UNIVERSITA' DEGLI STUDI DI UDINE

CORSO DI DOTTORATO DI RICERCA IN

SCIENZE E TECNOLOGIE CLINICHE

CICLO XXVII

PhD Thesis

NGS AND PERSONALIZED MEDICINE

SMN2 target re-sequencing in spinal muscular atrophy patients

Supervisor:	Candidate:
Prof. Francesco Curcio	Giorgia Dubsky de Wittenau

Anno Accademico 2014/2015

CONTENTS

ABBREVIATIONS

ACMG	**American College of Medical Genetics and Genomics standards**
DNA	**DeoxyriboNucleic Acid**
dNTP	**deoxyNucleoside TriPhosphate**
ddNTP	**dideoxyNucleoside TriPhosphate**
DMSO	**DiMethyl SulfOxide**
emPCR	**emulsion Polymerase Chain Reaction**
EMR	**Electronical Medical Record**
ESE	**Exonic Splicing Enhancer**
ESS	**Exonic Splicing Silencer**
GC	**Gene Conversion**
HGP	**Human Genome Project**
kb	**kilo bases**
LR-PCR	**Long Range PCR**
MLPA	**Multiple Ligation Probe Amplification**
MPS	**Massive Parallel Sequencing**
NAIP	**Neuronal Apoptosis Inhibitory Protein**
NGS	**Next Generation Sequencing**
PCR	**Polymerase Chain Reaction**
RNA	**RiboNucleic Acid**
SBE	**Single Base Extension**
SMA	**Spinal Muscular Atrophy**
SMN	**Survival MotorNeuron**
SMN-FL	**Survival MotorNeuron-full length**
SMNΔ7	**Survival MotorNeuron exon 7 deleted**
SNP	**Single Nucleotide Polymorphism**
snRNP	**small nuclear RiboNuclear Protein**
UPD	**UniParental Disomy**

INTRODUCTION

1 Human Genome Project and personalized medicine

1.1 Promises of the Human Genome Project

Over last hundred years, the scientific progress has been characterized by four main phases. The first identified the cellular basis of heredity: the chromosomes. The second defined the molecular basis of heredity: the DNA double helix. The third phase was the discovery of the biological mechanism by which cells read the information contained in genes and the invention of the recombinant DNA technologies of cloning and sequencing. Finally, the last was the deciphering of genes and then the entire genome, leading to the field of genomics (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001).

"Never would I have dreamed in 1953 that my scientific life would encompass the path from DNA's double helix to the 3 billion steps of the human genome. But when the opportunity arose to sequence the human genome, I knew it was something that could be done – and that must be done. The completion of the Human Genome Project is a truly momentous occasion for every human being around the globe" said Nobel Laureate James D. Watson.

The Human Genome Project (HGP) was an international scientific research project. Its primary goal was to determine the sequence of chemical base pairs contained in a haploid reference human genome with a clone-based sequencing strategy. The project was officially launched in 1990 as a 15-year program and

$3 billion plan for completing the genome sequence. Moreover, in 1998 Celera genomics founded a genome-sequencing facility. Their goal was to determine the genome sequence over 3 years with a random whole genome shotgun sequencing strategy. Finally, the significant progress reached in technologies for genome analysis allowed the project to be finished in 2003, two years ahead of schedule.

The Human Genome Sequencing consortium assembly was a composite derived from haploids of numerous donors, whereas the Celera version of the genome was a consensus sequence derived from five individuals. In 2007 Levy and his colleagues presented the diploid genome sequence of an individual human (Sterky F. and Lundeberg J., 2000; Venter et al., 2001; Levy et al., 2007).

Simultaneously, the single nucleotide polymorphism (SNP) consortium in 1999 aimed to create a SNP map and in 2001 created the first catalogue of 1.4 million SNPs in collaboration with Human Genome Sequencing consortium.

The completion of the Human Genome Project and mapping of the human genome single-nucleotide polymorphisms (SNPs) had a tremendous impact on our approach to medicine. Scientists began to speak about treatment tailored to patient with a specific genotype and the inter-individual variability in response to drugs.

This was the bedrock of pharmacogenomics; the emerging field of personalized medicine in which drugs and preventive strategies are specifically tailored to suit an individual's genetic profile (Chiche JD. et al., 2002).

1.2 Pharmacogenetics and pharmacogenomics

Pharmacogenetics is the study of how the actions of and reactions to drugs vary with the patient's genes.

The field of pharmacogenetics had its origin in the 1950s with the discovery of polymorphisms in single-gene controlled enzymes. These abnormalities in enzymes were found to predispose to unexpected adverse drug reactions. In 1957 the role of genetics in causing drug reactions were admitted and in 1959 the term pharmacogenetics was coined by Friedrich Vogel.

The development of this field over the years remained slow, there was no impact on clinical pharmacology, drug development and clinical medicine until the advent of DNA technology and in vitro molecular tests.

The term pharmacogenomics was introduced in the 1990s with the emergence of the HGP and the development of the genome sciences. Since many diseases develop as a result of a network of genes failing to perform correctly, pharmacogenomics can correlate genes with the individual responsiveness to a given drug (Emilien et al., 2000; Motulsky AG, Qi M, 2006).

Up to now, the medicine was largely based on the "one-dose-fits-all" model where patients with the same symptoms often were prescribed the same treatment at the same dose. Although therapeutic successes have occurred with this model, many patients were not responding to treatment. A more patient-centric approach of medical practice has been proposed to reduce unexpected manifestations. A patient's response to a drug is often linked to common genetic variations present in their genes. Knowing the types of genetic variations present in a patient can help predict the associated drug response. This would not only help physicians to individualize drug therapy, but would also help to improve effectiveness of the drug, decrease the chance of negative side effects and save healthcare costs (Abul-Huns et al., 2014).

1.3 D.NAMICA project

The advances in genetics and genomics translate to deeper insights in disease understanding and patient management. Moreover, the personalized approach creates huge amounts of data. All these data must be collected, synthesized, computerized and presented to clinicians.

D.NAMICA is cofounded by ERDF-European Regional Development Fund-Friuli Venezia Giulia Region Operational Program 2007-2013 (www.dinamica.it).

The project aim to create an integrated electronic medical record (EMR), in which specific software tools link clinical data together with "-omics" data. To this purpose, dilatative cardiomyopathy, hepatocellular carcinoma and spinal muscular atrophy have been chosen as pilot studies.

D.NAMICA highlights the awareness of improving genomic research and the importance of SNP discovery for clinical needs. Implementation of genotyping in disease characterization will facilitate the differential diagnosis and distinction of clinically overlapping phenotypes. Moreover, it is realistic to expect that the development of diagnostic or prognostic panels for the detection of new genetic polymorphisms that can be used as biomarkers will accelerate the translation of standardized research protocols from the bench to the patient bedside.

This will lead to the stratification of patients with high risk mutations and to improvement in counseling and patient management, with an early and own-tailored drug therapy. These fascinating opportunities arise from advances in DNA sequencing.

2 DNA sequencing

The HGP requested a dramatic increase in sequence throughout.

The sequencing techniques and especially the enzymatic chain termination method of Sanger have been further developed and adapted to different kinds of automation. Although DNA sequencing by the Sanger method is still regarded as the gold standard for DNA sequencing, the massive parallel sequencing (MPS) has been proposed in 2005 to elevate sequencing to a genome-wide scale (Haas et al., 2011; Sterky F. and Lundeberg J., 2000).

2.1 Sanger sequencing

In 1977, Sanger described a method for sequencing oligonucleotides via enzymatic polymerization; this method was known as dideoxynucleotide method. It generated more easily interpreted raw data and has become the most widely used sequencing technique. Briefly, a ^{32}P-labelled primer was annealed to a specific known region on the template DNA, which provided a starting point for DNA synthesis. Starting from the primer, deoxynucleoside triphosphates (dNTPs) were incorporated by DNA polymerase. This polymerization was extended until the enzyme incorporated a terminator or dideoxynucleoside triphosphate (ddNTP). Roughly, this method was performed in four different tubes, each containing the appropriate amount of one of the four terminators. In each tube, all generated fragments had the same 5'-end whereas the residue at the 3'-end was determined by the dideoxynucleoside used in that reaction. The mixture of different-sized DNA fragments was resolved by electrophoresis on

a denaturing polyacrylamide gel, in four parallel lanes. The pattern of bands showed the distribution of the termination in the synthesized strand of DNA (Sanger et al., 1977; Sterky F. and Lundberg J., 2000; Franca et al., 2002).

Up to now, the original technique took advantage of changes in the way of labelling the fragments, development of nucleotides and labels with new chemical properties, automation of signal detection and engineering of enzymes to obtain more reliable raw data.

As an alternative to radioisotopes, the progress led scientists to find new detection strategies, the chemiluminescent detection and fluorescent dyes. In the first method, the oligonucleotide bound the alkaline phosphatase enzyme. The enzyme catalyzed a luminescent reaction in presence of its substrate. Instead, the second method used fluorescent dye as primer-label or terminator-label.

Initially the fluorescently 5' modified-primer replaced the ^{32}P-labelled primer and each sample was run four times (once for each nucleotide) in both the sequencing reaction and the electrophoretic lane. This multi-run of the same sample made the technique slow. Aiming to reach a higher work rate, were developed a set of four different fluorescent dyes was developed that allowed all four reaction to be separated during the electrophoretic run in a single lane. Finally, the fluorescent moiety was linked to ddNTP instead of a primer. Each terminator binds a different fluorescent dye.

This method is currently used in DNA Sanger sequencing. Briefly, when the labelled terminator is incorporated, the enzyme terminates the extension at the same time. Thus in the same reaction there are four (one for each nucleotide)

terminated ladders with their own respective fluorescent label. This variation allows the single vial polymerization and single lane band separation.

If high throughput is desired, the four –dye system is preferable even if it suffers from the mobility shifts between the different dyes. Therefore, the use of more advanced base-calling algorithms is required to solve this drawback and to retain accuracy. This latter approach was first commercialized into an automatic sequencer by Applied Biosystems (Sterky F. and Lundberg J., 2000; Franca et al., 2002).

More than label and nucleotide chemistry, the DNA sequencing was enhanced by the technology advance. Thus automated DNA sequencers sprang up from the idea to combine together the electrophoretic step with the on-line detection of fluorescently dye-labelled fragments after excitation by a laser beam.

The ahead of time completion of HGP was obtained also by capillary electrophoresis. This technique was ten times faster than conventional gel electrophoresis due to narrow capillaries. Indeed small tubes efficiently dissipated the heat produced during electrophoresis; this allowed the use of a higher electric field. A further advance was the gel-filled capillary electrophoresis which provided high separation efficiency and good selectivity but also multiple injections on a single column (Sterky F. and Lundberg J., 2000; Franca et al., 2002).

To get a genome sequenced, it is essential to break it into pieces small enough to suit the sequencing technology. As shown in Figure 1, there are two major strategies for sequencing of large pieces of DNA. In the first method, the genomic DNA is randomly fragmented into smaller pieces, normally ranging

from 2 to 3 kb, and a library of M13 or plasmid sub-clones are generated. A large number of these clones are randomly isolated and sequenced using standard vector-specific primers. Due to the random nature of this process, the generated sequences overlap in many regions. The shotgun method usually produces a high level of redundancy which affects the total cost. This strategy depends enormously on computational resources to align and assemble all generated sequences to form contiguous stretches of contigs. A variation of this method introduced by Venter et al., 1996 allowed the sequencing of a whole genome at once.

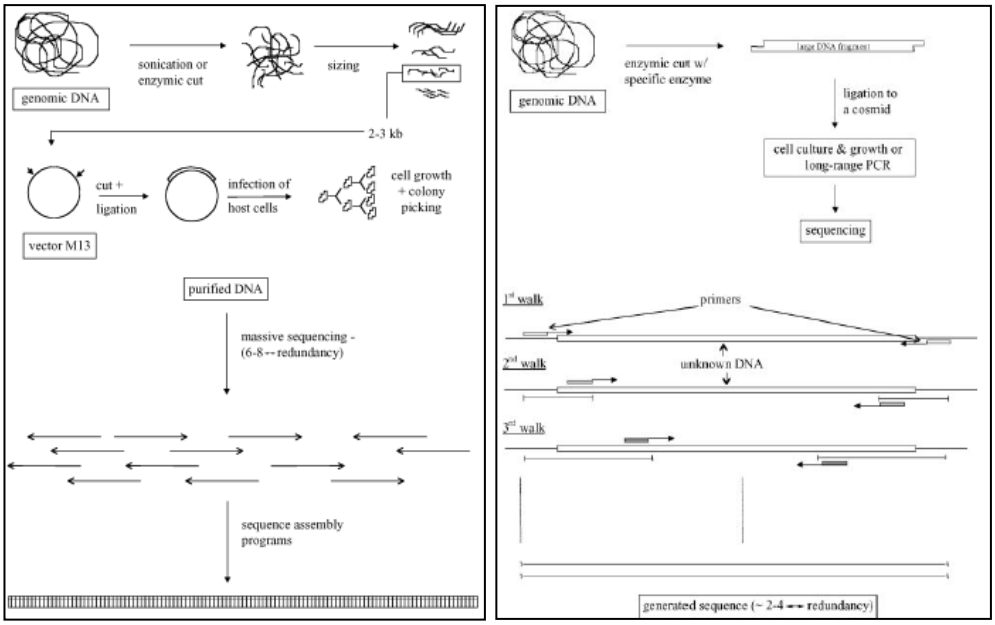


Figure 1 DNA sequencing approaches: shotgun and sequencing by primer walking (from Franca et al., 2002).

