PhD degree in Molecular Medicine

(Curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Disciplinary sector: Computational Biology Laboratory, IRCCS Eugenio Medea, Bosisio Parini

Computational Approaches in the Estimation and Analysis of Transcripts Differential Expression and Splicing: Application to Spinal Muscular Atrophy

By: Shikha Vashisht

IEO, Milan

Matricola R10339

Supervisor: Dr. Uberto Pozzoli
IRCCS Eugenio Medea, Bosisio Parini

Added Supervisor: Dr. Fabrizio Bianchi
IEO, Milan

Anno accademico 2015-2016

ABSTRACT

Spinal Muscular Atrophy (SMA) is among the most common genetic neurological diseases that cause infant mortality. SMA is caused by deletion or mutations in the survival motor neuron 1 gene (SMN1), which are expected to generate alterations in RNA transcription, or splicing and most importantly reductions in mRNA transport within the axons of motor neurons (MNs). SMA ultimately results in the selective degeneration of MNs in spinal cord, but the underlying reason is still not clear entirely. The aim of this study is to investigate splicing abnormalities in SMA, and to identify genes presenting differential splicing possibly involved in the pathogenesis of SMA at genome-wide level. We performed RNA-Sequencing data analysis on 2 SMA patients and 2 controls, with 2 biological replicates each sample, derived from their induced Pluripotent Stem Cell-differentiated-MNs. Three types of analyses were executed. Firstly, differential expression analysis was performed to identify possibly mis-regulated genes using Cufflinks. Secondly, alternative splicing analysis was conducted to find differentially-used exons (DUEs; using DEXSeq) as splicing patterns are known to be altered in MNs by the suboptimal levels of SMN protein. Thirdly, we did RNA-binding protein (RBP) - motif discovery for the set of identified alternative cassette-DUEs, to pinpoint possible mechanisms of such alterations, specific to MNs. The gene ontology enrichment analysis of significant DEGs and alternative cassette-DUEs revealed various interesting terms including axon-guidance, muscle-contraction, microtubule-based transport, axon-cargo transport, synapse etc. which suggests their involvement in SMA. Further, promising results were obtained from motif analysis which has identified 22 RBPs out of which 7 RBPs namely, *PABPC1, PABPC3, PABPC4, PABPC5, PABPN1, SART3 and KHDRBS1* are known for mRNAs stabilization and mRNA transport across MN-axon. Five RBPs from PABP family are known to interact directly with SMN protein that enhance mRNA transport in MNs. To validate our results specific wet-lab experiments are required, involving precise recognition of RNA-binding sites correspondent with our findings. Our work has provided a promising set of putative targets which might offer potential therapeutic role towards treating SMA.

During the course of our study, we have observed that current methods for an effective understanding of differential splicing events within the transcriptomic landscape at high resolution are insufficient. To address this problem, we developed a computational model which has a potential to precisely estimate the “transcript expression levels” within a given gene locus by disentangling mature and nascent transcription contributions for each transcript at per base resolution. We modeled exonic and intronic read coverages by applying a non-linear computational model and estimated expression for each transcript, which best approximated the observed expression in total RNA-Seq data. The performance of our model was good in terms of computational processing time and memory usage. The application of our model is in the detection of differential splicing events. At exon level, differences in the ratio of the sum of mature and the sum of nascent transcripts over all the transcripts in a gene locus gives an indication of differential splicing. We have implemented our model in R-statistical language.

PREFACE

This thesis will provide the detailed work that I have performed during four years of my PhD course. I have divided my work in four chapters: Introduction, Materials and Methods, Results and Discussion. In **Chapter 1**, a detailed introduction has been provided for our research area with current technologies to study the proposed biological problems. In **Chapter 2**, the details of implemented methods have been provided that are divided into two separate studies: (1) **Study of Alternative Splicing in SMA** and (2) **Development of Computational Model to Estimate Transcript Expression**. In this chapter, firstly, the materials and methods for study (1) have been described to investigate the alternative splicing patterns and their mis-regulations in SMA pathology. Secondly, materials and methods for the development of computational model are described. In **Chapter 3**, results obtained from the implemented methods are given in two separate sections, containing the results from: Study of Alternative Splicing in SMA and Development of Computational Model to Estimate Transcript Expression. In **Chapter 4**, the research problem, our implemented methods, major findings have been discussed and finally the obtained results have been concluded with future directions to subsequently enhance the current work in both studies. This chapter also contains two sections: Study of Alternative Splicing in SMA and Development of Computational Model to Estimate Transcript Expression.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Dr. Uberto Pozzoli who gave me such a wonderful opportunity to work in his laboratory. I thank him for his continuous guidance and support during the entire course of my PhD.

I am extremely thankful to Erika Molteni for the useful discussions, her valuable suggestions and continuous encouragement during the complete phase of my research. I especially thank her for all her great contributions and patience she has provided for reading through all my thesis drafts and always giving me useful suggestions, which added a good quality to the work. I thank all my lab mates, specially Giorgia Menozzi for being with me in all my good and bad times. I am thankful to her for motivating me, especially during my thesis writing journey.

I thank Dr. Stefania Corti and colleagues for providing us RNA-Seq data for the analysis. Most importantly, I would like to thank Cariplo foundation for funding my PhD research project. I thank both my internal and external supervisors for giving me their useful suggestions and recommendations to enhance the quality of my thesis.

Finally, my special thanks goes to my parents, my grandparents and my husband for having faith in my abilities and always giving me the shelter of unconditional love when I needed the most.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Dedicated to my grandfather...

In total my thesis contains 4 chapters, 42 figures, 6 tables and 300 references.

CHAPTER 1 - INTRODUCTION

Background

In this chapter, we intend to describe the relevant fundamentals of cellular processes on our area of study. The whole genetic stories revolve around the genetic information keeper DNA, which perform two essential tasks within every single cell: first is replication and second is transcription. During the transcription process DNA tends to convey its information as messages for producing an expression (as proteins). Messenger-RNA (mRNA) carries these messages which further undergo pruning and maturation, retaining only the coding sequences. Such post-transcriptional modifications of precursor-mRNA (pre-mRNA) involve several intricate mechanisms which require great specificity and fidelity. We are interested in studying the mechanisms ruling the preferential choice of some coding regions over the others and their aberrations, leading to fatal disorders.

We will start by describing post-transcriptional processes, their regulatory mechanisms and how mis-regulations in these mechanisms could impact our cellular systems. Further, we will describe the established methods to study these processes which include Next Generation Sequencing (NGS) methods and their key applications such as RNA-Sequencing technology to study transcriptome of a cell in a given time-point and transcription level variations between two different conditions can be analyzed. In order to perform such analysis, we have described computational tools suitable to quantify the expression levels of genes, transcripts and individual exons and to identify relative differences in their expressions levels between two different conditions. Further, to analyze the differential alternative splicing, we have introduced a novel computational model which has a potential to precisely estimate the “transcript expression levels” within a given gene locus by disentangling mature and nascent transcription contributions for each transcript at per base resolution.

1.1 Precursor messenger RNA Splicing

In the late 1970s, a surprising discovery revealed that the genes in eukaryotic cells are not continuous stretch of coding sequences (‘exons’) rather they are interrupted by very long intervening noncoding sequences, designated as ‘introns’[1,2]. Therefore, a journey from transcription to translation requires an additional albeit indispensable step, namely, ‘RNA splicing’, which help in the removal of such noncoding sequences from the Precursor messenger RNA (pre-mRNA) and ligate the coding sequences together, producing mature mRNA. The faithful processing of pre-mRNA

is very essential for the synthesis of functionally active protein in the cellular system. Therefore, in eukaryotic cells splicing is a critical step towards gene expression.

1.2 Biogenesis of snRNP Particles

Pre-mRNA splicing is carried out by an intricate macromolecular machinery called spliceosome which is made up of five small nuclear ribonucleoproteins (snRNPs) and more than 100 other essential proteins[3]. Wherein snRNPs are RNA-protein complexes fabricated of Uridine-rich small nuclear Ribonucleic acid (U snRNAs; non-coding class of RNAs) and large set of proteins. These non-coding species of RNAs are mainly classified as: (1) Sm snRNAs and (2) Sm-like (Lsm) snRNAs on the basis of sequence similarity and protein cofactors[4]. The Sm class contains U1, U2, U4, U4atac, U5, U11 and U12 snRNAs which are transcribed by RNA polymerase II, containing three essential recognition elements: 5′-trimethylguanosine (TMG) cap, Sm-protein-binding site (Sm-site) and 3′ stem–loop structure. Whereas Lsm class have U6 and U6atac snRNAs, containing 5'-γ-monomethylphosphate cap and a 3' stem–loop which are transcribed by RNA polymerase III. After the transcription, Sm-snRNAs moves out from the nucleus into the cytoplasm for post-transcriptional processing steps, however Lsm-snRNAs always stay inside the nucleus. Therefore, the whole snRNPs biogenesis process including all additional maturation steps is centered upon Sm-class snRNPs (**Figure 1.1**). The Sm-snRNPs biogenesis starts from the transcription of snRNA in the nucleus and the transcribed pre-snRNA binds with various export proteins to facilitate the transport within the cytoplasm through Nuclear Pore Complex (NPC). The export machinery is comprised of Phosphorylated adapter RNA export protein (PHAX), the export receptor Chromosome Region Maintenance-1 (CRM1 or exportin-1), the cap-Binding Complex (CBC; containing CBP80 and CBP20 domains, which specifically binds with PHAX) and GTP-bound RAs-related Nuclear protein (RAN GTP-ase). The PHAX protein shows distinctive behavior with its phosphorylation status such as its hyperphosphorylated form confines it within the nucleus whereas its hypophosphorylated form makes it cytoplasm-specific and favors pre-snRNA export[5,6]. Later, the export machinery disassociates from pre-snRNA followed by its assembly onto seven Sm-proteins (arranged as a ring like structure containing Sm B, D3, D1, D2, E, F and G to form a Sm-core RNP) with the help of Survival of Motor Neuron protein complex (SMN protein complex)[7]. The SMN complex is a macromolecular protein complex, containing self-oligomers of SMN protein associated with GEMIN2-8 proteins and unr-interacting protein (UNRIP). GEMINS have acquired their name from the dot-like sub-nuclear structures called ‘gems’ where they localize with the SMN protein.

The SMN protein complex specifically binds with snRNA's conserved regions (Sm-site and 3'-stem loop) which promotes the stable assembly of snRNA onto Sm-proteins and form a snRNP particle. Further, the 5'-mono-methylated cap (7-methylguanosine or m7G-cap) of snRNA is hypermethylated into tri-methylated cap (2, 2, 7-trimethylguanosine or m_3G-cap or TMG) by trimethylguanosine synthase-1 (TGS1) enzyme and 3'-end of snRNA is trimmed by exonuclease activity. These modifications facilitate the final re-import of processed snRNP particle into the nucleus with the help of nuclear import proteins, namely snurportin-1 (SPN) and importin-β (Imp-β)[8,9]. Within the nucleus snRNP first localized in the cajal bodies (CBs; sub-nuclear organelles) for the further maturation steps[10] and finally gets stored in "Interchromatin Granule Clusters" (IGC) or "nuclear speckles" for the pre-mRNA splicing activity[11] **(Figure 1.1)**.

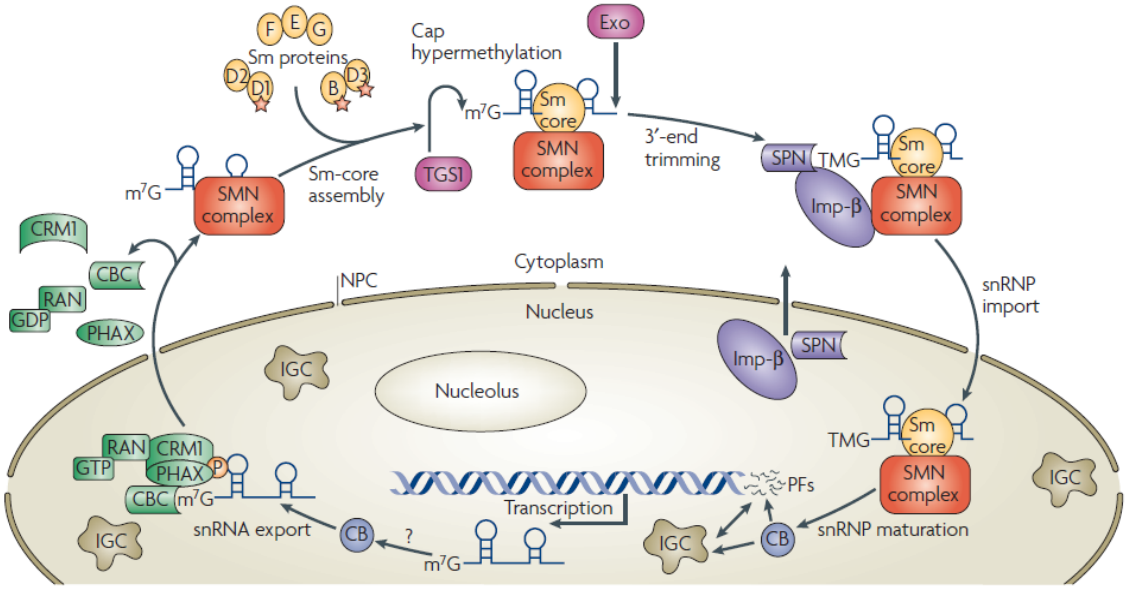


Figure 1.1: The transcription of small nuclear RNA and its processing into the functional small nuclear Ribonucleoprotein particles (snRNPs).

The process initiates with snRNA transcription in the nucleus, followed by its export within the cytoplasm by the help of various export proteins. Later, a series of post-transcription processing steps are carried out to generate a stable snRNP and further its re-import take place within the nucleus to follow the final maturation steps and utilize it for mRNA splicing mechanism. For the detailed explanation of this process read through the main text. **This figure has been taken from a review article published by Matera et al.**[4]

1.3 Spliceosome: A Highly Dynamic Machinery

Nuclear pre-mRNA splicing is mediated by an ordered assembly of spliceosome components which endure large number of structural rearrangements to attain catalytically active complex

form at the transcription site. Therefore, the spliceosomes are highly dynamic in their structural build. Basically they are divided into two distinct types: major and minor spliceosomes, on the basis of specific snRNPs in their core built.

- Major spliceosomes are composed of five main snRNPs i.e. U1, U2, U4, U5, U6 and several other ancillary proteins. This class of spliceosomes predominantly recognizes the canonical 5' and 3' splice-sites (GU-AG)[12].
- Minor spliceosomes also composed of five main snRNPs i.e. U11, U12, U4atac, U6atac and U5 with large number of other proteins and process the splicing of rare introns, containing non-canonical splice-sites (AU-AC splice-sites).

These machineries operate via multitude of RNA-protein, RNA-RNA, protein-protein interactions to enhance correct excise-and-ligate splicing reactions.

1.4 Major Spliceosome Mediated pre-mRNA Splicing

During the splicing reaction spliceosome complex performs several structural shifts at conformational and compositional levels, being highly dynamic in nature[13–15]. Such rearrangements take place between snRNAs, spliceosome proteins and pre-mRNA through their interactions with each other, forming an active spliceosome complex. Approximately 99.24% of the splice site junctions have 5'-GU and 3'-AG di-nucleotide consensus within intronic sequences[12,16]. Therefore, most of the eukaryotic intron sequences are spliced-out by major spliceosome machinery through specific selection and base-pairing with intronic consensus regions in an ordered fashion. Such regions mainly include 5'-splice site (5'-ss; donor) followed by the 3'-branch point (BPS; ~20-40 nucleotide long adenosine-rich sequence), 3'-polypyrimidine-tract (PPT) and 3'-splice-site (3'-ss; acceptor)[17]. Altogether, PPT, BPS and 3'-ss builds the 3'-intronic consensus (**Figure 1.2)**. In order to correctly differentiate the long intronic sequences (approximately 10^4 to 10^5 nucleotides) from the short exonic sequences (~300 nucleotides or less for internal exons), prior interactions between 5'-ss and 3'-ss surrounding the exons are necessary[17,18] that are generally required for the initial splice-site recognition. The weakly conserved sequences within and around exonic region helps in this process by building a stable multi-factorial complex with additional splicing factors located within ~300 nucleotides space (**Figure 1.2**). It is evident that for the initial splice site recognition U1 and U2 snRNPs forms complex around the exons via RNA-protein and protein-protein interactions (such as Serine/Arginine rich or SR proteins interact with U1 snRNP and U2 auxiliary factors or U2AFs

interact with U2 snRNP). This is recognized as an "exon definition", forming an "exon definition complex" **(Figure 1.2)**. Similarly, U1-U2 undergoes subsequent rearrangements and perform interactions by spanning intronic sequence, forming an "intron definition complex" **(Figure 1.2)**. These interactions bring 5'-ss, 3'-BPS and 3'-ss in close proximity to establish appropriate pairings within them. This process also helps to find precise location of an authentic exon which further mediates the assembly of spliceosome components to carry out splicing.

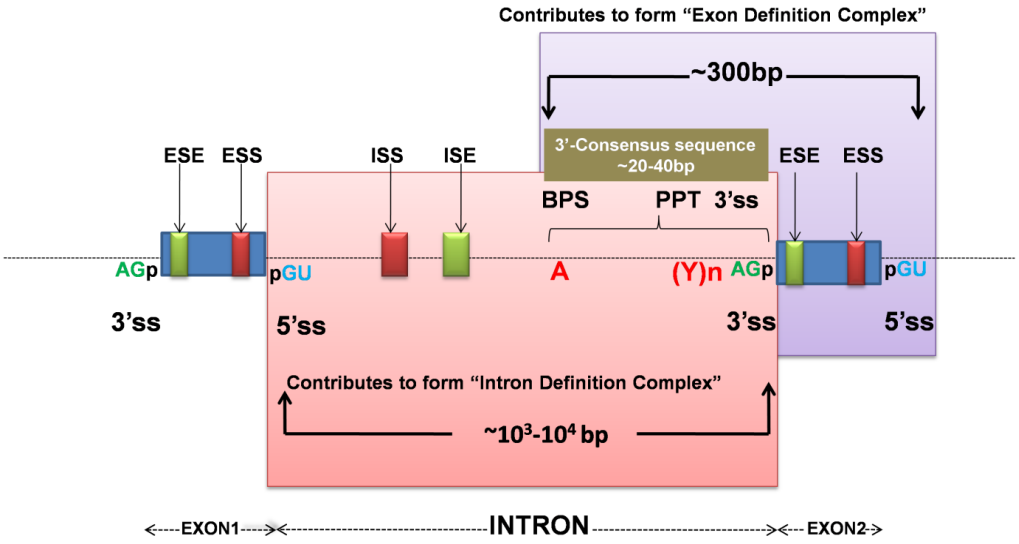


Figure 1.2: An organization of 5' and 3'-splice sites consensus with splicing regulatory elements (SREs).

The pre-mRNA comprises of short exons (represented with '*blue'* rectangles) interrupted with long intronic (represented with a '*black'* dotted line between EXON1 and EXON2) sequences. Within exonic and intronic sequences, "Exon Splicing Enhancer" (ESE; '*green'*) and Exon Splicing Silencer (ESS; '*red'*); "Intron Splicing Enhancer" (ISE; '*green'*) and Intron Splicing Silencer (ISS; '*red'*) elements can be present, respectively. The 5'-splice 5'-ss is located at EXON1-INTRON junction which consist of 'GU' dinucleotide consensus while the 3'-intronic region consists of 3'-BPS, 3'-PPT followed by the 3'-ss, consisting of 'AG' (within EXON2-INTRON junction), representing the 3'-consensus.

1.5 The Splicing Pathway

The complexity within splicing pathway arises due to the involvement of tremendous molecular participations and interactions within them (favors high precision)[19–21]. The mechanism initializes with an early recognition of the splice sites, where U1 snRNP base-pairs with 5'-ss in an ATP-independent manner. The 70kDa component of U1-snRNP interacts with SR proteins (trans-acting splicing factor) which are bound onto the Exon Splicing Enhancer (See **Figure 1.3** ESE; '*green'* rectangle localized within exonic region represented with '*blue*' rectangle) element within

the exons and stabilize protein-snRNA complex via protein-protein interactions. Further, the Splicing Factor 2 (SF2) binds with the 3'-BPS and subsequently U2AF joins the PPT or Y(n) sequence adjacent to 3'-BPS. U2AF is composed of 2 subunits: a larger subunit or U2AF 65kDa and a smaller subunit or U2AF 35kDa. The smaller subunit interacts with 3'-ss while the larger subunit interacts with PPT through its RNA Recognition Motif (RRM). All together the assembly of these spliceosome players completes the formation of E-complex[22] **(Figure 1.3)**. Further, U2 snRNP makes contact with BPS and forms a duplex through subsequent displacement of SF2. This step requires ATP hydrolysis and mediates the formation of A-complex **(Figure 1.3)**. Now, the pre-assembled snRNPs, containing U4/U6 and U5 joins above arrangement as U4/U6.U5 (a tri-snRNP complex) by base-pairing with 5'-ss, through an another ATP molecule breakdown and give rise to B-complex[23–25] **(Figure 1.3)**. But this complex is still catalytically inactive. Therefore, in order to generate a catalytically competent spliceosome, all of its elements undergo multiple conformational rearrangements via RNA-RNA, RNA-protein and protein-protein interactions. In this process, U6 immediately displaces from U4 and base-pairs with U2, that unravel the 5'-end of U6 and binds to 5'-ss. This arrangement facilitates the dissociation of U1 and U4, and the remaining components including U2, U5 and U6 form a catalytically active spliceosome B*-complex, which is ready to catalyze the first trans-esterification reaction of the splicing pathway **(Figure 1.3)**. In this step, BPS-2' hydroxyl group (2'-OH) attacks the phosphodiester bond at the 5'-ss, resulting in the cleavage of 5'-exon with free 3'-OH group and an intermediate 2'-5' branched lariat structure, forming C-complex **(Figure 1.3)**. Further, the remodeling of spliceosome components facilitate the second catalytic step of splicing, wherein 3'-OH of 5'-exon attacks the phosphodiester bond at the 3'-ss and performs second cleavage reaction[17,23,24,26]. Subsequently, ligation of the exons and removal of lariat structure takes place. The ligated exons get released from the spliceosome assembly by RNA helicase proteins (such as pre-mRNA splicing factor 22 or PRP22, PRP16, PRP17 and SLU7) which also trigger the spliceosome components disassembly and recycling[25,27–29,30,31] (**Figure 1.3**). Since the entire process undergoes several conformational changes, therefore specific proteins are also engaged to fulfill the high energy requirements such as proteins of DExD/H box family are essential for chaperoning the ATP hydrolysis[32,33]. Additionally, magnesium ions (mg^{2+}) are also required to stabilize the active RNA conformations during spliceosome assembly and splicing pathway[34,35].

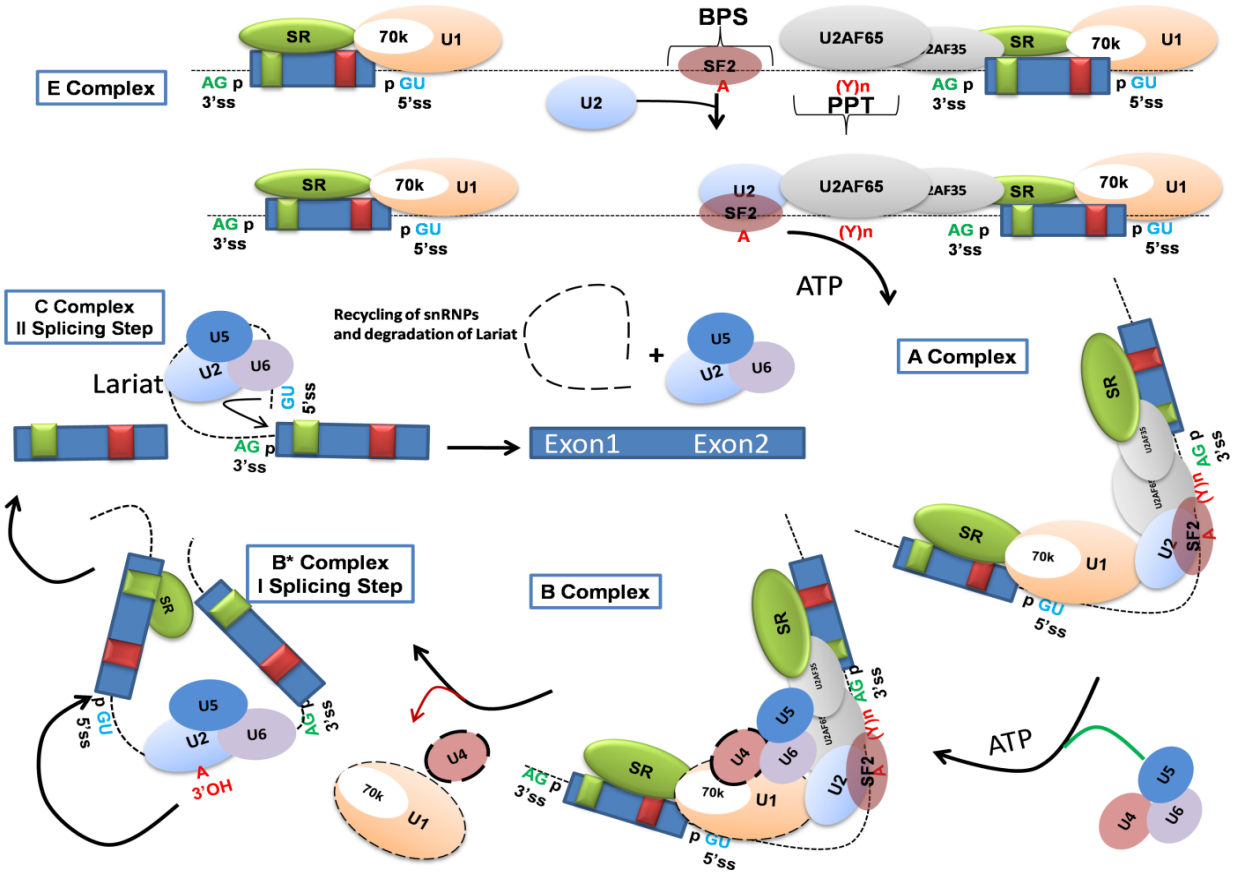


Figure 1.3: The schematic of the splicing pathway.

The pre-mRNA splicing initiates with the recognition of an authentic 5'ss and 3'-ss within the intronic sequence. In this schematic, the canonical splice sites 5'-GU (donor) and 3'-AG (acceptor) are shown which are recognized and spliced by major spliceosome components assembly **(See the main text for details)**.

1.6 Alternative Splicing: An Additional Dimension into Transcriptional Landscape

Spliceosome is not only capable in accurate recognition of the splice-sites but also has an ability to choose variety of different splice sites across pre-mRNA sequence and produce a large set of distinct transcripts from a single gene. This mechanism is known as alternative splicing (AS)[36–41]. AS is an important post-transcriptional phenomenon which serves as a pivotal medium for generating diverse set of proteins from a smaller number of genes. The human genome contains around 26,000 annotated genes, with approximately 2,33,785 exons and 2,07,344 introns. The mean length of a gene spans around 27kbp with a mean of 8.8 exons and mean length of 148.12 nucleotides (**Figure 1.4**). Mean length of intron is 3,365 nucleotides[42]. In general, every single gene has a capability to generate at least 2 – 3 alternative transcript forms, but some interesting exceptions such as *neurexin3* gene (a neuropeptide receptor in

humans) are also present which produces 1,728 different transcripts through AS mechanism[43]. More than ~95% of the human genes are known to produce alternative splice variants, suggesting its ubiquitous impact throughout the human genome[44]. Moreover, AS acts in a cell-type specific manner and it has been observed in all the tissues with relatively higher frequency in the nervous tissue[45].

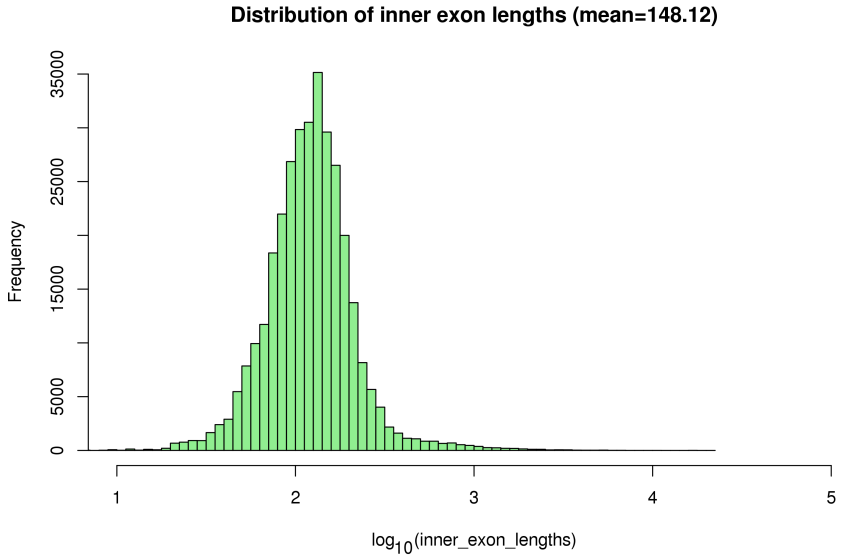


Figure 1.4: A bar plot illustrating internal exons length distribution in the human genome.
This distribution is based on the RefSeq hg19 gene-model annotations with "validated" and "reviewed" gene-tags. The total number of internal exons are 3,28,272 with their mean length of 148.12 bp.

1.7 Alternative Splicing Types

Alternative splicing can occur in five different types of patterns[46,47] (**Figure 1.5**):

- **Exon skipping:** It is the most frequent form of AS pattern in mammals. Wherein certain exon gets excluded (skipped) from one transcript but preserved in another transcript of a gene. Such exons are known as cassette exons.
- **Alternative 5'-Splice Site:** is a type of AS pattern where different 5'-donors gets utilized.
- **Alternative 3'-Splice Site:** is a type of AS pattern where different 3'-acceptors gets utilized.
- **Mutually exclusive exons:** is a type of AS pattern where selection of one of the two mutually exclusive exons take place by which only one of the two exons is retained per transcript but never both in a same transcript.

- **Intron retention:** During the pre-mRNA splicing process, if the intronic sequence remains unidentified by splicing factors then it gets retained in the mature mRNA and codes for a protein (mostly a non-functional protein). This AS event is less frequent as compared to other AS events among humans[48,49]. Braunschweig and colleagues have explained the specific essentiality of intron retention AS event in tuning expression of mammalian transcriptomes[50].

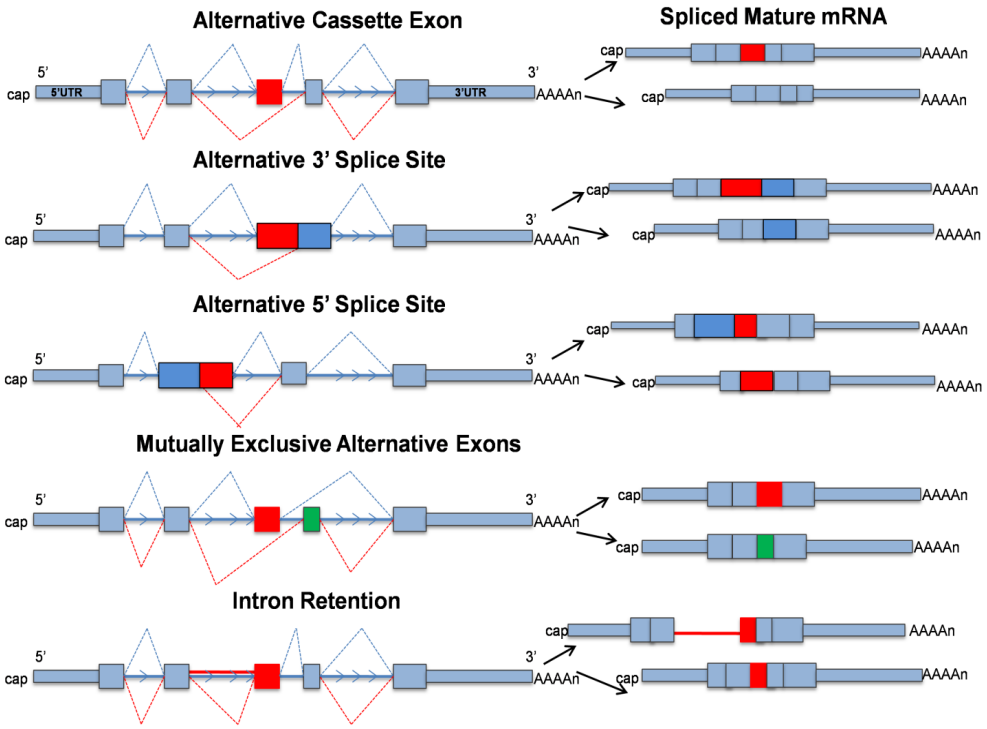


Figure 1.5: The representation of five canonical types of AS patterns in the eukaryotic genes.
In unspliced mRNAs (on the left hand side) the exons are represented with *'blue', 'red'* and *'green'* colored thick boxes and introns are represented with '*blue'* lines (in forward direction). The alternatively spliced transcripts are represented with ligated set of exons (on the right hand side). For each AS pattern two paths have been represented; path-1 and path-2 which are illustrated with *'red'* and '*blue'* colored dotted lines, respectively. In Alternative Cassette Exon or exon skipping AS event, skipping of an exon 3 is shown. In Alternative 3'-Splice Site selection event, an exon-3 has two alternative 3'-acceptor sites. In Alternative 5'-Splice Site selection event, an exon-2 has two alternative 5'-donor sites. In Mutually Exclusive AS event, exon-3 and exon-4 are mutually exclusive exons represented with '*red'* and *'green'* colored thick boxes, respectively. In intron retention event, the retained intron sequence is shown with '*red'* thick line.

1.8 Regulation of Alternative Splicing

AS must be highly regulated for accurate productivity and mis-regulations in AS patterns are known to be implicated in several fatal human diseases, in particular neurodegenerative disorders due to the longer length of genes expressed in this tissue[51]. More than 15% of the genetic diseases are known to be caused by subtle perturbations in AS patterns which are indeed very difficult to detect[52]. AS is mainly regulated by cis-acting regulatory elements (located within exonic or intronic sequences) by either favoring the exon inclusion or exclusion from the transcript. The relative location and function of these elements differentiate them into 'splicing enhancers' and 'splicing silencers'. For instance, if they are localized within exonic region they are called Exon Splicing Enhancers (ESEs) and Exon Splicing Silencers (ESSs) whereas if they are present within intronic region then they are called Intron Splicing Enhancers (ISEs) and Intron Splicing Silencers (ISSs), representing overall the class of Splicing Regulatory Elements (SREs). The SREs probe their effect by binding with trans-acting splicing factors[52,53] and regulates the splicing mainly during an initial step of spliceosome assembly which involves exon-definition complex and intron-definition complex formation.

The most extensively studied trans-acting splicing regulatory proteins are classified into three categories: (1) SR proteins[54–56], (2) heterogeneous nuclear ribonucleoproteins (hnRNPs)[57] and (3) tissue-specific RNA–binding proteins[58–61]. SR proteins chiefly contains one or two copies of RNA-Recognition Motifs (RRM) and the C-terminal arginine/serine dipeptide-rich (RS) domain[22,26,62,63]. The RRMs facilitates RNA-binding activity and RS domain helps in protein-protein interactions in a sequence specific manner. Few key examples of SR proteins include Splicing Factor 2 or Alternative Splicing Factor (SF2/ASF), U1-70k snRNP (70kDa) and U1 snRNP C proteins (U1-C). The family of SR proteins is highly conserved throughout metazoans[64] and are extensively studied to play essential roles in constitutive and alternative splicing splice-site selections. Notably, SR proteins are known to enhance the inclusion of exon (such as ASF1/SF2)[65], whereas hnRNPs are known to silence the exon inclusion (such as Polypyrimidine tract binding protein and RNA Binding Motif 5 protein)[66,67]. Further, in tissue-specific splicing regulatory factors, *Nova* is a very well-studied neuron-specific splicing regulator which can act as either splicing enhancer or silencer. It has been identified to regulate large set of genes in neurons and perform correlated functions implicated in pre-synaptic and post-synaptic neuronal activities[68]. Moreover, *Nova* was the first neuron-specific splicing regulatory factor discovered in mammals[61,69].

1.9 Implications of Alternative Splicing Disruptions in Human Diseases

AS demands highly controlled operations to yield genuine transcripts and, if diverts from normality it can result into serious health issues[70,71]. To date many human diseases are primarily caused by mis-regulations in AS patterns[72]. These diseases might be caused by the mutations, which in turn either formulate disruptions in the splicing of specific genes or interferes with the assembly of spliceosome machinery[73]. For instance, several genetic disorders are known to be caused by loss of spliceosome biogenesis or its function. In particular, neurodegenerative diseases which are mainly caused by the subtle variations in AS patterns. Due to the longer length of genes expressed in nervous tissue, neurons have higher rate of AS events with higher mis-regulation frequency[74–76]. Such as *Frontotemporal dementia with parkinsonism* associated with chromosome 17 is caused by mutations in microtubule-associated protein tau encoding gene (MAPT) which leads to the multiple mis-regulations in exon10 of tau gene[77]. *Alzheimer's disease* is also linked with abnormalities in the AS patterns of tau gene[78,79]. Another very common neurodegenerative disorder called *Spinal Muscular Atrophy* (SMA) is caused by the point mutations in exonic regulatory sequence of Survival Motor Neuron 1 gene (SMN1) that results in the loss of its function[80]. In my thesis work, one of our goal is to study SMA pathology in order to identify mis-regulations in the AS patterns of the SMA patients with respect to the healthy controls by analyzing transcriptomic data (using RNA-Sequencing technology) obtained from their Induced Pluripotent Stem Cells (iPSCs) derived Motor Neurons.

1.10 Spinal Muscular Atrophy (SMA)

SMA is characterized by the degeneration of alpha-motor neurons located in the anterior horn of spinal cord which are essential for the quick transmittance of nerve impulses to voluntary skeletal muscles **(Figure 1.6)**. It is a genetic disease which inherits in an autosomal recessive pattern and known to be a second leading cause of infant mortality after cystic fibrosis[81]. The estimated incidence of SMA is 1 in 10000 live births with the carrier frequency of approximately 1 in 50 individuals[82–85]. SMA manifests with high degree of genetic heterogeneity between patients, therefore it is classified into four phenotypes on the basis of age of onset and severity level that vary between acute to milder forms[84,86,87] **(Figure 1.7)**. Type I SMA (Werdnig-Hoffmann disease) is the most common and severe form of SMA which account for more than 50% of patients that are diagnosed with SMA[88]. The patients who suffer from severe form of SMA typically manifest pathology quite early in their life, usually before 6 months of age with a rapid and unexpected onset. The gradual loss of lower motor neurons causes failure of the

major body organs, especially the respiratory system. The children who are diagnosed with SMA type I never manage to sit unaided, they lack head control due to extremely poor muscle tone, and they experience bulbar denervation, accumulation of secretions in the lungs, causing respiratory distress which eventually leads to death within 2 years. The most fatal form of SMA type I is referred as SMA type 0 and is diagnosed in those who are born extremely weak (hypotonic) and merely survive few weeks even with intensive respiratory support. In type II SMA patients, onset of disease is usually noticed between 7 to 18 months, where children develop the ability to sit independently, but they are incapable to stand or walk, they also have respiratory troubles. Despite of fairly diminished life expectancy, they live well into adulthood[89]. SMA type III or juvenile form of SMA usually reveals after 18 months of age. They have an ability to walk without support, albeit many lose this ability later in their lives. The respiratory system involvement is less apparent, and their life expectancy is near to normal[90,91]. Type IV SMA patients have disease onset in their adulthood (> 18 years) where they experience very mild course of the pathology. This group of patients has the ability to walk in their adulthood and experience little to no respiratory and nutritional troubles. However, the severe benchmarks linked with the SMA pathology are changing with improved respiratory and nutrition care[92,93].

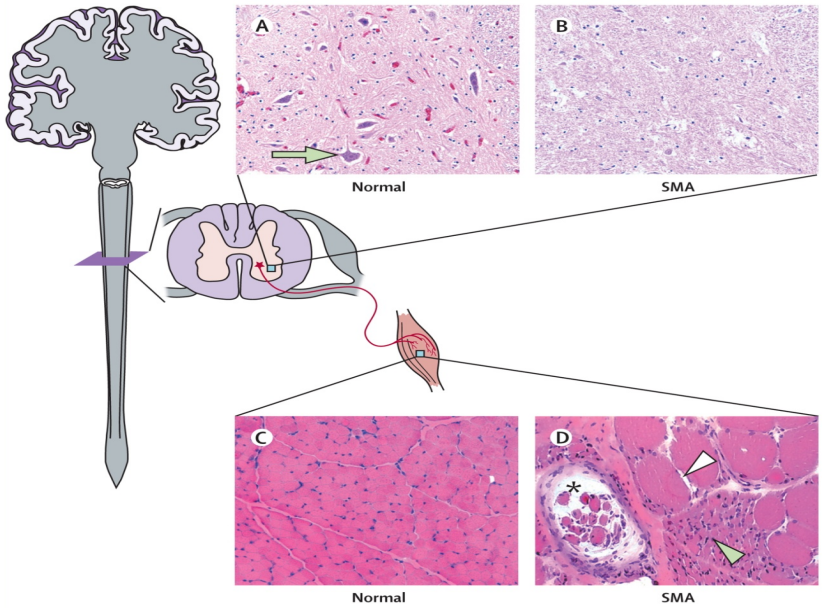


Figure 1.6: A cross section of the spinal cord, representing the morphological differences between SMA-patient and healthy control.

