https://doi.org/10.1016/J.MOLCEL.2022.02.001

Tan, D. Q., Li, Y., Yang, C., Li, J., Tan, S. H., Chin, D. W. L., Nakamura-Ishizu, A., Yang, H., & Suda, T. (2019). PRMT5 Modulates Splicing for Genome Integrity and Preserves Proteostasis of Hematopoietic Stem Cells. *Cell Reports*. https://doi.org/10.1016/j.celrep.2019.02.001

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*. https://doi.org/10.1186/s13059-018-1417-1

Ustianenko, D., Weyn-Vanhentenryck, S. M., & Zhang, C. (2017). Microexons: discovery, regulation, and function. *WIREs RNA*, *8*, 1418. https://doi.org/10.1002/wrna.1418

Vo, J. N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., Wu, Y. M., Dhanasekaran, S. M., Engelke, C. G., Cao, X., Robinson, D. R., Nesvizhskii, A. I., & Chinnaiyan, A. M. (2019). The Landscape of Circular RNA in Cancer. *Cell*, *176*(4), 869-881.e13. https://doi.org/10.1016/J.CELL.2018.12.021

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L.,

Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., … Kling, T. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics 2013 45:10*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Worman, H. J. (2012). Nuclear lamins and laminopathies. *Journal of Pathology*, *226*(2), 316–325. https://doi.org/10.1002/PATH.2999

Xiong, Z. M., Choi, J. Y., Wang, K., Zhang, H., Tariq, Z., Wu, D., Ko, E., Ladana, C., Sesaki, H., & Cao, K. (2016). Methylene blue alleviates nuclear and mitochondrial abnormalities in progeria. *Aging Cell*, *15*(2), 279–290. https://doi.org/10.1111/acel.12434

Young, S. G., Meta, M., Yang, S. H., & Fong, L. G. (2006). Prelamin A farnesylation and progeroid syndromes. In *Journal of Biological Chemistry* (Vol. 281, Issue 52, pp. 39741–39745). Elsevier. https://doi.org/10.1074/jbc.R600033200

Zhang, H., Xiong, Z. M., & Cao, K. (2014). Mechanisms controlling the smooth muscle cell death in progeria via down-regulation of poly(ADP-ribose) polymerase 1. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1320843111

Zhang, J., Kuo, C. C. J., & Chen, L. (2011). GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics*, *12*(1), 1–11. https://doi.org/10.1186/1471-2164-12-90/FIGURES/7

Zhang, X. H. F., Leslie, C. S., & Chasin, L. A. (2005). Dichotomous splicing signals in exon flanks. *Genome Research*, *15*(6), 768–779. https://doi.org/10.1101/GR.3217705