A second approach is the directed sequencing approach. In this approach the reaction is performed at a known position of the template with a first primer hybridizing to this known region. The second priming site is then chosen inside the newly generated sequence on the same strand with the same direction.

This method is also called primer walking and its major advantage is the reduced redundancy in comparison to the random approach. Moreover, the assembly of contigs is simpler since the exact position for each sequence reaction is known (Sterky F. and Lundberg J., 2000; Franca et al., 2002).

This very accurate method was described by Casals as follows: "one of the strengths of Sanger sequencing is the very low error rate; it is still considered the gold standard for nucleic acids sequencing. New mutations identified by Next-generation sequencing (NGS) technologies are validated using Sanger sequencing" (Casals et al., 2012).

2.2 Massive parallel sequencing

The decision of scientists to aim at the sequencing of the entire human genome, there was a huge impact on the development of techniques that allowed higher sequencing throughput and speed. As such, these efforts resulted in the ability to sequence in only some weeks what the Sanger technique did in several years.

A patent application by EMBL described a large-scale DNA sequencing technique without gels. This patented method used the extension of primers in "sequencing-by-synthesis, addition and detection of the incorporated base". The so-called "reversible terminators" were firstly proposed and described to reach speed and efficiency during polymerization (Ansorge W.J., 2009).

The DNA amplification without cloning and DNA sequencing without chain termination were the new concepts introduced by NGS. The next-generation high-throughput DNA sequencing techniques were selected by *Nature Methods* as the method of the year in 2007, due to the powerful way in which they

revolutionized the way of thinking of scientists and the broad range of their applications (Ansorge W.J., 2009).

2.2.1 Next-generation DNA sequencing platforms

Today, different combinations of methods for amplification, sequencing, and detection are used by next-generation sequencers from the three market leading providers, as briefly described subsequently.

Roche 454 sequencer was the first NGS system on the market in 2005.

It measures pyrophosphate that is released when single nucleotides are added to the nascent DNA chain, in a light-generating reaction known as pyrosequencing.

Briefly, the DNA fragments of a library (library preparation will be described in the next paragraph) are ligated through specific adapters to agarose beads and emulsion PCR (emPCR) is carried out for fragment amplification. Each agarose bead with millions of oligomers attached to its surface is loaded into a picotiter plate. Polymerase enzyme and primers are added to the beads and imaging of the light flashes from luciferase activity records which templates are adding that particular nucleotide. Moreover, the light emitted is directly proportional to the amount of nucleotides incorporated (Mardis E.R., 2008; Ansorge W.J., 2009; Haas et al., 2011). The aforesaid sequencing workflow is resumed in the upper panel of the Figure 2.

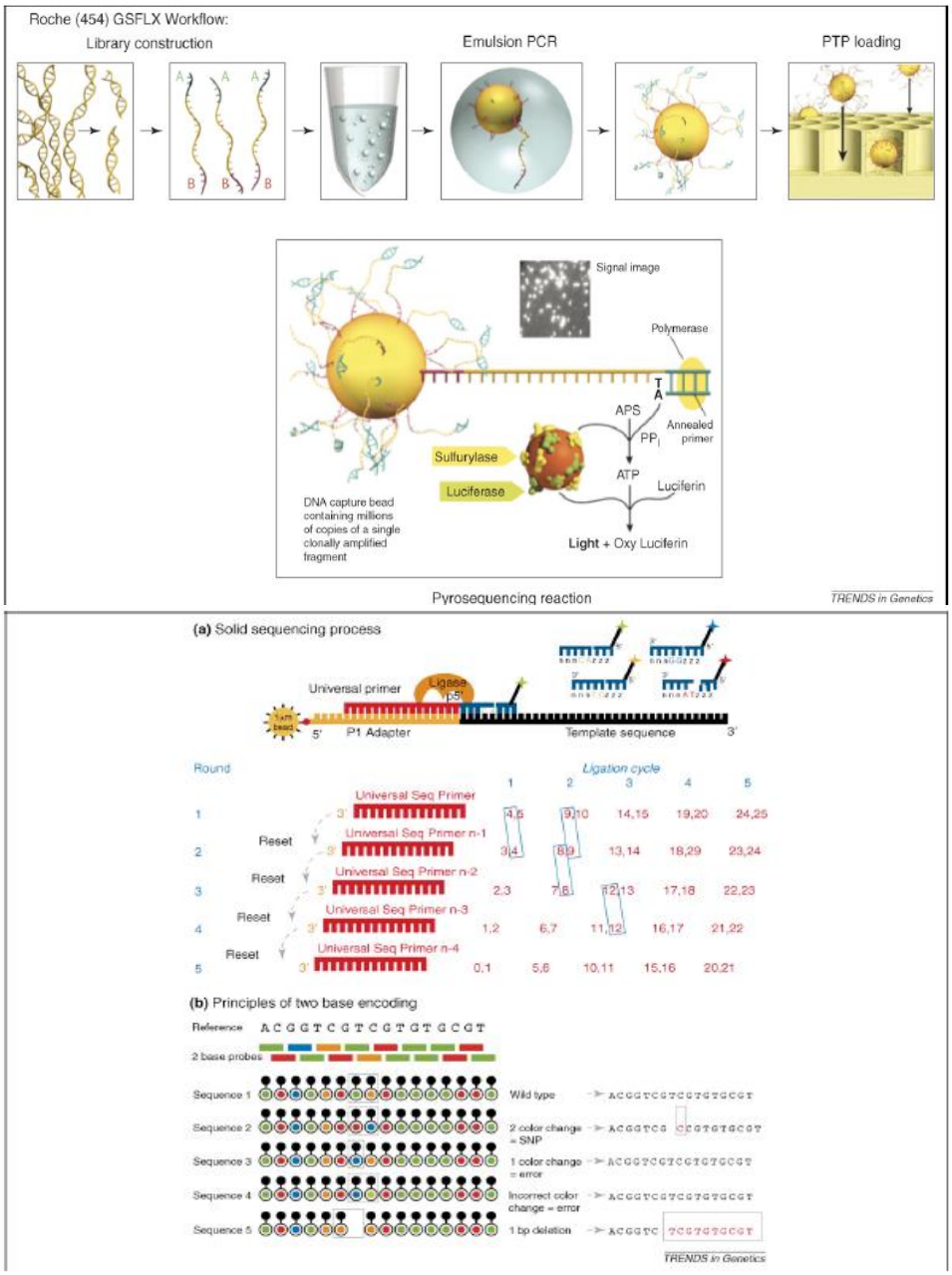


Figure 2 ***Upper panel****:* ***Roche 454 workflow.*** *Library construction ligates 454-specific adapters to DNA fragments and couples amplification beads with DNA in an emulsion PCR to amplify fragments before sequencing. The beads are loaded into the picotiter plate. The Pyrosequencing reaction occurs on nucleotide incorporation to report sequencing-by-synthesis.*
Lower panel: Sequencing-by-ligation, the SOLiD workflow. *A universal primer hybridizes to the SOLiD-specific adapter and four dye-labeled probes are added into the reaction. When one probe hybridizes to the target fragment, it is ligated to the universal primer and the fluorescence is detecteted. The sequence grows up two bases at a time and all new sequencing cycles use one-base smaller probes.*

SOLiD (**S**equecing by **O**ligo **Li**gation and **D**etection) **sequencers** achieved commercial release in 2007; it is the unique sequencing process catalyzed

by DNA ligase. This method is the commercial NGS system with the second highest market share. The workflow of this NGS technique is outlined in the lower panel of the Figure 2. Firstly, an adapter sequence at one end of each DNA fragment of the library, which is complementary to oligonucleotide, is laid on the beads surface. In the second step, all fragments on the beads are amplified by emulsion PCR and the beads are covalently bound on the surface of a glass support surface. The ligation-based sequencing uses a universal sequencing primer that is complementary to the SOLID specific adapters on the library fragments. Next, a mixture of octamers is also added into the sequencing reaction; one of the four fluorescent labels uniquely characterizes the doublet of fourth and fifth bases of the 8-bases fragment. When a matching octamer hybridizes to the DNA fragment sequence adjacent to the universal primer 3' end, the DNA ligase seals the phosphate backbone. After the ligation step, a fluorescent readout identifies the two bases. A subsequent chemical cleavage step removes the dye-carrying nucleotides that also block the further extension. Thereby removing the fluorescent group makes the DNA accessible again for a new ligation round. The process occurs in steps that identify the sequence of each fragment at five nucleotide intervals. Then, the sequencing is further continued in the same way with another primer, shorter by one base than the previous one (Mardis E.R., 2008; Ansorge W.J., 2009; Haas et al., 2011; Ropers H.H., 2012).

Finally the market leader, **Illumina sequencers** employs an ingenious procedure to generate millions of clonally amplified, single-stranded DNA fragments on the surface of a slide, or flow cell. The sequencing is based on reversible dye terminators that only allow one nucleotide to be added at a time. After each step,

fluorescence signals of all clones are monitored and stored as images (Haas et al., 2011; Ropers H.H., 2012). Extensive details on the Illumina sequencing are reported in the following paragraphs.

2.2.2 Illumina sequencing

One of the most established and widely-adopted NGS technologies, is the one developed by Illumina. The first Illumina sequencer, introduced in 2006, was the Genome Analyzer. This sequencer was based on the concept of "sequencing by synthesis" (SBS) to produce sequence reads of ~32-40 bp from tens of millions of surfaced-amplified DNA fragments simultaneously (Mardis et al., 2008). In particular, the Genome Analyzer uses a flow cell consisting of an optically transparent slide with 8 individual lanes on the surfaces to which are bound oligonucleotide anchors (Figure 3A). Template DNA is fragmented into lengths of several hundred base pairs and different method are used to allow the ligation of the DNA fragments to specific oligonucleotide adapters. These adapters are complementary to the flow-cell anchors in order to enable the ligation of the template DNA to the flow cell. DNA templates are then amplified in the flow cell by "bridge" amplification, which relies on captured DNA strands "arching" over and hybridizing to an adjacent anchor oligonucleotide. Multiple amplification cycles convert the single-molecule DNA template to a clonally amplified arching "cluster". For sequencing, the clusters are denatured, and a subsequent chemical cleavage reaction and wash leave only forward strands for single-end sequencing (Figure 3A). A primer complementary to the adapter sequences let the sequencing reaction start and the polymerase with

a mixture of 4 differently colored fluorescent reversible dye terminators are added. The terminators are incorporated according to the sequence complementarity in each strand in a clonal cluster. After incorporation, excess reagents are washed away and the fluorescence is recorded. With successive chemical steps, the reversible dye terminators are unblocked, the fluorescent labels are cleaved and washed away, and the next sequencing cycle is performed (Figure 3B).

This sequencing-by-synthesis process required approximately 2.5 days to generate read lengths of 36 bases and the overall sequence output was >1 Gb per analytical run (Bentley *et al.*, 2008). Subsequent to the first Genome Analyzer, new platforms with an improved throughput were launched by Illumina: the Genome Analyzer IIx, which can generate up to 90 Gb per analytical run with read lengths of 150 bp; and the latest version the ultra-high throughput HiSeq2500 that can produce 1000 Gb per run with read lengths of 150 bp.