A In the anterior horn region of the spinal cord tissue in healthy person has uniform distribution of the alpha-motor neuron cells (shown as a '*green*' arrow) while in **B** In the anterior horn region of the spinal cord of a SMA patient

represents noticeable degeneration of alpha-motor neuron cells. **D** Skeletal muscle of SMA patient shows hypertrophic fibers with wasted muscles (shown with '*white'* and '*green'* arrow heads, respectively), conversely **C** demonstrates healthy muscle with uniform distribution of muscle fibers. The area highlighted with '*black'* asterisk represents muscle spindles which remains unaffected and but more distinctly visible in SMA patient with respect to healthy control. **The figure has been taken from a review article published by Lunn et al.**[80].

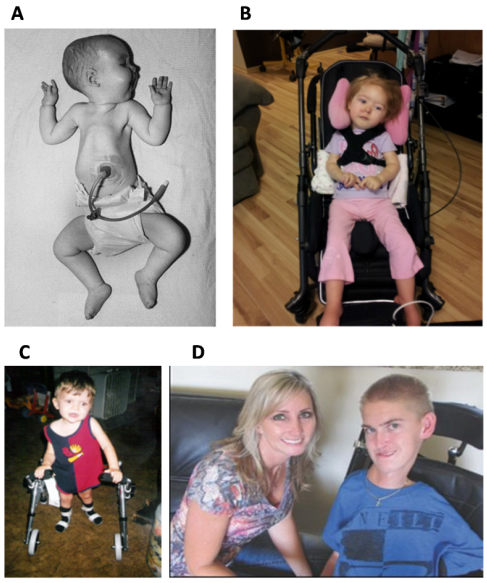


Figure 1.7: The classification of SMA phenotypes.

A An infant suffering from SMA type I, showing bell shaped lungs, and extremely poor muscle tone (hypotonic condition); **B** and **C** The children represent the SMA type II condition, **D** a patient with SMA type III/SMA type IV. **These images are taken from the google search about patients suffering from SMA.**

1.10.1 SMA Pathology Cause and Genes Involved

Brzustowicz et al. in 1990 investigated 13 clinically heterogeneous SMA groups in order to identify their genetic location in the human genome. They noticed albeit of the phenotypic variations in all groups the genetic location was same located on chromosome 5 at 5q11.2–q13.3 region[94] which was refined later[95–97]. Several investigations explained further the high instability within identified genomic region of SMA due to intrachromosomal rearrangements (such as duplications, gene conversions and deletions)[98,99]. SMA-specific region was further detailed by Lefebvre and colleagues as genetic and physical maps using pulsed-field gel electrophoresis (PFGE) coupled with single-stranded conformation polymorphism (SSCP) analysis[100] and successfully cloned the first novel gene responsible for SMA that was designated as *Survival Motor Neuron* gene (SMN). The 20kb SMN gene has 9 exons interrupted by 8 introns[101] and localized within a highly complex genomic region which exists as

large inverted duplicated region (500 kbp). Further, two homologous and inversely duplicated SMN genes were identified; a telomeric copy of SMN (SMN1) and a centromeric copy of SMN (SMN2) gene. Major cohorts of SMA patients (> 98%) were observed to have frameshift homozygous deletions within SMN1 gene[100,102,103] which makes it a SMA-determining gene. SMN1 and SMN2 genes have only five base pair of discrepancies. A synonymous mutation in exon 7 (nucleotide 6; TTC in SMN1 and TTT in SMN2; **Figure 1.8**), a single nucleotide variation at 3' non-coding region in exon 8 (nucleotide 1286; TGG in SMN1 and TGA in SMN2; **Figure 1.8**), and three single base substitutions within sixth and seventh intron **(Figure 1.8)**. Despite the slight differences between SMN1 and SMN2, the two genes do encode identical proteins but single nucleotide transition at position 6 of exon 7 (C to T) cause its skipping from SMN2 transcripts and produce a truncated SMN protein. SMN2 gene produces 90% of the truncated transcripts which encode non-functional *SMN-del7* protein that degrades rapidly[104] and only 10% of the full-length (FL) transcripts containing exon 7 and encode FL-SMN protein **(Figure 1.9)**. Therefore, SMN2 gene is insufficient to compensate the loss of SMN1 gene[100,105,106] which produce FL transcripts that encode a ubiquitous functional SMN protein (containing 294 amino acids). However, all patients retain at least one copy of SMN2 gene and it is evident that the SMA severity level is inversely correlated with SMN2 copy number. Milder patients generally have high copy number of *SMN2* gene, producing higher levels of the FL SMN protein than do the severely affected ones[107]. Therefore, SMA is caused by the deficiency of the SMN protein that is essential for carrying out the vital cellular activities.

A couple of studies explained the downstream effect of single nucleotide change on alternative splicing of exon 7 in SMN2 gene. Cartegni and colleagues[108] explained the disruptions in the AS patterns due to aberrant C-to-T transition in exon 7 of SMN2 gene[108,109]. Another study has provided a contrasting explanation which says, due to the single nucleotide change ESS is created that is bound by splicing silencers (such as hnRNP A1) at exon 7 of SMN2 and cause its skipping from majority of SMN2 transcripts[110–112].

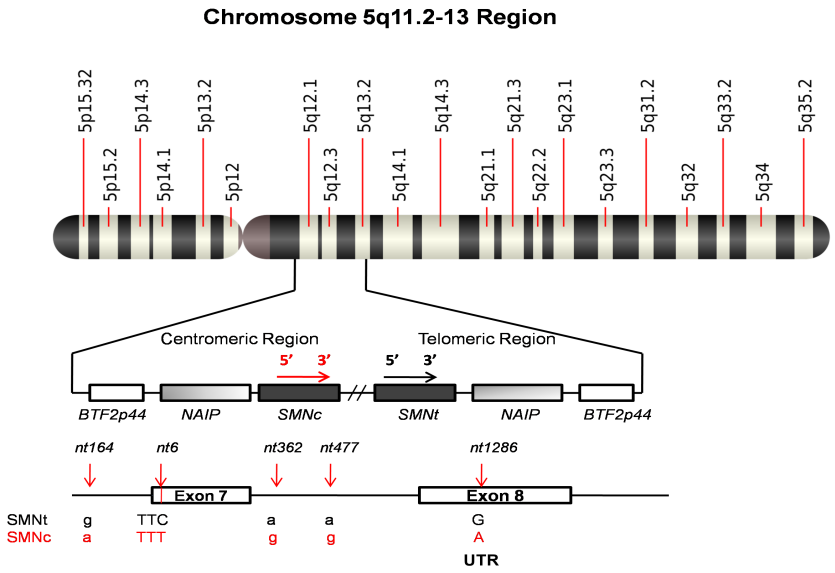


Figure 1.8: SMN genes located on chromosome 5 at chr5q11.2-13.3.

Two copies of SMN genes are mapped within chr5q11.2-13.3 region with other duplicated genes. The centromeric half represents the SMN2 gene (*SMNc*) location and the telomeric half represents SMN1 gene (*SMNt*) location. SMNc and SMNt genes differ by five nucleotide variations: (i) 'g-to-a' transition within intron 6 at nucleotide 164 (nt164); (ii) 'C-to-T' transition within Exon 7 at nt6; (iii), and (iv) two single nucleotide transitions from a-to-g' within intron 7 at nt362 and nt477, (v) 'G-to-A' transition within Exon 8 at nt1286. This position is located inside the 3'-UTR of Exon 8.

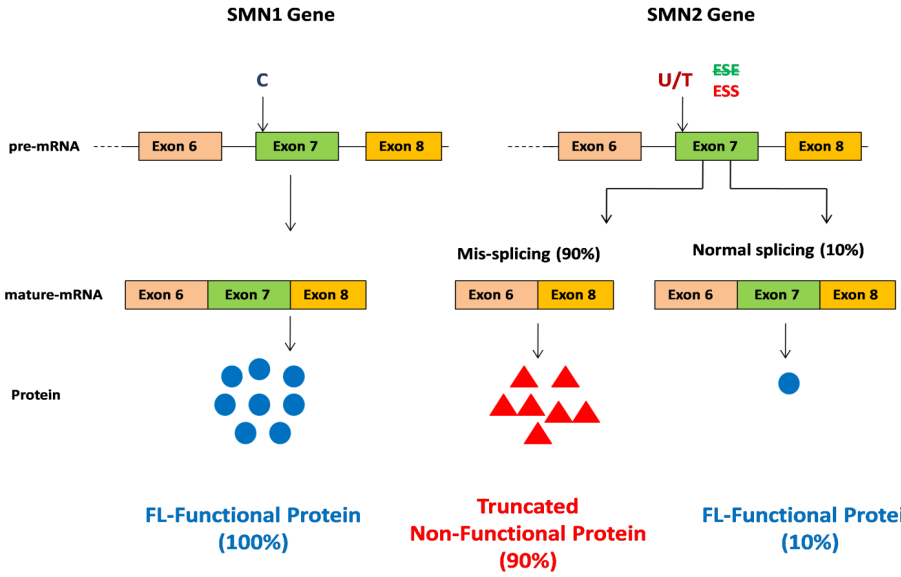


Figure 1.9: AS of SMN1 and SMN2 genes.

SMN1 gene has normal inclusion of exon 7 and produces full-length (FL) mature transcripts that encode functional SMN protein. In contrast, SMN2 gene undergoes C to T transition at position 6 of exon 7, causing its skipping due to

cis-acting ESE-SRE disruption which gets converted into ESS that is bound by splicing repressors (such as hnRNP A1) and promote its exclusion. SMN2 gene produce 90% non-functional SMN protein with only 10% FL-SMN protein.

1.10.2 SMN Protein, its Location and Function

Qing Liu and Gideon Dreyfuss reported the location of SMN gene within the novel nuclear structures which have been discovered incidentally by them while searching for hnRNP-interacting proteins (hnRNPRs) and named them as "gems"[113]. Gems are dot-like structures having granular appearance found to be enriched with SMN protein and often located in the close vicinity of cajal bodies (CBs). CBs are conserved subnuclear structures named after their discoverer, the Noble laureate Santiago Ramon y Cajal in 1903[114]. In addition to the nucleus, SMN is also present in the cytoplasm. Notably, gems and CBs share similarity in their physical properties including their copy number (2-6), size (very small in size 0·1–2·0 μm) and similar behavior towards metabolic cues[113,115]. Both structures are also shown to be regularly engaged in the state of assembly and disassembly during cell cycle, which further gives a notion to share strong functional relationships. CBs have been identified as one of the important processing locations of splicing snRNPs and other RNA processing factors. Therefore, they have a central role in the modification and maturation of splicing machinery. Moreover, SMN was also investigated to have a very important role in the assembly of snRNPs to form active spliceosome that perform pre-mRNA splicing in every cell[116–123] (**Figure 1.1**). Therefore, both of these nuclear structures are profoundly indispensable for post-transcriptional modifications of RNA. Later, Dreyfuss and colleagues collaboratively demonstrated that SMN protein is extremely unstable as a monomeric unit and forms discrete oligomers by self-associations. SMN oligomer recruits GEMIN2-8 and UNRIP, forming a macromolecular SMN-complex (40S to 80S). The SMN complex has evolved overtime by gradual block-wise addition of ancillary proteins[124]. Otter et al.[125] has presented a protein-protein interaction network of SMN complex. Miscellaneous studies shed light on the existence of self-association within SMN components, helping in strengthening the complex's stability through SMN self-oligomerization and eventually, its ubiquitous role in spliceosome assembly[116–120,126–132].

1.10.3 Motor Neuron Specific Functions of SMN Protein

Since SMA is specifically deleterious for motor neurons (MNs), therefore it is very important to study MN specific role of SMN protein and in this direction few hypotheses have been postulated, describing the SMA pathology. One hypothesis suggests the deficiency of SMN protein cause disruptions in spliceosome assembly and perturbs the splicing of selective set of

genes, which might be critical for MNs survival. Other studies suggests the essential role of SMN protein in neuronal processes such as mRNA transport to MN axons for their sustenance[133–139]. Given the fact that MNs-axons are quite long which makes their sub cellular RNA and protein localization considerably challenging and to perform this task large set of proteins works together where SMN plays a key role. In this support, a study has suggested, apart from the regular proteins involved in classical the SMN complex, SMN also partners with RBPs, including *hnRNP R, hnRNP Q, TDP 43, FMRP, HuD*[134]. Further, Wilfried Rossoll et al.[135] has provided an evidence for the direct interactions between SMN and *hnRNP R* RBP, which colocalizes in the axons and axon-terminals of MNs. *hnRNP R* is also known to interact directly with 3'-UTR of *β-actin* mRNA that assist its transport through axons and helps in the axonal growth. It has also been questioned that SMN protein might have significant roles in the development, maturation and stability of neuromuscular junctions (NMJs) and its lower levels promote the SMA pathology[140–144]. Various studies have presented contradictory findings by further investigating on this open question, involving NMJ involvement in SMA development. They have found the specific loss of central synapses, mainly proprioceptive inputs onto MNs somata and proximal dendrites which takes the precedence in the loss of MNs of mice SMA mice[145–147]. Moreover, the investigations regarding selective defects in NMJs they have found mostly all hind-limb muscles were fully innervated and also capable of eliciting muscle contractions in the studied mice models. Another latest study has shown the direct interaction of SMN with Ubiquitin-like modifier Activating enzyme 1 (*UBA1*) in neurons. The deficiency of SMN protein cause mis-regulations in the splicing patterns of *UBA1* and also reduces the *UBA1* expression levels which results in the perturbations of the key cellular mechanism of protein homeostasis, causing neurodegerenation[148]. All aforementioned studies have provided the remarkable understanding of the pathology; yet, the important questions still remain unclear such as which of the postulated studies are more relevant for targeting the disease or whether all applies equally with the disease progression.

1.10.4 Animal Models to Study SMA Pathobiology

In order to elucidate significant information towards the progression of SMA pathogenesis several SMA animal models were developed, including both vertebrates and invertebrates. The successful generation of SMA animal models was encouraged by a remarkable study which has presented the high conservation of the SMN protein during the evolutionarily processes among divergent species[149–151]. These studies also indicated the significant conservation of the

functional domains in SMN protein which are involved primarily in the RNA-binding processes[151] **(Figure 1.10)**. In addition to this, the effects of SMN mutation/depletion have been studied in *Caenorhabditis elegans*[152,153], *Drosophila melanogaster*[154], *Danio rerio*[140] and *Mus musculus*[155,156] by incorporating the point mutations in SMN1 gene or by creating complete knockdown variants, simulating one or more aspects of human SMA pathology. *Zebrafish* and *Drosophila* SMA models have shown similar effect in MN axons and NMJs. In *Zebrafish,* SMN mutant embryos exhibited failure in the axon pathfinding capabilities and SMN deficient *Drosophila* models developed specific perturbations in the pre-synaptic terminals ("NMJ boutons"). However, these models have also revealed some unique species-specific hallmarks such as *Zebrafish* has solely MN defects, whereas *C.elegans* SMA model has represented mainly muscular system involvements, and in *Drosophila* both MNs and muscular defects were noticed. Such deviations explain the species-specific effects of SMN mutations. Therefore, it is a very tough decision to consider one SMA disease-model for significantly recapitulating the genetics of human SMA pathology.

Furthermore, SMA mouse models were considered more reliable because of their close relatedness with humans at the genomic level. In mouse genome only SMN1 gene exists, while humans have SMN1 and SMN2 genes. Therefore, the complete homozygous deletion of SMN1 gene exon 7 (Smn-/-) would result an absolute lethality in mouse, whereas heterozygous mutations (Smn±) lead to the normal phenotypic development rather than SMA. Transgenic mouse lines were generated, containing disruptions in SMN1 gene and incorporating human SMN2 gene in variable copy number[155,156]. These models have certainly rewarded deep insights about the human SMA pathobiology on a behavioral as well as neurological level and might be utilized for the advanced development of neurodegenerative disorder therapeutics. However, as discussed earlier, in the current animal models (worms, flies or fishes) a common limitation is an absence of SMN2 gene[103] and they also require cumbersome knockdowns and overexpression experiment setups to study the disease mechanism. Although mice models have been widely accepted for studying SMA pathogenesis but they also need complicated transgenic activation of human SMN2 gene as a potential disease modifier. Therefore, it would be highly beneficial to utilize human cell based models to study such a complicated neuromuscular disorders.

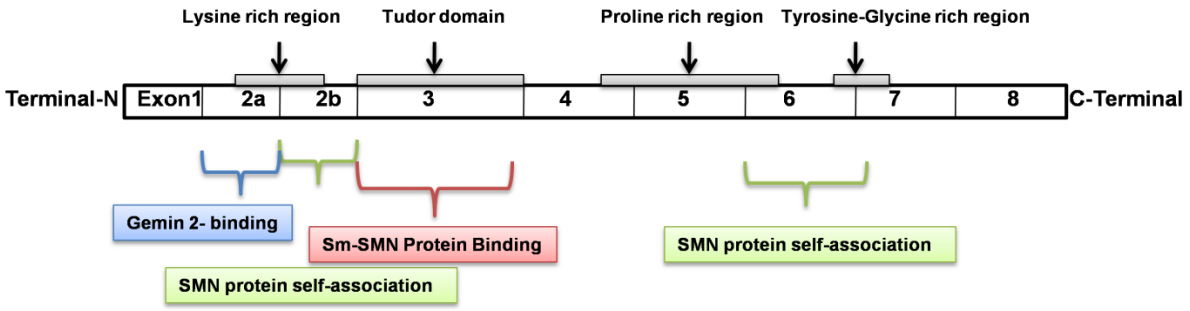


Figure 1.10: SMN protein and its functional domains.

The Lysine–rich (K-rich) region at "Exon-2b" and Tyrosine-Glycine (YG) domain at "Exon-6" act as a self-association domain which mediates the SMN protein to perform self-oligomerization that imparts stability and functionality to it. The Tudor domain is located within "Exon-3" of the SMN protein which promote Sm-Proteins binding on snRNA during snRNP biogenesis through Gemin binding that is supported by K-rich region at "Exon-2a". The Proline-rich (P-rich) region is present within "Exon-5". **The Figure is adapted from the review article published by Burghes et al.**[157]

1.10.5 Induced Pluripotent Stem Cells (iPSCs) Based SMA Models

Recently, a revolutionary method was presented by Takahashi and Yamanaka in 2006[158] in the area of stem cell biology to reprogram the mature somatic cells back into their pluripotent state which exhibits the embryonic stem cells like characteristics (ESCs). Therefore, the method was designated as induced Pluripotent Stem Cell (iPSC) technology. Initially, the experiments were performed on mouse adult fibroblast cells which were successfully converted into undifferentiated pluripotent stem cells using four essential transcription factors (TFs; *oct3/4, sox2, c-myc, klf4;* which were delivered by using retroviruses). They have identified these core set of TFs from the initial list of 24 key factors previously known in ESCs. These reprogrammed cells were called as induced Pluripotent Stem Cells (iPS cells). Furthermore, in 2007 they presented another work of iPS cells generation from human fibroblast cells into pluripotent state using above mentioned TFs[159]. Subsequently, another group presented a similar study for the generation of human iPS cells using lentiviral delivery system[160]. Most importantly they demonstrated the use of partly distinct TFs, including *OCT4, SOX2, NANOG, and LIN28.* The choice of these TFs was encouraged by a stem cell study which has identified *c-Myc* as an apoptotic factor for human ES cells[160]. Ebert and colleagues[161] were first to model SMA pathology by the use of iPSC technology. They isolated skin fibroblast cells from 3-year-old boy suffering from type I-SMA and reprogrammed them into iPS cells and differentiated into MNs, recapitulating SMA-specific characteristics. To investigate the changes in SMN protein levels of

iPS-SMA cells with respect to healthy controls, they similarly cultured iPS-wild type (iPS-WT) cells from his unaffected mother fibroblast cell lines. This study has significantly presented the decrease in SMN protein levels of iPS-SMA cells as compared to the iPS-WT cells. Recently, another study has identified the presence of certain downregulated genes/proteins with respect to the healthy controls by the virtue of iPSC technology[162]. This shows the reliability and success of iPSC-based models in studying neurodegenerative diseases.

Therefore, to investigate further in this direction, we hypothesized that due to the lower level of the SMN protein, there might be mis-regulations in the splicing patterns of specific set of genes and most importantly reduction in the mRNA transport; which is crucial for the survival of MNs. Towards our goal, we carried out the Genome-Wide RNA-Seq data analysis where the RNA samples were isolated from iPSCs derived MNs of two SMA patients and two healthy controls, with two biological replicates per sample. In doing so, skin fibroblast cells were isolated from SMA-patients and healthy controls that were reprogrammed into iPSCs[163] and further differentiated into MNs. This study may provide promising and important leads to understand the intricate pathological mechanisms of SMA in a detailed manner. We have organized this study into three main objectives which are explained below, probing different aspects of transcriptome data analysis.

- Firstly, we performed the global expression level analysis to identify differentially expressed genes (DEGs) and transcripts between two different conditions (SMA-patients and healthy controls).
- Secondly, we performed the exon-level analysis to identify relative changes in the expression at individual exon level which gives an account of differentially spliced transcripts between two different conditions.
- Thirdly, the identification of SREs and their specific RBPs from the list of differentially-used exons and their flanking upstream and downstream introns. Such analysis helps to pinpoint the specific splicing regulatory mechanisms and pathology related alterations in the splicing patterns, linked with MNs sustenance in SMA.

1.11 Transcriptome Analysis

In order to study the transcriptome of any organism, scientific efforts began with single candidate-gene based methods using the northern blotting techniques. Since this method was based on the prior knowledge of known transcripts and also required higher amounts of starting

material (but generate low-throughput), therefore it has low potential to discover novel transcripts. The development of nucleic acid amplification techniques such as reverse-transcription quantitative polymerase chain reaction (RT-qPCR) have raised the analysis potential. Only advanced methods, such as microarray technology and then Next Generation Sequencing technology (NGS) has allowed to visualize the wider picture of transcriptome landscape.

1.11.1 Use of Microarrays in the Analysis of Alternative Splicing Profiles

Microarrays have changed the single-gene expression prospective by providing a methodology to characterize and measure the expression levels of thousands of known genes or transcripts simultaneously on a single glass-chip experiment[164]. A typical microarray chip is composed of array of minute cDNA spots ('probes') which gets selectively hybridized to fluorescently labeled sequences of interest (unknown 'targets') and generates fluorescence intensity that is detected by a fluorescent detector. Using such multiplexing tool, expression quantification of multiple set of genes in a particular cell-type has gained applicability and most interestingly the differential expression analysis between two conditions[165]. It has also enabled the detection of AS events[166–169], non-coding RNA[170], Single Nucleotide Polymorphism[171] and so forth. Various types of microarrays are available on the basis of chip design and hybridization procedure. For example, spotted cDNA microarrays (oligonucleotide microarrays or Affymetrix GeneChips) are designed to quantify the relative differential expression of genes or transcripts between two conditions by measuring the abundance of mRNA transcripts in each sample. Further, advanced types of oligonucleotide microarrays include exon arrays, tiling arrays and exon-exon junction arrays which are capable to determine AS events and characterize the differences between spliced and unspliced mRNAs at genome-wide scale[172]. In contrast, tiling arrays have an advent to identify novel transcript variants (as they do not rely on priori transcripts information)[170]. Later, Yeakley et al. have presented bead-based fiber-optics microarrays, which operates without the need of prior laborious steps such as RNA purification or cDNA formation and it needs a small starting material (sub-nanograms of total-RNA)[166]. These microarrays can identify AS isoforms for a single gene and also can be used to differentiate closely related transcripts or genes.

➢ Advantages and Disadvantages

The economic value and the large-scale study attributes of microarray technology made it an attractive choice. The invention of tilling arrays offered the capability to perform studies at much wider level but they are unable to detect rare or novel transcriptional events, as prior sequence information is required to design tilling microarray chips. Moreover, the cost of tilling arrays is relatively higher than traditional microarrays, therefore it is not feasible to study large genomes with tiling arrays. Furthermore, in the presence of highly-related sequences, the risk of cross-hybridization is also a concern. In microarrays, digital or numeric expression quantification is not possible, which accounts for the relative "number of transcript copies" as "read counts" expressed in the sample under investigation, rather microarrays abundance estimation relies upon measuring the probe hybridization intensity signals.

In the following sections, we intend to discuss the sequencing technologies (at single nucleotide resolution) from first through third generation sequencing methods.

1.11.2 First Generation Sequencing

i. Sanger Sequencing

First generation DNA sequencing method was developed by Frederick Sanger and colleagues in 1977 by the use of modified 2',3'-dideoxynucleteotide triphosphates (ddNTPs)[173]. These are altered forms of normal dNTPs where 3'-hydroxyl group is removed from deoxyribose sugar which blocks the DNA chain elongation in-vitro. Therefore, it is also called *chain termination sequencing* method. The working principle is "base-per-base reading by non-reversible termination" of DNA polymerization reaction which means whenever any modified base is incorporated, the reaction is terminated and synthesis of new sequencing reaction begins. Likewise, multiple stretches of sequenced molecules (with different lengths) are synthesized. Subsequently, denaturation of the resultant molecules is performed with their mass-based sorting using gel electrophoresis (mass of the sequenced molecules represents their point of termination). Finally, autoradiography and gel imaging techniques are used to visualize the DNA bands and examining DNA sequence, respectively. Furthermore, various improvements within the incorporation of ddNTPs (using radioactive substances or fluorescent dyes) impart partial automation to Sanger sequencing. Later, gel electrophoresis method was replaced with capillary-based electrophoresis and commercialized by Applied Biosystems (ABI 370)[174,175]. This

system can generate 6 to 8 Mb of DNA sequences per day with the read length ranges between 600 to 900 bases. The cost of Sanger sequencing per 1Mb is around $500.

- **Advantages and Disadvantages**

Sanger sequencers give an advantage of longer read lengths but its major limitation is low-throughput with high cost. Additionally**,** the quality of first few bases is inferior due to the presence of primer sequence. The average error rate is high which is either due to the contamination with bacterial vectors used for fragment amplification or due to general sample contaminations or due to the presence of low-complexity regions (repeated regions).

1.11.3 Use of Expressed Sequence Tags to Study Alternative Splicing

Expressed sequenced tags (ESTs) are short sequences (200-800 nucleotides) of mRNA transcripts, representing transcriptionally active regions at given time-points in a cell[176]. Initially, Adams and coworkers described the use of ESTs for the characterization of the human genes[177]. Subsequently, Mironov et al. contributed to analyze the prevalence of AS in human genes by mapping ESTs onto known human gene sequences. They identified about 35% alternatively spliced genes with higher modulation frequency in 5' UTRs[178]. Further, Brett and colleagues have identified 38% human genes undergo AS with higher occurrences of exon skipping AS event[179]. Modrek et al. performed the mapping of "~2.1 million human mRNA and EST sequences" at genome-wide level[180] and they discovered 42% of the genes are alternatively spliced. Overall EST based studies determined approximately 30,000 AS events in the human genes[177–183].

- **Limitations**

The data obtained from EST sequencing are error-prone in nature (~1/100). The presence of vector genome contamination, partial transcripts processing and high redundancy with low quality regions near 5' and 3' ends of the fragments, can lead to AS prediction bias. A comprehensive review on the use and limitations of ESTs data has been provided by Nagaraj[184].

1.11.4 Second Generation Sequencing Methods

To address aforementioned limitations of the automated Sanger sequencing method, novel and more efficient methods were major milestone. Few years later, this revolutionary achievement was attained with the development of novel technology, having extreme power and efficiency of generating high-throughput data with a bonus of cost reductions and high accuracy. Moreover,

this new arena of sequencing has reflected the previous designation of high-throughput sequencing to next-generation sequencing technology (NGS) which implement the massive parallel sequencing procedure and produce high-throughput data with time and cost reductions.

All of the NGS approaches intend to discard the use of bacterial vectors for the cloning of target DNA which help in bias reductions. NGS platforms mainly involves the template fragmentation, adaptor ligation, substrate binding (NGS platform specific) and PCR amplification (for increasing the overall signal) followed by sequencing itself. Typically, NGS platforms differs by sequencing procedures involving type of enzyme used, underlying principles to generate reaction signals with their efficient recording/imaging (base-call). The generated read lengths are also system specific but all have shorter read lengths with respect to Sanger sequencing methods (except third generation sequencing methods).

We will begin with the briefings of available NGS methods to-date with their strengths and weaknesses. Most importantly, all NGS technologies are equally capable for sequencing any biopolymer (DNA or RNA). Herein, methods are explained with DNA as a starting material. However, for sequencing RNA an additional step of reverse transcription is required to obtain cDNA. Our main goal is to sequence RNA molecules, therefore we have also discussed RNA-sequencing procedure with a selected NGS platform (Illumina Genome Analyzer) in section-1.12 of this chapter. Second generation sequencing methods, including Sequencing by Synthesis (SBS) and Sequencing by Ligation (SBL) mainly follows reversible chain termination sequencing technique (also called Cyclic Reversible Termination or CRT sequencing). Most of the sequencing platforms in this category follows the general sequencing library preparation steps as described below with platform specific sequencing procedure:

A. Library preparation
   (a) Sample fragmentation into small fragments.
   (b) Adapter ligation on both sides of the fragments. The ligation can be performed either before or after the denaturation step of the fragments.
   (c) The adapter-ligated fragments are immobilized onto solid surface (magnetic beads or glass-slide) which contain primer-oligos (covalently attached) complementary to the adapter sequences of fragment.
   (d) The fragment enrichment is performed by using emulsion-PCR or clonal amplification.

B. Sequencing, which is NGS platform specific. During the sequencing reaction (at single base resolution), most of the NGS platforms follow a fixed cycle with the following steps:
   (a) DNA polymerase incorporate the fluorescent-labeled single base to the fragment (template).
   (b) Then, 3'-OH end of the incorporated base is blocked with a reversible terminator to inhibit further base incorporation. The reversible terminators consist of a 'cleavable fluorophore' attached with nucleobase of the incorporated nucleotide and 'small reversible moiety' which perform capping of 3'-OH group of incorporated base.
   (c) Laser excitation of attached fluorophore moiety and detection.
   (d) Reactivation of polymerase reaction by 3'-OH group unblocking.
   (e) The cycle is repeated millions of billions times in parallel.

This cycle diverges by detection chemistry, which rely on the type of used 3'-OH blocker. Mostly, all SBS approaches use 3'-blocked terminators[185]. In contrast, other more efficient terminators have also been developed, namely 3'-unblocked or virtual terminators[186]. Virtual terminators keep the 3'-OH group unmodified which allows the incoming nucleotide to interact naturally with the active-site of DNA polymerase enzyme.

(i) 454/Roche Sequencing Technology

In 2005, 454/Roche Company was the first to commercialize the NGS based sequencing platform, based on *pyrosequencing* principle. The system relies upon the release of *pyrophosphate* (PPi) during the DNA polymerization reaction. Wherein with every base incorporation, a PPi molecule is released which emits light, that is detected by the light sensitive cameras and gets recorded. The system is parallelized on a picotiter plate which contains around 2 million wells. where each well has a capacity to hold single-stranded sample fragment attached with streptavidin coated beads. The library preparation involves the aforementioned steps where DNA fragments are of 300-800bp size[187]. To initialize the sequencing procedure, the bead covered with amplified fragments are added onto the picotiter plate. The wells are prepared with an enzymatic mixture of DNA polymerase, ATP *sulfurylase and luciferase.* The loaded plate is placed inside the sequencer and during each sequencing reaction four bases are provided. With every base incorporation, a PPi molecule is released and ATP *sulfurylase* convert it into ATP molecule, which acts as a substrate to *luciferase* that catalyzes the conversion of *luciferin* into *oxyluciferin* (a light emitting reaction). The emitted light gets captured by high-resolution charge-coupled camera (CCD) and recorded as a nucleotide peak (pyrogram:

a graph representing light intensities). Further, *apyrase* enzyme degrades all the unpaired nucleotides to prevent any background noise and next reaction is performed. Notably, dATPs can be confused with *luciferase* enzyme which can interfere with the synthesis mechanism, therefore dATP modification into deoxyadenosine-5'-(α-thio)-triphosphate (dATPαS) is performed. In 2008, Roche/454 FLX Titanium system has offered read length up to 600 bp with total output of 700 Mb per run (accuracy of 99.9% per base-call). This technology has been implemented successfully in several genomics and transcriptomics sequencing projects[188], including the discovery of rare transcripts, novel genes, *de novo* transcriptome assembly of non-model organisms[189–194].

- **Advantages and Disadvantages**

High speed and longer read length are the major advantages of this platform. The main limitation of this platform is signal intensity drop over the sequencing run due to enzymatic activity reduction which results in poor-quality base-calls. Moreover, its sequencing cost is higher with low- throughput.

(ii) Illumina/Solexa Genome Analyzer

Illumina genome analyzer was released in 2006, which has presented *"bridge amplification"* or "*clonal amplification"* method (supplanted the emulsion PCR amplification method) for fragment amplification. During the amplification step, the denatured and adapter ligated ssDNA fragments are immobilized on the flow-cell (by hybridizing with primer sequences that are complementary to the adapters). Bridge amplification is performed to generate single fragment clusters wherein bridge-like structure is formed when 3'-end of the template fragment bends toward its complementary primer sequence gets hybridize with it. Subsequently, all clusters and all fragments per cluster are sequenced in parallel by SBS approach. These steps are repeated until the complete sequence is read in a massively parallel environment per flow-cell.

Bridge amplification method has the capacity of generating 800-1000 K clusters/mm^2 with sufficient flow cell loading. Illumina technology has been adapted by researchers for different purposes such as *de novo* genome[195,196] and transcriptome assembly[197], re-sequencing of complete genomes to identify *de novo* mutations in pathologies[198–200] and study genome-wide genes expression in normal and diseased conditions[201]. The platform offers ultra-high throughput within budget, constant improvements in read length[202].

- **Advantages and Disadvantages**

Ultra-high throughput with deeply sequenced reads at low cost is the great achievement of this platform. The observed problem in this technology is "leading and lagging strand dephasing". In sequencing CRT cycles, dephasing (in Illumina) refers to a condition where multiple copies of the template DNA fragment within a cluster move out of synchronization. If a DNA fragment exists in 10,000 copies within a cluster, then all of these fragment copies incorporate fluorescent-labeled nucleotide which gets recorded and imaged. Further, this nucleotide is unblocked for the next nucleotide incorporation reaction. If the unblocking does not take place on certain fragment, then that fragment would not be able to incorporate next nucleotide. However, it might get unblocked in the next cycle and in this case it will be lagging behind from rest of the leading fragment copies within the cluster. As the CRT cycles progresses, more and more fragment copies within the cluster move out of synchronization, resulting in deletions within the sequenced fragments.

(Iii) Life Technologies SOLiD

In 2007, Life Technologies launched the third NGS platform which is based on Sequencing by Oligo Ligation Detection (SOLiD) method. It uses ligase enzyme instead of DNA polymerase[203,204] which tend to read two nucleotides together per reaction and add more accuracy.

The sequencing library is prepared with aforementioned steps and ligation based sequencing is performed wherein a universal primer sequence is annealed (complementary) to the adapter sequence at one end of the fragment. Further, a set of four fluorescently labeled di-base probes (usually placed within 8-mers) and DNA ligase enzyme is introduced automatically by the system where probes compete for the ligation, and target fragment's complementary probe gets base-paired and ligated by ligase which result in fluorescence emission that is captured by the detectors and gets recorded. Subsequently, the non-ligated probes are discarded and fluorescence moieties are clipped-off in order to reactivate the 5'-phosphate group. In every ligation reaction, the di-base specificity is re-verified for base-1 (in next reaction) and base-2 (in previous reaction) to minimize the error-rate. After one-time set of cycles, the universal primer sequence is re-set by n-1 (where 'n' is the length of primer). Approximately 5-10 cycles (with each time primer reset) are performed (with a series of ligation sub-cycles) for each template fragment, producing color-space sequencing data. Several studies have used this platform for genome-wide sequencing studies[205–208].

- **Advantages and Disadvantages**

This platform has offered very high accuracy (99.99%) with low cost but the read length is still short (up to 75bp) as compared to other competitive NGS platforms.

(iv) Ion Torrent Sequencing Technology

Life technologies have presented an innovative sequencing method based on the "semiconductor chip technology". It is based on the ion detection (proton ions or H^+)[209,210]. This technology provides high speed, reduced costs and simple instrument design with respect to other NGS platforms. The technique uses the concept of proton ion (H^+) release during the phosphodiester bond formation in DNA polymerization reaction and detects the change in pH of solution with every new base addition in sequencing reaction. If H^+ ion concentration is high, it results in pH drop and vice versa. The sequencing chip is designed with two separate layers: first layer functions as a sample loader (with billions of tiny wells), and the second layer senses the pH change ("ion sensitive layer"). Currently, the ion torrent systems have provided a very high sequencing capacity ~100 Gb per run with very high accuracy (99.99%), which can be used to sequence whole genome very efficiently.

- **Advantages and Disadvantages**

Ion Torrent has high error-rate for homopolymers detection, as it only relies on sensing and measuring the pH changes. For instance, when no base is incorporated no voltage is detected whereas, if two identical bases are incorporated (one after the other) the detection is two times higher, but it is very difficult to capture changes in signal intensity accurately, corresponding large number of incorporated identical bases.

1.11.5 Third Generation Sequencing

(i) Single Molecule Real Time Sequencing Technology (SMRT)

In 2011, a new concept of NGS sequencing was commercialized by Pacific Biosciences (PacBio) to perform real time single molecule sequencing[211]. The SMRT works on the mechanism of detecting real time nucleotide incorporation events without DNA synthesis termination which reduces time and increases read length. During the library preparation no PCR amplification is required. They harness the power of DNA polymerase enzyme by using two proprietary technologies. Firstly, the use of four colored phospholinked nucleotides to visualize the activity of DNA polymerases which carry the fluorescent label on the terminal phosphate rather than on the base. As a result, enzyme cleaves away the fluorescent label as

part of the base incorporation step and leave behind a completely natural DNA strand. Secondly, the use of nanophotronic visualization chamber called Zero Mode Waveguide (ZMW; 70nm wide cylindrical metallic chamber with $20x10^{-21}$ zeptoliters detection volume)[212]. The nucleotides diffuse in and out of ZMW in microseconds. When the polymerase encounters the correct nucleotide it takes several milliseconds to incorporate it, and during this time its fluorescent label gets excited and emits light, which is captured by the sensitive detector. The whole process repeats several times, building the desired sequence length.

- **Advantages and Disadvantages**

SMRT provides the real-time single molecule sequencing utility with longer read length (2500bp-15000bp). This methodology gives an opportunity to understand the natural behavior of the DNA-polymerase during the base incorporation reaction. The PCR free sequencing adds high fidelity and confidence for reading single unique molecules. The error rate is high in terms of indels (15%) and base substitutions (1%).

1.12 Implications of NGS Technology for Sequencing RNA

1.12.1 RNA-Sequencing Technology (RNA-Seq)

The sequencing of RNA molecules requires an additional step of reverse transcription during the RNA-Seq library preparation[213]. Steps for the library preparation and sequencing of RNA samples are described as follows:

A. Isolation of total RNA sample and its purification to remove any DNA contaminants. Further, depletion of ribosomal-RNA is performed followed by poly-A enrichment step (to obtain only mature mRNA). The ribosomal-depleted RNA without the poly-A enrichment can be used to study wide range of other existing RNA-species and also unspliced mRNA (or nascent mRNA), which is useful to study mRNA-splicing.
B. The purified samples are fragmented into small fragments followed by the size selection of the fragments (using gel electrophoresis).
C. Reverse transcription of the RNA fragments into cDNA is performed.
D. Further library preparation steps and sequencing step varies for specific NGS platform:
   (a) Amplification of cDNA molecules into multiple identical sets or ensembles.
   (b) Execution of sequencing and detection at the single nucleotide resolution.
   (c) Bioinformatics analysis of generated RNA-Seq data is one of the major time consuming tasks which requires rational pipelines and further validations of the putative results.

Note, if the chosen technique involves the single molecule sequencing application, then the fragment amplification step is not involved.

Recently, the Association of Biomolecular Resource Facilities next-generation sequencing (ABRF) has presented a comprehensive comparison in the performance of all key NGS platforms. They analyzed RNA-Seq data by using 454/Roche, Life Sciences Ion torrent, Illumina HiSeq and PacBio SMRT, and identified a high correlation between these platforms for the transcript expression analysis[214].

1.13 Computational Challenges in NGS Data Analysis

The analysis of NGS data starts from the alignment of reads onto the reference genome, which is the most challenging step due to the short length of the reads (except PacBio). Prior to the alignment step, filtration of raw reads is required to reduce bias which might originate either during library preparation or during sequencing. The aligned reads can be analyzed to estimate relative gene or transcript abundance levels in the sample[215]. Various read alignment free methods are available which estimate relative gene or transcript abundances by completely avoiding the time consuming read alignment step using reference transcriptome indexes. These methods provide fast speed over alignment-based methods without compromising with the accuracy. Further, to validate the putative findings suitable wet-lab experiments should be designed. To carry out NGS data analysis of large genomes (for instance *Homo sapiens*), the computational systems must have sufficient memory and processors (CPUs; to parallelize the tasks).