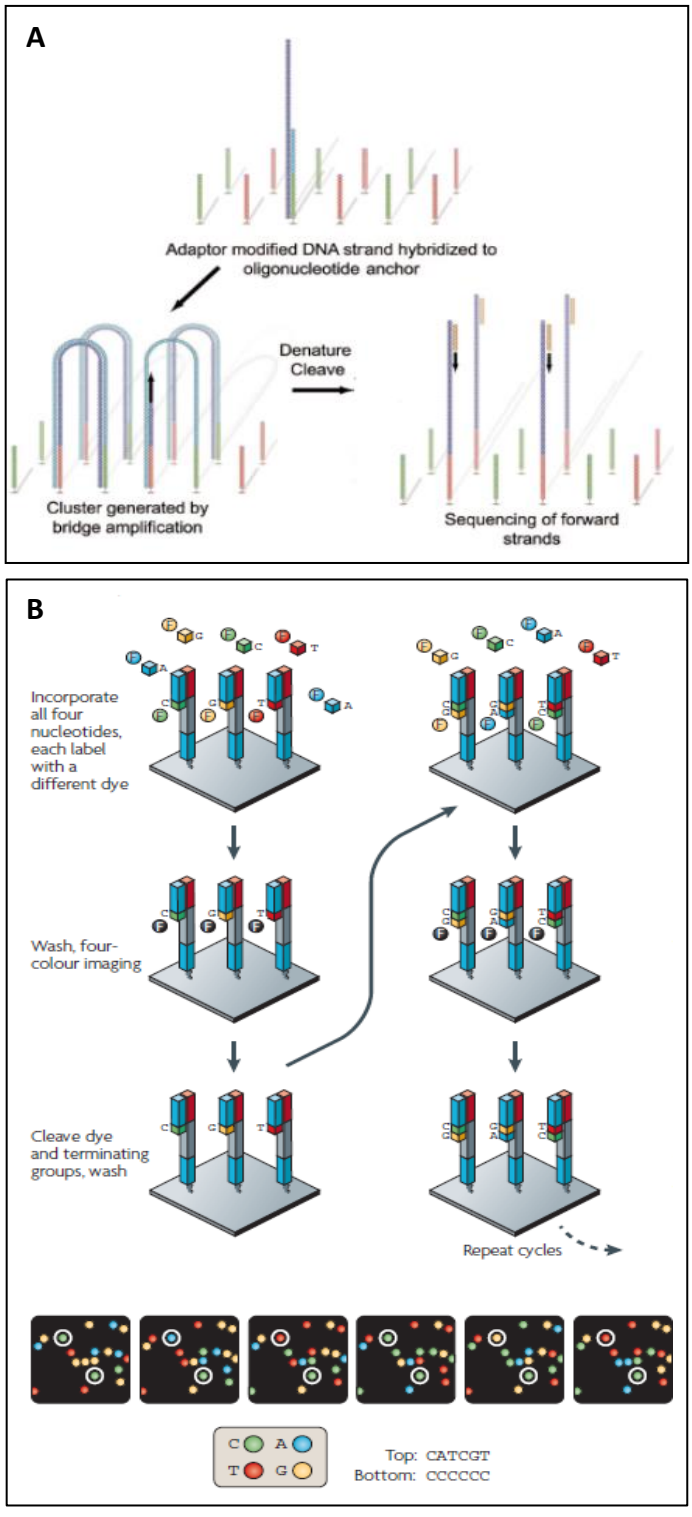


Figure 3 Illumina sequencing. Panel A) Adapter modified, single-stranded DNA is added to the flow cell and immobilized by hybridization. Bridge amplification generates clonally amplified clusters. Clusters are denatured and cleaved; Panel B) The sequencing is initiated with addition of primer, polymerase and 4 reversible dye terminators. The fluorescence is recorded after incorporation. Finally, the fluorophore and the block are removed before the next synthesis cycle.

Illumina also offers the possibility to sequence both ends of the template molecule. Such paired-end sequencing provides positional information that facilitates alignment and assembly (Campbell et al., 2008). An interesting application of Illumina NGS is the "multiplex sequencing", in which different DNA samples are sequenced in the same lane. Considering the high throughput of the latest sequencing machine, multiplexing is very useful when targeting specific genomic regions. In the multiplexed sequencing method, DNA libraries are tagged with a unique DNA sequence, or index, during sample preparation. Multiple samples are then pooled into a single lane on a flow cell and sequenced together in a single run. Samples are then divided by tag sequences that uniquely identify each of them during the analysis step.

Before sequencing, there are three important steps in preparing DNA: the nucleic acid fragmentation into the desired length, the adapter addition to the target fragment and the quantification of final library product for sequencing. Usually, library preparation takes advantage of physical or enzymatical fragmentation of DNA. The first consists of acoustic shearing and sonication of nucleic acids, while the second group is characterized by the use of non-specific endonuclease cocktails or transposase tagmentation reactions. The library size is mainly determined by the desired insert size because the length of adaptor sequences is a constant. This size must be compatible with limitations of Illumina instrumentation and sequencing application. Mainly, the optimal size of a library is the one that is suitable to cluster generation. Indeed, shorter products amplify more efficiently than longer products; such longer library inserts generate a more diffuse and homogenous-distribution of clusters on the flow cell surface.

As example of a sequencing application, paired-end sequencing requires a proper size to sequence without overlapping read pairs. In later years, commercial kits as Nextera (Adey et al., 2010; Caruccio N., 2011) and Haloplex (Mamanova et al., 2011; Mertes at al., 2011) allow the simultaneous occurrence of DNA fragmentation and adaptors connection (Head et al., 2014). Finally, DNA sample preparation ends with the quantification of the obtained library, particularly in multiplex sequencing that require the equimolar pooling of all samples (Head et al., 2014).

2.2.3 Impact of NGS

Sanger sequencing of the target region is still an accurate and cost effective way to obtain a molecular diagnosis in single gene disorders. Up to now, the identification of known alteration by Sanger sequencing is the main molecular test to support clinical suspicions. However, in 2012 Zhang and colleagues described how nowadays the selection of candidate gene(s) for sequence analysis is extremely difficult, as most inherited disorders exhibit genetic and clinical heterogeneity. The NGS technologies offer the way to reduce the stepwise sequencing and the time for diagnosis.

The broadest application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease, also allowing large-scale comparative and evolutionary studies to be performed. In particular, previous studies indicate that NGS can reliably detect variants with a frequency below 1 % (Druley et al., 2009).

The cost of genome sequencing is becoming low enough to make personal genomics a reality. In fact, the first human genome was about $3 billion, while the James Watson's genome was completed for less than $1 million. From 2008, the cost of sequencing dropped faster than what would have been expected from Moore's law (as shown in Figure 4) in contrast to storage, which is decreasing in line with this law.

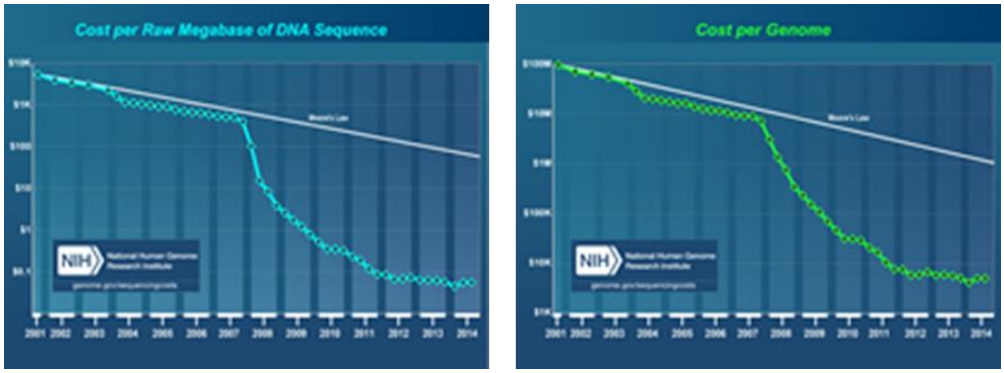


Figure 4 Cost per Mb and per genome calculated by NHGRI. Decreasing cost of sequencing in last 13 years compared with the expectation if it had followed Moore's law.

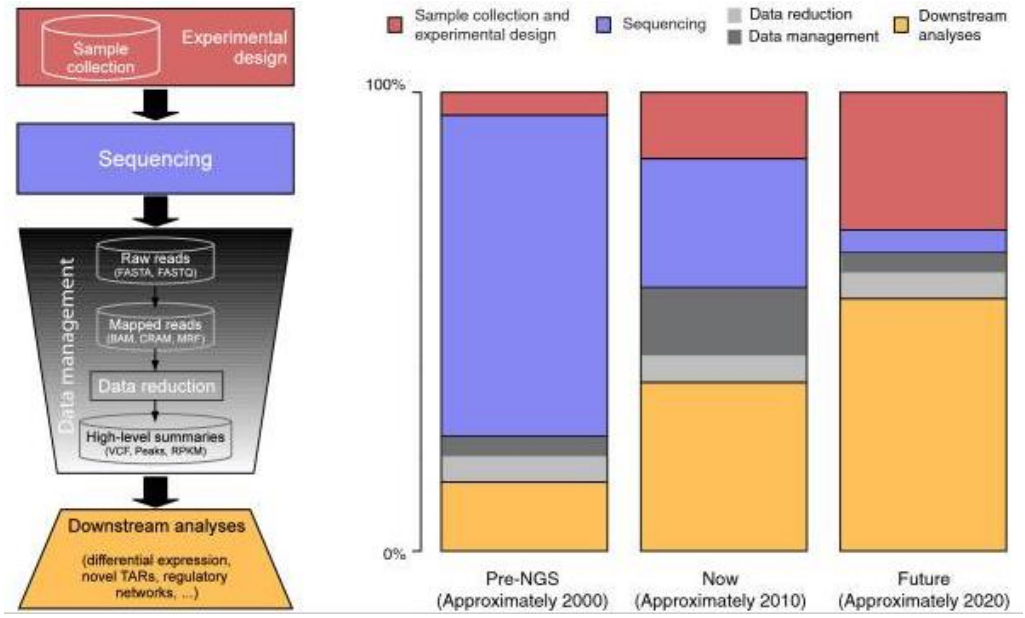


Figure 5 Contribution of each step to the overall cost of a sequencing project across time. Sboner et al., 2011.

The bottleneck is the data storage, downstream analyses and interpretation of results. As represented in Figure 5, in the future these activities will be the major contributor to the overall cost of a sequencing project (Sboner et al., 2011).

Moreover, careful considerations about privacy of the data and network bandwidth should be done before entering clinical diagnostic, especially for whole human sequencing (Sboner et al., 2011).

The high throughput of NGS allows targeted re-sequencing of a number of selected regions in a large number of pooled individuals (Ingman M. and Gyllensten U., 2009). Thus, in the contest of clinical applications it is often desirable to enrich a group of genes known to be related with a particular phenotype instead of whole human genome (Zhang et al., 2012).

Researchers have already produced panel kits that identify mutations in known genes for deafness, neurofibromatosis, retinitis pigmentosa, Marfan syndrome, X-linked intellectual disability, hereditary spastic paraplegias, mitochondrial disorders, dilated cardiomyopathy and many other recessive childhood pathologies (Ropers HH., 2012; Zhang et al., 2012).

3. *SPINAL MUSCULAR ATROPHY*

Spinal Muscular Atrophy (SMA) is a neurodegenerative disease that represents the most common genetic cause of infant mortality. After Cystic Fibrosis it is the most common autosomal recessive disorder in humans, with a carrier frequency of approximately 1 in 40 (Wirth et al., 2006) and therefore an incidence of 1 in 6000 in the human population (Pearn J., 1978).

It is characterized by progressive loss of α-motor neurons (see Figure 6) that take origin from anterior horn cells of the spinal cord, resulting in atrophy of the proximal muscles of the limbs and trunk that can lead either to paralysis or death (Gubitz et al., 2004).

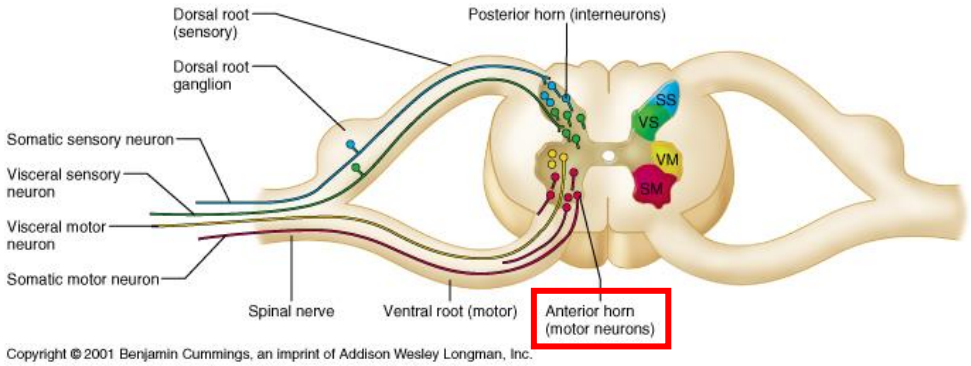


Figure 6 Schematic representation of motor neurons of spinal cord.

There are numerous other forms of spinal muscular atrophy which are genetically distinct and often affect different subsets of neurons and muscle. They include autosomal dominant forms of the disease (Sambuughin et al., 1998; Van der Vleuten et al., 1998), X-linked forms (Fischbeck et al., 1991), recessive forms that affect the distal muscles (Viollet et al., 2002), and a severe form of SMA (SMARD) with respiratory distress (Grohmann et al., 2001).

3.1 Classification and clinical diagnosis

Spinal muscular atrophy is clinically distinguished and classified in four phenotypes that reflect the different age of onset and severity of disease (Munsat TL. and Davies KE., 1992; Zerres K. and Rudnik-Schoneborn S., 1995):

TYPE I (Werdnig-Hoffmann disease) [# OMIM 253330]

This is the most severe form of SMA, also known as acute spinal muscular atrophy. The onset ranges from birth to age six months. Patients are never able to sit without support, to raise legs or lift their heads up and present symmetric muscle weakness, lack of motor development and hypotonia. It is the most common genetic cause of infant mortality in northern Europeans. Most deaths occur at two years for respiratory failure (Thomas NH. and Dubowitz V., 1994).

TYPE II (Dubowitz disease) [# OMIM 253550]

This is known as chronic spinal muscular atrophy and is considered an intermediate form of SMA. It is characterized by an age of onset between six and twelve months with a development usually normal before that time. Patients are able to sit but are unable to stand or walk unaided at any time. Therefore they often develop spinal deformity. Death occurs usually between ten and forty years of age (Russman BS., 2007).

TYPE III (Kugelberg-Welander) [# OMIM 253400]

Also known as juvenile spinal muscular atrophy, it is considered the mildest form. This type of SMA is characterized by a rather variable age of onset mainly in the first two decades, while only few patients start with symptoms between twenty

and thirty years. Patients show muscular weakness but are able to stand and to walk without support and show prolonged survival rates (Zerres et al., 1997).

TYPE IV [# OMIM 271150]

The type IV is usually associated with onset at older than ten years and is considered the adult form of SMA. It is characterized by a mild weakness and features similar to Type III (Brahe et al., 1995; Clermont et al., 1995). Patients present a normal life expectancy.

Prognosis depends on the phenotypic severity, ranging from high mortality within the first year for SMA type I to no mortality for the chronic and later-onset forms. In the past decades clinical diagnosis was based on muscular biopsy and electromyography that recorded electrical activity produced by skeletal muscles. Nowadays both techniques are considered invasive and currently used only to confirm clinical diagnosis as well as molecular techniques that do not instead involve any risk to the patient.

3.2 Genetics of spinal muscular atrophy

Spinal muscular atrophy is a recessive disease hence two copies of the mutated gene are necessary to have pathological traits: one inherited from the mother and one from the father. Such patients are genotypically homozygous and take origin most of the times from heterozygous parents who possess one normal allele and one mutant allele. In most cases, the parents are perfectly healthy because the mutant allele has no adverse effect when a normal allele is also present. The parents are said to be carriers of the disease. Therefore there is one out of four

or twenty-five percent chance to have an affected child with each pregnancy and a two out four chance to generate a healthy carrier child (see Figure 7).

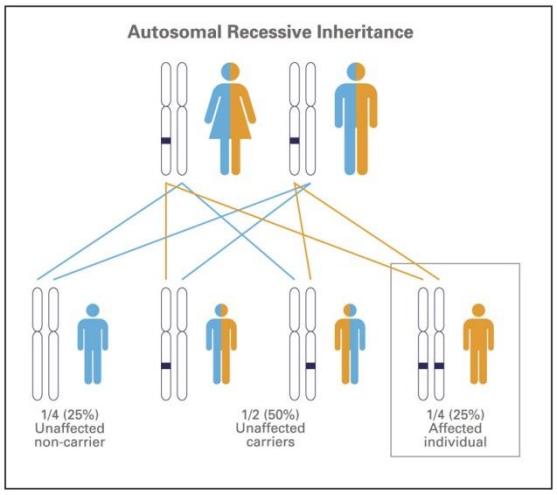


Figure 7 Schematic Representation of Autosomal Recessive Inheritance

3.3 Gene detection and localization

The identification of genes involved in SMA was complicated by the highly complex and unstable nature of the genome where they localize. Therefore, phenotype changes severity from real severe form in type I to very mild within type III and type IV (Pearn J., 1980).

By means of linkage analysis, all forms of SMA were mapped to chromosome 5q11.2 - q13.3 (Brzustowicz et al., 1990; Melki et al., 1990; Clermont et al., 1994). This region was found to be characterized by the presence of low copy repeat elements (Kleyn et al., 1993; Melki et al., 1994).

The further construction of a YAC contig encompassing the SMA locus enabled the detailed physical mapping of this region and led to the identification of a large inverted duplication of a 500 kb element within this region. As Figure 8 below

shows, this duplication was then divided in a telomeric and a centromeric element, named respectively E^{tel} and E^{cen} (Melki et al., 1994).

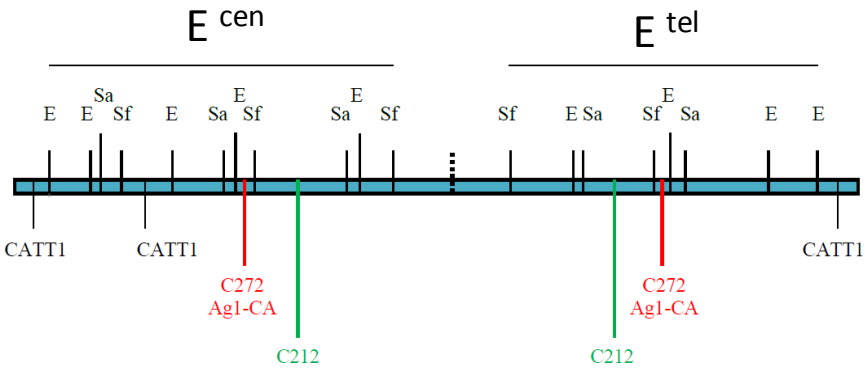


Figure 8 Schematic representation of DNA and microsatellite markers (indicated respectively above and below the genomic map of the 5q13 region) providing evidence for an inverted duplication of an element

A critical function for the survival motor-neuron gene (*SMN*) was strongly suggested by the fact that this gene was lacking or interrupted in most of the patients, suggesting that it represents the SMA-determining gene. Further studies supported the relevance of the SMN protein for motor neuron function and its pathogenetic role in the neuronal degeneration associated with SMA.

Another duplicated gene, the Neuronal Apoptosis Inhibitory Protein gene (*NAIP*), is present as a telomeric functional copy, a pseudo *NAIP* copy, and several truncated copies in the SMA region. It is deleted in 45% of Type I patients and in 18% of Types II and III, but also in 2% of unaffected carrier individuals. Thus, homozygous loss of *NAIP* is not sufficient to cause the disease. Other genes were finally mapped to this area: *GTF2H2*, which encodes the p44 subunit of transcription factor TFIIH (Bergin et al., 1997) and *SERF1*, small EDRK-rich factor 1, which is known to interact with RNA (Scharf et al., 1998). One copy of

the p44 gene is deleted in at least 15% of all SMA cases although its role in the disease is still unknown.

In conclusion, this duplicated region contains at least four genes and repetitive elements (see Figure 9) which makes it prone to rearrangements and deletions.

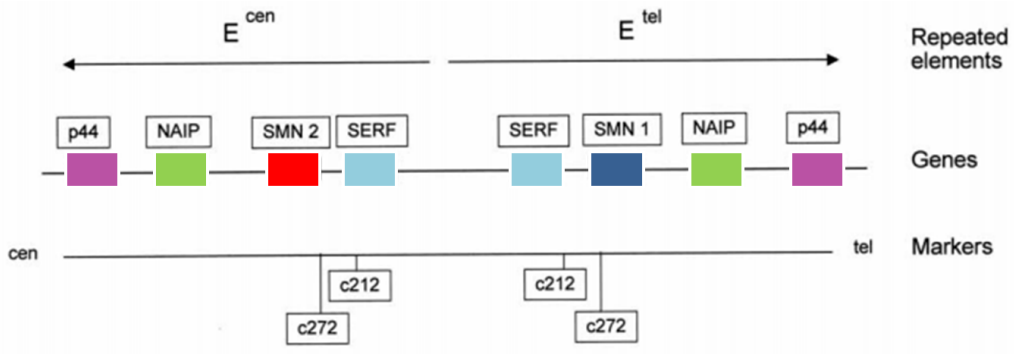


Figure 9 Scheme of the SMA locus which consists of two inverted repeated elements each containing four genes

3.4 *SMN* genes

All SMA subtypes have been mapped to chromosomal region 5q11.2-13.3 which shows a complex genomic structure, including a 500-kilobase duplication and inversion (Lefebvre et al., 1995). The *SMN* [Entrez GeneID n. 6606] has been shown to be the primary SMA-determining gene. It exists as two highly homologous copies:

- the telomeric copy (*SMNt* or *SMN1*) [OMIM 600354]
- the centromeric copy (*SMNc* or *SMN2*) [OMIM 60127]

Both genes are ~ 27 kilobase long and comprised of nine exons with only five base-pair differences within their 3' ends. Despite the high homology, only *SMN1* is necessary for the survival of motor neurons. The only critical difference between the two SMN genes is the C→T base change inside exon 7 which affects

the splicing pattern of the genes (Monani et al., 1999). Althought both SMN genes are transcribed, their transcripts and proteins result different (Bürglen et al., 1996).

Approximately 94% of individuals with clinically typical SMA are homozygously deleted for the telomeric copies hence these patients lack both copies of *SMN1* (Wirth B., 2000). Loss of *SMN1* can occur by deletion or by conversion to *SMN2*, namely Gene Conversion (GC) (Bussaglia et al., 1995; Cobben et al., 1996; Hahnen et al., 1996). The other 5% of affected individuals are known as compound heterozygotes and present deletion of one *SMN1* allele and a point mutation on the other *SMN1* allele.

The SMN2 gene can be present in zero to six copies per genotype. 5-9% of the healthy population carries a homozygous deletion of *SMN2;* however no pathology is associated with that genetic feature. The severity of SMA has been proven to be influenced by the number of *SMN2* copies: about 70% of type I SMA patients carry two *SMN2* copies, 82% of type II SMA patients have three *SMN2* copies, whereas type III patients have with very few exceptions a minimum of three to four *SMN2* copies (Feldkötter et al., 2002; Mailman et al., 2002; Ogino et al., 2003). Since each SMA patient retains at least one *SMN2* copy and since the number correlates with the phenotype, *SMN2* is considered as an interesting gene target for therapy (Wirth B., 2002).

The observation that complete knockout of SMN in any organism is embryonically lethal confirms that some SMN protein, either deriving from the *SMN1* or the *SMN2* gene, is required for survival of all cells. In order to increase SMN level, many animal models have been developed in past years

(Van Meerbeke JP. and Sumner CJ., 2011). In particular, many mouse models expressing a range of SMN protein levels and extensively covering the severe and mild types of SMA have been developed. Neurological and physiological manifestations of the disease support the relevance of these models (Bebee et al., 2012).

3.4.1 Allelic dispositions of *SMN1* and *SMN2*

The *SMN* can be present at different allelic dispositions, their disposition determine the classification of patients.

Healthy individuals are characterized by two SMN1 copies and variable number of SMN2 (as shown in **Error! Reference source not found.**).

Affected individuals (see **Error! Reference source not found.**) instead have no SMN1 genes but they do have variable SMN2 copy genes; the *SMN2*-copy number is often related to the severity of the phenotype. Loss of *SMN1* can occur either by deletion or by conversion to *SMN2*. Gene conversion has been postulated as a major mutational mechanism in SMA, leading to a replacement of *SMN1* by *SMN2* (Hahnen et al., 1996; Burghes AH., 1997; Campbell et al., 1997; DiDonato et al., 1997).

Moreover, there are rare cases of asymptomatic individuals who lack both copies of *SMN1* but without any symptoms (Brahe et al., 1995; Cobben et al., 1995; Hahnen et al., 1995; Wang et al., 1996; Somerville et al., 1997). These cases suggest that other phenotype modifiers could be involved.