In the following sections, we intend to introduce widely used bioinformatics tools to pre-process and analyze the NGS (RNA-Seq) data.

1.13.1 Raw RNA-Seq Read Files and their Pre-processing

Raw sequencing reads are stored as *Fastq* files or NCBI Sequence Read Archive (SRA) file formats. The *Fastq* files are text based files where each read is represented with four sections, containing the information of nucleotide sequence and their quality scores (*Phred* scores)[216] encoded with *ASCII* characters. Whereas, if the raw reads are archived in SRA format[217] then, the SRA file is first extracted into *Fastq* format using "SRA-toolkit (fastq-dump)" (http://www.ncbi.nlm.nih.gov/books/NBK158900/). The *Phred* score (Q) provides per-base quality values in terms of the probability of base-call being wrong during sequencing.

$$Q = -10log_{10}P$$

In Illumina, sequencing can be performed in two different modes with distinct library preparation procedures: Single-End (SE) and Paired-End (PE). In SE sequencing, the cDNA fragments are read from one end only whereas in PE sequencing, fragments are read from both the ends with predetermined distance within two reads. SE mode generates one *Fastq* file and PE mode generates two *Fastq* files (for "read1" and "read2").

Beginning from the sample isolation, library preparation, until sequencing, the whole process is prone to some errors, including low quality base-calls towards 3'-end of the read, presence of adapter and primer sequences, GC-content bias and PCR duplicates. Therefore, it is very important to quality check the data before downstream analysis. Various programs help in this task such as FastQC[218], RNA-SeQC[219], SAMstat[220] and on the basis of quality check report further processing can be performed whenever required. For instance, removal of low-quality reads and adapter or primer sequences using software such as Cutadapt[221], Trimmomatic[222], FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), htSeqTools[223], TrimGalore[224].

1.13.2 Read Alignment

The quality checked reads are aligned onto the reference genome or transcriptome. During short-read alignment variety of parameters should be considered which includes number of sequenced reads, read length, reference genome size, accurate splice-sites detection and SE and PE reads. The aligners vary on the basis of alignment algorithm which includes: Hash-tables and Burrows-Wheeler Transform (BWT) with backward search. Hash-table based aligners use "seed-and-extend" strategy. Such as MAQ[225], SeqMap[226], RMAP[227], RazerS[228]. In contrast, the BWT based aligners align the entire reads to the reference genome unlike "seed-and-extend" approach by first indexing and storing the reference genomes as compressed suffixes (including prefix-suffix tree, suffix array and Ferragina-Manzini or FM index)[229]. The widely used BWT based software include Bowtie[230], Bowtie2[231], BWA[232], SOAP2[233], GSNAP[234] and MapSplice[235]. Dobin et al.[236] has presented an ultra-fast read aligner, namely Spliced Transcripts Alignment to a Reference (STAR), which combines the seed generation and their search within uncompressed suffix arrays of reference genome (providing high alignment speed).

1.13.3 Read Alignment File Format and Visualization

Mostly aligners store the alignments in the Sequence Alignment/Map format (SAM) or Binary Alignment/Map format (BAM; compressed format of SAM). SAM is a tab-delimited text file which provide the detailed read alignment information in two sections: the header section and the alignment section. The header section lines start with '@' symbol. In the alignment section, each alignment line is represented by 11 main fields. These fields describe the read alignment (**Table 1.1**). The BAM format compresses the SAM format by BGZF (Blocked GNU Zip Format) compression and indexes them to reduce the space on the hard disk which also provides random quick alignment access. The genome-wide per-base coverage of aligned reads can be visualized using Integrative Genomics Viewer (IGV) (http://www.broadinstitute.org/igv/), GenomeView[237], LookSeq[238], BamView[239], and MagicViewer (http://bioinformatics.zj.cn/magicviewer/) tools.

Table 1.1: The alignment section within SAM file format.

Every aligned read is represented with 11 mandatory fields reported below.

Column	Field	TYPE	DESCRIPTION
1	QNAME	STRING	Query Name for the Template
2	FLAG	INTEGER	bitwise FLAG
3	RNAME	STRING	Reference Sequence Name
4	POS	INTEGER	Leftmost mapping Position (1-based)
5	MAPQ	INTEGER	Mapping Quality
6	CIGAR	STRING	CIGAR string represents sequence mapping
7	RNEXT	STRING	Reference Name of the Next read
8	PNEXT	INTEGER	Position of the Next read
9	TLEN	INTEGER	Length of the Template (observed)
10	SEQ	STRING	Seqment Sequence
11	QUAL	STRING	per-base phred quality scores in ASCII characters

1.14 Expression Quantification and Differential Expression Analysis using RNA-Seq Data

The expression estimation from the aligned reads is one of the basic routines using RNA-Seq data. This application can be extended to identify the relative differences in the expression levels between different biological conditions. To perform these tasks most widely used tools include Cufflinks[240], DESeq[241] and edgeR[242]. The gene or transcript expression estimations are given as counts which corresponds to the number of reads overlapping at a given genomic locus. The read counts are required to be normalized to scale the differences in gene or transcript length and library size per sample (total number of reads in a sample or sequencing depth). Several normalization criterions are present such as Read Per Kilobase per Million

mapped reads (RPKM), Fragments per Kilobase per Million mapped fragments (FPKM) and Transcripts Per Kilobase Million (TPM).

To enhance the speed of RNA-Seq data analysis which mainly involves the transcript abundance estimation analysis, several alignment-free methods have been designed recently, such as RNA-Skim[243], sailfish[244], salmon[245], kallisto[246].

On the other hand, if the study involves the identification of the localized events such as relative changes in the expression at an exon-level (differential-usage of the exons) between two conditions, "exon-centric" methods are required such as DEXSeq[247], DSGseq[248], Diffsplice[249], MATS[250], SpliceR[251] and so forth.

1.15 Selection of Bioinformatics Tools for RNA-Seq Data Analysis

1.15.1 Study of Alternative Splicing in SMA

The goal of this study is to identify mis-regulations in the AS regulatory patterns (controlled by specific SREs and RBPs) and disruptions in the mRNA transport of MNs, due to the lower levels of SMN protein in SMA patients with respect to the healthy controls by analyzing RNA-Seq data using a rationally designed pipeline (**See Chapter 2; Materials and Methods**). To perform this study, we have selected two widely used bioinformatics programs: Cufflinks and DEXSeq.

(i) Cufflinks Tools and Limitations

Cufflinks tools have the capability to quantify the expression level of genes and transcripts in RNA-Seq data samples. The quantified samples can be further used to identify the differentially expressed genes (DEGs), transcripts and differential AS events between different conditions using Cufflinks-Cuffdiff2 tool[252]. Despite the broad spectrum of the Cufflinks tools, we observed its two limitations. Firstly, Cuffdiff2 tool applies a highly conservative approach which results in the skipping of more complex phenomenon within the transcripts of the same gene. Such as in order to identify the differential AS events, it initially groups together the transcripts with same Transcription Start Site (TSS) and performs their statistical testing only, while the transcripts with different TSS in a gene remain untested. Suppose, we have a gene with 2 transcripts (having different TSS**; Figure 1.11**) and if we identify any relative changes in the expression levels of this gene then it could be either due to transcriptional regulation or due to alternative splicing. Such genes will be discarded from Cuffdiff2 analysis. Though it is quite challenging to disentangle such events, but we need to uphold alike cases and improve computational

methods which have the capability to disentangle them accurately. Therefore, to understand the regulatory mechanisms of such genes, we need to analyze them by applying more generic and less restrictive computational methods (discussed in next section). Secondly, the dependency of cufflinks on the gene annotations, which are rather far from complete and not even fully consistent within different source databases. Moreover, this is a major limitation for all the current tools as they heavily rely upon the gene annotations.

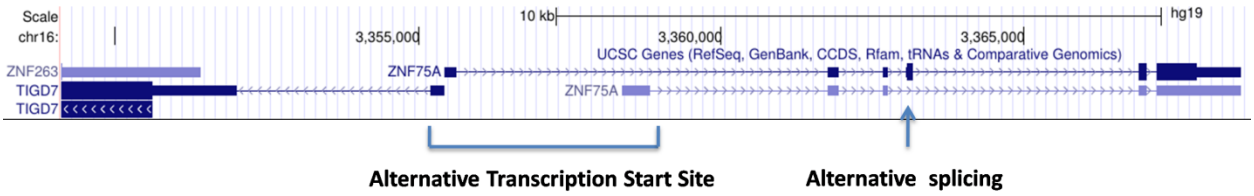


Figure 1.11: An example demonstrating the stringency of Cufflinks-Cuffdiff2 pipeline.

The gene ZNF75A expresses two different transcripts with the different TSS and also undergoes the AS event of exon skipping in transcript 2 (shown with '*Blue'* arrow).

(ii) DEXSeq Tool

We decided to use a more generic tool called DEXSeq which is designed specifically to identify relative changes in the expression at individual exon-level in a whole gene (Differentially-Used Exons) between different conditions. DEXSeq tests for all the exons present in a gene in a similar way without any prior assumption about same TSS or different TSS. Therefore, it produces a list of exons which might be resulted either due to changes at transcriptional level or at AS level. In addition, DEXSeq tests the relative changes in the read counts at single exon-level of the whole gene, but it does not estimate significant changes in isoform proportions.

1.16 Development of Computational Model to Estimate Transcript Expression

1.16.1 Motivation

The currently available computational methods are insufficient to understand differential splicing events at high resolution within the transcriptomic landscape. In order to address this problem, we designed a computational model which has a potential to precisely estimate the "transcript expression levels" within a given gene locus by disentangling mature and nascent transcription contributions for each transcript at per base resolution. Recently, Madsen et al. has presented an idea of estimating intron read coverage across the transcripts to measure the acute transcriptional activity of a cell in a steady-state[253]. They used a very simple approach by summarizing the intronic reads for each transcript within unique regions which are not-

overlapping with any other exons of the gene. In this study, we have presented a non-linear computational model which estimates the expression levels (intronic and exonic read coverages) by precisely measuring the contributions of mature and nascent transcription at per-base resolution for each transcript in a given gene locus. We examined the performance of our model at a genome-wide level by analyzing total RNA-Seq samples. With this, we take one step forward to study the transcriptome complexity in terms of differential splicing with more details. The application of our model in detecting differential splicing events. At exon level, differences in the ratio of the sum of mature and the sum of nascent transcripts over all the transcripts in a gene locus gives an indication of differential splicing.

1.16.2 Background

The computational methods are improving to get deep insights about the general and tissue-specific up-regulation or down-regulation of certain transcripts in a specific condition of the cell at steady-state. For instance, certain cell-types preferentially express some specific isoforms for a given gene locus over the other isoforms with respect to the standard gene annotation models. Therefore, from the list of possible isoforms for a given gene some will be plausible while others will not be plausible in a specific cell-type. Further, the presence of the specific isoforms can also be differentiated on the basis of their expression level. Sometimes similar set of plausible isoforms show higher expression within a specific biological condition whereas the lower expression is shown in another biological condition. Such up and down regulation of the isoforms expression can usually be noticed within pathological and normal conditions of the organism. Another crucial aspect within the transcriptional landscape involves the AS and its regulatory mechanisms as described previously. These mechanisms focus upon either the selective inclusion or exclusion of selective exon(s) within transcripts. Thus, enhancing or silencing their expression levels and as a result optimize the overall expression levels of the transcripts in a given gene locus. The existing five canonical AS types within eukaryotic cells should be accurately modeled in order to assign the correct expression contributions (weights) for every possible transcript in a gene. In doing so, we must first determine the number of plausible transcripts in a given gene locus to model precisely their expression estimations, but it is not a trivial task to accurately define the existence of some transcripts over the others. Therefore, we first introduce possible method to determine plausible isoforms within a given gene locus and modeled their expression levels by disentangling the mature and nascent transcription contributions (**See Chapter 2; Materials and Methods**).

1.16.3 Mature and Nascent Transcription

The transcription takes place in the nucleus of the cell and once it is completed most of the mature transcripts (mRNA) are exported inside the cytoplasm by the addition of 5'-cap and 3'-polyA-tail which prevents any damage and degradation. At the steady-state, when the total-RNA extraction is performed from the cells it contains both cytoplasmic-RNA as well as nuclear-RNA. Within the cytoplasm, high proportion covers the mature transcripts while in nucleus the transcripts could be at any stage of transcription that is whether RNA polymerase has just completed the transcription or still ongoing. The former will result in FL transcripts while the latter are still transcribing and are designated as nascent transcripts. Therefore, the total-RNA samples comprise a mixture of mature and nascent transcripts in variable proportions, depending upon the length of the transcript which can be modeled as intronic and exonic read coverages at per-base resolution. Furthermore, nascent transcription in combination with the co-transcriptional splicing mechanism provides the inference of AS in the transcripts[254]. The co-transcriptional splicing mechanism states that the mRNA transcription and the splicing of the transcribed parts of transcript operates side-by-side, giving rise to nascent and partially spliced mRNA expressions. During the start of each intron the nascent expression tends to accumulate and as the transcription of one intron between two authentic exons completes immediately splicing machinery also takes part and nascent expression shows gradual declination and ultimately lost and exonic coverages accumulates as mature mRNA expression. The nascent transcription gives the account for on-going transcription in a cell at a given time which shows the "saw-tooth" trend in the intronic-read coverage due to actively growing transcript lengths towards the 3'-end of the transcript[254]. This trend repeats for every intron due to the presence of co-transcriptional splicing mechanism. Therefore, there is an interplay between transcription and splicing regulatory factors in order to accomplish efficient mRNA transcription and its processing. Ameur et. al[254] have presented the idea of studying nascent transcription with co-transcriptional splicing phenomenon using the advent of total RNA-Seq data. We exploited intronic read coverage to precisely estimate the expression levels and therefore, assigning the accurate expression weights to each possible isoform of a gene.

2.1 Study of Alternative Splicing in SMA

The wet-lab procedure from reprogramming of the fibroblast cells into SMA-iPSCs and control-iPSCs, their differentiation into spinal motor neurons (MNs) until the RNA sample collection and RNA-sequencing process was performed by our colleagues (Corti S et al.[1]). We performed the complete computational analysis of the RNA-Sequencing data by devising the integrative computational pipelines to get deep insights about SMA pathogenesis at genome-wide level.

2.1.1 Reprogramming of Skin Fibroblast Cells into iPSCs

We reprogrammed the skin fibroblast cells into iPSCs from two SMA patients and two healthy controls, with two biological samples per sample using oriP/EBNA1-based episomal vectors by the nucleofection of episomal plasmid combinations (NHDF kit VPD-1001 with U-20 program, Amaxa) (**Figure 2.1A**). The used vectors and plasmid combinations (pEP4EO2SEN2K, pEP4EO2SET2K and pCEP4-M2L) were described previously[255]. The complementary DNAs (cDNAs) for the open reading frames of the human genes *OCT4, SOX2, NANOG, LIN28, c-Myc*, and *KLF4* were derived through direct PCR of human stem cell cDNA. After transfection, the fibroblast cells (1×10^6 cells per nucleofection) were plated onto Matrigel (BD Biosciences) covered 3 × 10–cm dishes, containing fibroblast culture medium, which was changed every day. After 4 days of transfection, the fibroblast culture medium was replaced with human Embryonic Stem cell (hESC) culture medium (mTeSR, Stemcell Technologies Inc.) for 8 to10 days. After 18 days of transfection, it was possible to identify the first colonies with an iPSC-like morphology. Within 18 to 20 days after transfection, the 3×10–cm dishes of reprogramming culture were stained with alkaline phosphatase (Millipore) to identify the eventual presence of human iPSC colonies. Between 25 to 30 days, the other two 10-cm dishes were passed to fresh10-cm Matrigel-covered dishes (1 ml each plate) at a ratio of 1:3. Further, to analyze and expand the reprogrammed cells, the iPSC colonies were picked that were morphologically more similar to hESCs. The efficiency of skin fibroblast cells reprogramming was about 3-6 colonies per 10^6 fibroblast cells.

[1] Dino Ferrari Centre, Neuroscience Section, Department of Pathophysiology and Transplantation, University of Milan, Neurology Unit, IRCCS Foundation Ca' Granda Ospedale Maggiore Policlinico, Milan 20135, Italy.

2.1.2 Differentiation of SMA-iPSCs and Control-iPSCs into Spinal Motor Neuron

We generated spinal motor neurons (MNs) from iPSCs following a multistage differentiation protocol developed for hESCs[256] (**Figure 2.1B**). To produce the MNs from SMA patients-iPSCs and controls-iPSCs, the cells were cultured with neuronal medium that is comprised of Dulbecco's modified Eagle's medium/F12 (Gibco, Invitrogen) supplemented with MEM (minimum essential medium) nonessential amino acids solution, N2 and heparin (2 mg/ml; Sigma-Aldrich). After 10 days, retinoic acid (RA; a caudalizing factor) was added (0.1 mM; Sigma-Aldrich) to promote the neural caudalization. At day 17, the posteriorized neuroectodermal cells were collected. After the isolation, the neuroectodermal clusters were then suspended for a week in the same neuronal medium supplemented with RA (0.1 mM) and sonic hedgehog (SHH) (100 to 200 ng/ml; R&D Systems Inc.). At day 24, other growth factors like Brain-Derived Neuronal Factor (BDNF), Glial-Derived Neuronal Factor (GDNF), and insulin-like growth factor-1 (IGF-1; 10 ng/ml; PeproTech) were added. After 4 to 5 weeks under differentiation conditions, the cells that expressed MN–specific transcription factors (TFs) such as HB9, ISLET1, and OLIG2 (spinal cord progenitor marker) and pan-neuronal markers such as TuJ1, Neurofilament, and MAP2 were generated. Most of these HB9/ISLET1-positive neurons expressed choline acetyltransferase (ChAT) and were SMI32 positive (MN-specific marker), demonstrating a MN phenotype. The invitro differentiation protocol yielded a mixed cell population which also contained non-MN cells. Given the limited availability of surface markers to isolate MNs and purify them further, a physical strategy based on gradient centrifugation was applied.

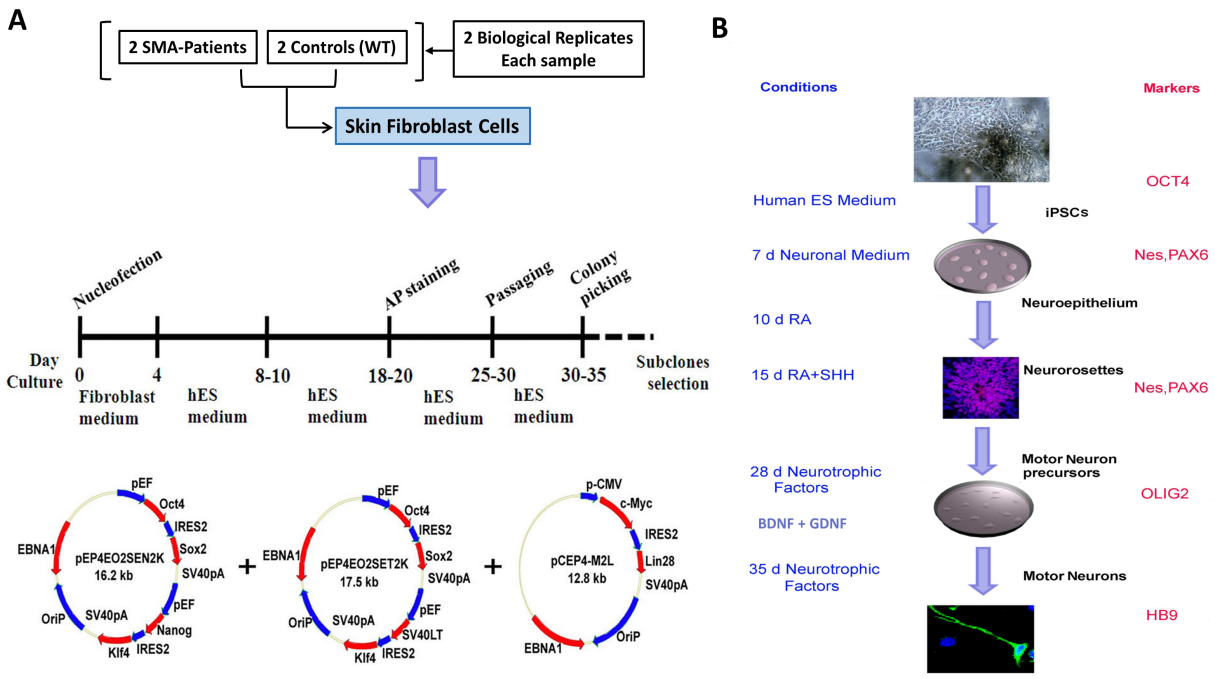


Figure 2.1: An experimental setup for the reprogramming of human skin fibroblast cells into iPS Cells using combination of reprogramming factors and their differentiation into MNs.

A The skin fibroblast cells isolated from SMA-patients and healthy controls (wild type or WT) are reprogrammed into iPS cells using the method of nucleofection of fibroblast cells with non-viral, non-integrating episomal vectors (oriP/EBNA1) derived from Epstein-Barr virus, encoding combinations of reprogramming factors. Transgenes and other vectors features are shown with red and green arrows, respectively. **B** The iPS cells are differentiated into MNs by multistage differentiation protocol. Wherein the iPS cells are plated with neuronal medium, containing Dulbecco's modified Eagle's medium/F12, supplemented with MEM nonessential amino acids solution, N2, and heparin. Further, RA and SHH are added to promote the neural caudalization and ventralization. Next, neurotrophic growth factors such as such as BDNF, GDNF are added and within 35 days under differentiation conditions, MNs are generated expressing MN-specific factors (HB9 in '*green*'). **The procedure for the iPSCs generation and MNs differentiation were previously mentioned by Corti et al.**[163] **and this figure has been adapted from the same article.**

2.1.3 RNA Sample Isolation and Library Preparation

Total RNA samples were extracted from SMA-patients iPSCs and controls iPSCs derived MNs using the RNeasy mini kit Qiagen. The integrity of isolated RNA samples was tested using Agilent 2000 analyzer (Agilent Technologies). In addition, total-RNA was treated with DNAse to remove any DNA contaminants during the isolation process, as described by Qiagen. Concentrations were determined using a Nanodrop spectrophotometer (Wilmington, DE). Only samples with ratios between 1.8 and 2.0 were used**.** TruSeq RNA library preparation kit was

used for RNA-Seq library preparation, including the poly-A enrichment step to select only mRNA molecules **(Figure 2.2)**.

2.1.4 RNA-Sequencing Data

Ultra-deep RNA-Sequencing was performed on total 8 samples using Illumina HiSeq 2000 platform (**Figure 2.2**). Two samples were from 2 SMA-iPSC derived MNs (with 2 biological replicates per sample). SMA patient-1 (replicate-1) consist of 156,662,164 PE reads (2 x 101nt length); SMA patient-1 (replicate-2) consist of 193,040,114 PE reads and SMA patient-2 (replicate-1) consist of 159,771,568 PE reads; SMA patient-2 (replicate-2) consist of 168,518,554 PE reads. Two samples were from 2 Controls-iPSC derived MNs (with 2 biological replicates per sample). Control-1 (replicate-1) consist of 139,791,014 PE reads; Control-1 (replicate-2) consist of 209,121,716 and Control-2 (replicate-1) consist of 205,784,853 PE reads; Control-2 (replicate-1) consist of 166,029,806 PE reads.

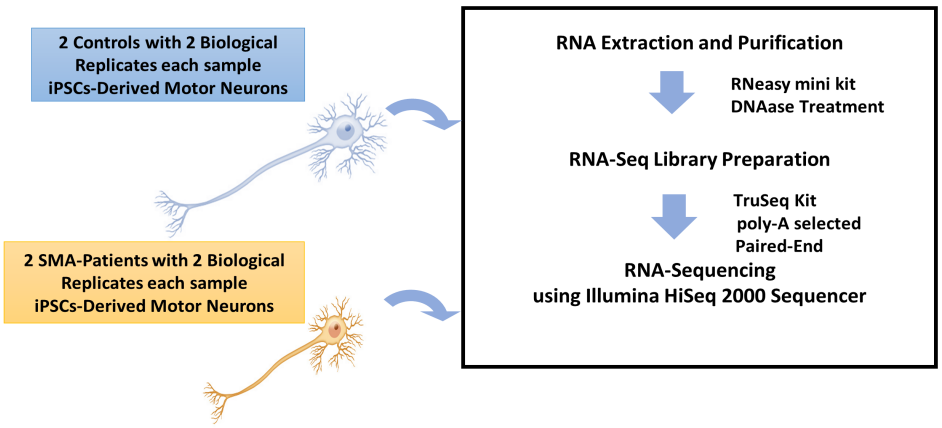


Figure 2.2: RNA isolation and RNA-Sequencing procedure.

Total-RNA samples isolated from SMA-iPSC derived MNs and control-iPSC derived MNs and purified by removing ribosomal-RNA and DNA contaminants. Paired-End, Poly-A selected RNA-sequencing libraries are prepared which are sequenced on Illumina HiSeq 2000 platform.

2.1.5 Pre-Processing of the Reads: Quality Check

Firstly, we performed the quality check of our dataset using FastQC software[218]. The raw RNA-Seq read files were analyzed for any possible bias including poor quality reads (*Phred score* or Q < 30), GC-content bias, duplicated sequences, presence of adaptor sequences or primer sequences. We assessed the read quality in terms of the read length distribution, *Phred* score distribution, and nucleotide frequencies obtained from FastQC statistics.

2.1.6 Read Alignment

After the quality check, we mapped reads to the reference genome (hg19). To assess the performance and accuracy of two widely used splice-aware aligners: TopHat2[257] and STAR[236] on our dataset, we set-up two pipelines (Pipeline-I and Pipeline-II). In Pipeline-I **(Figure 2.3)**, we used TopHat2 aligner with "--no-mixed", "--transcriptome-index" parameters. The other default parameters were kept as such, as they were designed to align PE RNA-Seq reads obtained from Illumina platform. The transcriptomic index was built using Ensembl gene-model annotations (Homo_sapiens.GRCh37.75.gff; hg19 assembly)[258]. Further, in Pipeline-II **(Figure 2.3)**, we used STAR aligner and align our RNA-Seq dataset with the following selected parameters: "outFilterIntronMotifs RemoveNoncanonical", "outSAMstrandField intronMotif", and "outSAMtype BAM SortedByCoordinate". The average percentage of uniquely aligned reads in all the samples for TopHat2 was 89.33% and for STAR was 92.46%.

The chosen parameters are described below:

TopHat2 Read Alignment:

1) **--transcriptome-index**: This option has been provided by TopHat2 to build the transcriptomic sequence file from the given gene annotation file (GTF file). Bowtie2 generates the index from this known transcript information. This operation helps in accelerating the read alignment step, as the reads are firstly aligned onto the transcriptome and then the unaligned reads are aligned onto genome to find unannotated splice site junctions.
2) **--no-mixed**: This option has been specifically designed for PE reads. By using this, the aligner only reports those alignments where both "read 1" and "read 2" can be mapped as a pair.

STAR Read Alignment:

1) **outFilterIntronMotifs RemoveNoncanonical:** The option helps in filtering the alignments by motifs. We applied "RemoveNoncanonical" option to consider only canonical junctions and discard alignments with non-canonical junctions.
2) **outSAMstrandField intronMotif:** This option is useful to make alignments compatible with Cufflinks-Cuffdiff2 tools for their downstream analysis. Such as, for unstranded RNA-Seq datasets, similar tools require a 'XS' flag to represent aligned read strand attribute and STAR aligner set this parameter by assigning 'XS' tag to all the splice

junctions. While in the case of strand-specific RNA-Seq datasets, there is no need to set this parameter.

3) **outSAMtype BAM SortedByCoordinate:** The option sorts the resulted read alignments by their coordinates.

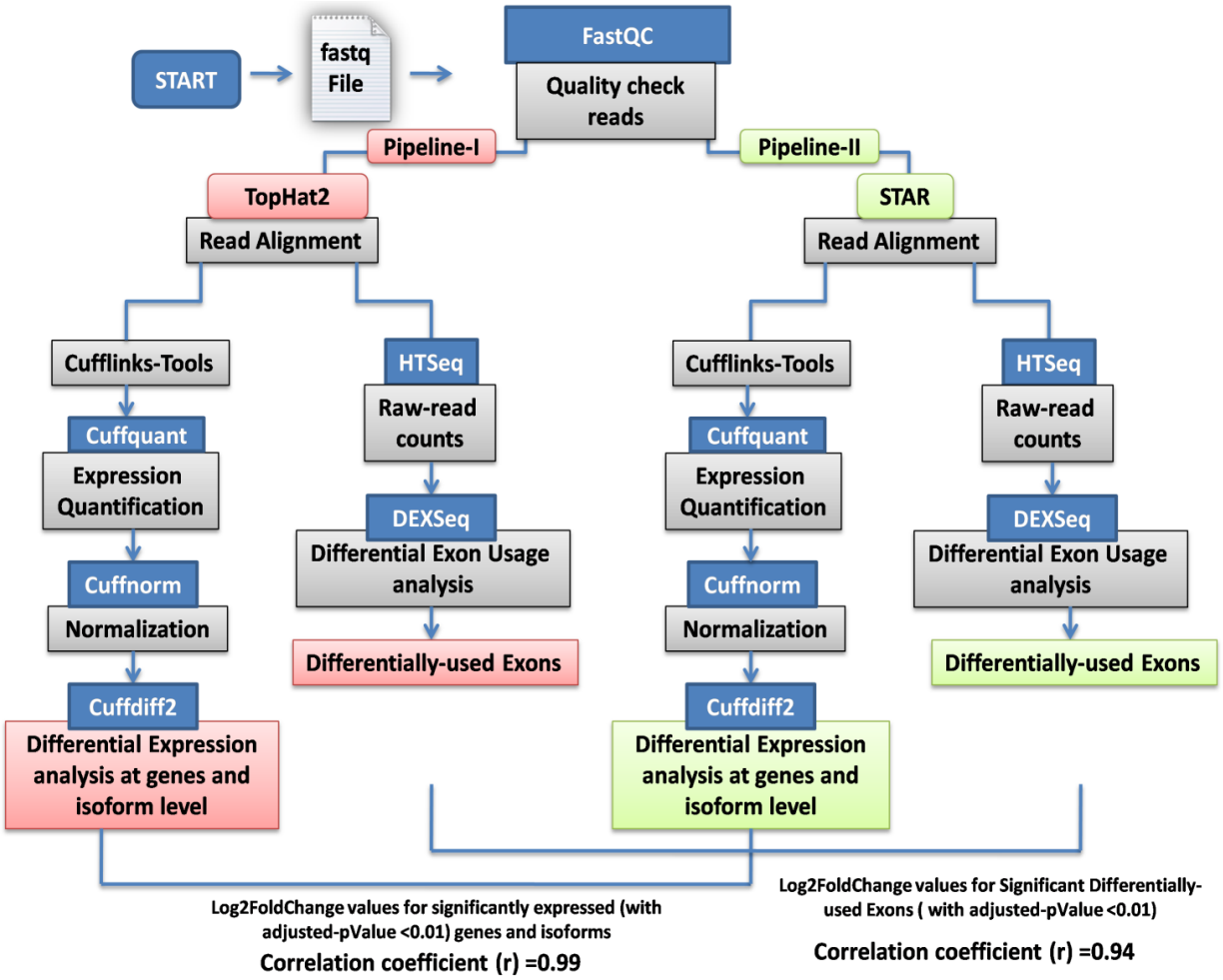


Figure 2.3: A rational approach for the selection of computational pipeline to analyze RNA-Seq data.
Starting from the quality checked reads performed on FastQC tool, the read alignment is performed onto the reference genome using two different splice-aware aligners: TopHat2 and STAR. The aligned reads are analyzed applying two pipelines: Pipeline-I and Pipeline-II. Within Pipeline-I, TopHat2 aligned samples are analyzed with Cufflinks tools to quantify the expression of genes and isoforms at genome-wide level and identify Differentially Expressed Genes and isoforms between SMA-patients and controls. These reads are also analyzed for identifying the Differentially-Used Exons between SMA-patients and controls using DEXSeq tool. Similarly, in Pipeline-II the STAR aligned reads are analyzed by Cufflinks and DEXSeq tools. The outcomes of the two pipelines are correlated by computing Pearson correlation coefficients.

2.1.7 Gene and Transcript Expression Level Quantification and Differential Expression Analysis Between Two Conditions

The alignment files (in BAM format) from both of the aligners were fed into Cufflinks tools[240]. We estimated the gene and transcript expression levels using "Cuffquant" on each individual sample with "-u" flag which support multi-read correction (those reads mapped at multiple loci within the genome). Further, we performed the normalization of the quantified expressions in order to scale the expressions levels for varying gene or transcript lengths and differences in the sequencing depth of the libraries using "Cuffnorm". To obtain the information about the similarities and dissimilarities in expression profiles within and between samples, we performed hierarchical (or unsupervised) clustering using the normalized expression levels of genes and isoforms (or transcripts). Subsequently, to determine the relative changes in the expression levels between different conditions (i.e. patients and controls) normalized gene and isoform level expressions were utilized using "Cuffdiff2"[252].

2.1.8 Differential Exon-Usage Analysis

To detect the differential usage of the exons between patients and controls, we applied DEXSeq[247] tool, on both pipeline-I and pipeline-II. This tool requires raw read-counts for the analysis, that we had obtained by applying HTSeq tool[259] on each sample in our dataset. HTSeq quantifies the aligned reads overlapping with each gene locus (gene/transcripts/exons) using pre-processed reference annotations (hg19 Ensembl). The pre-processing of the gene annotations is required to remove the redundancy in the transcripts as in gene-model annotations many features (exons) recur more than once, therefore DEXSeq collapse such information into "counting bins". These "counting bins" represent unique exons which overlaps with each other completely for each gene. If overlapping exons in two or more transcripts of a gene have different boundaries, then algorithm enforce the exons to split into parts which are referred as "exon-parts". To execute the preprocessing of reference annotations, DEXSeq package provides a python script. Further, raw read counts were analyzed for differential-exon usage by firstly, normalizing all the samples for their sequencing depth between and within samples; and biological variation (dispersion estimation) within samples (using Cox-Reid likelihood estimation method). Then, each counting-bin (exon or exon-part) in a whole gene was tested to determine the relative changes in the expressions (differential exon usage) between different conditions (patient versus control).

2.1.9 Execution of Computational Pipeline-I and Pipeline-II

To investigate on the performance and accuracy of both aligners, we applied the downstream analysis steps in the two pipelines: Pipeline-I and Pipeline-II **(Figure 2.3)**. In Pipeline-I, we used read alignments obtained from TopHat2 to be analyze by Cuffdiff2. In contrast, in Pipeline-II, we used read alignments obtained from STAR and we analyzed them by Cuffdiff2.

On the other hand, within Pipeline-I, Tophat2 obtained read alignments were also analyzed to identify Differential-Exon Usage (DEUs) using DEXSeq. Similarly, in Pipeline-II, similar analysis was performed on STAR obtained read alignments. Further, in order to choose the most robust pipeline out of the two, we intended to compare their outcomes by computing the Pearson correlation coefficient (r) values. In Cuffdiff2 analysis, we compared the $\log_2$ fold change ($\log_2$FC) values for the significantly differentially expressed genes resulted from Pipeline-I and Pipeline-II. Subsequently, we compared the $\log_2$FC values for significantly differentially used exons resulted from Pipeline-I and Pipeline-II. From these comparisons, we obtained very high correlations between both pipelines. Finally, we decided to use Pipeline-II because STAR is much faster than TopHat2, without losing accuracy.

$\log_2$ fold change is defined as:

$$log_2FC = log_2 \frac{(Condition2)}{(Condition1)} \qquad (1)$$

Where, Condition*1* **=** Control-FPKM expression values, and

Condition2 **=** Patient-FPKM expression values

2.1.10 A Rational Strategy for the Selection of Computational Tools to Estimate Expression Levels using Simulation Method

Nowadays, various computational methods are available for the downstream analysis of RNA-Seq data-sets, starting from aligning the reads until answering a study-specific question/s. For instance, quantification of expression levels, differential expression analysis at different genomic levels for genes, isoforms and exons, or investigation of the AS variations can be performed by implementing a variety of statistical methods. We found this task quite challenging and critical for accurately interpreting the RNA-Seq data. We started our experimental validation with the set of simulations. Simulated reads were generated on the basis of real data in order to mimic the biological phenomenon, underlying the quantification of the expression levels of genes and

isoforms. This helps to avoid under- or over-estimations within simulated read expression levels with respect to the real world data. We performed the read simulations using RNA-Seq by Expectation-Maximization simulation software (RSEM)[260]. RSEM is mainly designed to quantify the gene or isoform abundances in the RNA-Seq data (SE or PE) without relying upon the reference genome. Another utility in RSEM is RNA-Seq read simulation that can be guided by prior estimated gene or isoform abundances and modeled parameters. In our experiment, these parameters were obtained by providing the real RNA-Seq data. We took advantage of both of the utilities and generated simulated RNA-Seq PE reads. We applied our devised Pipeline-II on the simulated reads and obtained the gene and isoform expression estimations. Further, we computed the Pearson correlation coefficient between the expression values of simulated reads and expression estimations from RSEM-estimated model on the real data-set. The higher correlation coefficient between their expression estimations has reflected the great reliability of the computational pipeline we applied (**Figure 2.4**).

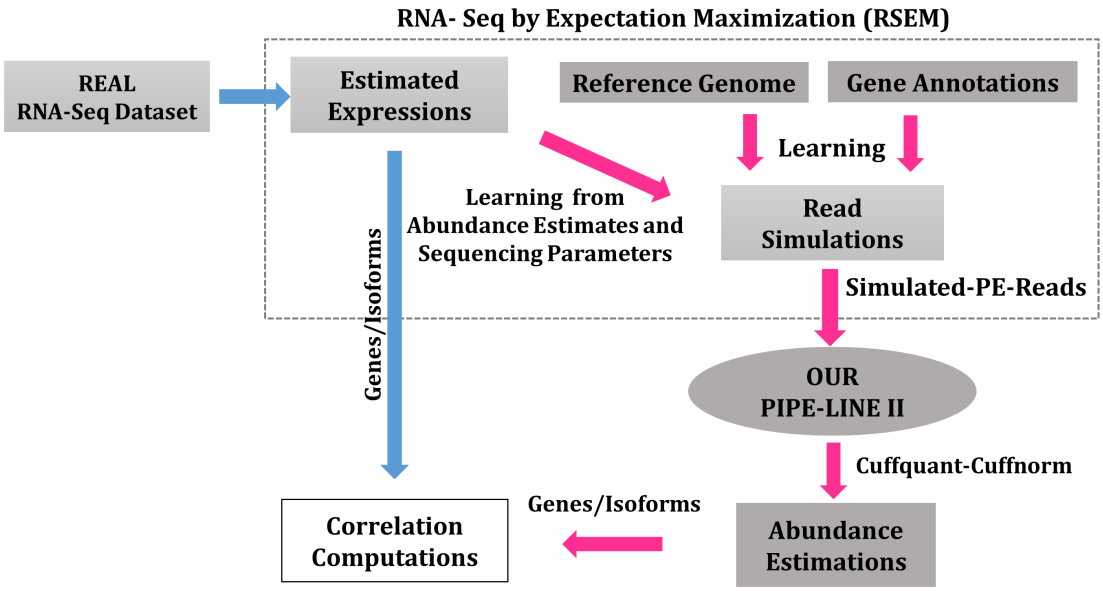


Figure 2.4: Simulation of PE reads using RSEM tool.

Firstly, the transcript level expressions are estimated from the real RNA-Seq dataset using RSEM. The read simulation requires the reference genome sequence file and known gene annotation file to learn the simulation algorithm and additionally to guide the simulation procedure, parameters from estimated expression obtained from real dataset can be used. Simulated PE reads are analyzed with Cuffquant-Cuffnorm in Pipeline-II (**described above in Figure 2.3**). The outcomes from the RSEM estimated expressions and transcript expressions obtained from Pipeline-II are correlated by computing the Pearson correlation coefficient.

2.1.11 Filtration of DEXSeq Obtained Exons

DEXSeq performs on the mathematical model of the gene annotations by collapsing the genomic features into counting-bins (exons and exon-parts). Because the tested set of "exon-parts" from DEXSeq analysis has no real existence and interpretation of these partial exons is highly complex. Therefore, we filtered-out "exon-parts" from the downstream analysis. In doing so, the DEXSeq results (obtained from the pipeline-II) were refined by following three criteria **(Figure 2.4)**:

- Firstly, we overlapped the resulted list of exons and exon-parts identified from DEXSeq with a list of known Alternative Cassette Exons (ACEs) expressed in human brain tissue taken from a recent study[51] and obtained a filtered list of DEXSeq results for ACEs only **(DEXSeq-ACEs; Figure 2.5 FILTER-1)**. These filtered set of exons represent only "exon-skipping" events that remained intact during the DEXSeq analysis.
- Secondly, the above shortlisted exons were filtered to remove first (Transcription Start Site or TSS exon) and last (Polyadenylation Site or PAS exon) exons from the corresponding transcripts using Ensembl hg19 gene annotations. This is because any relative changes in the expression of first exon of the transcript mainly occurs due to the transcriptional regulatory events that are controlled by transcription regulatory factors instead of AS regulatory events that are controlled by splicing regulatory factors **(Figure 2.5 FILTER-2)**. Therefore, by applying this filter we have focused only on the internal DEXSeq-ACEs of the transcripts.
- Lastly, these exons were epurated by filtering out those which were present in the list of significantly differentially expressed transcripts obtained from the Cuffdiff2 analysis. This is because these exons were resulted due to overall relative expression changes at the whole transcript level rather than changes at an individual exon level **(Figure 2.5 FILTER-3)**. Therefore, this step has provided a highly promising set of core DEXSeq-ACEs which are controlled by splicing regulatory factors. Now, these exons were divided into two parts **(Figure 2.6)**:
  1) Significant DEXSeq-ACEs: We obtained this list by putting a threshold on qvalue < 0.01 (significance level associated with each tested-exon for the differential-usage between different conditions in DEXSeq analysis). These significant DEXSeq-ACEs are Differentially-Used Alternative Cassette Exons (DUACEs) which show statistically significant relative changes in the expression at individual exon level due

to AS (exon-skipping) event between SMA-patients and healthy controls. Further, this list of significant DUACEs were divided on the basis of up-expression (or Enhanced) and down-expression (or Silenced) in SMA-patients with respect to controls using log_2FC values obtained from DEXSeq analysis for each tested exon (**Figure 2.6**).

a) Enhanced DUACEs ($log_2FC > 0$; in SMA-patients)
b) Silenced DUACEs ($log_2FC < 0$; in SMA-patients)

2) Non-Significant DEXSeq-ACEs: We obtained this list by putting a threshold on absolute ($log_2FC < 0.01$) and qvalue > 0.05. There exons were used as Control set of DEXSeq-ACEs for the downstream analysis.

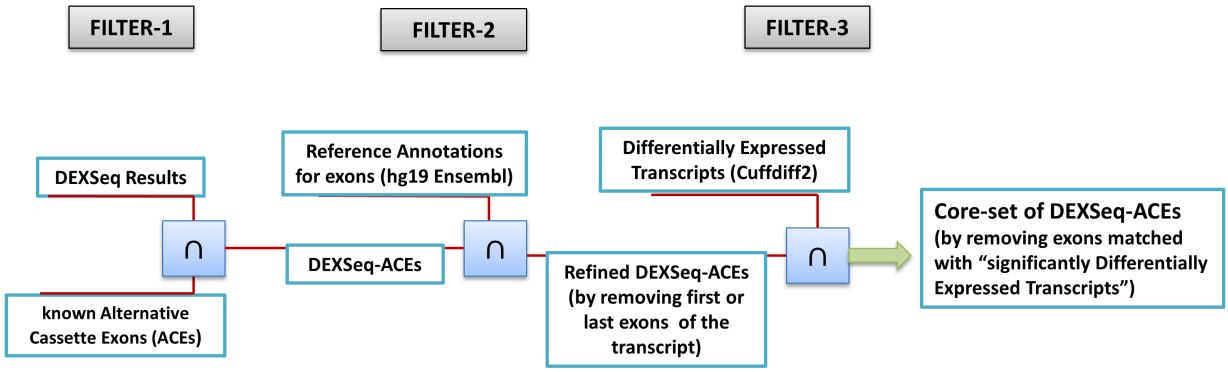


Figure 2.5: Filtration of DEXSeq identified exons by applying logical juxtaposition.
See main text for the details.

2.1.12 Identification of Motifs and RNA-Binding Proteins

The list of core-DEXSeq-ACEs were analyzed further to identify Splicing Regulatory Elements (SREs) which are known to localized within exonic and intronic sequences of the transcripts (**Figure 2.6**). These elements are short stretch of nucleotide sequences or motifs (consensus sequence) that regulates splicing by recruiting RNA-Binding Proteins (RBPs). The purpose of this analysis was to pinpoint the key mechanisms underlying mis-regulations in mRNA-splicing, specific to MNs. In doing so, the sequence level analysis was performed on the list of significant DUACEs (enhanced and silenced) and non-significant DEXSeq-ACEs with their upstream and downstream flanking introns.

Firstly, the genomic sequences were retrieved from enhanced, silenced and control exons and their flanking upstream and downstream introns using GeCo++ library[261]. The upstream and

downstream introns were restricted to the fixed length of 150bp and first 10bp were subtracted from each side of flanking intron of the corresponding exon. This was because the initial bases of the intronic region (on each side of the exon) consists of splice-sites (5' and 3') where the motif finding signal are relatively stronger than other locations of the intronic sequence. Therefore, to avoid such signal masking effect during motif discovery within SREs, first few bases should be excluded towards upstream and downstream introns.

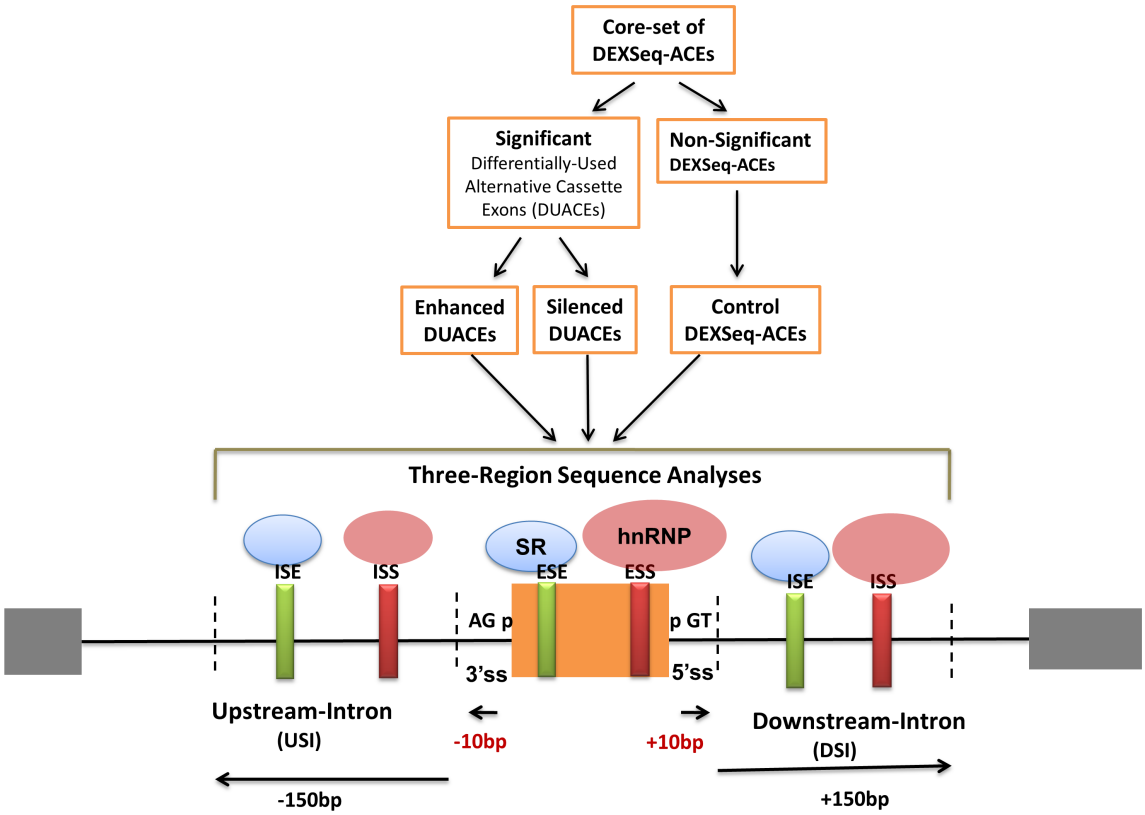


Figure 2.6: Categorization and three-region sequence level analysis of core DEXSeq-ACEs.

The core-set of DEXSeq-ACEs are categorized into significant and non-significant DEXSeq-ACEs. The significant DEXSeq-ACEs are DUACEs, having statistically significant relative changes in the expression at individual exon level between SMA-patients and healthy controls. On the basis of expression level either up or down in SMA-patients, the significant DUACEs are divided into Enhanced and Silenced DUACEs. All Significant DUACEs and Control DEXSeq-ACEs (Non-Significant) are used for three-region sequence level analyses to identify motifs (or SREs; shown as ISE and ISS in introns; ESE and ESS in exons, represented with '*green*' and '*red*' colored thick bars, respectively) and RBPs (SR or hnRNPs, represented with '*light blue*' and '*pink*' colored eclipses, respectively) that specifically binds on SREs. Three-regions consist of exons (represented with '*orange*' colored rectangle) and their flanking introns (Upstream-Intron or USI and Downstream-Intron or DSI, represented with '*black*' colored line) which has a fixed length of 150bp (marked with '*black*' colored dotted lines and long '*black*' colored arrows on both sides) and first 10bp

(GT at 5'ss and AG at 3'ss represented with short '*black*' colored arrows) are deducted on each side from intronic sequence analysis.

Overall, we have 9 sequence files (**Figure 2.7**), containing 3 files from the enhanced DUACEs (first file for upstream-intron of enhanced DUACEs, second file for enhanced DUACEs themselves and third file for downstream-intron of enhanced DUACEs); 3 files from silenced DUACEs (first file for upstream-intron of silenced DUACEs, second file for silenced DUACEs themselves and third file for downstream-intron of silenced DUACEs); and 3 files from control DEXSeq-ACEs (first file for upstream-intron of control DEXSeq-ACEs, second file for control DEXSeq-ACEs themselves and third file for downstream-intron of control DEXSeq-ACEs). However, for the motif discovery analysis, we used only enhanced and silenced DUACEs that comprised of 6 sequence files. Motif discovery was performed using MEMERIS software (Multiple EM for Motif Elicitation in RNAs Including Secondary Structures)[262]: an extension of MEME motif finder[263] which uses the concept of position specific scoring matrices (PSSMs) to identify biological patterns within the nucleotide sequences. MEMERIS also presents a unique approach of pattern finding by utilizing the information of secondary structures which guide motif search within the single-stranded regions of RNA. It intends to pre-compute the probability values for single-strandedness in a position-specific manner by applying two methods: (i) **PU:** the probability that all the nucleotides are unpaired, and (ii) **EF:** the expected fraction of nucleotides which do not make base-pairs, within a corresponding subsequence or substring.

The chosen parameters for MEMERIS execution include: "-w 5 -dna -pi 0.01 -mod zoops -nmotifs 5 -minsites 10 -secstruct".

Where,

1) **-w:** is the length of the motif to be searched in a given nucleotide sequence.
2) **-dna:** represents the nucleotide sequence.
3) **-pi:** refers to the "Pseudocount" value which helps in adjusting the influence of single-strandedness that is determined from EF or PU above described methods for a given motif length (-w) per sequence.
4) **-mod:** the given nucleotide sequence can be modeled by different models. Such as sequence containing either zero, one or more than one motifs which are non-overlapping in position specific scoring matrix (PSSM). PSSM gives the probability distribution values at every position within the given sequence. MEMERIS provides three models for motif

searching: OOPS, ZOOPS and TCM. In OOPS, model sequences are queried to find only "One motif Occurrence Per Sequence". In ZOOPS, model sequences are queried to find "Zero or One motif Occurrence Per Sequence". In Two Component Mixture (TCM) model sequences are queried to find Zero or more occurrences per sequence. We chose ZOOPS motif search model within the sequences which also allowed for an absence of the motif occurrence in the sequence.

5) **-nmotifs:** defines number of motifs to be searched.
6) -**minsites:** defines the minimum number of occurrences of the motifs per sequence.
7) **-secstruc:** retrieves the probability values to estimate single-strandedness of the sequences. These values are pre-computed by launching a Perl script "GetSecondaryStructureValues.perl" included in MEMERIS package. We used "-w 5" and method "EF".

Likewise, we identified total 30 novel motifs using 6 sequence files (5 motifs per sequence file from the motif search). Further, we intended to remove similar motifs which were having more than 60% similarity in their PSSMs. Only 6 *out of* 30 motifs were found to be unique (unrelated) on the basis of this criterion. Subsequently, enrichment analysis was performed for the 6 unique motifs. We quantified the occurrences of each motif in all 9 sequence databases (Enhanced, Silenced and Control exon and intron sequence files See **Figure 2.7**) using MEME suite program[264], namely, Find Individual Motif Occurrence (FIMO)[265]. FIMO search for all the provided motifs to find matches within the sequences of all files individually. Altogether, we have performed 54 analyses wherein we have 6 motifs which were searched in 9 sequence databases individually (**Figure 2.7**). Further, in order to determine if the overrepresentation differences for searched motifs within all sequence files (corresponding to specific region i.e. DSI, or Exon or USI and specific condition i.e. Enhanced or Silenced or Control) have any statistical significance or not, we applied the non-parametric Wilcoxon test. In total, we performed 54 pairwise comparisons and performed Wilcoxon test to obtain statistical significances.

Next, we analyzed 6 unique motifs to find known set of RNA-binding proteins (RBPs) using a computational tool, TOMTOM[266]. TOMTOM compares PSSM of user's identified motifs (query) within a database of known PSSMs (target), which are associated with specific RBPs. As a "target" motif database, we used "Ray 2013 *Homo sapiens* (DNA-encoded)"[267] which consists of 102 motifs of 7 – 8bp length.

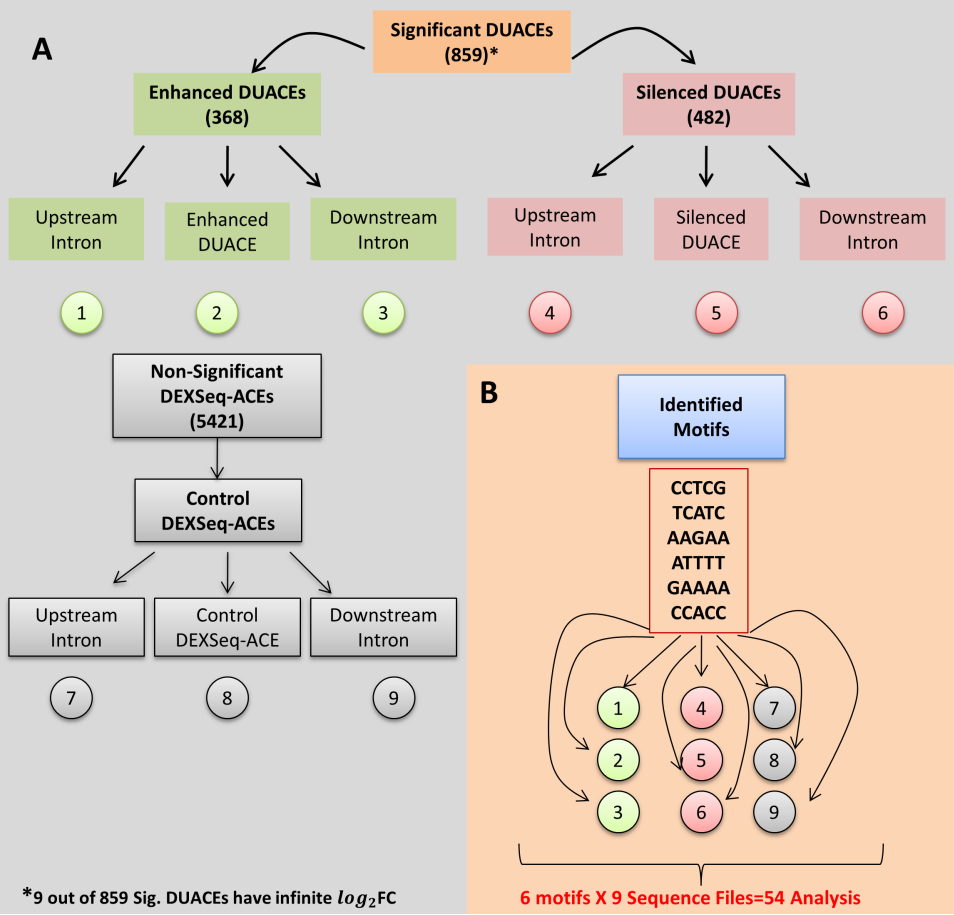


Figure 2.7: Motif identification and motif enrichment analysis.

A The motifs identification is performed on files labeled with circles '1' to '6' derived from the "Significant set of Enhanced and Silenced DUACEs". Each Enhanced DUACE has been divided into three regions: 1. Upstream Flanking Intron, 2. DUACE itself, and 3. Downstream Flanking Intron. **B** Every motif is searched within all the sequence files labeled with '1' to '9' which provides the occurrences (enrichment or overrepresentation) of each motif within all the sequence files or sequence databases. In total 54 analyses has been performed, including 6 motifs that have been searched in 9 sequence files, individually.

*Nine Significant DUACEs were having infinite log_2FC values which we excluded from the further analysis.

2.1.13 Validation of Significant DEGs and DUACEs with Functional Annotation Analysis

Gene Ontology (GO) enrichment analysis allows to determine characteristic biological attributes in a given gene set. It is based on the premise that functionally related genes should accumulate together in the corresponding GO category. We used Database for Annotation, Visualization and Integrated Discovery (DAVID[268]), a web-based functional

annotation analysis tool. It helps in identifying the enriched or over-represented gene ontology terms (GO), covering three domains: “Biological Process (BP)”, “Cellular Component (CC)” and “Molecular Function (MF)”. The enrichment of biological pathways and gene-set disease associations can also be obtained from this analysis. Firstly, we performed the DAVID analysis on our list of significantly Differentially Expressed Genes (DEGs) (qvalue < 0.05) that are derived from Cufflinks-Cuffdiff2 pipeline. In our analysis, the complete set of 1,858 significant DEGs were provided as “target genes”, and 63,651 genes including both significant and non-significant served the purpose of “background genes”.

Furthermore, we evaluated the genes corresponding to our set of DUACEs. Total of 859 genes, corresponding to the list of DUACEs were provided as “target genes” and the list of unique 14,758 genes, including both significant DUACEs and non-significant core DEXSeq-ACEs acted as “background genes” for functional annotation analysis.

2.2 Development of Computational Model to Estimate Transcript Expression

2.2.1 Re-Construction of Isoform Paths in a given Gene Locus

In a given gene locus, we can have multiple exon combinations giving rise to multiple different isoforms. Of course not every possible path exists in reality, therefore some isoforms will only represent hypothetical combinations which do not exist in reality and some will be actually coherent with the given annotations and also validated with the experimental data obtained from RNA-Seq technology. The isoform paths can be re-constructed by different approaches which are explained in the following sections.

2.2.2 Generation of Isoform Paths on the basis of Junction Information using Graphs

To construct all the possible isoform paths by considering all the possible exon/intron combinations in a given gene locus, the idea of Directed Acyclic Graphs (DAGs; **Figure 2.8**) can suffice the need. In mathematical terms, DAGs consist of finite set of vertices and edges with directional information without forming any directed cycle. Formally, graphs can be represented as $G = (V, E)$ where, V represents the set of 'vertices' or 'nodes' and E represents the connections or edges between any two vertices. In such graphs, the vertices and edges are arranged in an ordered fashion. To represent a given set of isoforms with the help of graphs, we considered the 5' and 3' splice site junctions as 'vertices' and introns or the exons of the isoform represented the 'edges' of the graph. By drawing all possible combinations of edges with the identified set of junctions, multiple isoform paths can be obtained. Later, these paths can be validated with the experimental data (for example, RNA-Seq data) by estimating the expression for each possible path at per-base resolution and preserving only the ones that actually are present in the analyzed samples.

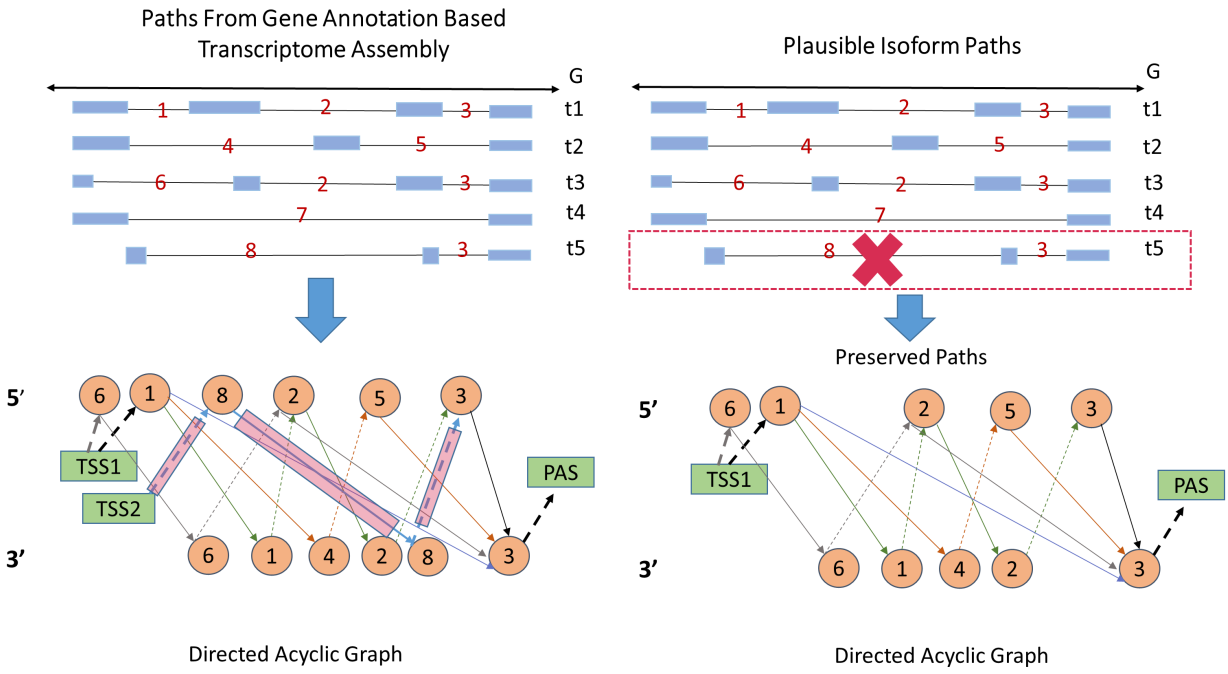


Figure 2.8: A representation of Directed Acyclic Graph (DAG).

A hypothetical gene *'G'* consists of five isoforms (t1 – t5 paths from gene annotations based transcriptomic assembly obtained from Cufflinks tool) where exons are represented with *'blue'* colored thick rectangular boxes and introns are represented with thin *'dark grey'* lines. Unique splice-site junctions are numbered from 1 through 8 (digits in '*red*' color). Directed acyclic graph is generated, comprising all the 5 isoforms (shown with different colored directed lines) from gene 'G', where splice-sites junctions form the *vertices* of the graph (represented as 5'-splice site and 3'-splice site, with *'orange'* colored circles); *edges* of the graph can be either exons or introns. Exons of the isoform are shown with dotted directed lines (where the direction is from 3'-5' splice-site) and introns of the isoform are shown with continuous directed lines (from 5'-3' direction). TSS and PAS of the isoforms are represented with *'green'* colored rectangles. If any junction is not supported by the junction information obtained from the RNA-Seq read alignments, then that complete isoform path is discarded. Path t5 is excluded from the list if splice-site junction-8 is not supported by junction information and 4 paths will be persevered.

2.2.3 Generation of Isoform Paths on the basis of Known Gene Annotations

Another simpler approach is to consider only the known gene annotation information and retrieve the isoform paths with already annotated exon/intron combinations. This approach will restrict the analysis to only known set of transcripts but makes the things much easier in terms of computation time and memory usage.

2.2.4 Generation of Isoform Paths with a Combined Approach

In this approach, two sources can be combined. In doing so, the set of paths are obtained from Cufflinks generated transcriptome assembly based on RNA-Seq data. Further, these paths are validated from the exon-intron junction information received from RNA-Seq data alignments. Only those paths will be preserved which are justified by junction information and the rest will be discarded. We used this approach to start with the set of plausible isoform paths in a given gene locus.

2.2.5 Obtaining the Defined Set of Information from Total RNA-Seq Data

The information about the exon-intron junctions was retrieved from the aligned total RNA-Seq samples which were stored in the BAM files. The "CIGAR" string information was utilized for defining the presence of Exon-Intron junctions. These junctions denote the location of 5' and 3' splice sites which were supported by total number of junctions within the given locus. As mentioned earlier, this information was utilized to filter the list of isoform paths obtained from Cufflinks transcriptome assembly. Only those paths from Cufflinks were considered valid which were supported by the above identified exon-intron junctions with enough supporting junction counts (junction count > 2) in the total RNA-Seq data. Ultimately, the filtered set of isoform paths were assigned to a given gene locus which were called plausible isoform paths, according to the processed samples.

2.2.6 MODEL

In a genomic locus for which a set of isoforms is given we obtain from the aligned fragments a vector

$$b = \begin{bmatrix} b(1) \\ b(2) \\ \vdots \\ b(L_g) \end{bmatrix}$$

containing the number of fragments covering each genomic position. The purpose of our model is to obtain an estimate of nascent and mature transcription levels which best approximate the measured *b* (which was computed at per-base resolution within the given gene locus by applying BamTools API[269]).

2.2.6.1 Coverage Probability along a Transcript: *CPT(X)*

Reads coverage deriving from an RNA transcript of length L_t is not uniform along the transcript itself (**Figure 2.9**). Given a discrete fragment length distribution *F(l)* (which can be obtained from the data) the probability of observing a fragment of length l_i at position x is affected by the distance from the transcript borders and can be expressed as:

$$p_i(x) = \begin{cases} x\dfrac{F(l_i)}{l_i}, & x < l_i \\ F(l_i), & l_i \leqslant x \leqslant L_t - l_i \\ (L_t - x)\dfrac{F(l_i)}{l_i}, & x > L_t - l_i \end{cases}$$

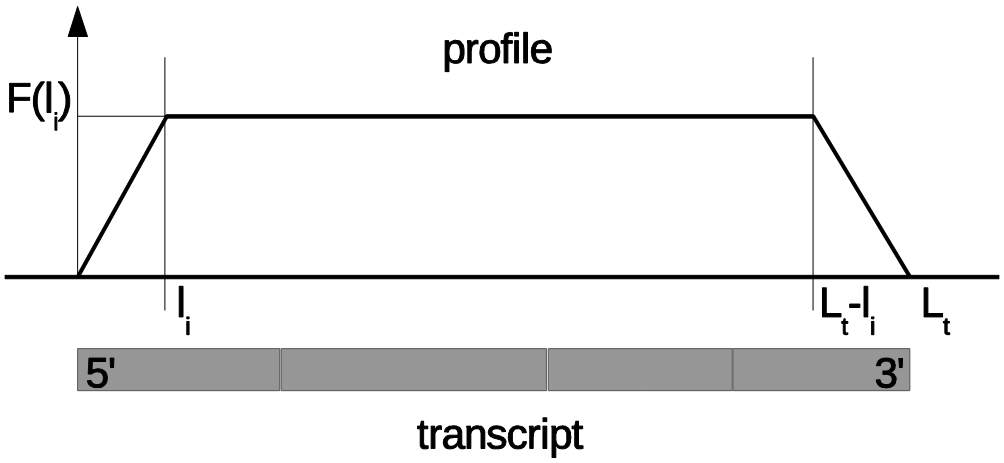


Figure 2.9: Mature transcript profile at per base resolution.

The profile describes the probability of observing a fragment of length l_i at any position along the transcript of length L_t. This probability is affected by the distance from the transcript borders (which are marked by 5'-start of the mature transcript and 3'-end of the mature transcript) that is because less number of fragments tend to overlap the *start* and *end* bases along the transcript. $F(l_i)$ is the relative frequency of the fragment of length l_i obtained from the distribution of relative frequencies of all fragment lengths in the analyzed total RNA-Seq data (**See Figure 3.15 in Results**; **Chapter 3**). The probability values remain constant in-between the transcript borders i.e. when any position along the transcript is equal to the fragment length l_i until it is equal or smaller than L_t-l_i.

By summing over the N_f possible fragment lengths:

$$CPT(x) = \sum_{i=1}^{N_f} p_i\,(x)$$

we obtain a the *CPT(x)* profile which describe the probability of observing a fragment of any length at each position along the transcript (**Figure 2.10**).

Where, N_f is the number of possible fragment lengths (obtained from fragment length distribution library of total RNA-Seq data) observed at position x along the transcript.

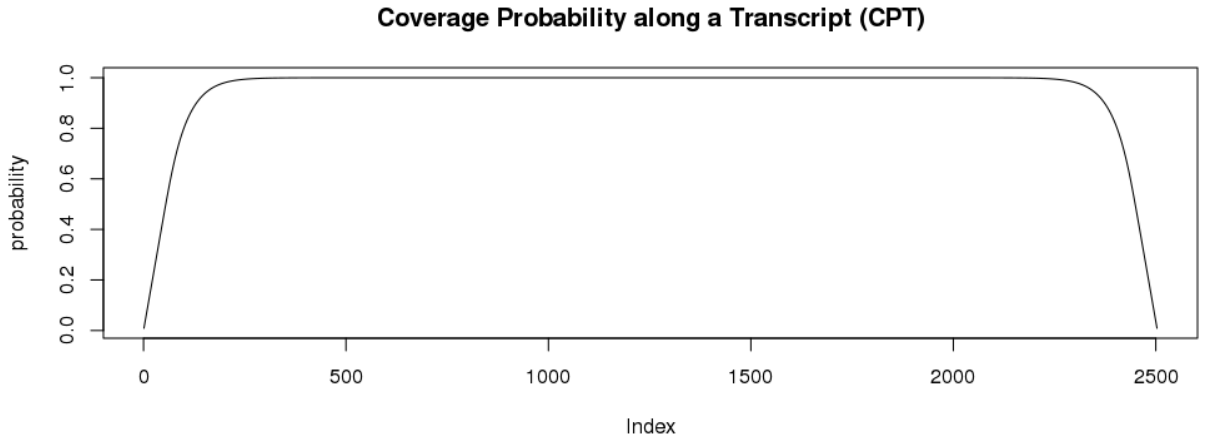


Figure 2.10: An example of Coverage Probability along a Transcript (CPT).

CPT gives the *probability* (along the *y-axis*) of observing a fragment of any length at each *position* (or *index* along *x-axis* of the mature transcript) along the transcript. For instance, at *position* or *Index* x=1 along the transcript of length L_t, CPT (x) is computed by summing over all the probabilities for N_f possible fragment lengths in the library at position x.

2.2.6.2 Genomic Profile for a Mature Transcript

A mature transcript has introns already spliced out. For a mature transcript, by mapping to the genomic locus its positions and corresponding *CPT(x)* values and setting to 0 intronic positions, we obtain a genomic mature probability profile $p_m(x)$ of length L_g (**Figure 2.11**).

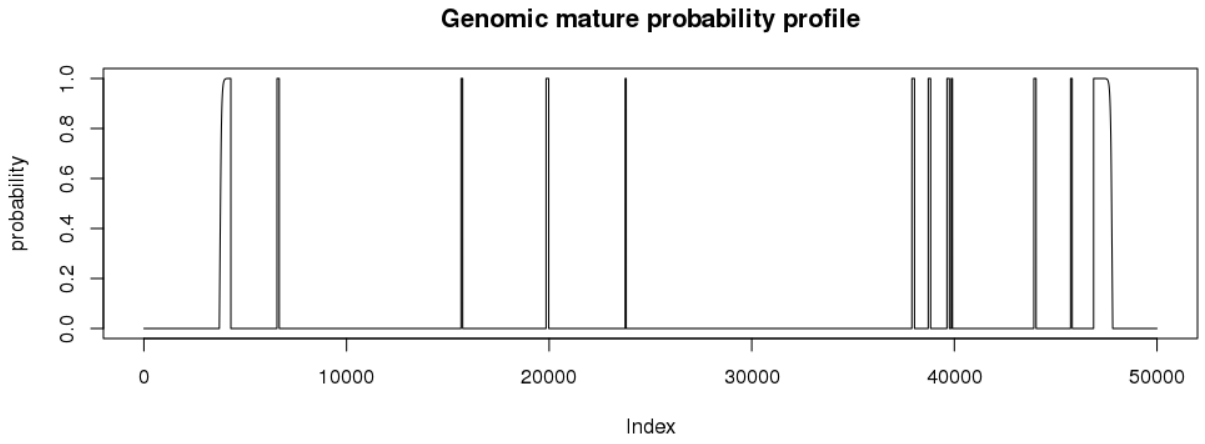


Figure 2.11: Genomic probability profile for a mature transcript.

The genomic coverage probability profile of the mature transcript where the transcript coordinates are mapped onto the reference genome to obtain corresponding genomic coordinates (represented as *Index* or *position* along the *x-axis;* genomic coordinates for 'Start' or TSS and 'End' or PAS position of the transcript). Genomic coverage probability profile (along the *y-axis*) accumulates across all the transcribed exons (represented as *high bars*) and across introns (represented with *a line* between narrow *long bars* shows no coverage across the genomic index) there is no coverage as in the mature transcript the introns are spliced out.

2.2.6.3 Genomic Profiles of Nascent Transcripts

In the simplifying hypotheses of constant transcription velocity, we can expect to observe in our library any partially transcribed fraction of a transcript with the same probability, depending on nascent transcription rate only. Since splicing is co-transcriptional, we consider the excision of an intron to occur immediately after completion of its downstream exon.

A transcript fraction with a still unspliced intron will not contain any previous intron and therefore will contribute to the genomic coverage profile in all transcribed exons and in the unique unspliced intron only (**Figure 2.12 and Figure 2.13**).

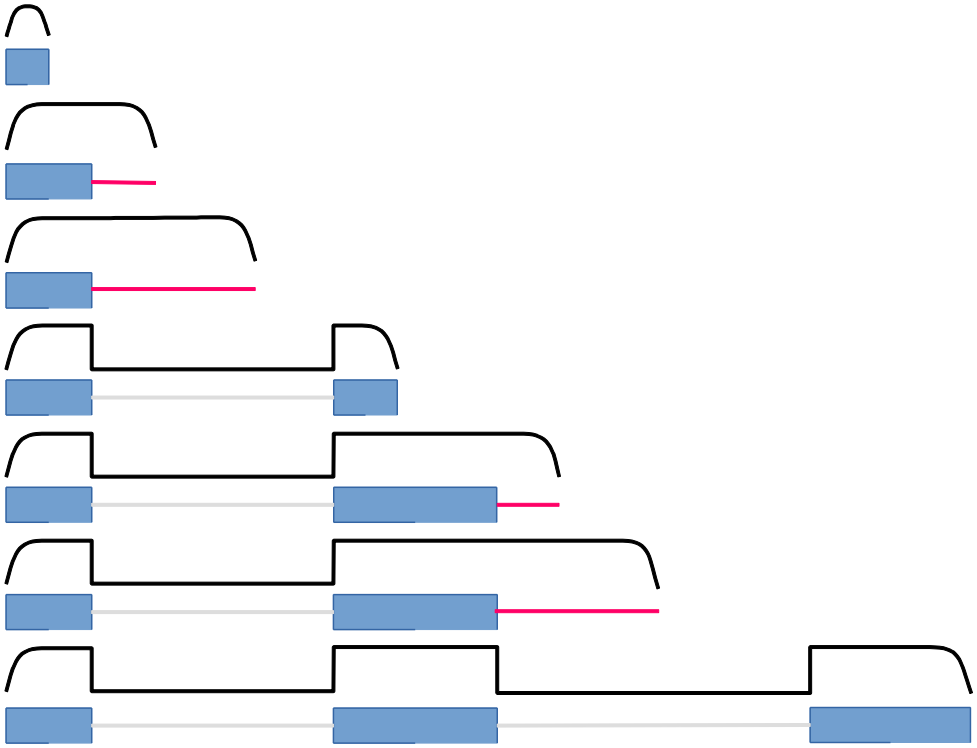


Figure 2.12: Example of genomic probability profiles for nascent transcripts at different stages of transcription.

Coverage probability profile start to accumulate gradually as the transcription proceeds. The transcription of the first exon gives a partially transcribed transcript or nascent transcript (represented with small *'blue'* square and coverage is shown with a '*black curve'* over it) which transcribes further the remaining first exon and partially transcribed intron (represented with *'red'* line). Across the transcribed exons, coverage probability profile tend to accumulate in every nascent transcript while the coverage profile is unique for every unspliced intron as splicing is co-transcriptional and according to our simplifying hypothesis, splicing takes place immediately after the last base of the downstream exon is transcribed as a result there will be no coverage for the previous intron (represented with *'grey'* line) and coverage probability profile will be unique for every intron (*saw-tooth*) across the transcript.

Therefore, the full genomic nascent probability profile $p_n(x)$of length L_g is obtained by averaging the genomic mapped *CPTs* deriving from each possible transcript fraction.

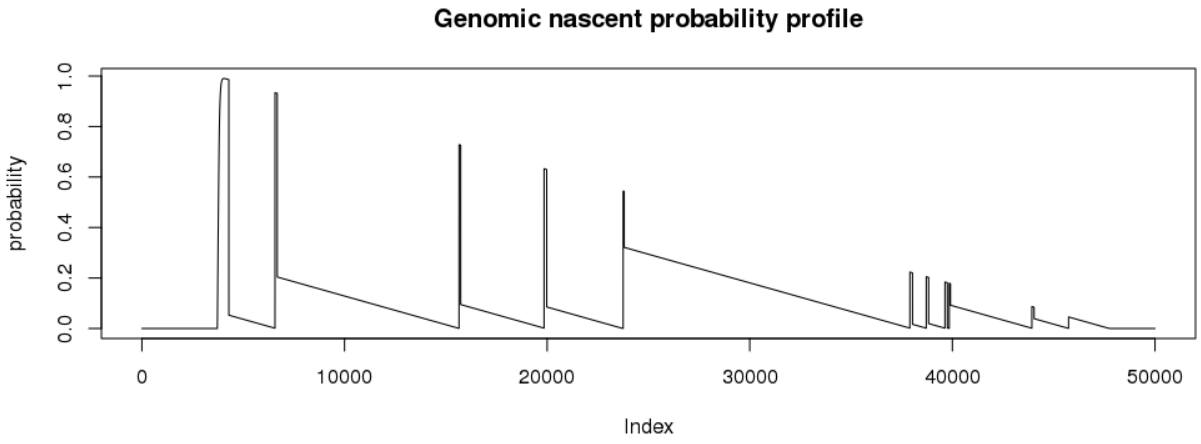


Figure 2.13: Genomic probability profile for a nascent transcript.

Along *x-axis* fragment length distribution probability profile and along *y-axis* genomic positions (*Index*) of the transcript. For all the transcribed exons the coverage profile accumulates (*high peaks* in the plot) across the nascent transcript and for the introns, coverage probability profile is unique for each unspliced intron (saw-tooth shaped coverage probability profile) along the nascent transcript

In a gene locus of length L_g we obtain both the mature and nascent probability profiles for each of the N_t transcripts spanning the locus and we organize the profiles in two matrices:

$$M_m = \begin{bmatrix} p_{m1}(1) & p_{m2}(1) & \cdots & p_{mNt}(1) \\ p_{m1}(2) & p_{m2}(2) & \cdots & p_{mNt}(2) \\ \vdots & \vdots & \cdots & \vdots \\ p_{m1}(L_g) & p_{m2}(L_g) & \cdots & p_{mNt}(L_g) \end{bmatrix}$$

and

$$M_n = \begin{bmatrix} p_{n1}(1) & p_{n2}(1) & \cdots & p_{nNt}(1) \\ p_{n1}(2) & p_{n2}(2) & \cdots & p_{nNt}(2) \\ \vdots & \vdots & \cdots & \vdots \\ p_{n1}(L_g) & p_{n2}(L_g) & \cdots & p_{nNt}(L_g) \end{bmatrix}$$

For the N_t transcripts spanning the locus we define:

$$E_m = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_{N_t} \end{bmatrix} \text{ and } E_n = \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_{N_t} \end{bmatrix}$$

Here, for the kth transcript m_k and n_k represent the number of mature and nascent molecules respectively. Then we model the vector b of observed fragments counts through the following equation:

$$b = M_m E_m + M_n E_n + \varepsilon$$

We could estimate E_m and E_n from the data by minimizing the error ε:

$$\min_{E_m \geq 0; E_n \geq 0} \|b - M_m E_m - M_n E_n\|$$

Where, $\mathbf{M_m}$ = Matrix of fragment length distribution probability profile for each "Mature transcript", $\mathbf{M_n}$ = Matrix of fragment length distribution probability profile for each "Nascent transcript". $\mathbf{E_n}$= Estimate of nascent transcript expression and $\mathbf{E_m}$ = Estimate of mature transcript expression

In this way, E_m and E_n estimates would be independent. Though, the number of mature transcripts copies depend on the rate of nascent transcription and therefore it cannot be considered independent form the number of nascent transcripts copies. We can write:

$$E_m = \begin{bmatrix} \alpha_1 n_1 \\ \alpha_2 n_2 \\ \vdots \\ \alpha_{N_t} n_{N_t} \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & a_{N_t} \end{bmatrix} E_n = AE_n$$

Where,

$$\alpha = diag(A) = \frac{E_m}{E_n}$$

represents the vector of ratios between mature and nascent fragments for each transcript.