Figure 10 Representation of SMN1 and SMN2 allelic disposition in healthy individual.

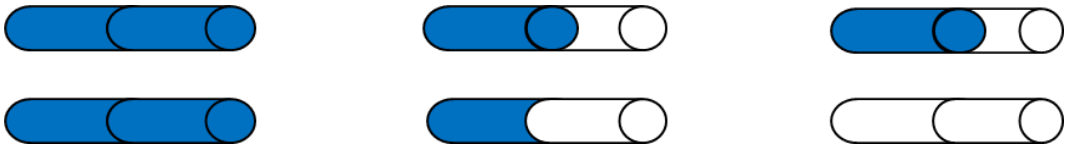


Figure 11 Representation of SMN2 allelic distribution in SMA affected individual.

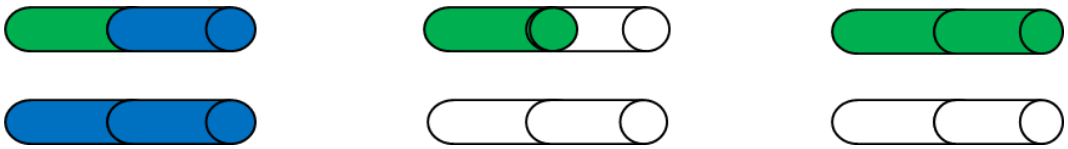


Figure 12 Representation of SMN1 and SMN2 allelic distribution in healthy SMA carriers.

The healthy carrier has one *SMN1* copy and variable number of *SMN2* (Figure 12). The frequency of one-copy carriers in the general population is estimated at about 1/40 (Wirth et al., 2006).

However, the probability of having an affected offspring where one of the parents is a carrier, having one *SMN1* copy, and the other parent having two SMN1 genes is not zero. In fact, the following cases do complicate carrier diagnosis in some SMA families:

- Individuals presenting the "2/0 genotype", where both SMN1 alleles map on the same chromosome 5. The frequency of such silent carriers, presenting both SMN1 genes in *cis*, is about 2-5% of SMA carriers (Alias et al., 2014; Luo et al., 2014). The recurrence risk is approximatively 25% with each pregnancy as for two one-copy carrier parents.

- Unusual mechanisms like Uniparental Disomy (UPD) or de novo mutation.

The UPD takes place when both homologues of a chromosomal region/segment are inherited from only one parent (Engel E., 1980). The incidence of UPD of any chromosome is estimated to be about 1:3500 live births (Robinson WP., 2000). Two different types of UPD can be distinguished: uniparental heterodisomy (heteroUPD) in which both homologues are inherited from one parent, and uniparental isodisomy (isoUPD) which represents the inheritance of two copies of a single parental homologue. Whereas the former is due to non-disjunction in meiosis I, the latter is the consequence of a non-disjunction in meiosis II. Both forms can cause disordered imprinting and isoUPD may also result in a child inheriting a homozygous mutation from a heterozygous parent, and thus being affected by a recessive condition (Poke et al., 2013). For instance, more than 50 patients with different recessive disorders due to isodisomy have been reported (Engel E., 2006). It may be considered as a possible mechanism involved in SMA when conducting prenatal testing and genetic counseling for this disorder (Brzustowicz et al., 1994). Moreover, isoUPD may involve the whole chromosome or may be segmental, where the homozygosity usually goes from the recombination point to the telomere. Different mechanisms can lead to UPD, either during mitotis or meiosis (Yamazawa et al., 2010).

De novo SMN1 mutations and rearrangements happen into one allele, resulting in the homozygous absence of the *SMN1* gene (2% of SMA patients). Thus, only one parent is a *SMN1*-deletion carrier. The majority of the reported *de novo* mutations are paternal in origin deletions, the consequence of a crossover that results in the

loss of the *SMN1* gene. Another possibility is that the *SMN1* is lost due to a *de novo* gene conversion event, where *SMN1* is converted into *SMN2* (Smith et al., 2007). There are basically three mechanisms that can determine de novo mutations in SMA: unequal crossing-over between homologous chromosomes, intrachromosomal deletion or gene conversion.

Because of its duplicated structure, this region seems susceptible to these types of mechanisms. It appears that unequal recombination causing larger deletions is associated with more severe phenotypes, whereas intrachromosomal deletions and gene conversions are associated with milder SMA (Wirth et al., 1997).

It is indeed very important for parents who have been shown to carry two SMN1 genes to undergo extended analysis to distinguish between the "2/0 genotype" and the above mentioned unusual mechanisms, as it results in a different impact on the potential risk in future pregnancies. A molecular approach on parents and blood relatives of SMA patients, combining dosage and linkage analysis, could be useful to characterize an individual with two SMN1 gene copies as a "2/0" carrier or as a *de novo* mutation, thus optimizing genetic counselling. Moreover, VNTR analyses (e.g. microsatellite or minisatellite) with markers flanking the SMA locus are currently used in uniparental disomy detection.

A comprehensive testing procedure for SMA carrier testing requires indeed a multi-modal approach that involves SMN gene-dosage, linkage analysis and genetic risk assessment (Smith et al., 2007).

3.4.2 Differences between *SMN1* and *SMN2*

The two SMN copies (*SMN1* and *SMN2*) display a high level of homology that includes intronic and promoter sequences (Monani et al., 1999). Telomeric SMN (SMNt or *SMN1*) and centromeric SMN (SMNc or *SMN2*) genes differ by five nucleotide changes (Bürglen et al., 1996) as shown in Figure 13. Three changes are located within introns 6 and 7, while two are located in exons 7 and 8.

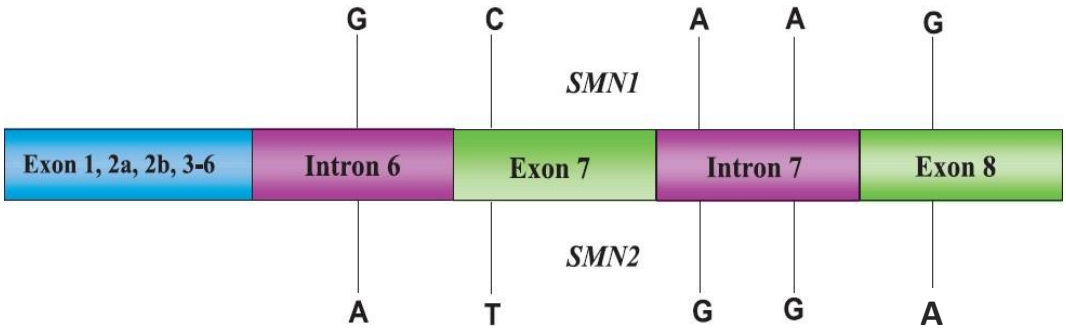


Figure 13 Nucleotide differences between telomeric (SMN1) and centromeric (SMN2) copy of SMN gene.

Neither of the two exonic base differences changes the predicted aminoacid sequence because substitutions are all conservative. Although translationally silent, the C → T transition in exon 7 of *SMN2* leads to frequent exon 7 skipping during splicing of *SMN2*-derived transcripts so that less full-length protein is expressed. For that reason, *SMN2* cannot rescue the pathological phenotype because is not sufficient to fully compensate the loss of *SMN1* in SMA patients (Lorson et al., 1999; Monani et al., 1999).

The exonic base-pair exchanges allow to clearly distinguishing *SMN1* from *SMN2* and are currently used for direct diagnosis of SMA. The *SMN2* is dispensable because approximately 5% of normal individuals lack both copies (Lefebvre et al., 1995).

3.4.3 Alternative *SMN2* splicing

Both *SMN1* and *SMN2* maintain identical coding sequences. The full-length cDNAs of the two genes are identical except for single nucleotide differences in exons 7 and 8, yet their transcriptional products are not the same. In particular, the silent cytosine-to-thymine (C → T) transition within exon 7 (position +6) induces the alternative splicing event common to the majority of SMN2-derived transcripts. SMN1 produces a majority of the full-length cDNA whereas SMN2 produces mostly transcript lacking exon 7 (see Figure 14 below).

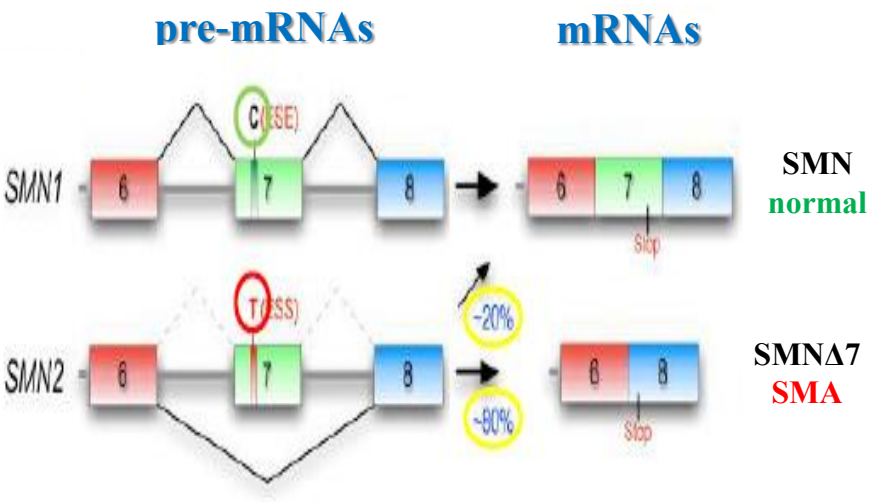


Figure 14 Alternative splicing of the SMN2 gene.

Exon 7 is a highly regulated region comprised of 54 nucleotides and contains the translation termination signal for all full-length products, whereas the translational termination of the exon-skipped product is at the 5' end of exon 8.

As a consequence of the alternative splicing, two different transcripts are encoded by *SMN2*. This gene produces 80-90% of truncated protein SMNΔ7 and 20-10% of full-length protein, which is typically associated with milder phenotype, such as SMA Type II and Type III, when the copy number of the SMN2 gene is increased (McAndrew et al., 1997; Feldkötter et al., 2002). This makes *SMN2* a genetic modifier of disease severity and an attractive molecular target in therapeutic strategies.

It should not be assumed that all SMN2 genes are equivalent, and sequence changes found within this gene should be further investigated for potential positive or negative effects on its transcription and post-transcriptional RNA processing (Prior et al., 2009).

SMNΔ7 protein is not sufficient to rescue the pathological phenotype because does not oligomerize efficiently, is unstable and leads to motor neuron cell death in patients lacking of both *SMN1* copies. Two possible explanations are proposed for the alternative *SMN2* splicing, both represented in Figure 15. The C →T change in *SMN2* exon 7 respectively:

- disrupts an exonic splicing enhancer (ESE) which is normally required for splicing and intron removal and is bound by the splicing factor ASF/SF2. The efficient binding of ASF/SF2 to *SMN1* exon 7 but not to *SMN2* exon 7 causes the latter to be skipped (Cartegni L. and Krainer AL., 2002);

- creates an exonic splicing silencer (ESS) to which a splicing repressor, hnRNP A1, binds. Binding of the repressor to *SMN2* exon 7 but not to *SMN1* exon 7 induces skipping of this exon from a majority of the transcripts from the former gene.

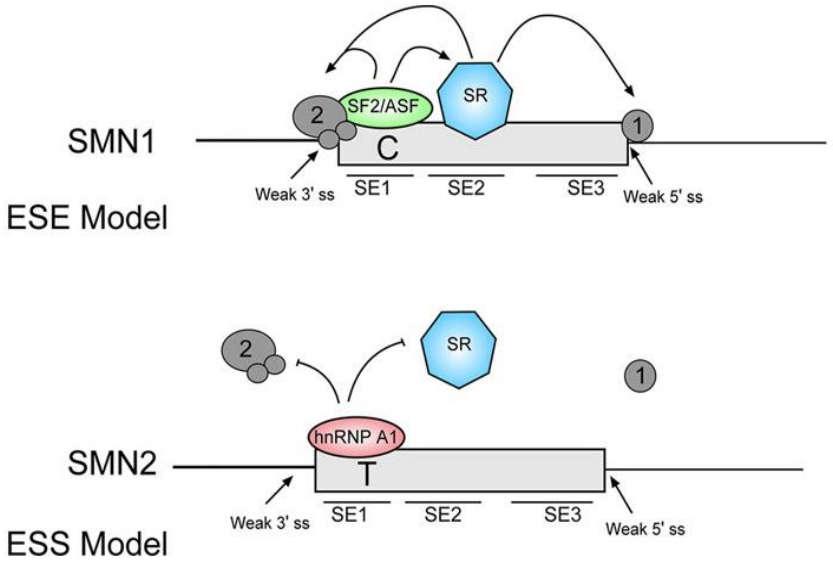


Figure 15 Exon 7 ESE and ESS model of SMN2 pre-mRNA splicing.