We rewrite our model equation as:

$$b = (M_m A + M_n)E_n + \varepsilon$$

and therefore we will solve:

$$\min_{A \geq 0; E_n \geq 0} \|b - (M_m A + M_n)E_n\|$$

2.2.6.4 Model Parameters Identification Procedure

To solve this nonlinear problem (i.e. the model equation is not linear in the parameters) we adopted an alternating gradient descent method. To this purpose, according to the minimization problem:

$$\min_{A \geq 0; E_n \geq 0} \|b - (M_m A + M_n)E_n\|$$

we define the following objective function:

$$f(A, E_n) = \sqrt{(b - M_m A E_n + M_n E_n)(b - M_m A E_n + M_n E_n)^T}$$

and we calculate the derivatives:

$$\frac{\partial f(A, E_n)}{\partial E_n} = \frac{{E_n}^T K^T K}{f(A, E_n)} - \frac{b^T K}{f(A, E_n)}$$

$$\frac{\partial f(A, E_n)}{\partial A} = \frac{{E_n}^T \left(A^T {M_m}^T + {M_n}^T\right) M_m X}{f(A, E_n)} - \frac{b^T M_m X}{f(A, E_n)}$$

Where, for convenience we defined $K = M_m A + M_n$

and,

$$X = \begin{bmatrix} n1 & 0 & \cdots & 0 \\ 0 & n2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & n_{N_t} \end{bmatrix}, diag(X) = E_n$$

At each step we alternate between updating A and E_n by using the following multiplicative update rules:

$$A' < -A \frac{{E_n}^T \left(A^T {M_m}^T + {M_n}^T\right) M_m X}{b^T M_m X}$$

and,

$$E_n{}' < -E_n \frac{E_n{}^T K^T K}{b^T K}$$

after updating both A and E_n.

In the tests we performed, iterations were stopped when the following criterion was met:

$$\|E_n{}' - E_n\| < 1 \cdot 10^{-3}$$

3.1 Study of Alternative Splicing in SMA

3.1.1 MNs Generated from SMA Patient iPSCs Present Reduced Cell Survival in Culture

The iPSCs generated from SMA patients and healthy control fibroblasts with non-viral and non-integrating methods showed pluripotency markers and were able to differentiate into MNs using established protocols[163] (**Figure 3.1A**). After 4–5 weeks under differentiation conditions, cells were generated that expressed MN-specific transcription factors (TFs), such as spinal cord progenitor markers (such as HB9, ISLET1, and OLIG2) and pan-neuronal markers (such as TuJ1, Neurofilament, and MAP2). Majority of these HB9/ISLET1-positive neurons expressed Choline Acetyl Transferase (ChAT) and were positive for the MN marker SMI-32, demonstrating a MN phenotype (**Figure 3.1B**). The *in vitro* differentiation protocol yielded mixed cell population that also included non-MN cells. Given the limited availability of surface markers to isolate MNs and purify them further, we applied a physical strategy based on gradient centrifugation. After cells were selected using this method, immunocytochemistry analysis showed that the percentage of $ChAT^+SMI32^+$ cells derived from healthy control iSPCs (WT-iPSCs) was $88.4 \pm 8.3\%$, and $87.6 \pm 7.7\%$ for cells derived from SMA-iPSCs. Further, astrocytic cells were quantified wherein less than 1% of the differentiated cells from iPSCs expressed the astrocyte marker GFAP. We also observed that SMA iPSC-derived MNs exhibit reduced survival in long-term culture, and in this study at 10 weeks, we observed a reduction of MNs number in the SMA-iPSC cultures compared with WT-iPSCs ($p < 0.001$; **Figure 3.1C**).

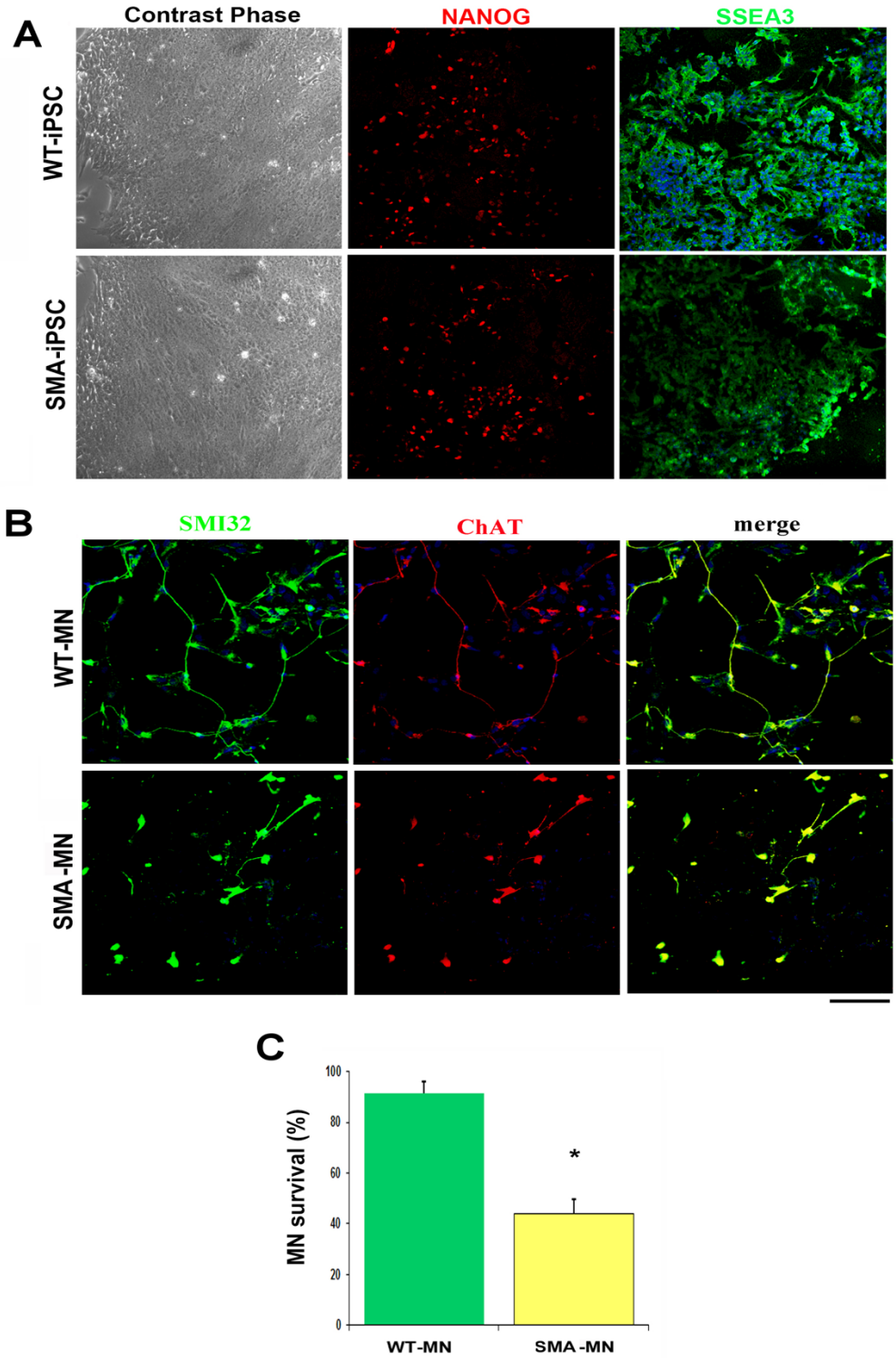


Figure 3.1: Reprogramming of skin fibroblast cells of SMA-patient and healthy control (WT) into iPSCs and their differentiation into MNs using non-viral and non-integrating method.

A Immunocytochemical characterization of iPSC clones derived from SMA patient (SMA-iPSCs) and healthy control

(WT-iPSCs) which expresses pluripotency TF NANOG (‘*red*’ colored iPSC colonies) and stem cell surface marker SSEA3 (‘*green*’ colored iPSC colonies). **B** SMA-iPSCs and WT-iPSCs differentiation into spinal MNs expresses MN-specific markers including SMI32 (‘*green*’) and ChAT (‘*red*’). MNs (‘*yellow’*) shows merging of ‘*green*’ and ‘*red*’ colors which represents double-positive MNs. **C** Quantification of MNs at 10 weeks after differentiation from iPSCs show reduced number of SMA-iPSCs derived MNs (‘*yellow*’). with respect to WT-iPSCs derived MNs (‘*green’)* (one-way ANOVA with Tukey’s post hoc test resulting *p < 0.001, 10 weeks).

3.1.2 Quality of RNA-Seq Samples and Read Mappability

The sequencing quality of RNA-Seq samples was high, including two SMA-patients and two healthy controls with two biological replicates each **(Figure 3.2 and Figure 3.3)**. Total 130-190 million reads (91 to 94% reads) were uniquely mapped across all samples onto the reference genome, revealing high mappability of our samples **(Table 3.1)**. Some reads (4-6%) were mapped at multiple genomic locations. Very few percentage of reads (0.3-0.13%) were mapped onto too many loci and others remain unmapped (2-3%). The aligned and indexed read files were visualized using Integrative Genomic Viewer (IGV)[270]. Great coherence was shown for the read densities (or coverage) within and between the sample conditions **(Figure 3.4)**.

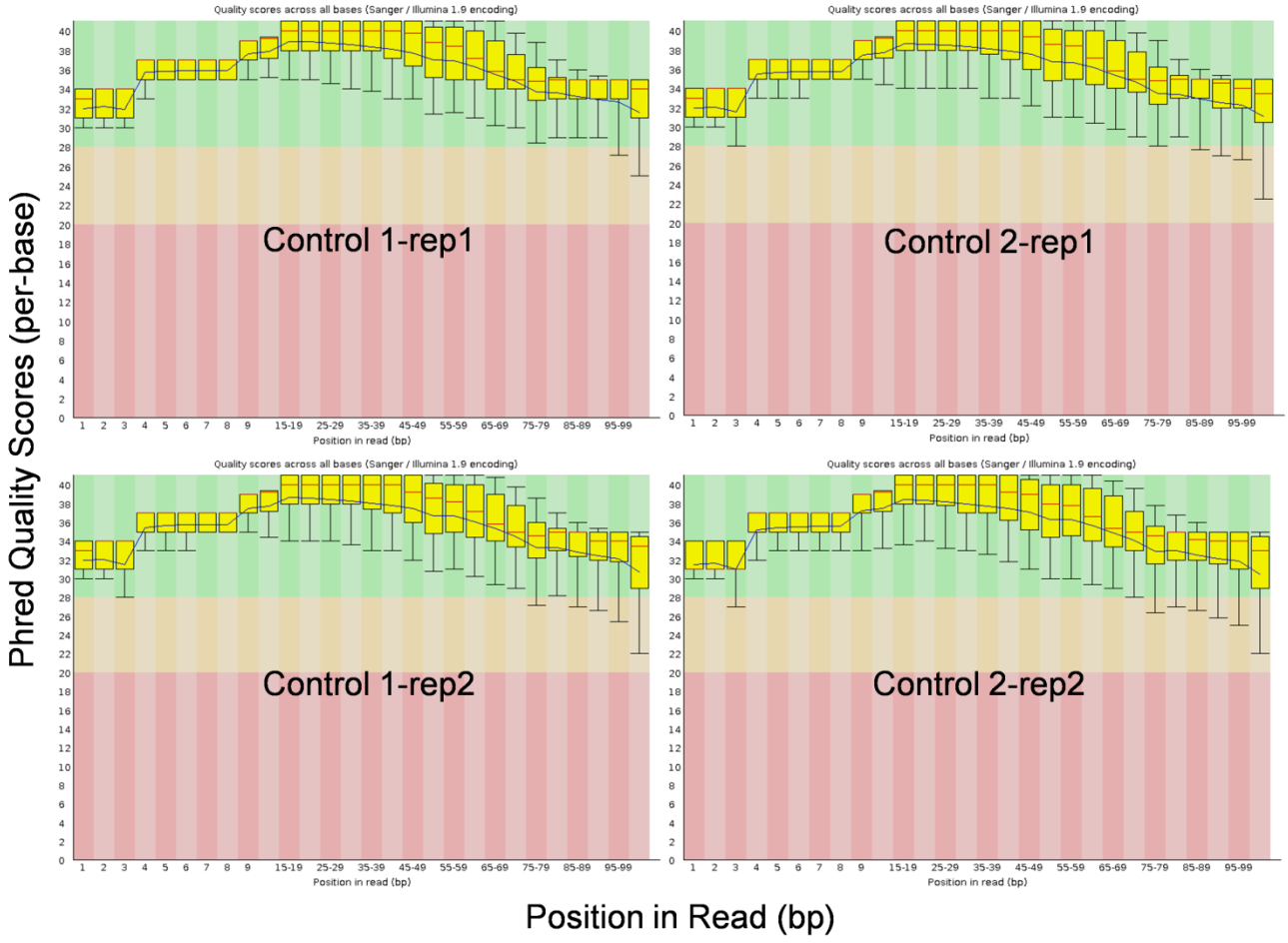


Figure 3.2: Box-plots from FastQC quality check for controls RNA-Seq samples.

The analysis is performed on two control samples with two replicates per sample. Control samples and replicates are shown as Control 1-rep1, Control 1-rep2 and Control 2-rep1, Control 2-rep2. Along *x-axis* and *y-axis* read per base positions and their *Phred* quality scores are plotted, respectively. Mean of the quality scores is shown with '*blue'* line (showing trend in read quality with an initial rise, then remains constant and finally decline towards the end of the read) and Median is shown with '*red'* line in all samples.

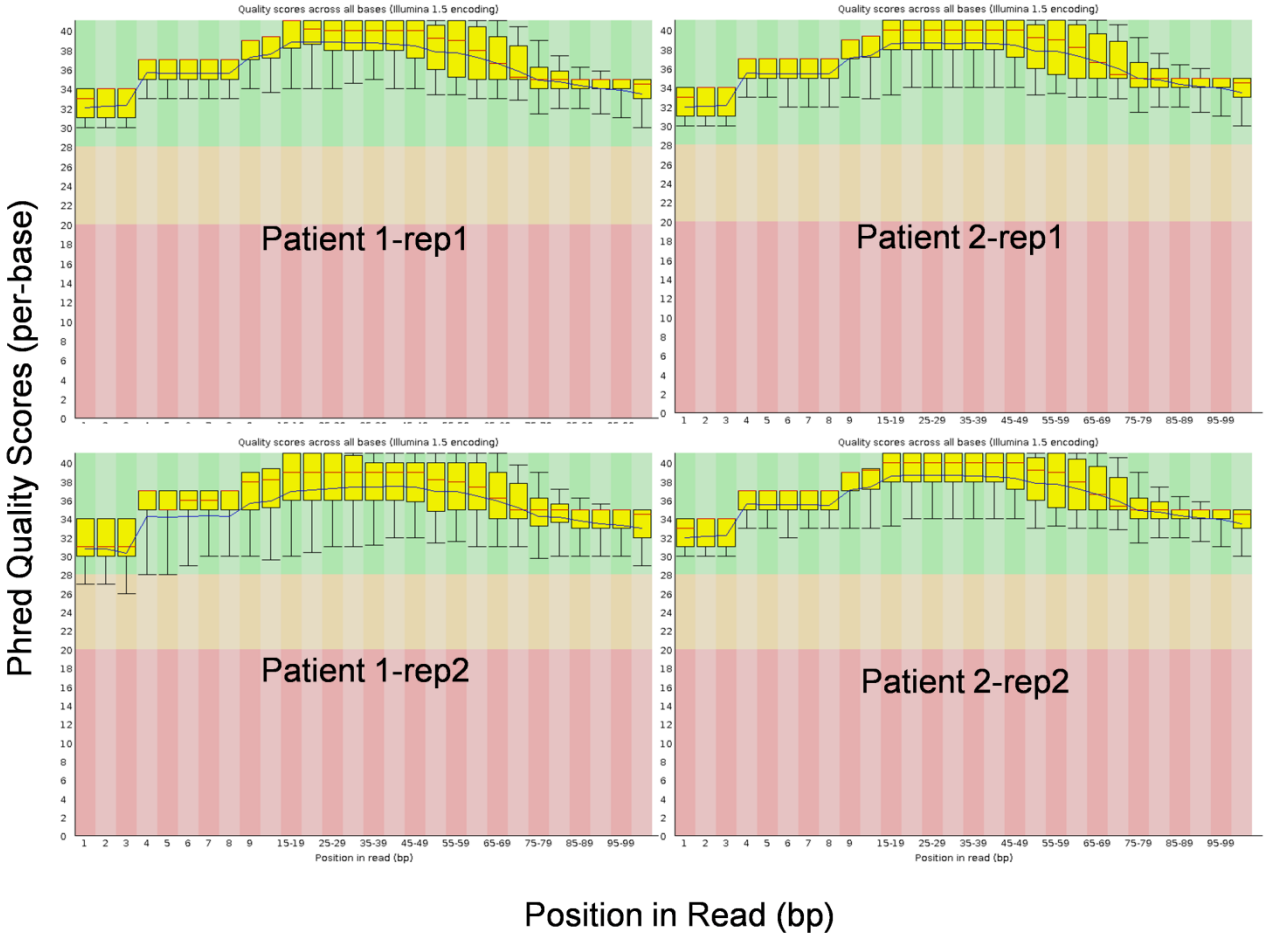


Figure 3.3: Box-plots from FastQC quality check for SMA-patient RNA-Seq samples.

The analysis is performed on two SMA-patient samples with two replicates per sample. SMA-patient samples and replicates are shown as Patient 1-rep1, Patient 1-rep2 and SMA-Patient 2 are shown as Patient 2-rep1, Patient 2-rep2. Along *x-axis* and *y-axis* read per base positions and their *Phred* quality scores are plotted, respectively. Mean value for the quality scores is shown with *'blue'* line and Median is shown with *'red'* lines in all the samples.

Table 3.1: The read alignments from RNA-Seq samples using STAR aligner.

Column 1 reports the sample condition and biological replicate; column 2 contains tags; in column 3 Total Paired-End (PE) reads per sample are given; column 4 contains the total number of uniquely mapped reads onto the reference genome; column 5 gives the percentage of uniquely mapped reads, column 6 gives the total number of multi-mapped reads with their percentages in column 7.

RNA-Seq Sample	Tag	Total PE Reads	Uniquely Mapped Reads	Uniquely Mapped Reads (%)	Multi-Mapping Reads	Multi-Mapping Reads (%)
Control 1-rep1	C11	139,791,014	130,816,448	93.58	6,044,955	4.32
Control 1-rep2	C12	209,121,716	192,292,558	91.95	11,266,224	5.39
Control 2-rep1	C21	205,784,853	192,250,393	93.42	9,307,034	4.52
Control 2-rep2	C22	166,029,806	155,794,899	93.84	7,291,950	4.39
Patient1-rep1	P11	156,662,164	146,786,589	93.71	6,469,136	4.13
Patient1-rep2	P12	193,040,114	178,406,512	92.42	9,178,705	4.75
Patient2-rep1	P21	159,771,568	145,378,464	90.99	9,188,251	5.75
Patient2-rep2	P22	168,518,554	152,992,352	90.79	10,886,377	6.46

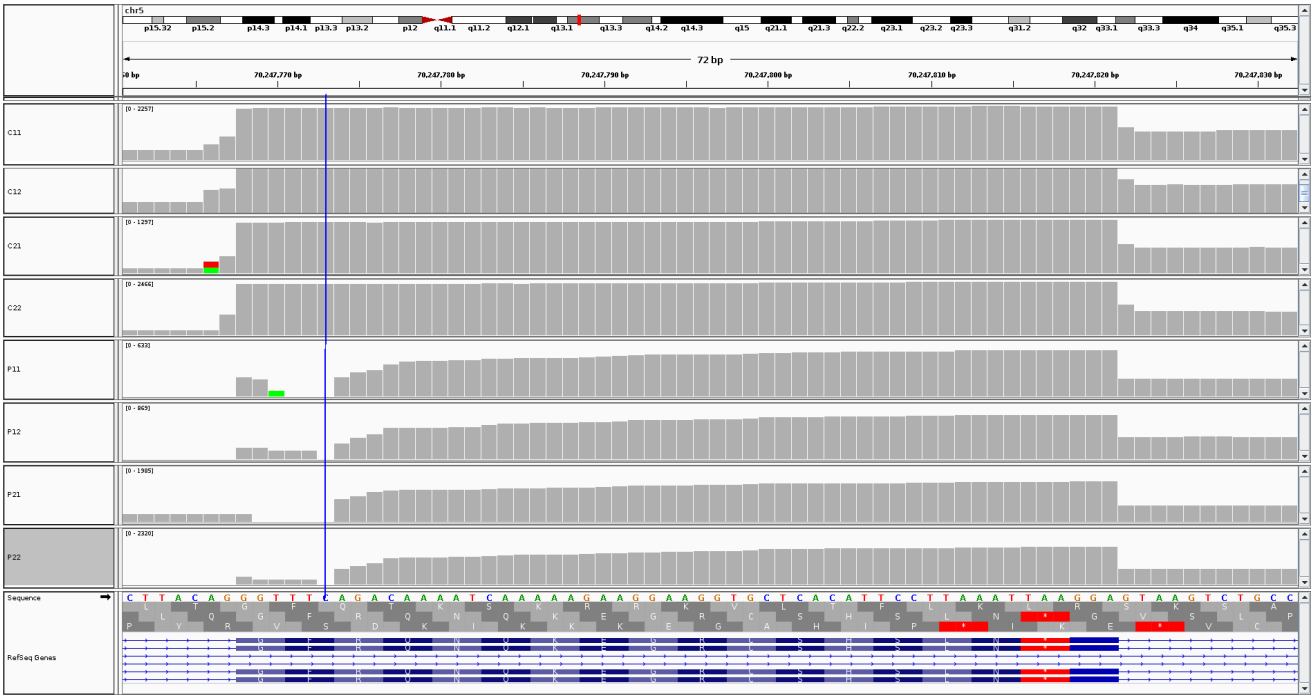


Figure 3.4: Visualization of read coverage with Integrative Genomics Viewer (IGV) at per-base resolution.
The read coverage is shown for position 6 of exon-7 in SMN1 gene (Cytosine or C; marked with a *'blue'* line) within SMA-patients and controls with their biological replicates. The IGV tracks labeled with C11, C12; C21, C22 represent Control 1 and Control 2, respectively; the tracks labeled with P11, P12; P21, P22 represent SMA-patient 1 and SMA-patient 2, respectively.

3.1.3 Similarities and Dissimilarities in the Expression Profiles of Two Different Biological Conditions

High-throughput RNA-Seq samples from 2 SMA patients and 2 healthy controls, with 2 biological replicates each sample, were processed. To investigate the potential relationship in terms of overall expression levels within samples (biological replicates) and between samples (two different conditions, i.e. patients and controls), we performed the hierarchical clustering by calculating the Euclidean distances between gene and isoform expression levels in all samples. All 8 samples were found to be clustered in pairs with their own biological replicates as expected (**Figure 3.5;** P11-P12, P21-P22 and C11-C12, C21-C22) which indicated the similarity in their expression profiles. Control samples were clustered together that indicated similar expression profiles in both controls and within their replicates. Moreover, this cluster was far away from SMA-patient-1 which suggested very different expression profiles between two different conditions. The control samples cluster was less distant from SMA-patient-2 cluster

(**Figure 3.5**; cluster containing, P21 and P22), albeit certainly distinct. Clusters containing SMA-patient-1 (P11 and P12) and SMA-patient-2 (P21 and P22) samples were not clustered together that suggested the presence of heterogeneity in SMA-patients.

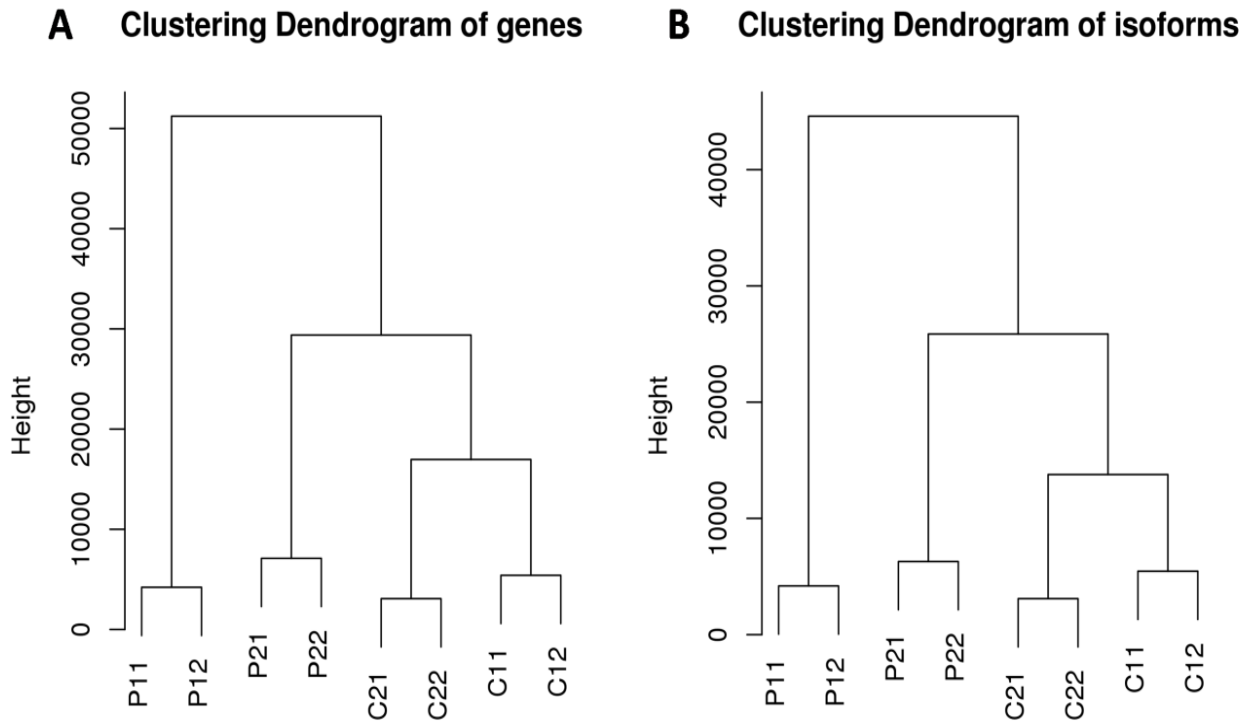


Figure 3.5: The hierarchical clustering of gene and isoform expression levels in SMA-patients and healthy controls and their biological replicates.

A Clustering dendrogram for Genes expression levels obtained by calculating the Euclidean distances between SMA-patients and controls using "ward.D" method, **B** Clustering dendrogram for Isoforms expression levels obtained by calculating the Euclidean distances between SMA-patients and controls using "ward.D" method. SMA-patient-1 replicates correspond to P11, P12; SMA-patient-2 replicates correspond to P21, P22; Control-1 replicates correspond to C11, C12; Control-2 replicates correspond to C21, C22. Along the *y-axis* "Height" represents the distances in expression levels (**A** for Genes, and **B** for Isoforms) of analyzed samples.

3.1.4 Correlation in Fold Change Values of Significantly Differentially Expressed Genes and Isoforms using Pipeline-I and Pipeline-II

We were interested in comparing two splice-aware aligners, namely TopHat2 and STAR. Given the fact that, Tophat2 and Cufflinks tools together have established a gold-standard for analyzing RNA-Seq data, therefore we wanted to investigate if we change the aligner then how this will impact the cufflinks results, will they be still coherent or not? In doing so, we designed

two workflows: Pipeline-I and Pipeline-II (See Materials and methods: Study-A). To evaluate their performance, we compared the outcomes of downstream analysis steps in both pipelines. First, we compared the *Log_2FC* values by computing the Pearson correlation (r) coefficient for the significantly differentially expressed genes and isoforms (adjusted_p-value or qvalue < 0.01) resulted from Cuffdiff2 analysis using both pipelines. We obtained a very strong correlation between both pipelines with r = 0.98 for significantly differentially expressed genes (**Figure 3.6 A**) and r = 0.99 for significantly differentially expressed isoforms (**Figure 3.6 B**). This indicated that both aligners were equally robust and accurate. STAR aligner was having very high speed over TopHat2. Given that our RNA-Seq samples were highly deeply sequenced which took much longer in the alignment step with TopHat2 while very less with STAR without losing accuracy. Therefore, we opted for STAR over TopHat2.

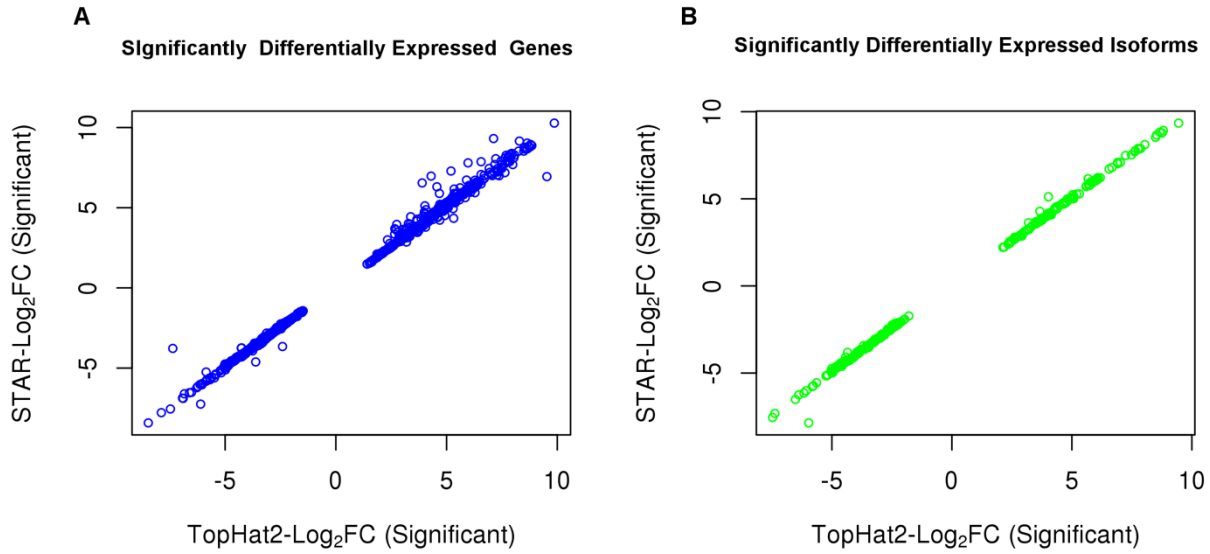


Figure 3.6: Scatter plots showing correlation between the log_2 fold change (*Log_2FC*) values for significantly differentially expressed genes and isoforms obtained from cuffdiff2 using Pipeline-I and Pipeline-II.

A Scatter plot representing *Log_2FC* values (represented with *'blue'* circles) for significantly differentially expressed genes (adjusted_p-value < 0.01) obtained from cuffdiff2 tool applied on Tophat2 (along the *x-axis*) and STAR alignments (along the *y-axis*)., **B** Scatter plot representing *Log_2FC* values (represented with *'green'* circles) for significantly differentially expressed isoforms (adjusted_p-value < 0.01) obtained from cuffdiff2 tool applied on Tophat2 (along the *x-axis*) and STAR alignments (along the *y-axis*).

3.1.5 Correlation between two pipelines within fold change values of significantly Differentially-Used Exons

The identification of accurate splice-site junctions (both known and novel) during read alignment onto the reference genome is critical for identifying relative changes in the expression at individual exon-level between two different conditions. We aligned the reads using two aligners: TopHat2 and STAR in order to see the impact of different aligners on the results of DEXSeq analysis in pipeline-I and pipeline-II. We compared DEXSeq results obtained from both pipelines. In doing so, we selected the significant Differentially-Used Exons with qvalue < 0.01 and computed the Pearson correlation coefficient for their Log_2FC values. We achieved a good correlation between them with r = 0.94 (**Figure 3.7**). Consequently, this has confirmed that both aligners have equivalent accuracy in splice-site alignment from RNA-Seq data. We preferred STAR for its ultra-high speed over Tophat2.

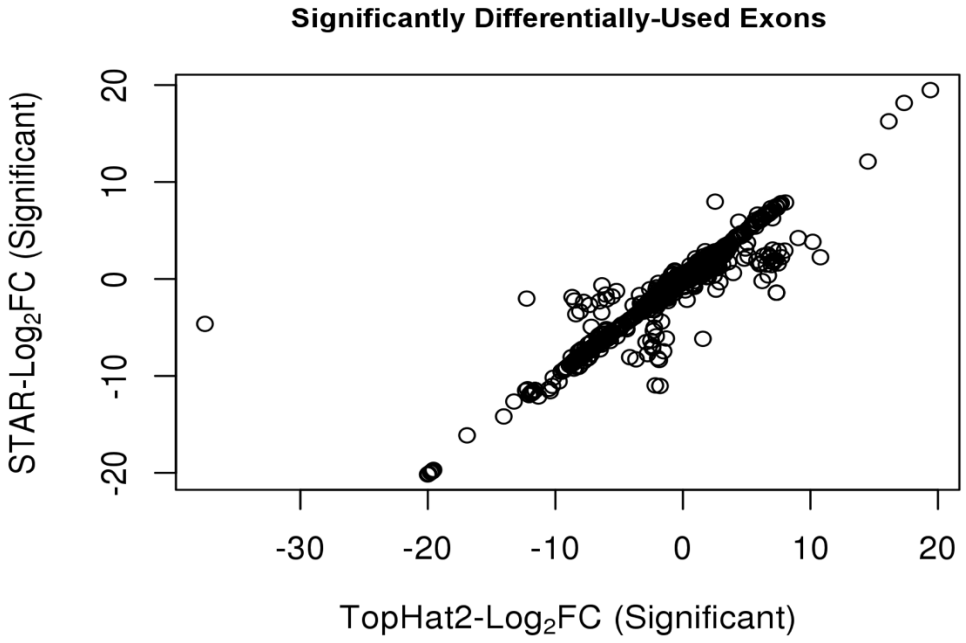


Figure 3.7: Scatter plot showing correlation between *Log2FC* values for significantly Differentially –Used Exons obtained from DEXSeq tool using pipeline-I and pipeline-II.

Scatter plot representing Log_2FC values (represented with *'black'* circles) for significantly Differentially–Used Exons (adjusted_p-value < 0.01) obtained from DEXSeq tool applied on Tophat2 (along the *x-axis*) and STAR alignments (along the *y-axis*).

3.1.6 Simulation of RNA-Seq Reads

To evaluate the performance of applied pipeline, we run simulations study on the basis of estimated parameters obtained from the real RNA-Seq data. We applied RSEM tool to quantify transcript abundances from real-RNA-Seq data and allowed the RSEM-simulator to simulate PE reads learned from the real data-based estimated parameters. Simulated PE reads were aligned by STAR aligner (**see Materials and Methods, Figure 2.3**) and transcript expressions were quantified using Cuffquant-Cuffnorm programs in pipeline-II. In order to determine the correlation between RSEM estimated transcript expressions and expressions estimated from simulated reads, we computed the Pearson correlation coefficient between them. We obtained correlation coefficient, r = 0.94, that has validated the used analysis pipeline-II **(Figure 3.8)**.

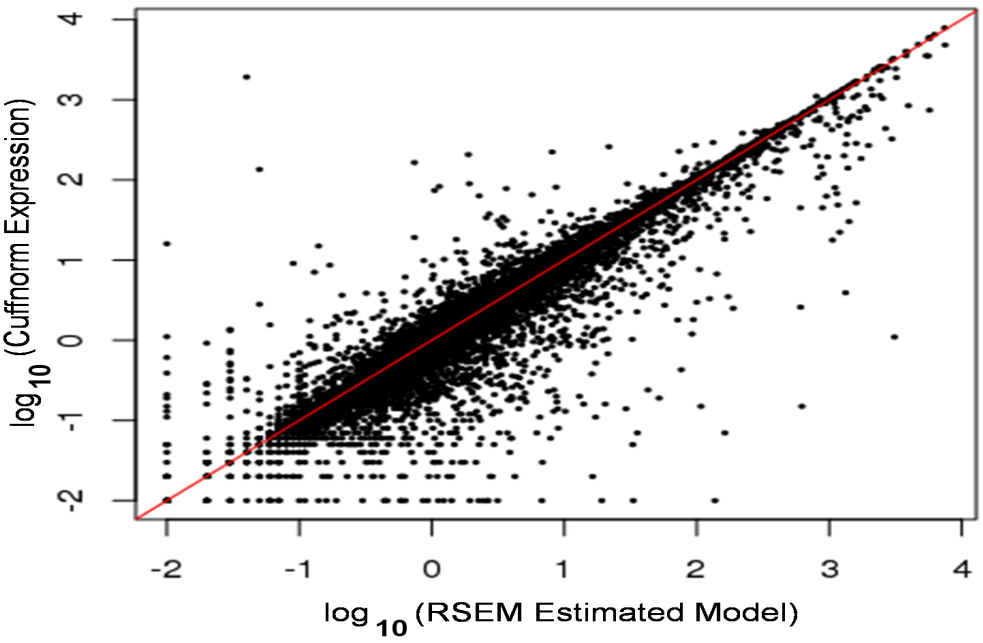


Figure 3.8: A scatter plot showing correlation between expression levels obtained from read simulations analyzed by pipeline-II and RSEM quantified isoform expressions.

The scatter plot (*'red'* line indicates the *goodness-of-fit*) between the expression estimations obtained from simulated PE-reads applying Cuffquant-Cuffnorm (along the *y-axis,* in logarithmic scale) and RESM based transcript expression quantifications (along the *x-axis,* in logarithmic scale).

3.1.7 Functional Annotation Analysis: Differentially Expressed Genes (DEGs)

We performed the gene ontology (GO) enrichment analysis on significant DEGs using DAVID[268] software, to find the correlation between the set of DEGs and key processes implicated in the pathobiology of SMA. SMA is primarily a disease of degeneration of MNs in the spinal cord and

cause muscular atrophy. In this study, we expected to observe good correlation between the significant DEGs and neuromuscular system related processes. Indeed, the enrichment analysis provided well supported results in terms of significantly enriched GO terms related to neuromuscular systemic processes.

In the Biological Process (BP) category we found 32 overrepresented GO terms, including "regulation of neurotransmitter transport" (GO:0051588), "neuron development" (GO:0048666), "neuron differentiation" (GO:0030182), "axon guidance" (GO:0007411), "neuron projection development" (GO:0031175), "synaptic transmission" (GO:0007268), "muscle contraction" (GO:0006936), "BMP signaling pathway" (GO:0030509), "transmission of nerve impulse" (GO:0019226), "axonogenesis" (GO:0007409) and others. These processes highlight the essential mechanisms related to the neuromuscular system development and maintenance. Additionally, such processes are known to be greatly hampered in SMA pathogenesis, due to insufficiency of SMN protein. Moreover, an axon specific isoform of SMN, namely axonal-SMN encoding a-SMN protein has been identified. This axon specific SMN protein is localized within the axonal structures of MNs in the spinal cord and supports axonogenesis[271].

In the Cellular Component (CC) category, we found 10 significantly enriched GO-terms: "synapse" (GO:0045202), "sarcomere" (GO:0030017), "actin cytoskeleton" (GO:0015629), "striated muscle thin filament" (GO:0005865), "synapse part" (GO:0044456), "myosin complex" (GO:0016459), "presynaptic membrane" (GO:0042734), "synaptosome" (GO:0019717), "contractile fiber part" (GO:0044449) and "focal adhesion" (GO:0005925). The identified terms pinpoint the involvement of cellular regions specific for building neuromuscular junctions and muscle contraction that are impaired in SMA patients.

Further, in the Molecular Function (MF) category, we found 7 enriched GO-terms: "actin filament binding" (GO:0051015), "cytoskeletal protein binding" (GO:0008092), "calcium ion binding" (GO:0005509), "glycosaminoglycan binding" (GO:0005539)," gated channel activity" (GO:0022836), "sequence-specific DNA binding" (GO:0043565) and "phosphatidylcholine-sterol O-acyltransferase activator activity" (GO:0060228). These functions show the binding of essential regulatory factors necessary for carrying out important neuromuscular processes.

Notably, in the biological pathway enrichment category from functional annotation analysis, we found key pathways pivotal for the neuromuscular processes such as "Muscle contraction" and "Synaptic Transmission". We identified "agrPathway:Agrin in Postsynaptic Differentiation"

pathway which though remained below significance level (p-value = 0.089) and several genes involved in this pathway are known to play essential roles in the development, maintenance and maturation of neuromuscular junctions[272] such as actin, alpha 1, skeletal muscle (ACTA1), epidermal growth factor receptor (EGFR), integrin, beta 1 (ITGB1), jun oncogene (JUN), laminin, alpha 3 (LAMA3), neuregulin 2 (NRG2) and paxillin (PXN).

All overrepresented GO-terms mentioned above have broadly supported the neuromuscular system specific processes. The p-values for these 'significantly enriched GO-terms' were in the range of 1e-3 and 1e-13. **Figure 3.9** illustrates all overrepresented GO-terms with their statistics.

To visualize and interpret the gene-set enrichment results obtained from DAVID tool, we used Enrichment map plugin in cytoscape software. It is a network based method to effectively explore the enrichment analysis results (**Figure 3.10**). In this map, we presented few enriched GO terms resulted from DAVID results which were related with each other through biological processes, molecular functions and cellular components. Such as in the central cluster of the enrichment map in **Figure 3.10** GO terms mainly represent processes related to neuron development connected with cell morphogenesis involved in neuron differentiation, neuron projection development, axonogenesis and axon guidance. All of these processes are vital for the neuron and its projection development (axon), which guide the signal transmission through the long axons until the synapse establishment with muscle fibers that exhibit muscles contraction and movement[133,142,273]. Second cluster has covered the specific processes involved in the release of neurotransmitters (with the help of calcium ion binding) and to establish the synapse formation between presynaptic membrane and postsynaptic membrane on the muscle fibers[143,274]. Specific GO terms related to the regulation of protein transport and protein localization across axons and axon terminals are essential for the development and sustenance of MNs. Lung development is linked with respiratory system development which is majorly impacted in SMA patients due to poor muscle tone[92,100,275]. These results indicated that our identified set of DEGs were enriched in terms which are linked with neuromuscular processes that gets disrupted in SMA pathology.

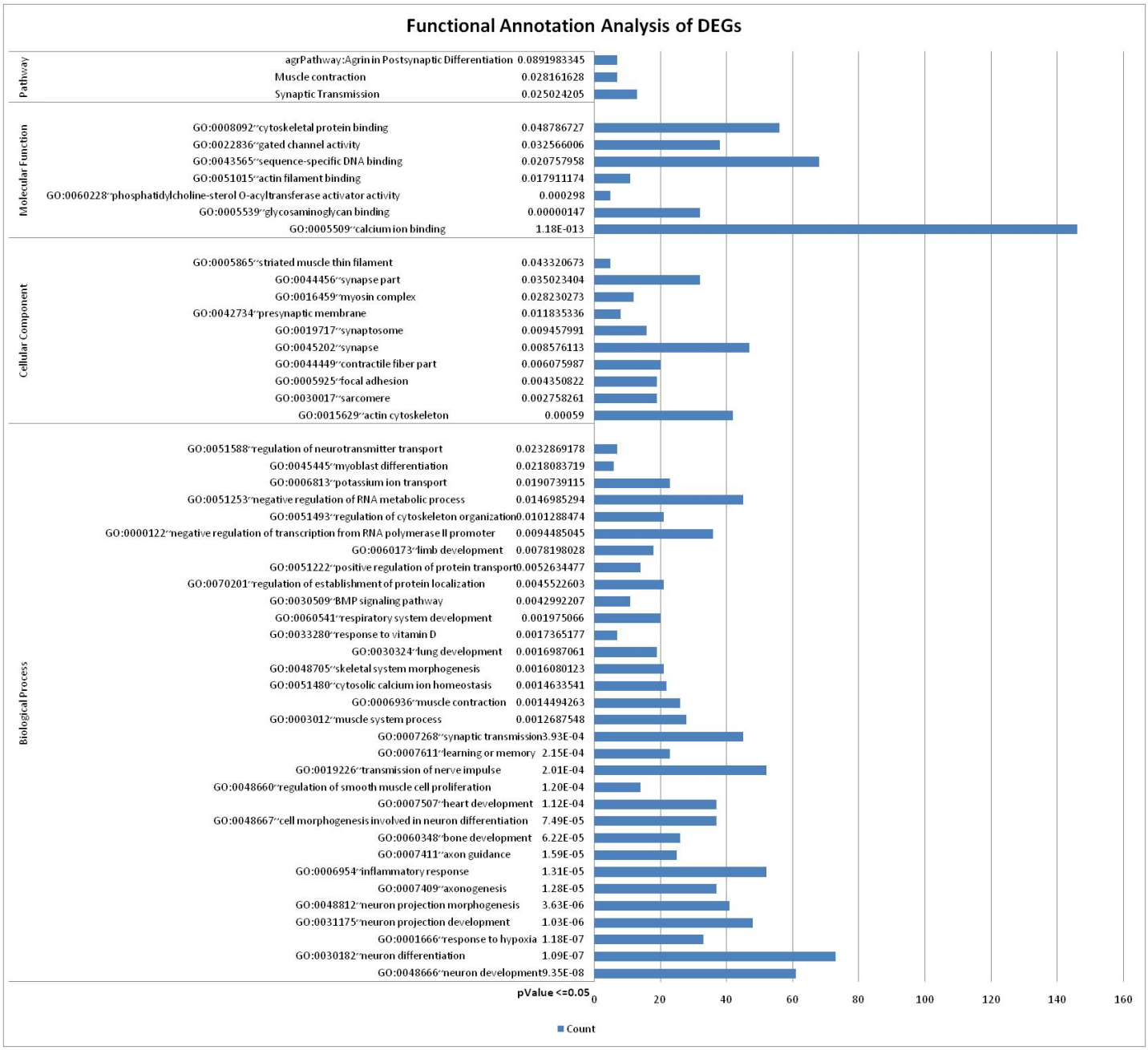


Figure 3.9: The functional annotation analysis of significantly 'DEGS' (qvalue < 0.05).

The enriched GO-terms (Biological Process, Cellular Component and Molecular Function) and pathways have significant p-value <= 0.05, except *agrPAthway/Agrin in Postsynaptic differentiation* pathway that has p-value = 0.089. The horizontal bars represent the "genes count" associated with the particular enriched GO-term or enriched pathway.

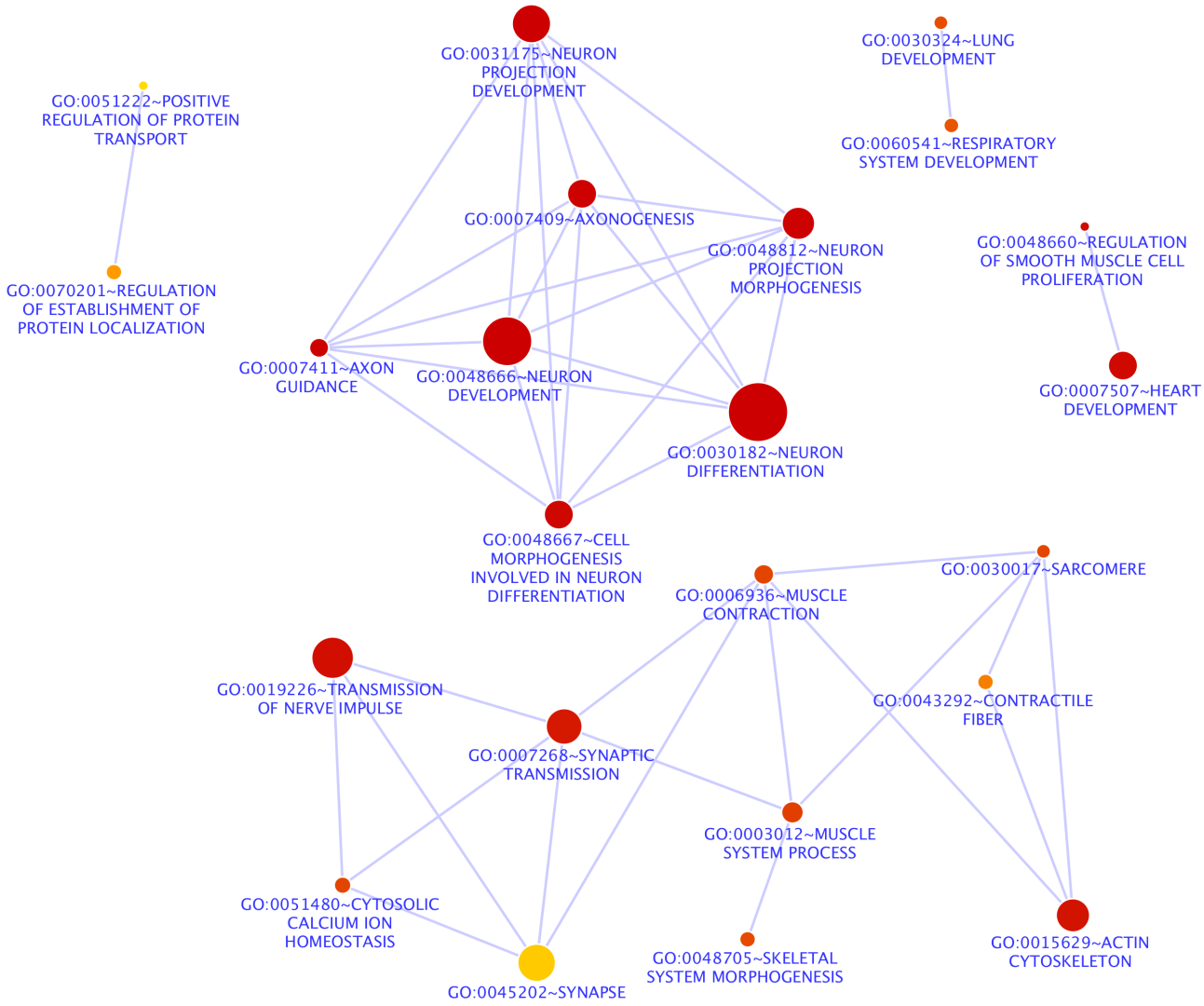


Figure 3.10: Network-based visualization of DEGs enrichment analysis results obtained from DAVID tool using Enrichment map (Cytoscape plugin).

In this map each node is representing the enriched GO term and color of the node indicates the level of GO term significance (i.e. darker color *'deep red'* means higher significance level while lighter color *'yellow'* indicates low significance level, however all the shown GO terms are statistically significant). The size of the node represents the number of genes enriched with each GO term. The edge (undirected *'light blue'* thin lines) between the nodes indicates the relationship between enriched GO terms with each other.

3.1.8 Functional Annotation Analysis: Differentially-Used Alternative Cassette Exons

We identified the relative changes in the expression at the level of individual exons between SMA-patients and healthy controls using DEXSeq[247] tool. The selection was restricted for the internal alternative cassette exons from the list of DEXSeq resulted exons. Further, on the basis of their statistical significances (qvalue < 0.01) we obtained a list of significant Differentially-Used Alternative Cassette Exons (DUACEs). To annotate the genes corresponding to these epurated exons and identify their biological relevance, we performed GO enrichment analysis using DAVID[268] software. From this analysis, we obtained the specific mechanisms which are responsible for overall MNs developmental processes, their maintenance and skeletal muscular system development (through interaction between nerves and muscle fibers).

From DAVID analysis, we identified total 30 significantly overrepresented GO-terms. Out of these, 14 GO-terms were represented in the BP category: ”axon cargo transport” (GO:0008088), “protein localization in organelle” (GO:0033365), “protein import” (GO:0017038), “protein targeting” (GO:0006605), “microtubule-based transport” (GO:0010970), “neuron development” (GO:0048666), “neuron projection morphogenesis” (GO:0048812), “axonogenesis” (GO:0007409), “regulation of transcription from RNA polymerase II promoter” (GO:0006357) and others (**Figure 3.11**). These terms have highlighted essential mechanisms such as axon protein transport in MNs that enhance their survival. SMN protein has been studied to play pivotal activities in such transports and its deficiency pinpoints the known aspects of disturbances in motoneuron vital axon-protein transport in SMA pathology[134,135].

Eleven GO-terms were overrepresented in CC category, including “microtubule cytoskeleton” (GO:0015630),” sarcolemma” (GO:0042383), ”sarcoplasm” (GO:0016528), ”sarcoplasmic reticulum” (GO:0016529), ”microtubule” (GO:0005874), ”calcium channel complex” (GO:0034704), “neuron projection” (GO:0043005), “axon” (GO:0030424), ”dendrite” (GO:0030425), “endoplasmic reticulum membrane” (GO:0005789) and “microtubule associated complex” (GO:0005875). These enriched terms are linked to neuron cell specific compartments (such as axon and dendrites) and striated muscle cell specific regions (such as sarcoplasm, sarcolemma or postsynaptic membrane and sarcoplasmic reticulum). Further, microtubule-based cytoskeleton structures support the RNA and protein transport from the motoneuron cell body (soma) to the axon terminals to establish synapses with the muscle fiber. Given the fact

that the distance between MN soma and axon terminal is very large (~ 1 meter) which makes the mRNA and protein transport highly challenging and to accomplish this task various cellular components and building blocks are engaged. Such as microtubules which plays very important role that also interacts with SMN protein to facilitate the transport of specific mRNAs across axons until the axon terminals. Lack of adequate levels of SMN protein possibly result in the reduction of mRNAs transport that might disrupt the synapse formation between neurons and muscles and cause SMA. These results provide promising insights and purport our data-set to correlate with SMA pathology and specific mechanisms within neuromuscular system. The GO term “endoplasmic reticulum (ER) membrane” was also identified which is consistent with a recent study by Lee L. Rubin et al.[276], that has highlighted the specific rise of ER stress in SMA, causing selective degeneration of motoneurons in SMA pathology[276].

Furthermore, we identified 5 enriched GO-terms in MF category: “microtubule binding” (GO:0008017), ”microtubule motor activity” (GO:0003777), “calcium ion binding” (GO:0005509), “motor activity” (GO:0003774), and “cytoskeletal” (GO:0008092). They primarily represent the protein transport that is mediated by precise binding of the proteins onto the microtubule cytoskeleton structures[273]. The release of calcium ions and their binding has specific roles in the release of neurotransmitters from the presynaptic membrane to the postsynaptic membrane for the signal transmission from neurons to the muscles[277].

We visualized the GO enrichment analysis results using network based map obtained from Enrichment map plugin of Cytoscape (**Figure 3.12**) where the central cluster represented the processes related to the microtubule-based intracellular cargo transport across the axons and axon terminals. Other smaller clusters represented the MNs cellular compartments including axons and dendrites. A cluster with muscle fibers cellular compartments such as sarcoplasmic reticulum, sarcoplasm and sarcolemma.

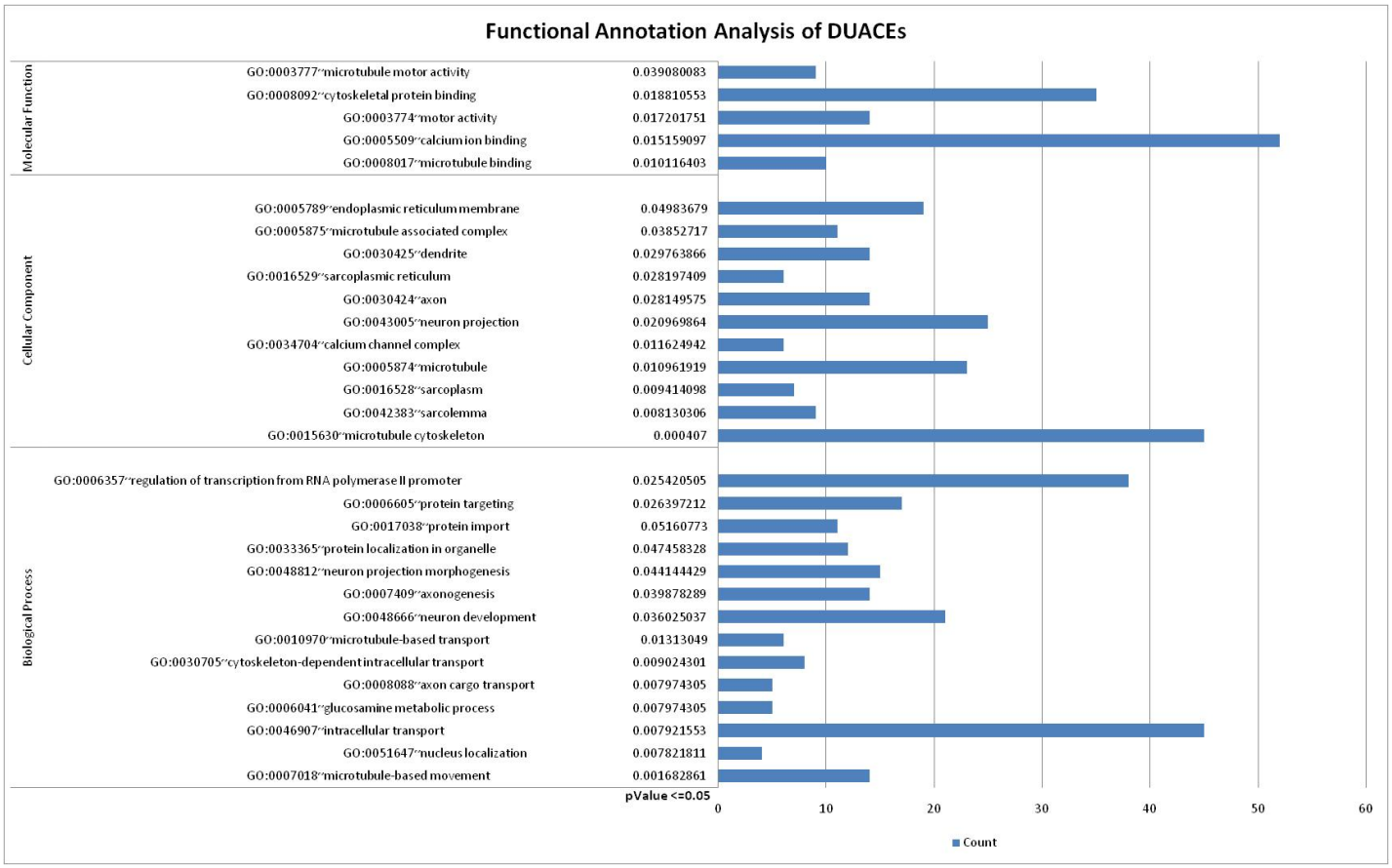


Figure 3.11: The functional annotation analysis of significantly Differentially-Used Alternative Cassette Exons (DUACEs; qvalue < 0.05) corresponding genes.

The enriched GO-terms (Biological Process, Cellular Component and Molecular Function) have significant p-value <= 0.05. The horizontal bars represent the "counts of the genes" associated with the particular enriched GO-term.

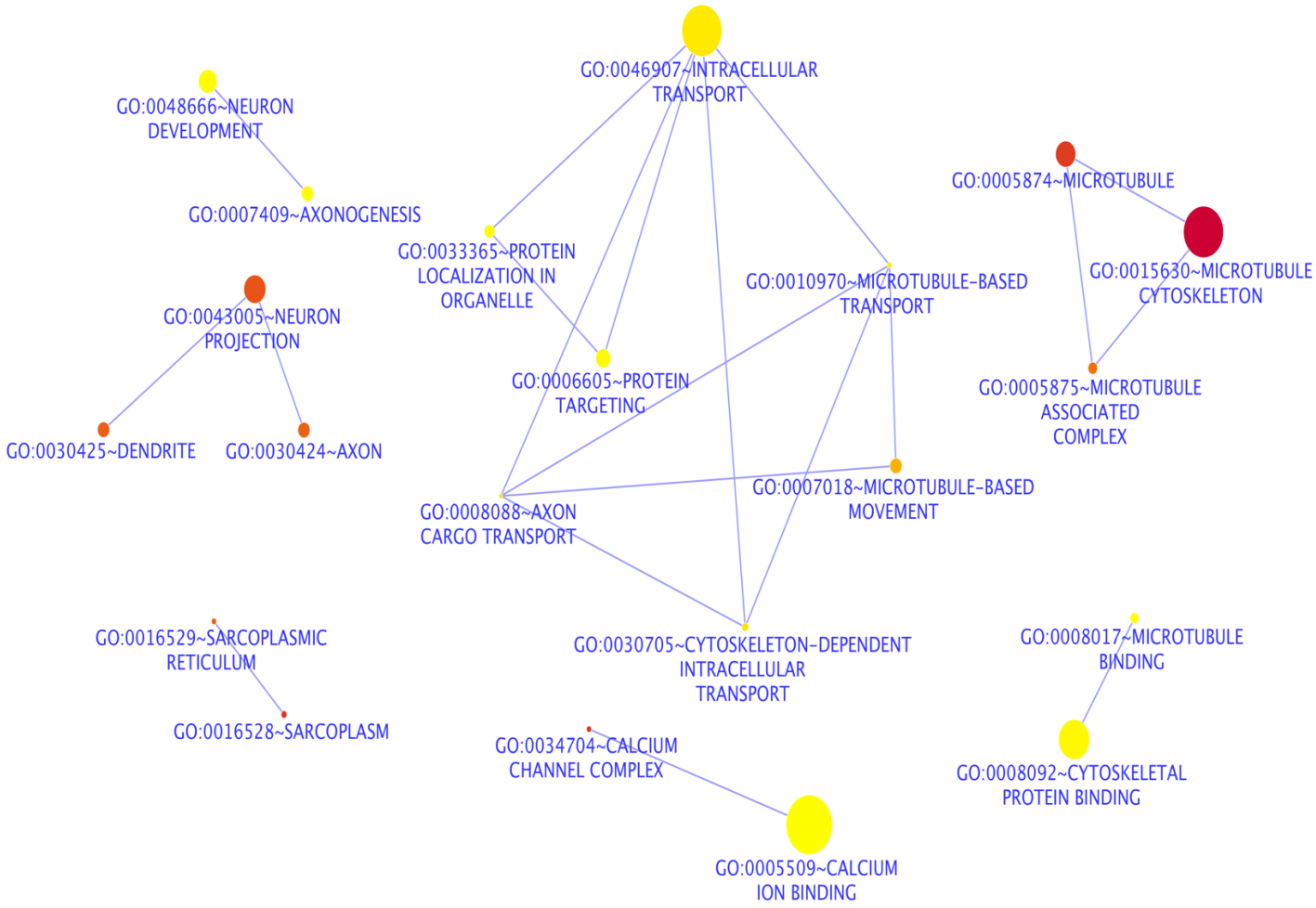


Figure 3.12: Network-based visualization of DUACEs enrichment results obtained from DAVID tool using Enrichment map (a Cytoscape plugin).

The terminology of the graph has been kept same as in **Figure 3.10**.

3.1.9 Identification of Splicing Regulatory Elements and RNA-Binding Proteins

Alternative splicing regulation represents a vital aspect of controlled transcriptional activity within every eukaryotic cell. The main controllers in this process includes cis-acting splicing regulatory elements (SRE; motifs) which acts by recruiting trans-acting splicing regulatory factors. In this study, we aimed to identify SREs located within the exonic sequences as ESEs and ESSs; or within the intronic sequences as ISEs and ISSs. Specific RNA-Binding Proteins (RBPs) bind on these sequence patterns (i.e. SREs or motifs) and regulate exclusion or inclusion of the exon within the transcript. The lower levels of SMN protein in SMA pathology might disrupt the splicing patterns of the specific set of transcripts by disturbing their SREs which makes them

inaccessible for specific RBPs. From the DEXSeq analysis, total 45,483 DEXSeq-ACEs were obtained (**Figure 3.13**), which were divided into two parts: Significant DUACEs and Non-significant DEXSeq-ACEs. In Significant DUACEs list, we have 859 exons out of which 368 DUACEs were having enhanced expression and 482 DUACEs were having silenced expression in SMA-patients with respect to healthy controls. In non-significant DEXSeq-ACEs list, we have 5,421 exons which were used as control in the analysis.

We applied sequence level analysis on the identified set of enhanced and silenced DUACEs along with their flanking introns using MEMERIS[262]. Total 30 motifs were discovered from 6 sequence files, wherein 3 files were from enhanced DUACEs, their upstream introns and downstream introns, 3 files were from silenced DUACEs, with their upstream introns and downstream introns. We filtered-out those motifs which were having more than 60% similarity in their PSSMs. As a result, 6 unique motifs were selected, having the following consensus: *"CCTCG", "TCATC", "AAGAA", "ATTTT", "CCACC",* and *"GAAAA".* Further, each sequence file was scanned to compute the occurrences of each identified motif using FIMO[265] tool by means of position specific frequency matrix. Overrepresentation of each motif was compared in all sequence files containing enhanced DUACEs, silenced DUACEs and control set of exons with their upstream intron and downstream intron sequences, in pairwise manner.

For instance, occurrences of motif-1 were computed within enhanced-upstream introns sequence file and occurrences of motif-1 in silenced-upstream introns sequence file and then their occurrences were compared with each other. Similarly, all possible pairwise comparisons were performed for each of the 6 motifs occurrences. The statistical significances were obtained for the differences of motif occurrences using Wilcoxon non-parametric statistical test. In all comparisons, we obtained certain regions with statistically significant differences in their occurrences in different condition which suggest overrepresentation of certain motif in one region over the another. These results are given in **Table 3.2**. Further, we computed the average occurrences of each motif per sequence length in all sequence files which were represented with bar plots in **Figure 3.14**.

In order to identify RBPs for the identified set of unique motifs we used TOMTOM tool which compare query PSSMs (identified motifs) with target PSSMs (Ray database[267]). TOMTOM has identified 22 similar PSSMs associated with 22 known RBPs in the target database. The resulted RBPs are given in **Table 3.3**.

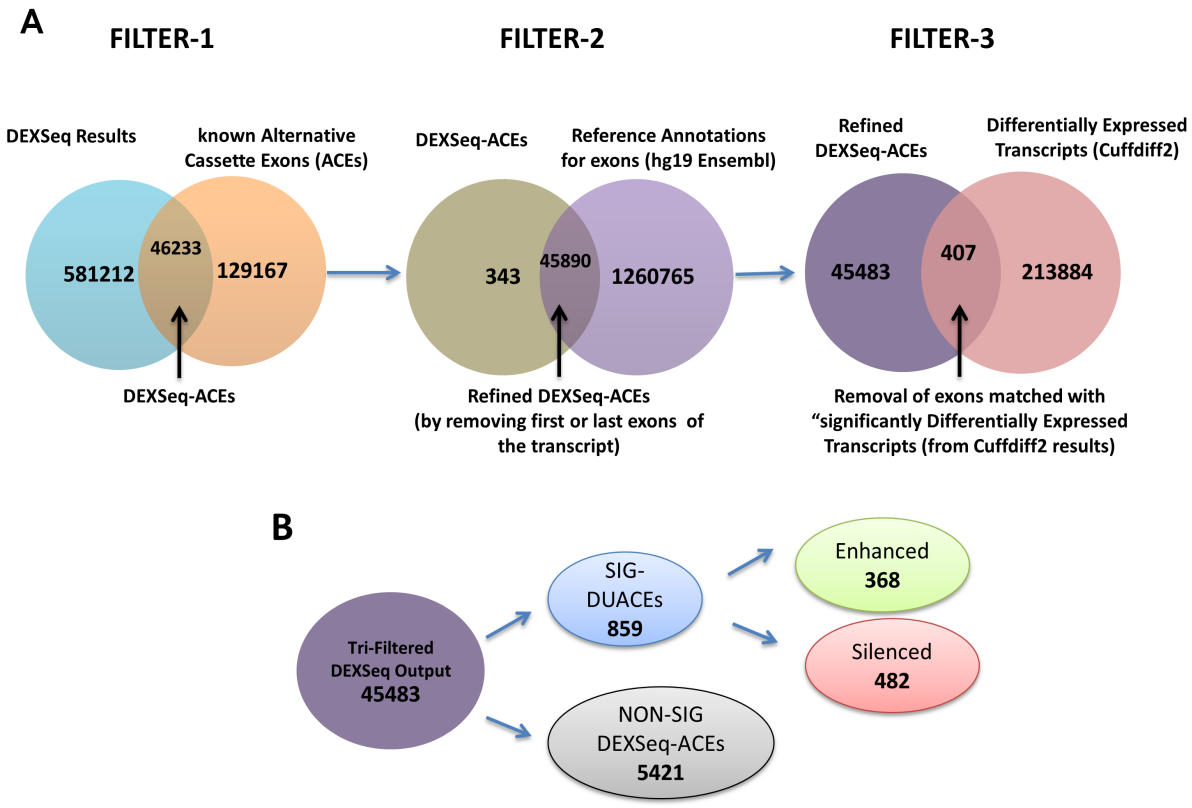


Figure 3.13: Filtration of DEXSeq results to obtain refined set of exons for SREs and RBPs identification. See main text for details.

Table 3.2: Pairwise overrepresentation comparisons for each motif within Enhanced, Silenced and Control sequence databases.

The significant differences among the analyzed regions for motif enrichment analysis are shown with the p-value<0.05. In this table, "#Motif" column represents the corresponding number of the identified motif; "Motif Region" column provides the information for the region where the motif was identified and the expression level of searched region. The region can be either 'Exon', 'Upstream Intron' or 'Downstream Intron' and expression level can be either 'Enhanced' or 'Silenced'. "Seq_Type" column represents the region where the given motif (column 1) was scanned. "Motif Occurrences Comparison" column provides the pairwise comparisons in their occurrences. For instance, the first row of the table gives the comparison between "ctrl-vs-silenced" that means "Control DUACEs Upstream Intron sequence file" and "Silenced Exon Upstream Intron sequence file" were scanned for "#Motif 1" and its occurrences were compared. For instance, their overrepresentation differences were statistically significant with p-value = 0.033 (Wilcoxon statistical test).

#Motif	Motif Region	Seq_Type	Motif Occurrences Comparison	Wilcoxon test (p-value)
1	Silenced DUACE Upstream Intron	upstream	ctrl-vs-silenced	0.032952787
3	Silenced DUACE	upstream	ctrl-vs-silenced	0.001653789
5	Enhanced DUACE Upstream Intron	upstream	ctrl-vs-silenced	0.023272979
2	Silenced DUACE Upstream Intron	upstream	enhanced-vs-silenced	0.035851368
4	Silenced DUACE	upstream	enhanced-vs-silenced	0.028702714
1	Silenced DUACE Upstream Intron	exon	ctrl-vs-enhanced	0.001364972
1	Silenced DUACE Upstream Intron	exon	ctrl-vs-silenced	0.006138266
2	Silenced DUACE Upstream Intron	exon	ctrl-vs-silenced	0.015140815
4	Silenced DUACE	exon	ctrl-vs-silenced	0.029128923
2	Silenced DUACE Upstream Intron	exon	enhanced-vs-silenced	0.037523772
1	Silenced DUACE Upstream Intron	downstream	ctrl-vs-silenced	0.022302108
4	Silenced DUACE	downstream	ctrl-vs-silenced	0.036428094
4	Silenced DUACE	downstream	enhanced-vs-silenced	0.023538089

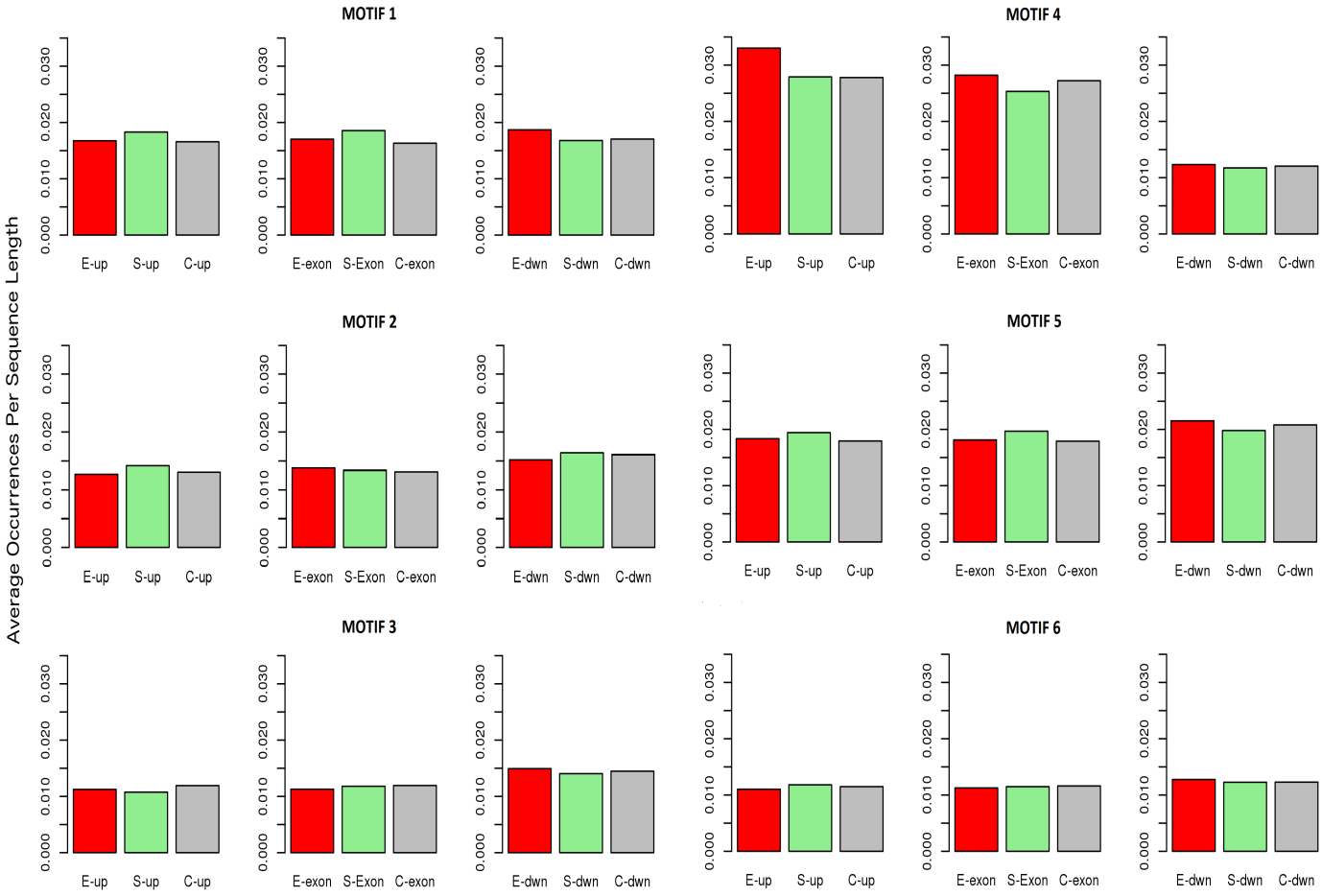


Figure 3.14: Bar plot representing the average occurrences of each motif per sequence length in all sequence files.

Enhanced Upstream Intron (E-up), Silenced Upstream Intron (S-up), and Control Upstream Intron (C-up); Enhanced DUACE (E-exon), Silenced DUACE (S-exon), and Control (C-exon); Enhanced Downstream Intron (E-dwn), Silenced Downstream Intron (S-dwn), and Control Downstream Intron (C-dwn) are represented along the x-*axis* of each bar with average occurrence of each motif per sequence length along the *y-axis.*

Table 3.3: Identified set of RNA Binding Proteins (RBPs).

In this table, the first column "#Motif" gives the information of identified motif; the second column "Query ID" represents the region and condition where the motif was identified; third column 'Target ID' represents the RBP ID in the target database; fourth column 'RBP' gives the Protein ID; fifth column gives the protein full names; and sixth column provides the significances of the identified RBPs with p-value <0.05.

#Motif	Query ID	Target ID	RBP	Protein Full Name	p-value
1	Silenced DUACE Upstream Intron	RNCMPT00045	PPRC1	Peroxisome proliferator-activated receptor gamma coactivator-related protein 1	0.0465279
2	Silenced DUACE Upstream Intron	RNCMPT00083	YBX1	Y-box-binding protein 1	0.0507951
3	Silenced DUACE	RNCMPT00019	SRSF10	Serine/arginine-rich splicing factor 10	0.0281991
3	Silenced DUACE	RNCMPT00043	PABPC4	Polyadenylate-binding protein cytoplasmic 4	0.0421641
4	Silenced DUACE	RNCMPT00025	HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2	0.0026912
4	Silenced DUACE	RNCMPT00167	HNRNPCL1	heterogeneous nuclear ribonucleoprotein C-like 1	0.0038673
4	Silenced DUACE	RNCMPT00032	HuR	ELAV-like protein 1 or human antigen R	0.0069628
4	Silenced DUACE	RNCMPT00012	CPEB2	Cytoplasmic polyadenylation element-binding protein 4	0.0196742
4	Silenced DUACE	RNCMPT00269	PTBP1	Polypyrimidine tract-binding protein 1	0.0209912
4	Silenced DUACE	RNCMPT00165	TIA1	TIA1 cytotoxic granule-associated RNA binding protein	0.0247981
4	Silenced DUACE	RNCMPT00159	RALY	RALY heterogeneous nuclear ribonucleoprotein	0.0330535
4	Silenced DUACE	RNCMPT00079	U2AF2	U2 small nuclear RNA auxiliary factor 2	0.0350781
4	Silenced DUACE	RNCMPT00158	CPEB4	Cytoplasmic polyadenylation element-binding protein 4	0.0372912
5	Enhanced DUACE Upstream Intron	RNCMPT00026	HNRNPK	Heterogeneous nuclear ribonucleoprotein K	0.0491327
6	Enhanced DUACE	RNCMPT00153	PABPC3	Polyadenylate-binding protein cytoplasmic 3	0.0031335
6	Enhanced DUACE	RNCMPT00171	PABPC5	Polyadenylate-binding protein cytoplasmic 5	0.0031335
6	Enhanced DUACE	RNCMPT00064	SART3	Squamous cell carcinoma antigen recognized by T-cells 3	0.0048223
6	Enhanced DUACE	RNCMPT00155	PABPC1	Polyadenylate-binding protein cytoplasmic 1	0.0048223

6	Enhanced DUACE	RNCMPT00157	PABPN1	Polyadenylate-binding protein nuclear 1	0.0205277
6	Enhanced DUACE	RNCMPT00043	PABPC4	Polyadenylate-binding protein cytoplasmic 4	0.0249685
6	Enhanced DUACE	RNCMPT00019	SRSF10	Serine/arginine-rich splicing factor 10	0.0281991
6	Enhanced DUACE	RNCMPT00169	KHDRBS1	KH domain-containing, RNA-binding, signal transduction-associated protein 1	0.0431269

3.2 Development of Computational Model to Estimate Transcript Expression

3.2.1 Total RNA-Seq Strand Specific Data

To examine the robustness and accuracy of our model, we downloaded the publicly available strand specific total RNA-Seq data with paired-end (PE) reads from NCBI Sequence Read Archive repository under accession SRP043027 (http://www.ncbi.nlm.nih.gov/Traces/sra/). We selected one deeply sequenced sample with four biological replicates (Strand specific U2OS_RZSS_R1-R4) under this experiment[278]. The library preparation was performed by using Illumina TruSeq kit and sequenced on Illumina HiSeq 2000 platform with 100bp read length in PE mode. Total sequenced library sizes range from 210-260 Million PE reads.

We assessed the quality of raw data with FastQC program[218] and trimmed the poor quality bases using TrimGalore application (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The trimmed reads were re-assessed for final quality check before the subsequent analysis. In TrimGalore, we used the following parameters: "--paired", "--phred33", "--fastqc", "--illumina", "--gzip". Where, "--paired" option gives the sequencing mode information that is PE in our data. "--phred33" encodes for the Phred quality score from Illumina sequencer where the quality scores range from 33-126 in ACSII characters. "--fastqc" commands for an automatic quality check after the completion of trimming procedure. "--illumina" gives the information for the used sequencer (such as Illumina for our samples). "--gzip" option performs the compression of the processed files.

The pre-processed samples were aligned using STAR aligner[236]. The uniquely mapped reads were processed while multi-loci mapped reads were discarded. Total 165-185 Million PE reads (~ 80%) were uniquely mapped onto the reference genome (hg19).

Further, we performed the transcriptome assembly using the aligned samples in order to build the sample-specific transcript assemblies. The Cufflinks tool[240] was applied on each of the four aligned samples and obtained four sample-specific transcript assemblies, stored in Gene Transfer Format (GTF). Consequently, the individual sample transcript assemblies were merged into single master transcriptome GTF assembly by applying Cuffmerge tool.

3.2.2 Analysis of Gene Loci with Our Model

We randomly selected a list of gene loci from the gene annotation file with transcript and exon coordinate information. The designed model can be applied on complete set of the genes present in an annotation file we provided, or individual gene can be analyzed by specifying its gene locus. We analyzed several gene loci and here we have presented the results of 8 genes with variable complexity in their number of exons, number of possible isoform paths, TSS (either same or different across plausible isoforms) and AS events **(Table 3.4)**. If PE reads alignment (in a fragment) extend across the given original genomic coordinates, then in those cases fragments were counted within the extended genomic interval. For each gene locus **(in Table 3.4**), its original genomic coordinates were provided to the exon-intron junction finding program (in the model) which provide its corresponding extended genomic interval and then calculate the number of supporting junctions from the aligned fragments (i.e. PE reads aligning across splice site junctions) in the data. For instance, in PAQR8 gene locus, the beginning position within the original genomic coordinates (Chr6:52226219-52272575) got extended (downstream for the genes with plus strand) from Chr6:52226219 to Chr6:52142695 while the ending position remain same. In BPGM gene locus, ending position within the original genomic coordinates (Chr7:134331560-134594387) got extended from Chr7:134364565 to Chr7: 134594387 (upstream) while beginning position remain unchanged. In SNRPC gene locus, both beginning and ending positons within the original genomic coordinate (Chr6:34725183-34741571) got extended (Chr6:34423975-34842151) and in FAM46C gene locus, the original coordinates (Chr1:118148556-118170994) remain unchanged.

Table 3.4: A list of processed gene loci.

Each gene locus is given with its corresponding Gene Id; Gene Locus; Extended Interval; Number of plausible isoform paths and Strand.

#	Gene Id	Gene Locus	Extended Interval	Number of Plausible Isoform paths	Strand
1	FAM46C	Chr1:118148556-118170994	Chr1:118148556-118170994	1	Plus
2	PAQR8	Chr6:52226219-52272575	Chr6:52142695-52272575	3	Plus
3	BPGM	Chr7:134331560-134364565	Chr7:134331560-134594387	4	Plus

4	SNRPC	Chr6:34725183-34741571	Chr6:34423975-34842151	4	Plus
5	LSM6	Chr4:147096837-147121152	Chr4:147096837-147121152	7	Plus
6	FOSB	Chr19:45971253-45978437	Chr19:45771161-45982034	11	Plus
7	TEP1	Chr14:20833826-20881588	Chr14:20808516-20881588	11	Minus
8	POT1	Chr7:124462440-124570067	Chr7:124462440-124570212	20	Minus

3.2.3 Isoforms Re-construction for Each Gene Locus

For a given gene locus, the isoform paths were re-constructed using both exon-intron junction information that was supported by the total RNA-Seq data and the information obtained from Cufflinks annotated isoform paths. The purpose of combining both information was to consider only those paths which were annotated and also well-supported by the real data. For instance, if some path was annotated within Cufflinks transcriptome annotations but missing in total RNA-Seq data then, in that case we chose to exclude that isoform path from our list. By doing so, the isoform path selection became more efficient through the reduction of unrealistic isoform paths as compared to keeping all the possible exon-intron combinations. The presented gene loci **(Table 3.4)** with varying number of isoform paths that range from 1-20 helped us to examine the efficiency of our model (with increased complexity in gene loci) in the expression estimation of each isoform path within all gene loci. In simplistic case like FAM46C gene, mature and nascent probability profiles were computed for one isoform path while in the complex cases like POT1 these profiles were computed for 20 isoform paths at per base resolution.