Although the mechanisms differ, they both reduce levels of the FL-SMN transcript from the *SMN2*. Currently it remains unclear which of the mechanisms best explains the alternative splicing event. They probably both contribute to the skipping of exon 7 from the SMN2 gene.

3.4.4 Allelic variants of *SMN1*

Doubts about *SMN1* as the SMA-determining gene have arisen but have been eliminated by the identification of several independent intragenic *SMN1* mutations

(Bussaglia et al., 1995; Lefebvre et al., 1995; Brahe et al., 1996; Bürglen et al., 1996; McAndrew et al., 1997; Talbot et al., 1997; Wang et al., 1998).

In particular, several missense mutation clusters have been described in and around the Y/G box (highly conserved tyrosine/glycine-rich sequence) included in the C terminus of SMN, such as p.S262I, p.Y272C, p.T274I, p.G275S, and p.G279V (see Figure 16 for a schematic representation). The p.Y272C, for example, is an A → G transition which causes a tyrosin (Y) to be substituted with a cystein (C) and is associated with SMA type I, the most severe form of the disease. This mutation reduces the half-life of SMN and the efficiency of oligomerization (Pellizzoni et al., 1999) and represents 20% of SMN1 gene mutations (Zapletalová et al., 2007).

Many studies revealed that the occurrence of different intragenic mutations depends on the ethnicity of SMA patients. For example, the c.399_402delAGAG is present only in the Spanish population and is associated with a large spectrum of phenotypes from type I to asymptomatic patients (Cuscó et al., 2003).

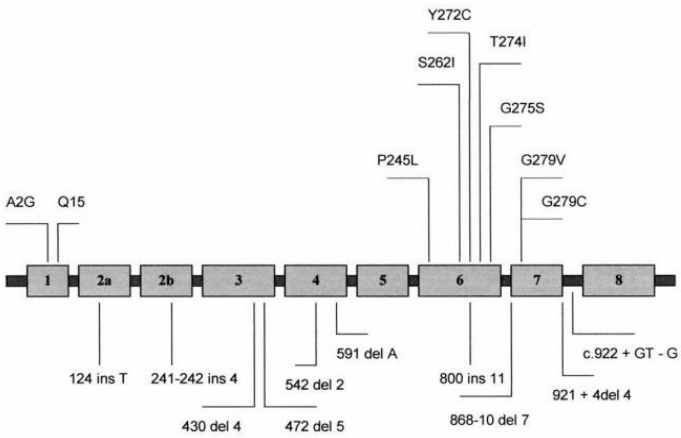


Figure 16 Location of mutations described in the SMN1 gene; ins: insertion, del: deletion.

3.4.5 Allelic variants of *SMN2*

As previously mentioned, not all *SMN2* copies result identical, sequence variations can influence the disease severity. As a consequence, SMA phenotype depends either on the number of SMN2 genes or on the *SMN2* sequence.

A homozygous substitution of guanine by cytosine at nucleotide 859 (c.859G>C) found in *SMN2* exon 7 at the +25 position which replaces a glycine by an arginine at codon 287 (p.G287R) leads to a significant change in aminoacid structure. It probably alters the three-dimensional conformation of the SMN protein and partially restores the normal exon splicing leading to a major full-length SMN2 transcript production. It is well established that the C → T at position +6 in *SMN2* exon 7 normally causes the production of lacking exon 7 SMN2 transcripts probably by disrupting an ESE or creating an ESS. The positive effect of c.859G>C probably occurs by creating a new ESE or disrupting an ESS (Bernal et al., 2010).

It is, therefore, very important to find and evaluate the *SMN2* gene variants because of the fundamental role of the centromeric *SMN* copy in pathological phenotype establishment.

3.5 The SMN protein

The SMN protein encodes a 38 kDa protein which has housekeeping functions and is essential for cell survival. It consists of 294 aminoacids and is ubiquitously expressed in all tissues, at particularly high levels in the spinal motor neurons. A tight correlation between the level of SMN protein and the severity of the

disease has been observed in tissues and cells derived from SMA patients: although individuals lacking the *SMN1* gene express vastly reduced levels of the protein, milder affected patients generally produce higher levels of the protein than severely affected ones (Burghes et al., 1997; Wirth et al., 1995; Lefebvre et al., 1997).

The SMN protein is a component of large macromolecular complexes that are found both in cytoplasmic and nuclear compartments and have different roles and functions (see Figure 17).

First of all, SMN protein is shown to interact with Sm proteins, core components of small nuclear ribonucleoproteins (snRNPs). Moreover, cytoplasmic SMN plays an essential role in snRNP biogenesis and is required for the transport of the snRNP complex into the nucleus leading to the assembling of an active splicing complex.

A number of observations indicate that SMN could also be involved in RNA metabolism and localizes also in the nucleus. In fact, nuclear SMN protein is tightly associated with a group of proteins termed *Gems* (or Gemins of Cajal Bodies) which form a large multifunctional complex and are associated with RNA metabolism (Liu Q. and Dreyfuss G., 1996). Besides, a clustering of missense mutations in the SMN gene has been described in SMA patients in a region of aminoacids containing a tyrosine/glycine-rich motif which is present in various RNA binding proteins (Talbot et al., 1997).

Nuclear SMN may also have a function in gene regulation at transcriptional level. This was understood observing that the papillomavirus nuclear transcription activator, E2, interacts with SMN *in vitro* and *in vivo* and that SMN enhances

the E2-dependent transcriptional activation of genes (Strasswimmer et al., 1999). Thus, SMN protein may also have a role in gene expression but the mechanism is still to be clarified.

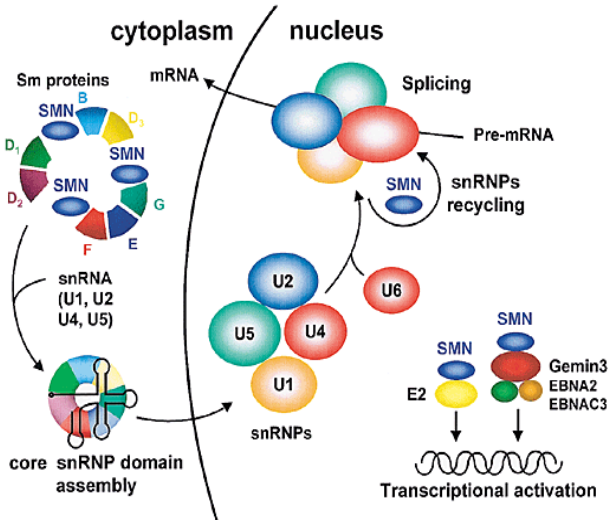


Figure 17 The known roles of the SMN protein.

The major function of SMN protein remains the promotion of snRNP biogenesis whereby the SMN complex interacts with several proteins such as the Sm proteins, which represent essential components of the splicing machinery. The ability of the SMN complex to facilitate and achieve an accurate snRNP assembly is based on its capacity to bind both Sm proteins and also snRNAs. These observations suggest that the SMN complex brings the protein and RNA components together for snRNP assembly, ensuring that Sm cores are only formed on the correct RNA molecules (Pellizzoni et al., 2002). In this process, the SMN protein acts like an ATP-dependent molecular chaperone that promotes the assembly of specific RNA and protein components.

In order to evaluate the role of SMN protein in snRNPs biogenesis pathway, several N-terminal deletion mutants of SMN were generated. Each of them

induced a wrong organization of snRNPs leading to aberrant splicing events (Pellizzoni et al., 1998). In particular, SMN protein plays a critical role in the cytoplasmic assembly of spliceosomal snRNPs, such as U1, U2, U4 and U5. In this process, the U snRNAs first bind the Sm proteins leading to formation of snRNPs which are thereafter imported into the nucleus where they're involved in pre-mRNA splicing. Moreover, the SMN protein is tightly associated with another protein, called SIP1, which has a similar cellular localization and tissue distribution to that of SMN (Liu et al., 1997) and a role in the assembly of snRNPs (Fischer et al., 1997).

Besides, as stated above, the SMN protein oligomerizes and forms a stable complex called the SMN complex, which include different Gemins: Gemin2 (formerly *SIP1*), Gemin3/DP103 (a DEAD-box RNA helicase), Gemin4, Gemin5/p175 (a WD repeat protein), Gemin6 and Gemin7. While Gemins 2, 3, 5 and 7 interact directly with SMN, Gemins 4 and 6 are indirectly associated because they require binding to Gemin 3 and 7, respectively, for the association with the SMN complex (Liu et al., 1997; Baccon et al., 2002).

Two contrasting views have emerged about the role of SMN protein in motor neuron degeneration and muscular atrophy. The first view postulates that SMA is a direct consequence of a defect in snRNP biogenesis and pre-mRNA splicing, whereas the second one suggests that inefficient small nuclear ribonucleoprotein assembly could cause inappropriate splicing of one or more motor neuron-specific messages crucial for the survival of these cells. For example, See and colleagues report an alteration of *Nrxn2* splicing in motorneurons from mouse model

of SMA. This may explain the pre-synaptic defects at neuro-muscular endplates in SMA pathophysiology (See et al., 2013).

In any case, SMN oligomerization and Sm protein binding are impaired in SMNΔ7 individuals. These findings directly link the molecular mechanism of SMA to a deficiency in the interaction of SMN with spliceosomal snRNP Sm proteins (Pellizzoni et al., 1999).

Anyway, further studies are needed to establish which hypothesis is more likely (Monani et al., 2005).

3.6 Genotype–phenotype correlations

No correlation exists between the loss of *SMN1* exon 7 and the severity of disease, thus the homozygous exon 7 deletion is observed in all phenotypes at about the same frequency (Prior TW. and Russman BS., 2000). The level of SMN protein can instead influence the severity of the disease. In fact, patients with SMA I-phenotype have as little as 9% of the normal amount of full-length SMN; those with SMA II have 14%, and those with SMA III, about 18%.

3.6.1 Role of *SMN2*

The incapability to reach normal levels of SMN protein in SMA patients is due to the aberrant splicing of the *SMN2*. The C → T substitution in exon 7 of the SMN2 gene does not affect expression of this gene at transcriptional level whereas it has a profound effect on the level of total SMN protein. In fact, SMNΔ7 transcript is translated but is presumably quickly degraded so that it cannot fully compensate

for the lack of functional SMN protein in affected individuals. *SMN2*, therefore, produces only a fraction of the SMN protein compared to its telomeric homolog, levels clearly insufficient for the health and survival of the motor neurons (Coovert et al., 1997; Lefebvre et al., 1997). 10% of SMN2 transcripts are correctly spliced and encode for a functional protein, identical to the one encoded by SMN1 gene. Therefore when the *SMN2* copy number is increased, more full-length transcripts are generated and result in the milder phenotypes (McAndrew et al., 1997; Mailman et al., 2002). Several studies have shown that the *SMN2* copy number modifies the severity of the disease. For example, patients with the milder type II or III SMA have been shown to have more copies of the SMN2 gene than type I patients. Others support the idea that the alternative splicing of SMN2 gene might have some negative impact on the disease phenotype suggesting a pro-apoptotic effect of the SMNΔ7 protein (Kerr et al., 2000; Vyas et al., 2002). However, the fact that eight *SMN2* copies in mice and humans completely rescue the phenotype (Monani *et al.*, 2000; Vitali et al., 1999) and recent data on phenotypic improvements in SMA via increased SMNΔ7 protein levels in double transgenic SMA-like mice strongly argue against a negative effect of the SMNΔ7 protein (Le et al., 2005).

Despite the crucial influence of the *SMN2* copy number on the phenotype in mild SMA, the *SMN2* does not fully explain variable expressivity whereas other SMA modifiers may be involved (Cobben et al., 1995; Hahnen et al., 1995; Wang et al., 1996).

3.6.2 Role of PLASTIN 3

Plastins are a family of actin-binding proteins that have a role in the organization of the actin cytoskeleton, cell migration and adhesion. Three different isoforms have been characterized in mice compared to only two in humans, all having a tissue-specific expression.

Plastin 3 (*PLS3; OMIM 300131*) has been shown to co-localize with SMN protein in granules throughout motor neuron axons and is thought to be a protective gene against SMA. In particular, in asymptomatic females an increased level of expression of PLS3 has been shown than in effected siblings with concordant *SMN1* and *SMN2* genotypes. This suggests that PLS3 may be a gender-specific SMA modifier (Oprea et al., 2008). Cases where females were more severely affected than males have also been described suggesting that other genetic or environmental factors may be involved (Cuscó et al., 2006; Bernal et al., 2011).