3.2.4 Fragment Length Distribution and Transcript Profile Generation

As the first step, we determined the distribution of relative frequencies for all the fragment lengths from total RNA-Seq data, wherein the maximum length of the fragment was 500bp and most frequently obtained fragments were of ~100bp length (**Figure 3.15**). By using relative frequencies from the fragment length distribution profile along with annotation based isoform structures, the mature and nascent probability profiles for each isoform in our dataset were computed. Wherein mature transcript probability profile describes the probability of observing a fragment of any length at each position along the transcript within a given gene locus. In the mature transcripts, these probabilities got affected by the distance from the transcript borders

which can be seen in the example of SNRPC gene (**Figure 3.17C**) where fragment length probability values progressively increase from the transcript start position until the given length of the fragment has reached and then it became constant and finally started to decrease progressively towards the end of the transcript (**for details see Materials and Methods section 2.2.6.1**). The nascent transcript probability profiles give the equal probability of observing any of the partial transcripts in the data library. In the nascent transcript probability profile, the read coverage across exons in partial transcripts got accumulated (estimated read coverages as peaks indicated in **Figure 3.17D**) while each unspliced intron along the partial transcript acquire unique probability profile (estimated read coverages as saw-tooth shapes indicated in **Figure 3.17D**) due to co-transcriptional splicing phenomenon.

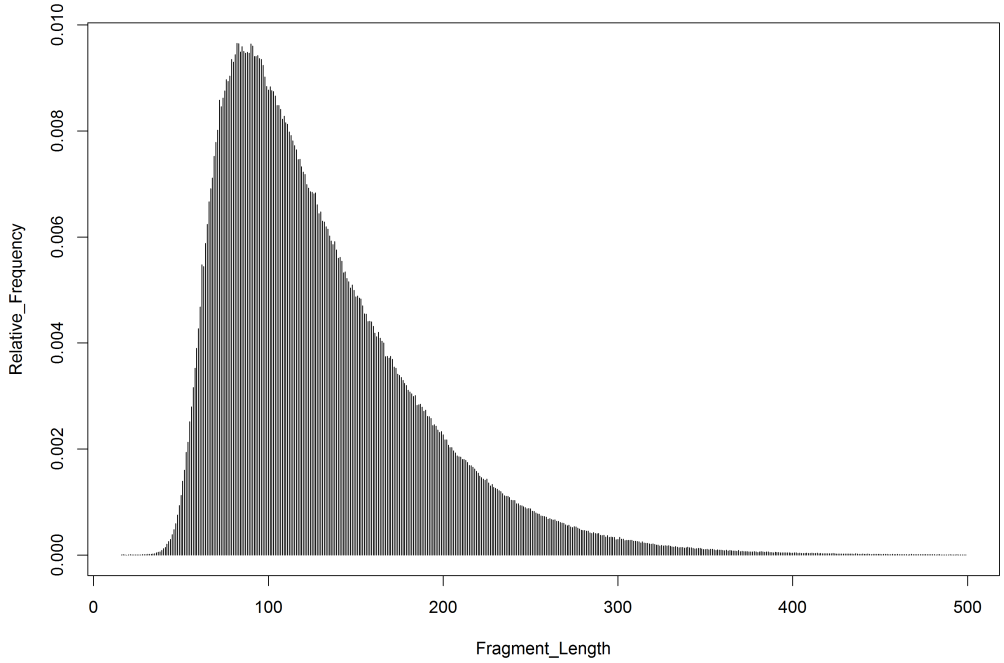


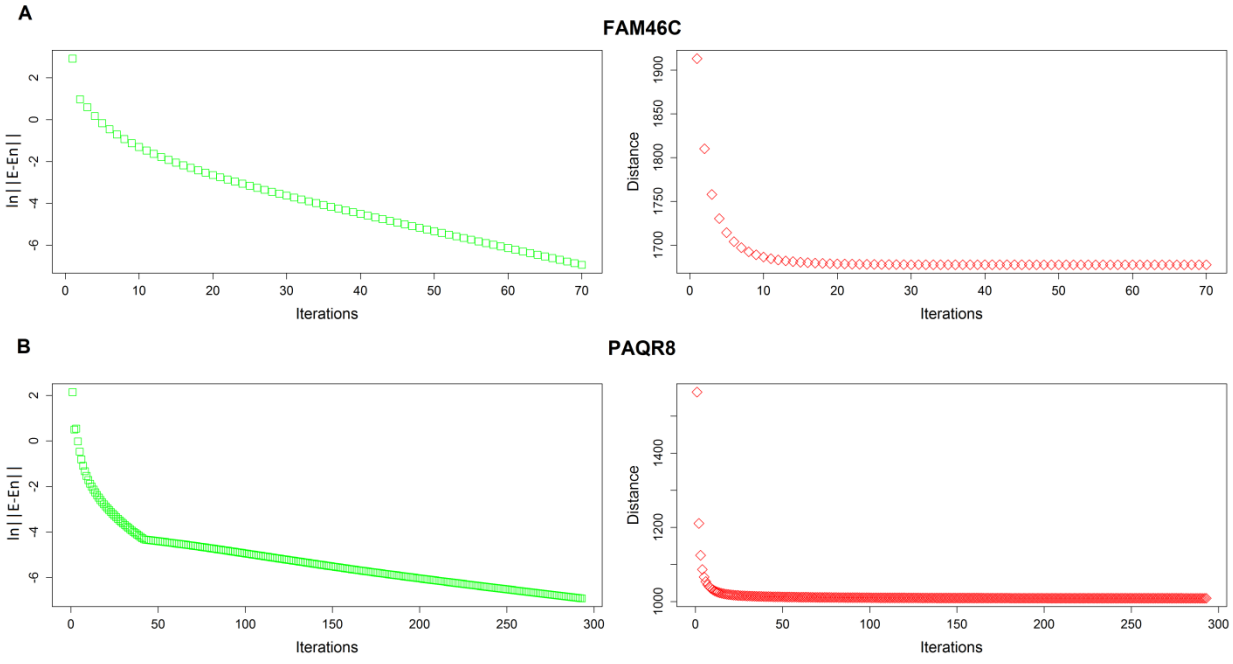
Figure 3.15: A fragment length distribution from total RNA-Seq data.

Along the *x-axis* fragment lengths are plotted with their relative frequencies on the *y-axis* from total RNA-Seq data.

3.2.5 Expression Estimation

We applied our model to each gene locus in our dataset and estimated the expression of each transcript (including mature and nascent transcription contributions) which best approximated the observed expression in the data. Iteration stopping threshold was set to 1 * 10-3. Convergence was reached for all loci in a reasonable CPU time. In **Figure 3.16** the convergence criterion (expressed in logarithmic scale) and the overall distance between our

coverage estimate and the measured one are reported for each iteration step. The number of iterations before convergence varied between 70 and 2500. The overall residual distance resulted always stable after relatively few iteration and well before convergence is reached. In **Table 3.5** we reported the results for gene loci we analyzed as well as the cufflinks FPKMs for comparison. The modeled expression estimates for mature and nascent transcripts with the observed read coverage have been represented for SNRPC gene in **Figure 3.17**. Wherein total estimated read coverage give the contributions of both mature and nascent transcription together which can be directly compared to the observed read coverage. In the result shown in **Figure 3.17**, total estimated read coverage was approximately equivalent to the observed read coverage as represented in the plots **Figure 3.17A and Figure 3.17B.** The contribution of mature transcription to the total estimated read coverage has been shown in **Figure 3.17C** likewise the contribution of nascent transcription to the total estimated read coverage has been shown in **Figure 3.17D**.



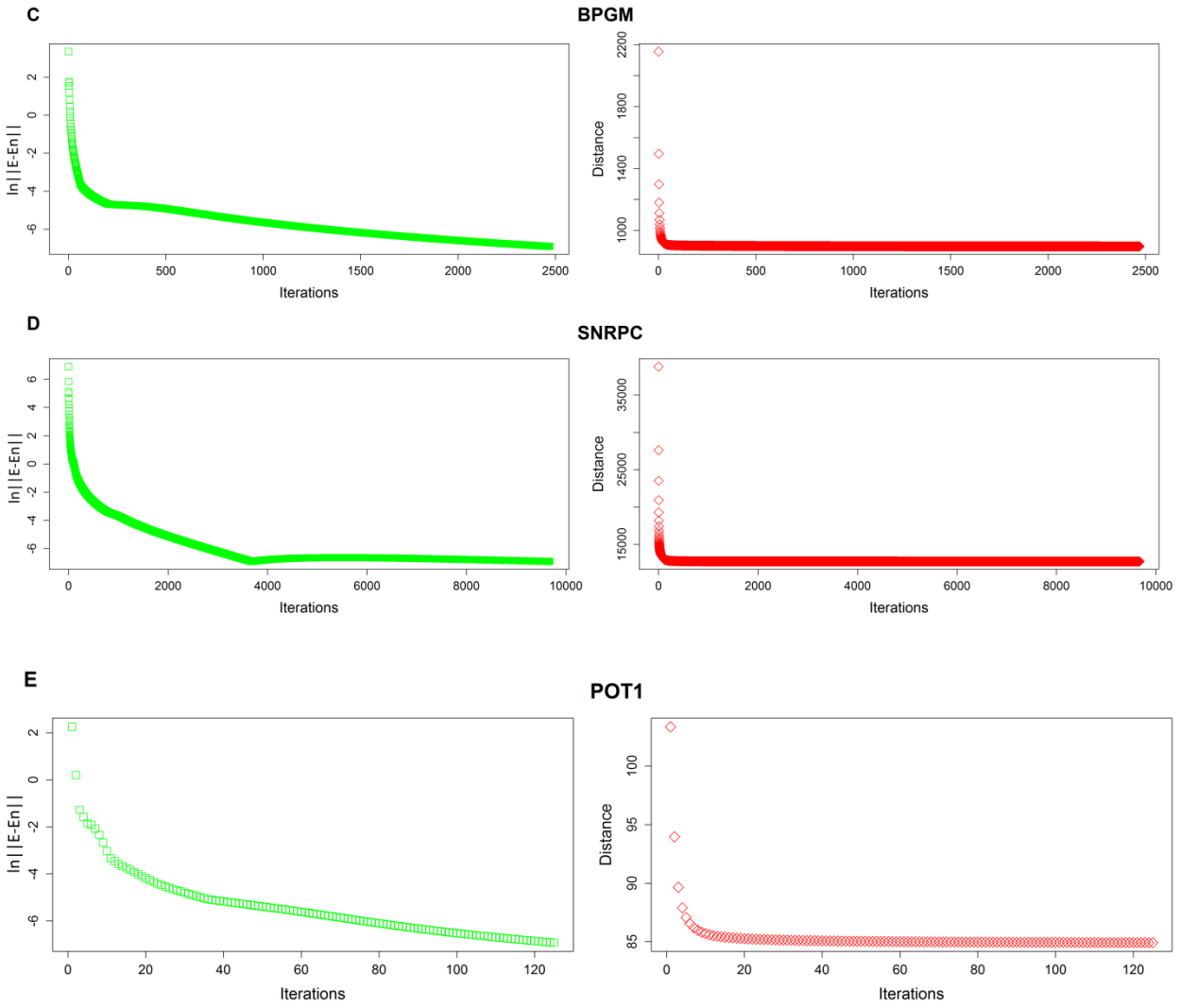


Figure 3.16: Scatter plots representing the speed of convergence and distance between the initial and minimal value obtained at successful convergence.

From **A** through **E** five genes with gene symbols *FAM46C, PAQR8, BPGM, SNRPC* and *POT1* have been represented wherein convergence speed is measured at each iteration step by computing the norm of difference between the two consecutive estimated expression values plotted as an iteration error (*y-axis*) for each iteration (*x-axis*) (shown with empty squares in *'green'* color on the left-hand side plots for every gene). The distance or residual distance indicates the difference between the estimated expression value and observed expression value at each iteration step until the minimal distance is achieved upon successful convergence. This distance has been plotted along the *y-axis* with each iteration step on the *x-axis* (shown with empty diamonds in '*red*' on the right-hand side plots for every gene).

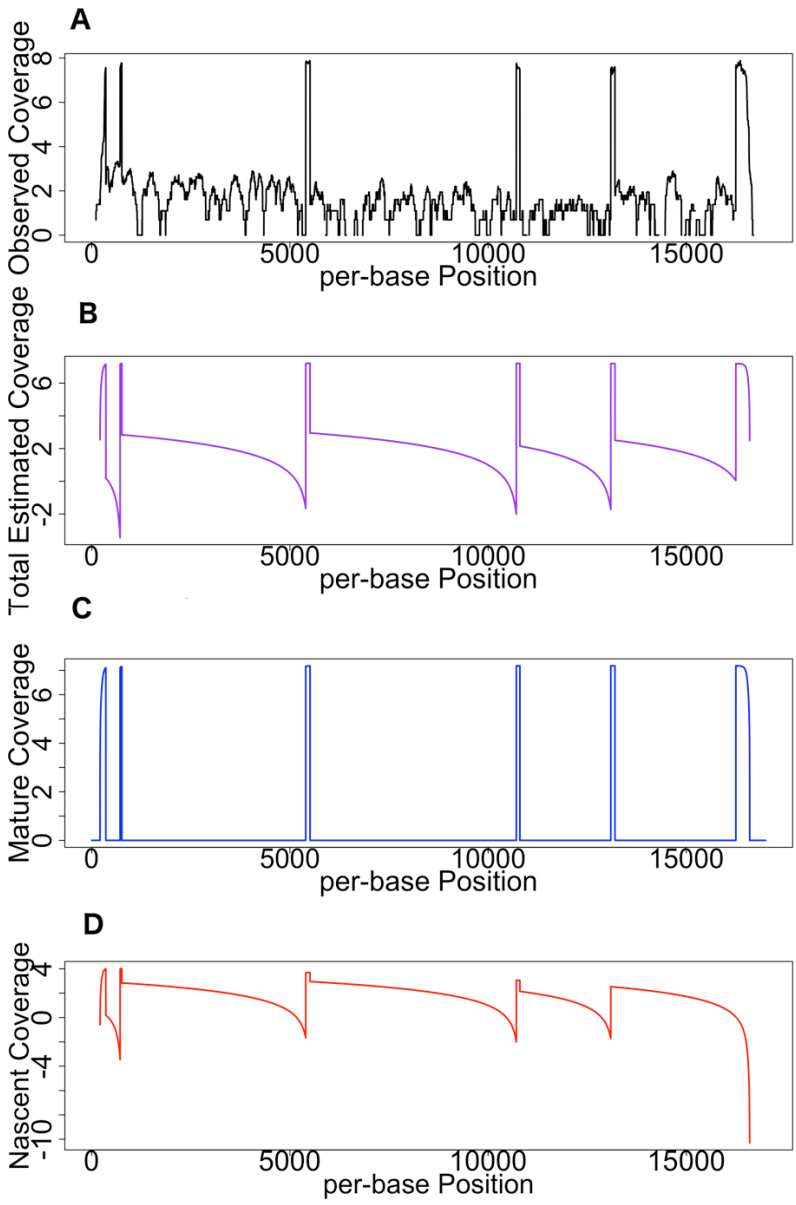


Figure 3.17: Modeled read coverages and observed read coverages for SNRPC gene locus from total RNA-Seq data.

A Plot represents the observed read coverage (total RNA-Seq data, *y-axis* in logarithmic scale) at per-base resolution (*x-axis*) for SNRPC gene locus. **B** Plot represents the total estimated read coverage (*y-axis* in logarithmic scale) contributed from mature and nascent RNA transcription at per-base resolution (*x-axis*). **C** Plot represents the contribution from estimated expression of mature RNA transcription ('Mature' on *y-axis* in logarithmic scale) at per-base resolution (*x-axis*). **D** Plot represents the contribution from estimated nascent RNA transcription ('Nascent' on *y-axis* in logarithmic scale) at per-base resolution (*x-axis*).

3.2.6 Generation of BED files and Visualization in the Genome Browser

After the successful application of the model to each gene locus, we generated BED files. These files were visualized using Integrative Genome Viewer (IGV)[270] (**Figure 3.18**). The re-constructed isoform paths were consistent with the annotated isoforms in the known gene annotations. In the results, every modeled transcript was annotated with three estimated measures: Mature RNA estimate (M), Nascent RNA estimate (N) and *α*, where *α* is the ratio between mature and nascent RNA estimates which give the account for accumulated mature transcript expression. As the contribution of mature RNA transcription is dependent upon the number of nascent transcripts (which make our computational model non-linear), the estimate of mature transcript was given by the multiplication of nascent transcript estimate and *α.*

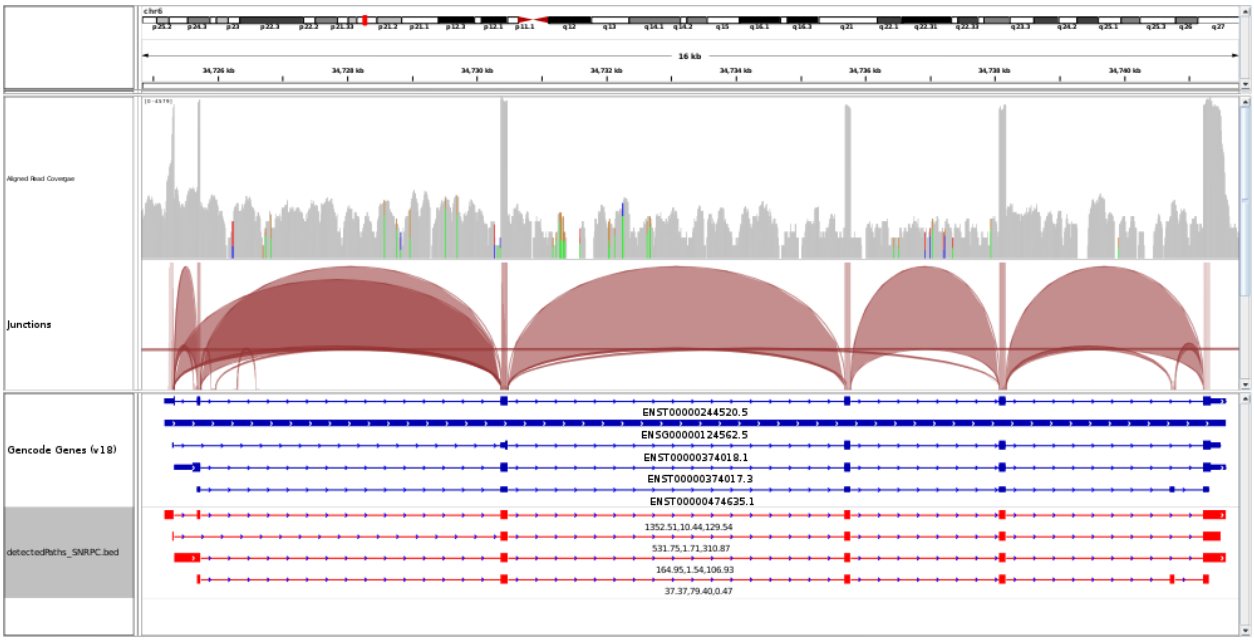


Figure 3.18: Visualization of modeled SNRPC gene locus on IGV.

The SNRPC gene has 4 isoforms as shown in the annotations (exons are represented with '*blue*' thick boxes and introns are represented with '*blue'* thin lines in forward direction). Below each isoform its corresponding ensembl transcript ID has been reported. Observed read coverage track (represented with '*grey'* peaks) have been reported which represents the total aligned read density within the given locus. Splice site junctions are also shown within each exon and intron boundaries (represented with *'brown'* bands), which are supported by read alignments along the splice site junctions in read alignment file. Our modeled isoforms are shown in the detected paths panel (exons are represented as '*red'* thick boxes and introns are represented as ''*red'* thin lines in forward direction). Below each detected isoform, 3 *comma* separated values have been given, wherein first value is for mature RNA estimate (M), second value is for Nascent RNA (N) and third value gives alpha (M/N) in the model. For SNRPC *isoform-1* estimated expression for mature transcript is 1352.51, for nascent transcript is 10.44 with alpha 129.54. In *isoform-2* estimated

expression for mature transcript is 531.75, for nascent transcript is 1.71 with alpha 310.87. In isoform-3 expression for mature transcript is 164.95, for nascent transcript is 1.54 with alpha 106.93. In Isoform-4, expression for mature transcript is 37.37, for nascent transcript is 79.40 with alpha 0.47.

Table 3.5: The expression estimations of the mature (M) and nascent (N) with alpha (a ratio between M and N) in 8 analyzed gene loci with variable number of plausible isoform paths.

#	Gene Symbol and Transcript Id	Gene and Transcript coordinates	Isoform paths	M	N	alpha	Cufflinks FPKM
1	**FAM46C**	**chr1:118148556-118170994**	**1**				
	ENST00000369448.3	chr1:118148556-118170994		3.12E+001	8.36E+000	3.74E+000	1.87382
2	**PAQR8**	**Chr6:52226219-52272575**	**3**				
	ENST00000512121.1	chr6:52226219-52268347		4.90E+000	1.42E-001	3.44E+001	0
	ENST00000442253.2	chr6:52226926-52272575		2.49E+001	3.30E+000	7.54E+000	1.49633
	ENST00000360726.3	chr6:52227244-52272489		1.31E-001	5.03E-002	2.60E+000	0
3	**BPGM**	**Chr7:134331560-134364565**	**4**				
	ENST00000344924.3	chr7:134331560-134364565		5.81E+001	2.33E+000	2.49E+001	3.42405
	ENST00000418040.1	chr7:134331563-134364560		1.74E-003	7.56E-004	2.31E+000	0
	ENST00000393132.2	chr7:134331583-134364565		2.13E-004	5.39E+000	3.96E-005	0
	ENST00000443095.1	chr7:134345173-134346493		5.63E+000	5.80E-002	9.70E+001	0
4	**SNRPC**	**chr6:34725183-34741571**	**4**				
	ENST00000244520.5	chr6:34725183-34741571		1.35E+003	1.04E+001	1.30E+002	113.989
	ENST00000374018.1	chr6:34725302-34741491		5.32E+002	1.71E+000	3.11E+002	0
	ENST00000374017.3	chr6:34725331-34741571		1.65E+002	1.54E+000	1.07E+002	0

	ENST00000474635.1	chr6:34725689-34741309		3.74E+001	7.94E+001	4.71E-001	0
5	**LSM6**	**Chr4:147096837-147121152**	**7**				
	ENST00000296581.5	chr4:147096837-147111196		4.68E+001	2.94E+001	1.59E+000	0
	ENST00000515311.1	chr4:147096855-147121152		3.47E+002	1.84E+001	1.89E+001	29.6428
	ENST00000503982.1	chr4:147096879-147104798		3.21E+001	1.10E-001	2.93E+002	0
	ENST00000502781.1	chr4:147096881-147111198		1.26E+001	2.28E+000	5.51E+000	4.13649
	ENST00000504181.1	chr4:147096900-147111339		9.42E-104	7.15E-089	1.32E-015	0
	ENST00000510331.1	chr4:147096922-147097838		6.02E+000	5.81E+001	1.04E-001	0
	ENST00000507449.1	chr4:147104075-147108703		1.01E+002	1.24E-001	8.12E+002	0
6	**FOSB**	**Chr19:45971253-45978437**	**11**				
	ENST00000443841.2	chr19:45971253-45978436		5.30E-025	8.45E-022	6.28E-004	0
	ENST00000417353.2	chr19:45971253-45978436		1.16E-006	9.66E-004	1.20E-003	0
	ENST00000585836.1	chr19:45971253-45978436		1.13E+000	6.56E-001	1.72E+000	0
	ENST00000353609.3	chr19:45971253-45978436		1.48E-014	4.02E-002	3.68E-013	0
	ENST00000591858.1	chr19:45971253-45978436		6.51E-011	3.52E+000	1.85E-011	0
	ENST00000590335.1	chr19:45971254-45975339		1.22E+000	8.92E+000	1.36E-001	0
	ENST00000592436.1	chr19:45971693-45976298		8.61E-054	2.57E-041	3.35E-013	0
	ENST00000592811.1	chr19:45973134-45977855		7.76E-138	2.47E-122	3.14E-016	0
	ENST00000586615.1	chr19:45973171-45978437		3.60E-145	5.69E-132	6.33E-014	0
	ENST00000589593.1	chr19:45973523-45975811		2.91E-090	7.56E-084	3.85E-007	0
	ENST00000587358.1	chr19:45974337-45976325		1.56E-168	1.14E-161	1.36E-007	0
7	**TEP1**	**Chr14:20833826-20881588**	**11**				

	ENST00000553365.1	chr14:20833826-20841259		7.55E-013	2.69E-001	2.81E-012	0
	ENST00000262715.5	chr14:20833826-20881580		3.30E-001	2.32E-002	1.43E+001	3.18327
	ENST00000555008.1	chr14:20835790-20859904		1.93E-001	1.27E-002	1.52E+001	0
	ENST00000556935.1	chr14:20836553-20881579		5.43E-009	8.37E-006	6.49E-004	0
	ENST00000555727.1	chr14:20836553-20881578		9.52E-001	1.74E-001	5.47E+000	4.772
	ENST00000553984.1	chr14:20837526-20841239		7.60E-020	8.89E-001	8.55E-020	0
	ENST00000545983.1	chr14:20839677-20850421		7.11E-034	6.86E-026	1.04E-008	0
	ENST00000556488.1	chr14:20841666-20846368		6.67E-001	4.07E-002	1.64E+001	0
	ENST00000471684.2	chr14:20841943-20846203		9.22E-059	6.50E-047	1.42E-012	0
	ENST00000557627.1	chr14:20868826-20872016		8.61E-019	1.42E-001	6.08E-018	0
	ENST00000556549.1	chr14:20876101-20881588		1.85E+000	6.05E-002	3.06E+001	0
8	**POT1**	**Chr7:124462440-124570067**	**20**				
	ENST00000430927.1	chr7:124462440-124467304		5.22E-001	8.98E-002	5.82E+000	0
	CUFF.12817.2	chr7:124462440-124570067		1.28E-002	4.54E-003	2.83E+000	0
	ENST00000357628.3	chr7:124462440-124570035		7.07E-002	1.53E-002	4.61E+000	0.98168
	ENST00000393329.1	chr7:124462441-124570037		1.03E-001	1.63E-002	6.33E+000	1.39618
	ENST00000436534.1	chr7:124462455-124469396		8.85E-169	6.68E-082	1.33E-087	0
	ENST00000609106.1	chr7:124463910-124569856		1.78E-017	1.11E-011	1.60E-006	1.25495
	ENST00000608057.1	chr7:124464016-124537238		1.73E-017	3.83E-011	4.51E-007	0
	ENST00000607932.1	chr7:124464016-124537238		2.67E-015	1.25E-009	2.14E-006	0
	ENST00000608200.1	chr7:124480710-124482886		1.71E-001	8.56E-003	2.00E+001	0
	ENST00000466483.1	chr7:124481035-124483303		1.43E-001	1.28E-001	1.12E+000	0
	ENST00000610141.1	chr7:124491862-124499104		1.38E-067	4.51E-046	3.05E-022	0
	ENST00000608126.1	chr7:124491980-124493581		0	3.17E-001	0	0

	ENST00000487564.1	chr7:124498835-124503439		2.42E-001	2.36E-002	1.02E+001	0
	ENST00000429326.1	chr7:124499032-124537256		2.06E+000	5.39E-002	3.82E+001	0
	ENST00000446993.1	chr7:124510973-124569998		1.74E-031	1.49E-029	1.17E-002	0
	ENST00000609702.1	chr7:124510999-124569881		8.45E-003	3.11E-004	2.72E+001	0
	ENST00000608261.1	chr7:124532320-124569879		3.50E-001	1.01E-002	3.45E+001	0
	ENST00000608437.1	chr7:124532756-124569879		2.31E-006	2.22E-007	1.04E+001	0
	ENST00000461288.1	chr7:124538315-124569856		9.92E-001	7.28E-003	1.36E+002	0
	ENST00000464453.1	chr7:124568975-124569840		4.79E-004	6.58E-002	7.29E-003	0

4.1 Study of Alternative Splicing in SMA

Lack of SMN protein leads to a fatal neurodegenerative disorder: SMA. This fact suggests highly important role of SMN protein in the MNs, coupled with its ubiquitous role in snRNPs biogenesis and spliceosome assembly. In our study, we focused our attention to investigate on the AS mechanisms within MNs, which might get disrupted due to the loss of SMN protein. In particular, our hypothesis states that, the loss of SMN protein might impact the 'AS patterns' of specific set of genes (which are probably linked with the survival of alpha-motor neuron cells in the spinal cord) and most importantly its loss might cause the drop in mRNAs transport within the axons of MNs.

Here, we have identified higher percentage of significantly down-regulated genes (58%) than up-regulated genes (42%), in SMA-patients with respect to healthy controls. This is expected to observe the down expression of the genes in SMA-patients, but it is in the disagreement with Rubin et al. work[287]. RBPs have been shown to interact with SMN protein and such RNA-protein interactions contribute to the enhanced mRNA stabilization within the cytoplasm which extend their life span and expression within the cell[279]. In general, Poly-A Binding Proteins (PABPs) binds on mRNA to stabilize them and moreover, they are potential SMN protein interactions within MNs[279]. In our results, we identified 5 PABP family RBPs out of 22 significantly enriched RBPs namely, PABPC1, PABPC3, PABPC4, PABPC5, PABPN1 (**Table 2.4**). These RBPs have been shown in earlier studies to enhance the mRNA stability until translation is initiated[279–284]. In healthy controls due to presence of normal SMN protein levels, SMN-specific RBPs interacts with it and successfully form a stable RNA-protein complex with the processed mRNA. As a result, the overall gene/transcript expression levels remain sufficiently high which has also been observed in our samples. Conversely in SMA, due to the lower levels of SMN protein, SMN-specific RBPs cannot interact with it, which might cause the destabilization of RNA-protein complex. Consequently, processed mRNA in MNs of SMA-patients tend to degrade comparatively early, effecting their overall expression levels, as also indicated in our results.

Our results from exon-centric analysis has identified approximately similar proportions of silenced exons (57%) and enhanced exons (43%) in SMA patients with respect to healthy controls, as obtained from DEG analysis. Here, we assume that the alternative splicing regulatory mechanisms responsible for the selection of certain splice sites (to perform exon inclusion) within a transcript may get mis-regulated due to SMN protein deficiency, which plays an essential role in the assembly of splicing machinery.

Further, results from our functional annotation analysis has revealed several key regulatory processes specific to neuromuscular system development and maintenance. Interestingly, we have found several over-represented terms having direct associations with the key regulatory mechanisms of motor neuron axon, protein transport and localization towards the end terminals of axons to facilitate their growth. These facts have been previously validated by many experimental studies in SMA animal models[133–136,138,145,285], describing the importance of SMN protein in transport activities within motor neurons. SMN protein has been shown to actively interact and associate with the cytoskeleton (neurofilaments) of motor neurons to aid such axon cargo transportation that is essential for the development/growth of axons and ultimately sustenance of motor neuron cells. While their impaired association has been observed in SMA pathogenesis[136]. We obtained similar terms related to microtubule-based movement of mRNA and proteins which are also supported by actin filament binding. Further, Giavazzi and colleagues have observed the specific rise of SMN protein levels during the development stages of human central nervous system, specifically in the process of axonogenesis and axon sprouting[273]. We have also identified biological processes linked with neuron development and axonogenesis in our analysis. Furthermore, an enrichment of cytosolic calcium ion homeostasis was also found, which is consistent with a study published by Ruiz et al.[277], determining the abnormal accumulation of the calcium ions in nerve fiber terminals of SMA mouse models with respect to their control experiments. An interesting study by Kong and colleagues[143] has determined the specific reduction of synaptic vesicles in SMA mice model which impact the neurotransmission across pre-synaptic terminals and affect NMJs maturation. These findings corroborate our results, reporting the enrichment of mis-regulated genes involved in impulse transmission coupled with neurotransmitter release and its regulation. Recently, a study by Rubin and colleagues[276] has identified the “hyperactivation of ER stress pathway”, resulting into motor neurons degeneration in SMA patients with respect to healthy controls. Consistent to this, we have also found enrichment of gene related to endoplasmic reticulum membrane in our data. Many authors[140–144,153] have worked upon finding the responsible mechanisms behind NMJs disruptions in SMA pathology and the role of agrin protein have been highlighted for proper NMJs development and their maturation during synapse establishment. In SMA pathology, the expression levels of agrin are found to be greatly reduced, resulting in NMJs impaired physiology[274]. In agreement to these finding, we have identified agrin pathway in our analysis, but below the significance levels.

From our motif analysis study, the identified a significant set of PABP family RBPs have been observed to bind on similar cis-acting binding site (Motif 6) localized within enhanced DUACE sequences. This finding supports the idea of preferential role of mRNA and RBPs interactions, contributing to mRNA stabilization within cytoplasm and protection from any possible degradations (such as mRNA uridylation) until translation is initiated[286]. Furthermore, in SMA-patients, HNRNPCL1 RBP has been found to be significantly under-expressed with respect to the controls. In our results, this protein has been identified to bind within silenced DUACE sequences, that might indicate the negative regulation of AS mechanisms in SMA-patients. However, till date, no specific study has investigated in-depth the splice site selection regulatory mechanisms for this protein. The identified HNRNPC RBP is known to mediate the exon skipping by binding to YBX-1 and HNRNPL splicing factors. YBX-1 RBP also has a role in the AS regulation of pre-mRNA. Another study by Nasrin et al.[287] has reported the exon-10 skipping event in Muscle specific Receptor tyrosine Kinase (MuSK), a postsynaptic transmembrane molecule, due to the up-regulation of YBX-1, HNRNPL and HNRNPC trans-acting splicing factors. Akten and colleagues have reported SMN-HuD (Hu Antigen D or ELAV Like Neuron-Specific RNA Binding Protein 4 or ELAVL4 complex) interaction with CPG15 protein (alias Neuritin 1 or NRN1) that mediates the axon growth in MNs[288]. Interestingly, in our data we have observed nearly no expression for NRN1 gene (near zero read coverage) in SMA-patient1 samples (P11 and P12). Further, in this context we have identified another Hu family RBP, namely HuR which has been demonstrated in many studies to be involved in mRNA stabilization by binding specifically to AU-rich elements (ARE) present within 3'-UTR of the transcripts. Farooq and colleagues have exploited HuR RBP for enhancing the SMN-mRNA stabilization and SMN protein expression regulations[289]. HuR is also known to interact with acetylcholinesterase (AChE) during the differentiation and development of muscles (myogenesis)[290–292].

Earlier work by Storbeck et al. has identified SRSF10 (isoform 1) as a 'splicing enhancer' trans-acting factor which specifically recognizes and binds to the GAA-rich regions within mRNA[293]. This SR family splicing factor has been shown to revert the AS pattern of SMN2 gene towards the formation of full-length SMN transcripts by enhancing the inclusion of exon-7[294]. The positive activity of SRSF10 has offered potential therapeutic benefits to SMA-patients by significantly increasing the levels of functional SMN protein[295,296]. Here, similar to these previous findings, we have identified SRSF10 RBP which pinpoints its relevance in SMA pathogenesis.

Additionally, our RNA-Seq data have revealed the clear distinction for the expression level of SMN1 gene between SMA patients and controls; specifically, SMA patients have no read coverage within exon-7 at nucleotide position 6 (**Figure 3.4 in Chapter 3 - Results section 3.1.2**). This indicates the great reliability of the procedure being used for the generation of MNs, specific to SMA patients and healthy controls, using iPSC technology that has added a great potential to identify patient specific targets for such complex neurodegenerative disorders. Further, we obtained consistent variability in the overall expression levels of genes and isoforms within and between biological replicate samples from SMA-patients and controls (**Figure 3.5 in Chapter 3 - Results section 3.1.5**), which supports our data and analysis procedure.

In future, the significant set of RBPs we have identified requires subsequent wet-lab experimental validations to determine their exact binding sites within mRNA, which guides the selection of specific splice-sites during pre-mRNA splicing. UV cross-linking and immunoprecipitation (CLIP) method[68] and individual nucleotide resolution CLIP (iCLIP) method[297] has been introduced to determine RNA-RBP interactions and to identify exact binding sites of RBPs on pre-mRNA. In this study, RNA samples were poly-A selected which restricts the analysis for mature mRNA only, therefore in future experimental plans, total RNA-Sequencing of the samples should be performed. The data generated from total RNA-Sequencing can be analyzed in detail using the computational model we have developed in Study-B which has a potential to precisely estimate the transcript expression level in a given gene locus and effectively detect differential splicing events between two conditions.

4.2 Development of Computational Model to Estimate Transcript Expression

In this study, we have presented a non-linear model for estimating the expression levels of each transcript within a given gene locus by disentangling the contributions of mature and nascent RNA transcription at a steady-state. Given the high complexity within gene loci, we assume that by quantifying these two phenomenon, the precise weightage of each possible transcript for a gene can be estimated from total RNA-Seq data. The mature RNA measures are dependent upon the nascent RNA levels which make the system non-linear while estimating their contributions in a given gene locus.

In the comparison study of estimated expression values from our model with cufflinks quantified FPKM expressions in analyzed genes, we found most of the isoforms which have higher expression estimate in our model also remains consistent with the expressions obtained from cufflinks tool (corresponding isoforms are represented with *'grey'* highlighted rows in **Table 3.5**).

To further verify the accuracy of our modeled estimations we would like to set up simulation experiments to compare the estimated expressions with simulated read expressions. The simulation of nascent transcripts is rather complex than simulating only mature transcripts. This is because in nascent transcripts two mechanisms has to be considered side-by-side that are on-going transcription and partially spliced regions of transcribing transcript after every step of complete transcription of single intron and its neighboring exons. More specific techniques are available to measure the nascent transcription of cells, giving the account of transcriptional activity. Such as Global Run-On (GRO) Seq[298] and RNA Polymerase II (RNA-II) Chip-Seq[299,300], which can be run in parallel to compare the obtained expression estimations with our modeled estimations.

Further, we would also like to apply our method without providing any gene annotations by the combination of approach of Directed Acyclic Graphs (DAGs) and exon-intron junction information from the total RNA-Seq data. The possible paths and their exponential increase by considering all the possible combinations with all types of alternative splicing events makes everything quite complex but can be handled if we combine the junction information and per-base read coverages to include only plausible isoform paths and exclude the unrealistic paths from the further expression modeling.

Currently in our model we intend to refine the already annotated paths by combining the exon-intron junction information from the given data obtained by computing the coverages at per-base resolution and in future we will apply our model to predict the isoform paths without providing any annotations but the real challenge is to correctly identify the transcription start site (TSS) and polyadenylation site (PAS) in a given gene locus. We attempted to devise a method for predicting the precise TSS and PAS by observing the Border-effects at start and end-site of the transcripts. Such effects arise due to variations in the Fragment Length Distributions (FLD) at the start and end-sites of the transcripts. Later, we investigated that for more complex gene loci where the TSS and PAS containing exons are very small their FLD profiles spreads from first exon to the neighboring exons. Therefore, we are still working to devise other strategies to tackle these issues.

Another improvement we are considering is in the selection of iteration termination criterion to further enhance the speed of convergence. The complete algorithm has been implemented in R-programming language and in future we would like to implement it in C++ to improve the processing time and convergence speed of highly complex gene loci or whole genome so that we can compare the results with above mentioned methods. Most importantly, we want to examine the performance of our model in studying the differential splicing between two or more conditions for which we need total RNA-Sequencing experiments. In conclusion, our method gives the promising results with accurate estimation of isoform expression levels within reasonable computational processing time.

References

1. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci.* **74,** 3171–3175 (1977).
2. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12,** 1–8 (1977).
3. Will, C. L. & Lührmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3,** 1–2 (2011).
4. Matera, A. G., Terns, R. M. & Terns, M. P. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.* **8,** 209–20 (2007).
5. Ohno, M., Segref, A., Bachi, A., Wilm, M. & Mattaj, I. W. PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. *Cell* **101,** 187–198 (2000).
6. Segref, A., Mattaj, I. W. & Ohno, M. The evolutionarily conserved region of the U snRNA export mediator PHAX is a novel RNA-binding domain that is essential for U snRNA export. *RNA* **7,** 351–60 (2001).
7. Terns, M. P. & Terns, R. M. Macromolecular complexes: SMN--the master assembler. *Curr. Biol.* **11,** R862-4 (2001).
8. Palacios, I., Hetzer, M., Adam, S. A. & Mattaj, I. W. Nuclear import of U snRNPs requires importin beta. *EMBO J.* **16,** 6783–92 (1997).
9. Huber, J. *et al.* Snurportin1 , an m 3 G-cap-specific nuclear import receptor with a novel domain structure. *EMBO J.* **17,** 4114–4126 (1998).
10. Staněk, D. & Neugebauer, K. M. The Cajal body: A meeting place for spliceosomal snRNPs in the nuclear maze. *Chromosoma* **115,** 343–354 (2006).
11. Lamond, A. I. & Spector, D. L. Nuclear speckles: a model for nuclear organelles. *Nat. Rev. Mol. Cell Biol.* **4,** 605–612 (2003).
12. Burset, M., Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28,** 4364–4375 (2000).
13. Jurica, M. S. & Moore, M. J. Pre-mRNA splicing: Awash in a sea of proteins. *Mol. Cell* **12,** 5–14 (2003).
14. Nilsen, T. W. RNA-RNA interactions in the spliceosome: Unraveling the ties that bind. *Cell* **78,** 1–4 (1994).
15. Madhani, H. D. & Guthrie, C. Dynamic RNA-RNA interactions in the spliceosome. *Annu. Rev. Genet.* **28,** 1–26 (1994).
16. Modrek, B. & Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **30,** 13–19 (2002).
17. Reed, R. Initial splice-site recognition and pairing during pre-mRNA splicing. *Curr. Opin.*

Genet. Dev. **6,** 215–220 (1996).

18. Robberson, B. L., Cote, G. J. & Berget, S. M. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* **10,** 84–94 (1990).

19. Reed, R. Mechanisms of fidelity in pre-mRNA splicing. *Curr. Opin. Cell Biol.* **12,** 340–345 (2000).

20. Konarska, M. M. & Sharp, P. A. Interactions between Small Nuclear Ribonucleoprotein Particles in Formation of Spliceosomes. *Cell* **49,** 763–774 (1986).

21. Brody, E. & Abelson, J. The 'spliceosome': yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. *Science* **228,** 963–7 (1985).

22. Staknis, D. & Reed, R. SR proteins promote the first specific recognition of Pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. *Mol. Cell. Biol.* **14,** 7670–7682 (1994).

23. Sun, J. S. & Manley, J. L. A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing. *Genes Dev.* **9,** 843–854 (1995).

24. Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: Motors, clocks, springs, and things. *Cell* **92,** 315–326 (1998).

25. Burge, C. B., Tuschl, T. & Sharp, P. A. Splicing of Precursors to mRNAs by the Spliceosomes. *Cold Spring Harb. Monogr. Arch.* **37,** 525–560 (1999).

26. Fu, X. D. The superfamily of arginine/serine-rich splicing factors. *RNA* **1,** 663–680 (1995).

27. Company, M., Arenas, J. & Abelson, J. Requirement of the RNA helicase-like protein PRP22 for release of messenger RNA from spliceosomes. *Nature* **349,** 487–493 (1991).

28. Cordin, O., Hahn, D. & Beggs, J. D. Structure, function and regulation of spliceosomal RNA helicases. *Curr. Opin. Cell Biol.* **24,** 431–438 (2012).

29. Will, C. L. & Lührmann, R. Protein functions in pre-mRNA splicing. *Curr. Opin. Cell Biol.* **9,** 320–328 (1997).

30. Jones, M. H., Frank, D. N. & Guthrie, C. Characterization and functional ordering of Slu7p and Prp17p during the second step of pre-mRNA splicing in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 9687–91 (1995).

31. Schwer, B. & Guthrie, C. A conformational rearrangement in the spliceosome is dependent on PRP16 and ATP hydrolysis. *EMBO J.* **11,** 5033–9 (1992).

32. Small, E. C., Leggett, S. R., Winans, A. A. & Staley, J. P. The EF-G-like GTPase Snu114p Regulates Spliceosome Dynamics Mediated by Brr2p, a DExD/H Box ATPase. *Mol. Cell* **23,** 389–399 (2006).

33. Schwer, B. & Gross, C. H. Prp22, a DExH-box RNA helicase, plays two distinct roles in yeast pre-mRNA splicing. *EMBO J.* **17,** 2086–2094 (1998).

34. Butcher, S. E. The spliceosome and its metal ions. *Met Ions Life Sci* **9,** 235–251 (2011).

35. Sontheimer, E. J., Sun, S. G. & Piccirilli, J. A. Metal ion catalysis during splicing of

premessenger RNA. *Nature* **388,** 801–805 (1997).

36. Black, D. L. Finding splice sites within a wilderness of RNA. *RNA* **1,** 763–71 (1995).

37. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.* **72,** 291–336 (2003).

38. Cooper, T. a & Mattox, W. The regulation of splice-site selection, and its role in human disease. *Am. J. Hum. Genet.* **61,** 259–266 (1997).

39. Graveley, B. R. Alternative splicing: Increasing diversity in the proteomic world. *Trends Genet.* **17,** 100–107 (2001).

40. Ladd, A. N. & Cooper, T. a. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.* **3,** 8 (2002).

41. Smith, C. W. J. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25,** 381–388 (2000).

42. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

43. Rowen, L. *et al.* Analysis of the human neurexin genes: alternative splicing and the generation of protein diversity. *Genomics* **79,** 587–597 (2002).

44. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

45. Grabowski, P. J. Splicing regulation in neurons: Tinkering with cell-specific control. *Cell* **92,** 709–712 (1998).

46. Sammeth, M., Foissac, S. & Guigo, R. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.* **4,** (2008).

47. Breitbart, R. E., Andreadis, A. & Nadal-Ginard, B. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.* **56,** 467–95 (1987).

48. Kan, Z., States, D. & Gish, W. Selecting for functional alternative splices in ESTs. *Genome Res.* **12,** 1837–45 (2002).

49. Ner-Gaon, H. *et al.* Intron retention is a major phenomenon in alternative splicing in Arabidopsis. *Plant J.* **39,** 877–85 (2004).

50. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24,** 1774–1786 (2014).

51. Yan, Q. *et al.* Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 1502849112- (2015).

52. Matlin, A. J., Clark, F. & Smith, C. W. J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6,** 386–98 (2005).

53. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an

integrated splicing code. *RNA* **14,** 802–13 (2008).

54. Zhou, Z. & Fu, X. D. Regulation of splicing by SR proteins and SR protein-specific kinases. *Chromosoma* **122,** 191–207 (2013).

55. Long, J. C. & Caceres, J. F. The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J* **417,** 15–27 (2009).

56. Änkö, M.-L. Regulation of gene expression programmes by serine–arginine rich splicing factors. *Semin. Cell Dev. Biol.* **32,** 11–21 (2014).

57. Martinez-Contreras, R. *et al.* hnRNP proteins and splicing control. *Adv. Exp. Med. Biol.* **623,** 123–147 (2007).

58. Darnell, R. B. RNA Protein Interaction in Neurons. *Annu. Rev. Neurosci.* **36,** 243–270 (2013).

59. Han, H. *et al.* MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* **498,** 241–5 (2013).

60. Wang, E. T. *et al.* Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell* **150,** 710–724 (2012).

61. Li, H. *et al.* Dynamic expression pattern of neuro-oncological ventral antigen 1 (Nova1) in the rat brain after focal cerebral ischemia/reperfusion insults. *J. Histochem. Cytochem.* **61,** 45–54 (2013).

62. Xiao, S. H. & Manley, J. L. Phosphorylation of the ASF/SF2 RS domain affects both protein-protein and protein-RNA interactions and is necessary for splicing. *Genes Dev.* **11,** 334–344 (1997).

63. Tacke, R. & Manley, J. L. Determinants of SR protein specificity. *Curr. Opin. Cell Biol.* **11,** 358–362 (1999).

64. Manley, J. L. & Tacke, R. SR proteins and splicing control. *Genes Dev.* **10,** 1569–1579 (1996).

65. Bonnal, S. *et al.* RBM5/Luca-15/H37 Regulates Fas Alternative Splice Site Pairing after Exon Definition. *Mol. Cell* **32,** 81–95 (2008).

66. Sharma, S., Kohlstaedt, L. a, Damianov, A., Rio, D. C. & Black, D. L. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nat. Struct. Mol. Biol.* **15,** 183–91 (2008).

67. Izquierdo, J. M. *et al.* Regulation of Fas alternative splicing by antagonistic effects of TIA-1 and PTB on exon definition. *Mol. Cell* **19,** 475–84 (2005).

68. Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302,** 1212–5 (2003).

69. Darnell, R. B. & Posner, J. B. Paraneoplastic syndromes and the nervous system. *N. Engl. J. Med.* **3,** 287–288 (2003).

70. Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1792,** 14–26 (2009).

71. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465,** 53–59 (2010).

72. Faustino, N. A., Cooper, T. a & Andre, N. Pre-mRNA splicing and human disease. *Genes Dev.* **17,** 419–437 (2003).

73. Singh, R. K. & Cooper, T. A. Pre-mRNA splicing in disease and therapeutics. *Trends Mol. Med.* **18,** 472–482 (2012).

74. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol* **5,** R74 (2004).

75. Johnson, M. B. *et al.* Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. *Neuron* **62,** 494–509 (2009).

76. Courtney, E., Kornfeld, S., Janitz, K. & Janitz, M. Transcriptome profiling in neurodegenerative disease. *J. Neurosci. Methods* **193,** 189–202 (2010).

77. D'Souza, I. *et al.* Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 5598–603 (1999).

78. Ballatore, C., Lee, V. M.-Y. & Trojanowski, J. Q. Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nat. Rev. Neurosci.* **8,** 663–672 (2007).

79. Hernández, F. & Avila, J. Tauopathies. *Cell. Mol. Life Sci.* **64,** 2219–2233 (2007).

80. Lunn, M. R. & Wang, C. H. Spinal muscular atrophy. *Lancet* **371,** 2120–33 (2008).

81. Wirth, B. An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinal muscular atrophy (SMA). *Hum. Mutat.* **15,** 228–237 (2000).

82. Pearn, J. Incidence, prevalence, and gene frequency studies of chronic childhood spinal muscular atrophy. *J. Med. Genet.* **15,** 409–413 (1978).

83. Pearn, J. H., Gardner-Medwin, D. & Wilson, J. A clinical study of chronic childhood spinal muscular atrophy. *J. Neurol. Sci.* **38,** 23–37 (1978).

84. Pearn, J. Classification of spinal muscular atrophies. *Lancet (London, England)* **1,** 919–922 (1980).

85. Ogino, S., Leonard, D. G. B., Rennert, H., Ewens, W. J. & Wilson, R. B. Genetic risk assessment in carrier testing for spinal muscular atrophy. *Am. J. Med. Genet.* **110,** 301–307 (2002).

86. Munsat, T. L. International SMA Collaboration. *Neuromuscul. Disord.* **1,** 81 (1991).

87. Russman, B. S. Spinal muscular atrophy: clinical classification and disease heterogeneity. *J. Child Neurol.* **22,** 946–951 (2007).

88. Markowitz, J. a. Spinal Muscular Atrophy in the Neonate. *J. Obstet. Gynecol. Neonatal Nurs.* **33,** 12–20 (2004).

89. Fidziańska, A. Spinal muscular atrophy in childhood. *Semin. Pediatr. Neurol.* **3,** 53–58 (1996).

90. Garvie, J. M. & Woolf, L. A. Kugelberg-Welander Syndrome (Hereditary Proximal Spinal

Muscular Atrophy). *Br. Med. J.* 1458–1461 (1966).

91. Kugelberg, E. & Welander, L. Heredofamilial juvenile muscular atrophy simulating muscular dystrophy. *AMA. Arch. Neurol. Psychiatry* **75,** 500–9 (1956).

92. Gregoretti, C. *et al.* Survival of patients with spinal muscular atrophy type 1. *Pediatrics* **131,** e1509-14 (2013).

93. Oskoui, M. *et al.* The changing natural history of spinal muscular atrophy type 1. *Neurology* **69,** 1931–1936 (2007).

94. Brzustowicz, L. M. *et al.* Genetic mapping of chronic childhood-onset spinal muscular atrophy to chromosome 5q11.2-13.3. *Nature* **344,** 540–1 (1990).

95. Brzustowicz, L. M. *et al.* Fine-mapping of the spinal muscular atrophy locus to a region flanked by MAP1B and D5S6. *Genomics* **13,** 991–998 (1992).

96. Wang, C. H. *et al.* Refinement of the spinal muscular atrophy locus by genetic and physical mapping. *Am. J. Hum. Genet.* **56,** 202–9 (1995).

97. Clermont, O. *et al.* Use of genetic and physical mapping to locate the spinal muscular atrophy locus between two new highly polymorphic DNA markers. *Am. J. Hum. Genet.* **54,** 687–94 (1994).

98. Melki, J. *et al.* Gene for chronic proximal spinal muscular atrophies maps to chromosome 5q. *Nature* **344,** 767–768 (1990).

99. Schmutz, J. *et al.* The DNA sequence and comparative analysis of human chromosome 5. *Nature* **431,** 268–274 (2004).

100. Lefebvre, S. *et al.* Identification and characterization of a spinal muscular atrophy-determining gene. *Cell* **80,** 155–165 (1995).

101. Burglen, L. *et al.* Structure and organization of the human survival motor neurone (SMN) gene. *Genomics* **32,** 479–482 (1996).

102. Burghes, A. H. When is a deletion not a deletion? When it is converted. *Am. J. Hum. Genet.* **61,** 9–15 (1997).

103. Rochette, C. F., Gilbert, N. & Simard, L. R. SMN gene duplication and the emergence of the SMN2 gene occurred in distinct hominids: SMN2 is unique to Homo sapiens. *Hum. Genet.* **108,** 255–266 (2001).

104. Vitte, J. *et al.* Refined characterization of the expression and stability of the SMN gene products. *Am. J. Pathol.* **171,** 1269–80 (2007).

105. Lorson, C. L., Hahnen, E., Androphy, E. J. & Wirth, B. A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 6307–11 (1999).

106. Lorson, C. L. & Androphy, E. J. An exonic enhancer is required for inclusion of an essential exon in the SMA-determining gene SMN. *Hum. Mol. Genet.* **9,** 259–265 (2000).

107. Gavrilov, D. K., Shi, X., Das, K., Gilliam, T. C. & Wang, C. H. Differential SMN2 expression associated with SMA severity. *Nat. Genet.* **20,** 230–231 (1998).

108. Cartegni, L. & Krainer, A. R. Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1. *Nat. Genet.* **30,** 377–84 (2002).

109. Cartegni, L., Hastings, M. L., Calarco, J. A., de Stanchina, E. & Krainer, A. R. Determinants of exon 7 splicing in the spinal muscular atrophy genes, SMN1 and SMN2. *Am. J. Hum. Genet.* **78,** 63–77 (2006).

110. Kashima, T. & Manley, J. L. A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat. Genet.* **34,** 460–463 (2003).

111. Kashima, T., Rao, N. & Manley, J. L. An intronic element contributes to splicing repression in spinal muscular atrophy. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 3426–3431 (2007).

112. Kashima, T., Rao, N., David, C. J. & Manley, J. I. hnRNP A1 functions with specificity in repression of SMN2 exon 7 splicing. *Hum. Mol. Genet.* **16,** 3149–3159 (2007).

113. Liu, Q. & Dreyfuss, G. A novel nuclear structure containing the survival of motor neurons protein. *EMBO J.* **15,** 3555–3565 (1996).

114. Lafarga, M., Casafont, I., Bengoechea, R., Tapia, O. & Berciano, M. T. Cajal's contribution to the knowledge of the neuronal cell nucleus. *Chromosoma* **118,** 437–443 (2009).

115. Carvalho, T. *et al.* The spinal muscular atrophy disease gene product, SMN: A link between snRNP biogenesis and the Cajal (coiled) body. *J. Cell Biol.* **147,** 715–727 (1999).

116. Liu, Q., Fischer, U., Wang, F. & Dreyfuss, G. The spinal muscular atrophy disease gene product, SMN, and its associated protein SIP1 are in a complex with spliceosomal snRNP proteins. *Cell* **90,** 1013–1021 (1997).

117. Lorson, C. L. *et al.* SMN oligomerization defect correlates with spinal muscular atrophy severity. *Nat Genet* **19,** 63–66 (1998).

118. Meister, G., Bühler, D., Pillai, R., Lottspeich, F. & Fischer, U. A multiprotein complex mediates the ATP-dependent assembly of spliceosomal U snRNPs. *Nat. Cell Biol.* **3,** 945–949 (2001).

119. Pellizzoni, L., Yong, J. & Dreyfuss, G. Essential role for the SMN complex in the specificity of snRNP assembly. *Science* **298,** 1775–1779 (2002).

120. Young, P. J. *et al.* The exon 2b region of the spinal muscular atrophy protein, SMN, is involved in self-association and SIP1 binding. *Hum. Mol. Genet.* **9,** 2869–77 (2000).

121. Eggert, C., Chari, A., Laggerbauer, B. & Fischer, U. Spinal muscular atrophy: the RNP connection. *Trends Mol. Med.* **12,** 113–121 (2006).

122. Gubitz, A. K., Feng, W. & Dreyfuss, G. The SMN complex. *Exp. Cell Res.* **296,** 51–56 (2004).

123. Pellizzoni, L. Chaperoning ribonucleoprotein biogenesis in health and disease. *EMBO Rep.* **8,** 340–345 (2007).

124. Kroiss, M. *et al.* Evolution of an RNP assembly system: a minimal SMN complex facilitates formation of UsnRNPs in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 10045–50 (2008).

125. Otter, S. *et al.* A comprehensive interaction map of the human survival of motor neuron (SMN) complex. *J. Biol. Chem.* **282,** 5825–5833 (2007).

126. Baccon, J., Pellizzoni, L., Rappsilber, J., Mann, M. & Dreyfuss, G. Identification and characterization of Gemin7, a novel component of the survival of motor neuron complex. *J. Biol. Chem.* **277,** 31957–31962 (2002).

127. Battle, D. J. *et al.* The SMN complex: An assembly machine for RNPs. in *Cold Spring Harbor Symposia on Quantitative Biology* **71,** 313–320 (2006).

128. Carissimi, C. *et al.* Unrip is a component of SMN complexes active in snRNP assembly. *FEBS Lett.* **579,** 2348–2354 (2005).

129. Carissimi, C., Saieva, L., Gabanella, F. & Pellizzoni, L. Gemin8 is required for the architecture and function of the survival motor neuron complex. *J. Biol. Chem.* **281,** 37009–37016 (2006).

130. Charroux, B. *et al.* Gemin3: A novel DEAD box protein that interacts with SMN, the spinal muscular atrophy gene product, and is a component of gems. *J. Cell Biol.* **147,** 1181–1193 (1999).

131. Charroux, B. *et al.* Gemin4: A novel component of the SMN complex that is found in both gems and nucleoli. *J. Cell Biol.* **148,** 1177–1186 (2000).

132. Gubitz, A. K. *et al.* Gemin5, a novel WD repeat protein component of the SMN complex that binds Sm proteins. *J. Biol. Chem.* **277,** 5631–5636 (2002).

133. Pagliardini, S. *et al.* Subcellular localization and axonal transport of the survival motor neuron (SMN) protein in the developing rat spinal cord. *Hum. Mol. Genet.* **9,** 47–56 (2000).

134. Rossoll, W. *et al.* Specific interaction of Smn, the spinal muscular atrophy determining gene product, with hnRNP-R and gry-rbp/hnRNP-Q: a role for Smn in RNA processing in motor axons? *Hum. Mol. Genet.* **11,** 93–105 (2002).

135. Rossoll, W. *et al.* Smn, the spinal muscular atrophy-determining gene product, modulates axon growth and localization of β-actin mRNA in growth cones of motoneurons. *J. Cell Biol.* **163,** 801–812 (2003).

136. Zhang, H. L. *et al.* Active transport of the survival motor neuron protein and the role of exon-7 in cytoplasmic localization. *J. Neurosci.* **23,** 6627–37 (2003).

137. Tadesse, H., Deschnes-Furry, J., Boisvenue, S. & Cote, J. KH-type splicing regulatory protein interacts with survival motor neuron protein and is misregulated in spinal muscular atrophy. *Hum. Mol. Genet.* **17,** 506–524 (2008).

138. Glinka, M. *et al.* The heterogeneous nuclear ribonucleoprotein-R is necessary for axonal β-actin mRNA translocation in spinal motor neurons. *Hum. Mol. Genet.* **19,** 1951–1966 (2010).

139. Carrel, T. L. *et al.* Survival motor neuron function in motor axons is independent of functions required for small nuclear ribonucleoprotein biogenesis. *J Neurosci* **26,** 11014–11022 (2006).

140. Boon, K. L. *et al.* Zebrafish survival motor neuron mutants exhibit presynaptic neuromuscular junction defects. *Hum. Mol. Genet.* **18,** 3615–3625 (2009).

141. Chan, Y. B. *et al.* Neuromuscular defects in a Drosophila survival motor neuron gene mutant. *Hum. Mol. Genet.* **12,** 1367–1376 (2003).

142. Kariya, S. *et al.* Reduced SMN protein impairs maturation of the neuromuscular junctions in mouse models of spinal muscular atrophy. *Hum. Mol. Genet.* **17,** 2552–2569 (2008).

143. Kong, L. *et al.* Impaired Synaptic Vesicle Release and Immaturity of Neuromuscular Junctions in Spinal Muscular Atrophy Mice. *J. Neurosci.* **29,** 842–851 (2009).

144. Murray, L. M. *et al.* Selective vulnerability of motor neurons and dissociation of pre- and post-synaptic pathology at the neuromuscular junction in mouse models of spinal muscular atrophy. *Hum. Mol. Genet.* **17,** 949–962 (2008).

145. Mentis, G. Z. *et al.* Early Functional Impairment of Sensory-Motor Connectivity in a Mouse Model of Spinal Muscular Atrophy. *Neuron* **69,** 453–467 (2011).

146. Gogliotti, R. G. *et al.* Motor Neuron Rescue in Spinal Muscular Atrophy Mice Demonstrates That Sensory-Motor Defects Are a Consequence, Not a Cause, of Motor Neuron Dysfunction. *J. Neurosci.* **32,** 3818–3829 (2012).

147. Ling, K. K. Y., Lin, M. Y., Zingg, B., Feng, Z. & Ko, C. P. Synaptic defects in the spinal and neuromuscular circuitry in a mouse model of spinal muscular atrophy. *PLoS One* **5,** (2010).

148. Wishart T.M Riessland M, Reimer M.M, Hunter G, Hannam M.L, Eaton S, Fuller H.R, Roche S.L, Somers E, Morse R, Young P.J, Lamont D.J, Hammerschmidt M, Joshi A, Hohenstein P, Morris G.E, Parson S.H, Skehel P.A, Becker T, Robinson I.M, Becker C.G, Wirth B &, M. C. a. Disrupted ubiquitin homeostasis and β-catenin signalling in spinal muscular atrophy. *J Clin Invest* **124,** (2014).

149. Miguel-Aliaga, I. *et al.* The Caenorhabditis elegans orthologue of the human gene responsible for spinal muscular atrophy is a maternal product critical for germline maturation and embryonic viability. *Hum. Mol. Genet.* **8,** 2133–2143 (1999).

150. Paushkin, S. *et al.* The survival motor neuron protein of Schizosacharomyces pombe: Conservation of survival motor neuron interaction domains in divergent organisms. *J. Biol. Chem.* **275,** 23841–23846 (2000).

151. Bertrandy, S. *et al.* The RNA-binding properties of SMN: deletion analysis of the zebrafish orthologue defines domains conserved in evolution. *Hum. Mol. Genet.* **8,** 775–782 (1999).

152. Briese, M. *et al.* Deletion of smn-1, the Caenorhabditis elegans ortholog of the spinal muscular atrophy gene, results in locomotor dysfunction and reduced lifespan. *Hum. Mol. Genet.* **18,** 97–104 (2009).

153. Sleigh, J. N. *et al.* A novel Caenorhabditis elegans allele, smn-1(cb131), mimicking a mild form of spinal muscular atrophy, provides a convenient drug screening platform

highlighting new and pre-approved compounds. *Hum. Mol. Genet.* **20,** 245–60 (2011).

154. Chang, H. C. H. *et al.* Modeling spinal muscular atrophy in Drosophila. *PLoS One* **3,** 1–18 (2008).

155. Hsieh-Li, H. M. *et al.* A mouse model for spinal muscular atrophy. *Nat Genet* **24,** 66–70 (2000).

156. Monani, U. R., Coovert, D. D. & Burghes, a H. Animal models of spinal muscular atrophy. *Hum. Mol. Genet.* **9,** 2451–2457 (2000).

157. Edens, B. M., Ajroud-Driss, S., Ma, L. & Ma, Y. Molecular mechanisms and animal models of spinal muscular atrophy. *Biochim. Biophys. Acta* **1852,** 685–92 (2015).

158. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126,** 663–676 (2006).

159. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131,** 861–872 (2007).

160. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318,** 1917–20 (2007).

161. Ebert, A. D. *et al.* Induced pluripotent stem cells from a spinal muscular atrophy patient. *NIH Public Access* **457,** 277–280 (2009).

162. Fuller, H. R. *et al.* Spinal Muscular Atrophy Patient iPSC-Derived Motor Neurons Have Reduced Expression of Proteins Important in Neuronal Development. *Front. Cell. Neurosci.* **9,** 1–15 (2016).

163. Corti, S. *et al.* Genetic correction of human induced pluripotent stem cells from patients with spinal muscular atrophy. *Sci. Transl. Med.* **4,** 165ra162 (2012).

164. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–70 (1995).

165. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–11 (2000).

166. Yeakley, J. M. *et al.* Profiling alternative splicing on fiber-optic arrays. *Nat. Biotechnol.* **20,** 353–358 (2002).

167. Johnson, J. M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302,** 2141–4 (2003).

168. Castle, J. *et al.* Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. *Genome Biol.* **4,** R66 (2003).

169. Wang, H. *et al.* Gene structure-based splice variant deconvolution using a microarry platform. *Bioinformatics* **19,** i315–i322 (2003).

170. Bertone, P., Gerstein, M. & Snyder, M. Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. *Chromosome Res.* **13,** 259–74 (2005).

171. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077–82 (1998).

172. Clark, T. A., Sugnet, C. W. & Ares, M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296,** 907–10 (2002).

173. Sanger, F., Nicklen, S. & Coulson, a R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–7 (1977).

174. Smith, L. M., Fung, S., Hunkapiller, M. W., Hunkapiller, T. J. & Hood, L. E. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.* **13,** 2399–412 (1985).

175. Smith, L. M. *et al.* Fluorescence detection in automated DNA sequence analysis. *Nature* **321,** 674–679 (1986).

176. Bonaldo, M. F., Lennon, G. & Soares, M. B. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6,** 791–806 (1996).

177. Adams, M. D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252,** 1651–6 (1991).

178. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res.* **9,** 1288–93 (1999).

179. Brett, D. *et al.* EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.* **474,** 83–86 (2000).

180. Modrek, B., Resch, A., Grasso, C. & Lee, C. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* **29,** 2850–2859 (2001).

181. Eyras, E., Caccamo, M., Curwen, V. & Clamp, M. ESTGenes: Alternative splicing from ESTs in Ensembl. *Genome Res.* **14,** 976–987 (2004).

182. Kan, Z., Rouchka, E. C., Gish, W. R. & States, D. J. Gene Structure Prediction and Alternative Splicing Analysis Using Genomically Aligned ESTs. *Genome Res.* **11,** 889–900 (2001).

183. Jang, W., Chen, H. C., Sicotte, H. & Schuler, G. D. Making effective use of human genomic sequence data. *Trends Genet.* **15,** 284–6 (1999).

184. Nagaraj, S. H., Gasser, R. B. & Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief. Bioinform.* **8,** 6–21 (2007).

185. Ju, J. *et al.* Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 19635–40 (2006).

186. Bowers, J. *et al.* Virtual terminator nucleotides for next-generation DNA sequencing. *Nat. Methods* **6,** 593–5 (2009).

187. Nakano, M. *et al.* Single-molecule PCR using water-in-oil emulsion. *J. Biotechnol.* **102,** 117–24 (2003).

188. Rothberg, J. M. & Leamon, J. H. The development and impact of 454 sequencing. *Nat. Biotechnol.* **26,** 1117–24 (2008).

189. Vera, J. C. *et al.* Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* **17,** 1636–47 (2008).

190. Emrich, S. J., Barbazuk, W. B., Li, L. & Schnable, P. S. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17,** 69–73 (2007).

191. Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L. & Schnable, P. S. SNP discovery via 454 transcriptome sequencing. *Plant J.* **51,** 910–8 (2007).

192. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci.* **104,** 14616–14621 (2007).

193. Korbel, J. O. *et al.* Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome. *October* **318,** 420–426 (2009).

194. Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444,** 330–6 (2006).

195. Dalloul, R. A. *et al.* Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol.* **8,** (2010).

196. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463,** 311–7 (2010).

197. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12,** 671–682 (2011).

198. Michaelson, J. J. *et al.* Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* **151,** 1431–1442 (2012).

199. Evrony, G. D., Cai, X. & Walsh, C. A. Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain. *Cell* **151,** 483–496 (2012).

200. Ku, C. S. *et al.* A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. *Mol. Psychiatry* **18,** 141–53 (2013).

201. Zhang, B. *et al.* Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell* **153,** 707–720 (2013).

202. Shokralla, S., Spall, J. L., Gibson, J. F. & Hajibabaei, M. Next-generation sequencing technologies for environmental DNA research. *Mol. Ecol.* **21,** 1794–805 (2012).

203. Tomkinson, A. E., Vijayakumar, S., Pascal, J. M. & Ellenberger, T. DNA ligases: Structure, reaction mechanism, and function. *Chem. Rev.* **106,** 687–699 (2006).

204. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science* **241,** 1077–1080 (1988).

205. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Meth* **5,** 613–619 (2008).

206. Valouev, A. *et al.* A high-resolution, nucleosome position map of C. elegans reveals a

lack of universal sequence-dictated positioning. *Genome Res.* **18,** 1051–63 (2008).

207. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6,** 377–382 (2009).

208. McKernan, K. J. *et al.* Sequence and Structural Variation in a Human Genome Uncovered by Short-Read , Massively Parallel Ligation. *Genome Res.* **19,** 1527–1541 (2009).

209. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475,** 348–352 (2011).

210. Pennisi, E. Genomics. Semiconductors inspire new sequencing technologies. *Science* **327,** 1190 (2010).

211. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323,** 133–8 (2009).

212. Levene, M. J. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science (80-.).* **299,** 682–686 (2003).

213. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63 (2009).

214. Li, S. *et al.* Multi-platform and cross-methodological reproducibility of transcriptome profiling by RNA-seq in the ABRF Next- Generation Sequencing Study. *Nat Biotechnol.* **32,** 915–925 (2014).

215. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth* **5,** 621–628 (2008).

216. Ewing, B., Hillier, L. D. & Wendl, M. C. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.* **8,** 186–194 (1998).

217. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic Acids Res.* **39,** 2010–2012 (2011).

218. Andrews, S. FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Institute* (2010).

219. Deluca, D. S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28,** 1530–1532 (2012).

220. Lassmann, T., Hayashizaki, Y. & Daub, C. O. SAMStat: Monitoring biases in next generation sequencing data. *Bioinformatics* **27,** 130–131 (2011).

221. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

222. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

223. Planet, E., Attolini, C. S.-O., Reina, O., Flores, O. & Rossell, D. htSeqTools: high-throughput sequencing quality control, processing and visualization in R. *Bioinformatics* **28,** 589–90 (2012).

224. Williams, C. R., Baccarella, A., Parrish, J. Z. & Kim, C. C. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC Bioinformatics* **17,** 103 (2016).

225. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18,** 1851–1858 (2008).

226. Jiang, H. & Wong, W. H. SeqMap: Mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24,** 2395–2396 (2008).

227. Smith, A. D., Xuan, Z. & Zhang, M. Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics* **9,** 128 (2008).

228. Weese, D., Emde, A. K., Rausch, T., Döring, A. & Reinert, K. RazerS - Fast read mapping with sensitivity control. *Genome Res.* **19,** 1646–1654 (2009).

229. Ferragina, P. & Manzini, G. Opportunistic data structures with applications. in *Proceedings 41st Annual Symposium on Foundations of Computer Science* 390–398 (IEEE Comput. Soc, 2000). doi:10.1109/SFCS.2000.892127

230. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).

231. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

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

233. Li, R. *et al.* SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25,** 1966–1967 (2009).

234. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).

235. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38,** 1–14 (2010).

236. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

237. Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. & Van De Peer, Y. GenomeView: A next-generation genome browser. *Nucleic Acids Res.* **40,** 1–10 (2012).

238. Manske, H. M. & Kwiatkowski, D. P. LookSeq: A browser-based viewer for deep sequencing data. *Genome Res.* **19,** 2125–2132 (2009).

239. Carver, T., Böhme, U., Otto, T. D., Parkhill, J. & Berriman, M. BamView: viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics* **26,** 676–7 (2010).

240. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).

241. Anders, S. *et al.* Differential expression analysis for sequence count data. *Genome Biol.*

11, R106 (2010).

242. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2009).

243. Zhang, Z. & Wang, W. RNA-Skim: a rapid method for RNA-Seq quantification at transcript level. *Bioinformatics* **30,** i283–i292 (2014).

244. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32,** 462–464 (2014).

245. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. *Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference*. *bioRxiv* (2015). doi:10.1101/021592

246. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

247. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–17 (2012).

248. Wang, W., Qin, Z., Feng, Z., Wang, X. & Zhang, X. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene* **518,** 164–70 (2013).

249. Hu, Y. *et al.* DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.* **41,** e39 (2013).

250. Shen, S. *et al.* MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.* **40,** 1–13 (2012).

251. Vitting-Seerup, K., Porse, B. T., Sandelin, A. & Waage, J. spliceR: an R package for classification of alternative splicing and prediction of coding potential from RNA-seq data. *BMC Bioinformatics* **15,** 81 (2014).

252. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2013).

253. Madsen, J. G. S. *et al.* iRNA-seq: Computational method for genome-wide assessment of acute transcriptional regulation from total RNA-seq data. *Nucleic Acids Res.* **43,** e40 (2015).

254. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18,** 1435–40 (2011).

255. Yu, J. *et al.* Human induced pluripotent stem cells free of vector and transgene sequences. *Science* **324,** 797–801 (2009).

256. Hu, B.-Y. & Zhang, S.-C. Differentiation of spinal motor neurons from pluripotent human stem cells. *Nat. Protoc.* **4,** 1295–304 (2009).

257. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

258. Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42,** 749–755 (2014).

259. Anders, S., Pyl, P. T. & Huber, W. Genome analysis HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

260. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

261. Cereda, M., Sironi, M., Cavalleri, M. & Pozzoli, U. GeCo++: a C++ library for genomic features computation and annotation in the presence of variants. *Bioinformatics* **27,** 1313–1315 (2011).

262. Hiller, M., Pudimat, R., Busch, A. & Backofen, R. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.* **34,** e117 (1-10) (2006).

263. Bailey, T. L. & Elkan, C. Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. *Proc. Second Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36 (1994).

264. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37,** 202–208 (2009).

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

266. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8,** R24 (2007).

267. Ray, D. *et al.* A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499,** 172–177 (2013).

268. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8,** R183 (2007).

269. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stŕmberg, M. P. & Marth, G. T. Bamtools: A C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27,** 1691–1692 (2011).

270. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

271. Setola, V. *et al.* Axonal-SMN (a-SMN), a protein isoform of the survival motor neuron gene, is specifically involved in axonogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 1959–64 (2007).

272. Boido, M. & Vercelli, A. Neuromuscular Junctions as Key Contributors and Therapeutic Targets in Spinal Muscular Atrophy. *Front. Neuroanat.* **10,** 1–10 (2016).

273. Giavazzi, A., Setola, V., Simonati, A. & Battaglia, G. Neuronal-Specific Roles of the Survival Motor Neuron Protein. *J. Neuropathol. Exp. Neurol.* **65,** 267–277 (2006).

274. Witzemann, V. Development of the neuromuscular junction. *Cell Tissue Res.* **326,** 263–271 (2006).

275. Tangsrud, S. E., Carlsen, K. C., Lund-Petersen, I. & Carlsen, K. H. Lung function

measurements in young children with spinal muscle atrophy; a cross sectional survey on the effect of position and bracing. *Arch. Dis. Child.* **84,** 521–4 (2001).

276. Ng, S. Y. *et al.* Genome-wide RNA-Seq of Human Motor Neurons Implicates Selective ER Stress Activation in Spinal Muscular Atrophy. *Cell Stem Cell* **17,** 569–584 (2015).

277. Ruiz, R., Casanas, J. J., Torres-Benito, L., Cano, R. & Tabares, L. Altered Intracellular Ca2+ Homeostasis in Nerve Terminals of Severe Spinal Muscular Atrophy Mice. *J. Neurosci.* **30,** 849–857 (2010).

278. Sigurgeirsson, B., Emanuelsson, O. & Lundeberg, J. Analysis of stranded information using an automated procedure for strand specific RNA sequencing. *BMC Genomics* **15,** 631 (2014).

279. Zhang, H., Xing, L., Singer, R. H. & Bassell, G. J. QNQKE targeting motif for the SMN-Gemin multiprotein complex in neurons. *J. Neurosci. Res.* **85,** 2657–2667 (2007).

280. Smith, R. W. P., Blee, T. K. P. & Gray, N. K. Poly(A)-binding proteins are required for diverse biological processes in metazoans. *Biochem. Soc. Trans.* **42,** 1229–37 (2014).

281. Burgess, H. M. & Gray, N. K. mRNA-specific regulation of translation by poly(A)-binding proteins. *Biochem. Soc. Trans.* **38,** 1517–22 (2010).

282. Burgess, H. M. *et al.* Nuclear relocalisation of cytoplasmic poly(A)-binding proteins PABP1 and PABP4 in response to UV irradiation reveals mRNA-dependent export of metazoan PABPs. *J. Cell Sci.* **124,** 3344–55 (2011).

283. Bhattacharjee, R. B. & Bag, J. Depletion of nuclear poly(A) binding protein PABPN1 produces a compensatory response by cytoplasmic PABP4 and PABP5 in cultured human cells. *PLoS One* **7,** e53036 (2012).

284. Apponi, L. H., Corbett, A. H. & Pavlath, G. K. Control of mRNA stability contributes to low levels of nuclear poly(A) binding protein 1 (PABPN1) in skeletal muscle. *Skelet. Muscle* **3,** 23 (2013).

285. McWhorter, M. L., Monani, U. R., Burghes, A. H. M. & Beattie, C. E. Knockdown of the survival motor neuron (Smn) protein in zebrafish causes defects in motor axon outgrowth and pathfinding. *J. Cell Biol.* **162,** 919–931 (2003).

286. Lim, J. *et al.* Uridylation by TUT4 and TUT7 Marks mRNA for Degradation. *Cell* **159,** 1365–1376 (2014).

287. Nasrin, F. *et al.* HnRNP C, YB-1 and hnRNP L coordinately enhance skipping of human MUSK exon 10 to generate a Wnt-insensitive MuSK isoform. *Sci. Rep.* **4,** 6841 (2014).

288. Akten, B. *et al.* Interaction of survival of motor neuron (SMN) and HuD proteins with mRNA cpg15 rescues motor neuron axonal deficits. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 10337–42 (2011).

289. Farooq, F., Balabanian, S., Liu, X., Holcik, M. & MacKenzie, A. p38 Mitogen-activated protein kinase stabilizes SMN mRNA through RNA binding protein HuR. *Hum. Mol. Genet.* **18,** 4035–4045 (2009).

290. Van Der Giessen, K., Di-Marco, S., Clair, E. & Gallouzi, I. E. RNAi-mediated HuR

Depletion Leads to the Inhibition of Muscle Cell Differentiation. *J. Biol. Chem.* **278,** 47119–47128 (2003).

291. Figueroa, A. *et al.* Role of HuR in skeletal myogenesis through coordinate regulation of muscle differentiation genes. *Mol. Cell. Biol.* **23,** 4991–5004 (2003).

292. Deschênes-Furry, J. *et al.* The RNA-binding Protein HuR Binds to Acetylcholinesterase Transcripts and Regulates Their Expression in Differentiating Skeletal Muscle Cells. *J. Biol. Chem.* **280,** 25361–25368 (2005).

293. Storbeck, M. *et al.* Neuronal-specific deficiency of the splicing factor Tra2b causes apoptosis in neurogenic areas of the developing mouse brain. *PLoS One* **9,** e89020 (2014).

294. Hofmann, Y., Lorson, C. L., Stamm, S., Androphy, E. J. & Wirth, B. Htra2-beta 1 stimulates an exonic splicing enhancer and can restore full-length SMN expression to survival motor neuron 2 (SMN2). *Proc. Natl. Acad. Sci. U. S. A.* **97,** 9618–23 (2000).

295. Brichta, L. *et al.* Valproic acid increases the SMN2 protein level: a well-known drug as a potential therapy for spinal muscular atrophy. *Hum. Mol. Genet.* **12,** 2481–9 (2003).

296. Saito, T. *et al.* A Study of valproic acid for patients with spinal muscular atrophy. *Neurol. Clin. Neurosci.* **3,** 49–57 (2015).

297. Huppertz, I. *et al.* iCLIP: Protein–RNA interactions at nucleotide resolution. *Methods* **65,** 274–287 (2014).

298. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322,** 1845–8 (2008).

299. Mokry, M. *et al.* Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Res.* **40,** 148–158 (2012).

300. Nielsen, R. *et al.* Genome-wide profiling of PPAR :RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes Dev.* **22,** 2953–2967 (2008).