Another study showed an inverse correlation between PLS3 expression in blood and SMA severity in older postpubertal female patients (Stratigopoulos et al., 2010).

Moreover, PLS3 seems to be extremely important for axonogenesis through increasing the F-actin level (Oprea et al., 2008). A potential association between PLS3 and SMN was observed in a zebrafish SMA model (Hao et al., 2012) in contrast with more recent data showing that the modifying effect of PLS3 is independent of SMN levels (Ackermann et al., 2013).

For all these reasons, PLS3 is suggested as a possible target for SMA therapy.

3.7 Therapeutic strategies

Since each SMA patient retains at least one *SMN2* copy and since the number correlates with the phenotype, *SMN2* is considered as an interesting gene target for therapy (Wirth B., 2002).

A large number of various drugs and molecules have been reported to increase the SMN protein level in vitro via transcriptional activation or by restoring correct SMN2 gene pre-mRNA splicing. This makes SMA one of the few genetic disorders in humans, in which activation of a copy gene opens a therapeutic approach. The first drugs to date are the histone deacetylase inhibitors valproic acid (VPA) (Hahnen et al., 2006) and phenylbutyrate (Andreassi et al., 2004). Motor neuron replacement by neural stem cells could be another therapeutic strategy. Neural stem cell transplantation can ameliorate the disease phenotype in a spinal muscular atrophy mouse model, showing increased survival rate and improved neuromuscular function (Corti et al., 2009).

Neuroprotective effects of transplanted stem cells were also described in several other models of neurodegeneration, including ALS, Purkinje neuron degeneration and retinal disease. Moreover, the possibility to generate induced pluripotent stem cells (iPS) from patients' fibroblasts represents another option to have genetically compatible neurons for damaged tissue repairing (López-González R. and Velasco I., 2012).

Last but not least, gene therapy represents a permanent solution for SMA through viral delivery and insertion of the entire SMN1 gene or cDNA sequence into the genome of patients with SMA. For instance, a lentivector expressing human SMN was successfully used to restore SMN protein levels in SMA type I fibroblasts,

reducing motor neuron death and increasing life expectancy (Azzouz et al., 2004). More recently, the results of intrathecal delivery of *SMN1* by scAAV9 reveal improvements in survival in SMA mice (Passini et al., 2014). Currently, Isis Pharmaceuticals is developing a specific oligonucleotide called ISIS-SMNRx designed to act against this strong splicing silencer to promote the inclusion of the missing exon 7. Preliminary results show a dose-dependent amelioration of muscle function (Zanetta et al., 2014). Therefore, gene therapy may offer one of the best therapeutic alternatives although the risks of a wrong delivered gene insertion or overexpression should always been considered.

3.8 Molecular diagnosis

Two different kinds of techniques are currently used for molecular diagnosis of spinal muscular atrophy.

The qualitative one is based on the RFLP approach combined to standard PCR assay to distinguish *SMN1* from *SMN2* taking advantage of the base differences in exons 7 and 8. It is currently used by most clinical laboratories and allows detecting homozygous deletions of *SMN1* exon 7 and exon 8 but cannot distinguish carriers with one copy of *SMN1* from normal individuals with two copies in *Trans* of the same gene (Van der Steege et al., 1995).

The quantitative methods instead allow the establishing of exact number of *SMN1* and *SMN2* alleles. Several techniques enable the determination of the SMN genes copy number in a reliable and reproducible manner despite the near-identity of their sequences. The most commonly used are Real-Time PCR (Mc Andrew et al., 1997; Cuscó et al., 2002; Feldkötter et al., 2002;

Gómez-Curet et al., 2007) and Multiple Ligation Probe Amplification (MLPA) (Scarciolla et al., 2006; Alias *et al.*, 2011).

The precise quantification of *SMN1* is fundamental for a complete and informative diagnosis inclusive of carrier risk assessment which is an essential component of clinical reports and genetic counselling. Quantitative techniques permit to distinguish between healthy, affected and carrier individuals, depending on the number of SMN1 alleles. Although the presence of two *SMN1* gene copies significantly reduces the risk of being a deletion carrier, false negative rate is still present. Such methods do not actually distinguish between healthy individuals with two *SMN1* copies and "2/0 genotypes" (2-5% of SMA carriers). Moreover, carrier diagnosis is complicated by *de novo* mutations (2% of SMA patients) and UPD. Therefore, only the detection of a homozygous *SMN1* deletion is reliable for prediction of disease and an accurate screening for mutations is necessary for an accurate risk assessment and genetic counselling.

Moreover, the establishment of the *SMN2* copies in patients with SMA is an essential prerequisite for therapy, especially given the recent identification of drugs able to increase full-length SMN2 mRNA.

AIMS

The aim of the project is the setting and application of NGS techniques to the SMN genes with the purpose of identifying SNPs and small indel mutations that can be related to different clinical features of SMA. The importance of SNPs characterization relies in the attested ability of certain known single base mutations to regulate the FL-SMN2 transcript level, hence the FL and functional SMN protein production. A future goal of this project is to use the information gathered on modulator polymorphisms for a more detailed diagnosis/prognosis for patients and for pharmacogenetic studies aimed at verifying the association of the selected polymorphism with the response to pharmacological treatment(s).

The project finally aims also to validate the quality aspects of the whole process of targeting NGS in order to suit diagnosis requirements.

The interpretation of the NGS results will be the essential step toward realization of personalized genomics and medicine.

MATERIALS AND METHODS

The workflow of NGS sequencing is characterized by 4 steps: selection of patients and gathering material, library preparation to suit samples for the sequencers, sequencing of samples and, finally, data analysis. Next paragraphs will explain the NGS sequencing workflow of this project in detail.

1.1 Specimen collection and DNA extraction

Individuals included were collected in two different nations: Italy (n=21) and Spain (n=20).

Both cohorts were composed of individuals with homozygous *SMN1* deletion and only 2 subjects of the Italian cohort were asymptomatic.

The genomic DNA was obtained with informed consent from the peripheral blood specimens of all tested individuals and isolated using the Purgene blood kit (Qiagen, Germany) according to the manufacturer's instructions.

Sample concentration was determined using the Quantifluor dsDNA System (Promega, Wisconsin) while the quality was determined on 1% Agarose gel (TBE 0.5 X) with Ethidium Bromide and Lambda DNA/HindIII as marker of molecular weight (Fermentas, Massachusetts). The gel image was acquired through a Gel Doc 2000 (Bio-Rad, California) and displayed on Quantity One (Bio-Rad, California).

1.2 SMN2 genotyping

A total of 50 ng of genomic DNA from all samples was used for Real time PCR and MLPA. The first method was used to clearly quantify the SMN2 copy number (Passon et al., 2009); instead the MLPA was used to detect samples with gene conversions (Passon et al., 2010). The MLPA PCR products were separated by capillary electrophoresis using the ABI 3730 DNA Analyzer (Applied Biosystems, California). The GeneScan-500 LIZ (Applied Biosystems, California) was used as size standard for the final products analysis. The Hidi-Formamide (Applied Biosystems, California) was used for denaturation of DNA prior to injection on capillary electrophoresis system. The results were finally analyzed with Peak Scanner Software (Applied Biosystems, California).

1.3 Primer design

All primers used in the project are designed by Primer3Plus (http://www.primer3plus.com). At the same time Primer Blast (http://www.ncbi.nlm.nih.gov/tools/primer-blast) was used to discard any primer that would anneal on Alu sequences (which are a family of repetitive elements in the human genome and highly represented in the SMN genes). Every satisfactory couple of primers was then tested for primer-primer and self-primer dimer formation through Oligo Property Scan (http://www.eurofinsdna.com).

First of all, primers were designed to obtain the entire *SMN2* sequencing with long range amplifications. Software programs were tuned as follows: optimal primer

would be 33 base pairs long, melting temperature of 65.0 °C and content in GC equal to 50% of all the bases of the primer (Table1).

Primer name	Primer sequence	Primer length (bp)	Tm (°C)	Annealing site (from ATG)	Gene region
LRSMN_F_003	TGGTCAACATCATCCCATTCTCCCCTTCCTCCA	33	65	- 1419	5' region
LRSMN_F_008	AGAACAGCATTCCCGTAGTCTAGATGAAGTC	31	68	+ 13381	intron 1
LRSMN_F_010	TTGGATTCTATTTGGACTTGTCTC	24	60	+ 10809	intron 1
LRSMN_F_012	CTTCCAAATCTCTACCCTCTATCCTTCACC	30	67	+ 17137	intron 2
LRSMN_R_004	AATCCAGCCAGGTAGTGTGGTGGCTTTGATGTT	33	65	+ 28776	3' region
LRSMN_R_007	GTGTCCTGTTTGAGACACAGAACCATACTAC	31	66	+ 14607	intron 2
LRSMN_R_011	CTAGAAAGGGACAAGCCTTAAGGTTCCA	28	67	+ 18601	intron 2
LRSMN_9R_12kb	GTAATATTAAGTATGTTCATGTTGTTGCGCA	31	64	+10946	intron1

Table 1 Set of designed oligonucleotides for Long Range PCR.

At the same time, primers were also designed for the Sanger sequencing of SMN genes, with default parameters of software (Table 2).

Primer name	Primer sequence	Primer length (bp)	Tm (°C)	Amplicon size	Sequencing target
SMNex1F	AAATGTGGGAGGGCGATAA	19	64,1	295	exon1
SMNex1R	CGGAAGAAGGGTGCTGAGA	19	65,6		
SMNex2aF	TGTGTGGATTAAGATGACTCTTGG	24	63,9	219	exon 2a
SMNex2aR	TCCTTTCCAAATGAATAACGAGA	23	63,4		
SMNex2bF	GACGGAGCCTTGAGACTAGC	20	63,1	389	Exon2b
SMNex2bR	ATGCATGTTCCTAAATAACAGAAA	24	60,3		
SMNex3F	TATCCTTCACCTCCCCACTG	20	63,9	390	exon 3
SMNex3R	TCGGTGGATCAAACTGACAA	20	64,2		
SMNex4F	TTCAATTTCTGGAAGCAGAGAC	22	62,3	383	exon 4
SMNex4R	CAAAAGTTTCATGGGAGAGC	20	61,2		
SMNex5F	TGGTTTTGAGTCCTTTTTATTCC	23	61,9	250	exon 5
SMNex5R	TGACATTTTACAATCCTCTATTCTGC	26	63,1		
SMNex6F	GCAAAAATACAATTAATTTCCAGCA	25	63,6	366	exon6
SMNex6R	TGCAAGAGTAATTTAAGCCTCAGA	24	62,9		
SMNe7F	TGTCTTGTGAAACAAAATGCT	21	61,1	1112	exons 7/8
SMN25R	CAATGAACAGCCATGTCCAC	20	64,1		
SMNex5R1	CCTTGGGCCAGATTCTAATG	20	57,3	282	exon 5
SMNex5R2	TAATGCCTTTCTGTTACCCAGA	22	56,5	343	exon 5

Table 2 Sanger sequencing primer. Each primer is used whether to amplify the target region or to sequence reaction.

Thus, the set of primers in table 2 was firstly designed to sequence all SMN exons and exon-intron junction for diagnostic purposes. Then those primers were also used to validate variants.

Finally, the SBE set of primers for variant validation are reported below (Table 3):

Primer name	Primer sequence	Length(bp)	Tm (°C)
sbe_f_69345130	TACTAAATACAAAAAATAGCTGAGC	25	54.8
sbe_f_69348033	GAGACTTCACCTCAAAAAAAAAAAAAAA	28	56.3
sbe_r_69348033	CAGGGTCTTAAAATCCTCACTTCCTTT	27	61.9
sbe_r_69348440	AACTGATTCTCCTGCCTCAGCCTCCCA	27	68
sbe_r_69355622	GAGATTGCAGTGAGTGGAGATCAGAG	26	64.8
sbe_f_69356085	AGGCAGGAGAATCACTTGAACCTGGGC	27	68
sbe_f_69356114	CAGAGGTTGCAGTGAGCCGAGATCATG	27	68
sbe_r_69357190	AACAAAAAAAACCAAATTTAGCTGGGCA	27	58.9
sbe_f_69357245	TGTTTGTTTTTGAGACAGAGTCTCACT	27	60.4
sbe_r_69357509	ACCTTGTGTCTACTAAAAATACAAAAATT	29	56.8
sbe_f_69357899	TTTCACCATGTTGGCCAGGCTGGTCTC	27	68
sbe_r_69358605	TTCTCCCCTCAAATTCCTGTGTTCAAA	27	61.9
sbe_f_69359017	GGAATTAAATTGTAGGGGCATTC	23	57.1
sbe_r_69359017	GTTCTAAGAATGAATGCCATCAAG	24	57.6
sbe_f_69359824	TGAGACTCCATCTCAGAAAAC	21	55.9
sbe_r_69359824	AAAAATACTAAAAGGAATACATTGTTTGTT	30	55.8
sbe_f_69360743	CTTTCCAGTATACACTGAAACTA	23	55.3
sbe_f_69362410	ATTACCAAGGGGGAAGAGAGC	21	59.8
sbe_r_69362410	AGCAAGTGCTCATCAACTGTT	21	55.9
sbe_f_69363717	AAAAAAATTTGCCGGGCGTGATGG	24	61
sbe_r_69363717	GGGCACCTGTAGTCCCAGCT	20	63.4
sbe_f_69364605	TGAATCTAAAATGATGTACCCTCTTAG	27	58.9
sbe_f_69367010	ATCTTGGCTCACAGCAAGCTCTGCCT	26	66.4
sbe_r_69367010	AGGCAGGAGAATGGCGTGAATCCAG	25	66.3
sbe_f_69367063	CCTCAGAGGTAGCTGGGACTACAG	24	66.1
sbe_r_69367063	AAAATTAGACAGGCGTGGTGGCAGGCA	27	66.5
sbe_f_69368270	AGTGGCGGGCAGTTGTAATCCCAGCTA	27	68
sbe_r_69368270	AGTGATTCTCCTGCCTCAACCTCTCAA	27	65
sbe_f_69368717	TAATCCCAGCACTTTGGGAGGCCA	24	64.4
sbe_r_69368717	CTGACCTCGTGATCCACCTGCC	22	65.8
sbe_f_69368815	AAAAGTACAAAAACAAATTAGCCGGGCAT	29	61
sbe_r_69368815	CTGGGACAAAAGGTGCCCGCCAA	23	66

sbe_f_69368911	AGCTTGCAGTGAGCCGAGATTGTGC	25	66.3
sbe_f_69370574	TGGCTCACGCCGGTAATCCCAACA	24	66.1
sbe_r_69370594	TCTGGAACTCCCGACTTCAGGTGATTC	27	66.5
sbe_r_69370742	TCACCGCAACCTCCGCCTTCTC	22	65.8
sbe_f_69370895	TTTTTTAAGATGGAGTTTTGCCCTGTC	27	60.4
sbe_r_69371063	CTTGGGGGACAGAGCAAGACACC	23	66
sbe_r_69371300	GGCTCACGCCTATAATCCCAGCACTTG	27	68
sbe_f_69371368	TAGTCTTGTATTTTTAGTAGAGTCGGG	27	60.4
sbe_r_69371368	AGAACAGCCTGACCAACATGGAGAAA	26	63.2
sbe_r_69371981	TTTAAAAATTAGGCCGGCGTGGTGGCTCA	29	66.7

Table 3 SBE primers. f=forward primer; r= reverse primer.

All primers were ordered at a 100µM concentration and lyophilized by Sigma Aldrich (Missouri).

1.4 Haloplex library preparation

The library preparation with Haloplex protocol (Agilent Technologies, California) did not need any sample preparation. The target region was sent to specialized bioinformatics. In order to separate the target region from the genomic one, the company's specialists decided the best combinations of enzymes. The sent target region was between coordinates 69.340.342-69.375.000 from ENSEMBL (http:// http://www.ensembl.org/index.html). The region contained also the *SMN2* (coordinates 69.345.350-69.374.349). The Haloplex design report sent back from the company showed the on-target sequence coverage. After the Haloplex design and the genomic quantification, eleven samples and one Enrichment Control DNA sample were prepared. Each sample was diluted to a final concentration of 5ng/µL, instead 45µL for the control sample was aliquoted in a separate tube. These genomic DNAs were digested in eight reactions, each containing two restriction enzymes. The DNA digestion was run into a thermal cycler for 30 min

at 37°C. At this point, 4 µL of the control DNA was transferred into a new tube and then incubated at 80°C for 5 minutes to inactivate the enzymes. This inactivated control was run into Bioanalyzer 2100. A high sensitivity DNA kit (Agilent technologies, California) was used, following the manufacturer's protocol. The Haloplex protocol continued only when the control DNA was properly digested. During the electrophoresis run of the control sample, a Hybridization Master Mix was prepared for each sample by combining 50µL of Hybridization solution and 20µL of Haloplex probe. A total of 10µL of indexing primer cassette to each tube containing hybridization master mix was added. The digested DNA samples were transferred into the hybridization reaction tubes. The control sample instead was filled with water. All samples were briefly vortexed and spinned. The hybridization reaction was performed in a thermal cycler at 95°C for 10 minutes and then at 46°C for 3 hours. The second attempt to obtain the Haloplex library was performed at 54°C for 3 hours. The circularized target DNA-Haloplex probe hybrids are captured on strepdavidin beads due to biotin modified probes. Firstly, 40µL (per sample) of beads were incubated for 5 minutes into a magnetic rack. The supernatant was changed with the same volume of capture solution. The ready-for-use beads were added into hybridized samples. Samples were incubated at room temperature for 15 minutes before incubation into the magnetic rack. For each sample, supernatant was discarded, the wash solution added, incubated for 10 minutes at 46°C and moved into the magnetic rack.

A total of 50µL of a previously prepared Ligation Master Mix (which contained 47.5µL of ligation solution and 2.5µL of DNA ligase) was then added to each

capture reaction tube. Beads were resuspended thoroughly before their incubation for 10 minutes at 55°C. When the 10-minute ligation period was complete, the captured DNA was eluted as follows. Each sample was incubated into the magnetic rack and the supernatant was discarded. When tubes were removed from the magnetic rack, samples were resuspended with SSC buffer and incubated again into the magnetic rack. When the solution was clear, the supernatant was discarded and samples were resuspended with 25µL of freshly-prepared NaOH 50mM. After thorough resuspension of beads, sample tubes were moved into the magnetic rack. The supernatant was collected for the next step of the protocol: PCR amplification of captured libraries. A total of 20µL of the supernatant was added to 30µL of PCR Master Mix. The first attempt was amplified with KAPA HiFi DNA Polymerase (Kapa biosystems, Massachussetes), the second indeed was amplified with the Herculase II fusion DNA polymerase (Agilent Technologies, California). The first PCR master mix was composed of 10 µL of KAPA HiFi fidelity Buffer 5X, 0.4 dNTPs 100mM, 1µL each primer at 25µM, 0.5µL of acetic acid 2M, 1U of KAPA HiFi HotStart Polymerase and water up to 30µL of final volume for each sample.

The second effort master mix was composed of 10µL of 5X Herculase II reaction Buffer, 0.4µL dNTPs 100mM, 1µL of each primer 25µM, 0,5µL Acetic acid 2M and 1µL Herculase II Fusion DNA polymerase and each sample was filled in with water. The first amplification program was: 95°C for 5 minutes, 23 cycles at 98°C for 20 seconds, 65°C for 15 seconds, 72°C for 30 seconds, final extension at 72°C for 5 minutes. The amplification program of the second effort evolved as follows:

98°C for 2 minutes, 23 cycles: 98°C for 30 seconds, 60°C for 30 seconds and 72°C for 1 minute; 72°C for 10 minutes and hold at 8°C.

Thus libraries were obtained per sample and were purified with AMPure XP beads (Beckman-Coulter, California). Beads were firstly used at a sample volume ratio of 1.5:1 and then modified to 3.5:1. After 5-minute incubation at room temperature with continuous shaking, samples were moved again into the magnetic rack and the supernatant was removed and discard. While samples were in the magnetic rack the beads were washed twice with freshly prepared 70% ethanol. When all disturbed beads settled down the supernatant was removed. Samples were air-dried with lid open at room temperature until the residual ethanol completely evaporated. Tubes were moved from the magnetic rack and for each sample beads were resuspended with 40μL of Tris-HCl (pH8.0). The last incubation into the magnetic rack allowed recovering of the supernatant containing the libraries and to discard the beads. Finally, aiming to validate the enrichment and quantify the enriched target DNA in each sample, a microfluidics analysis using Bioanalyzer 2100 with a High Sensitivity DNA kit (Agilent technologies, California) was run. In order to prepare this run, the samples were diluted to 400 ρg. The Bioanalyzer-measured concentration of 175-625 bp products in each sample were used to pool equimolar amounts of differentially indexed samples in order to optimize the sequencing capacity. These samples were ready to be sequenced by MiSeq sequencing platform (Illumina, California).

Illumina did not have an official recommendation for a minimum sequencing coverage level. Most users determined the necessary coverage level based on the type of study and on the size of reference genome. The general equation for

computing coverage was the following: $\mathbf{C = L*N/G}$; where **C** stands for *coverage*, **G** is the *haploid genome length*, **L** is the *read length* and **N** is the *number of reads*. A high coverage meant that each base is sequenced many times, and that is fundamental for a reliable base calling.

1.5 Long Range PCRs settings

The PCR mix and the thermal protocol were designed by reconsidering the protocol given for the AccuTaq LA DNA Polymerase (Sigma Aldrich, Missouri). The suggested thermal protocol was 98 °C for 30 seconds, 30 cycles: 94 °C for 10 seconds, 62 °C for 20 seconds, chosen temperature (it depends on primers' melting temperature) for 19 minutes, 68 °C for 10 minutes. A total of 200 ng of DNA were amplified in a final volume of 50 µL with 500 µM dNTPs mix, 2% DMSO, 400 nM for each primer, 0.05 U/µL of ACCUTaq Polymerase.

Each amplicon was run on 1% Agarose gel (TBE 0.5 X) with Ethidium Bromide and Lambda DNA/HindIII as marker of molecular weight (Fermentas, Massachusetts). The gel image was acquired through a Gel Doc 2000 (Bio-Rad, California) and displayed on Quantity One (Bio-Rad, California).

The fragments were purified using Wizard DNA Clean-up System (Promega, Wisconsin) according to the manufacturer's suggestion, considering a final elution volume of 20 µl. All purified amplicons were again run on a 1% Agarose gel (TBE 0.5 X) with Ethidium Bromide and Lambda DNA/HindIII as marker of molecular weight (Fermentas, Massachusetts).

All purified fragments were then quantified using Qubit Fluorimetric Quantitation (Life Technologies, California) according to the manufacturer's instructions.

To have the total coverage of the sequencing target region the fragments of the SMN gene amplified by Long Range PCR had to be reunited into a mix. In order to have equal sequencing coverage for all amplicons, they were to be equally represented in the mix. Therefore, several equimolar mixes were prepared on the base of the final amount of DNA amplicons desired and the estimated concentration of the LR PCR products.

1.6 Nextera library preparation

After the equimolar pooling of amplicons, each sample was Tagmented by Nextera sample preparation kit (Illumina, California). Each equimolar mix (a sample) was prepared considering the concentration of each fragment and a final volume of 20 µl, as requested by Nextera DNA Sample Prep Kit for library preparation. A total of 50ng of pooled PCRs were used for tagmentation step. To the pooled PCRs were added 25 µl of Tagment DNA Buffer and 5 µl di Tagmented DNA Enzyme1. All samples were firstly centrifuged 280xg for 1 min at 20°C and then moved into a thermalcycler for 5 min at 55°C. The tagmented samples were purified by QIAquick PCR Purification (Qiagen, Germany) instead of Zymo as suggested by Nextera protocol. A total of 100 µL of Buffer PB were added to the 20 µL tagmentation product and mixed. Each sample was applied to a QIAquick column that was previously placed in a 2 ml collection tube. The columns were centrifuged for 1 minute at 14000 rpm. After discarding the flow-through, each QIAquick column was placed back into the same collection

tube and 0.75 ml of Buffer PE were added. The columns were centrifuged for 1 minute at 14000 rpm. The flow-through was discarded and each QIAquick column was placed back into its tube for an additional centrifugation for 2 minutes at 14000 rpm. The QIAquick column was then placed in a clean 1,5 ml tube. Finally, 11 μL of water were added to the center of the QIAquick column and it was left open for 5 minutes. The columns were centrifuged for 1,30 minutes at 14000 rpm. The final eluted volume should be around 10 μL containing purified tagmented PCR products. The tagmentation was checked with a microfluidics run into a Bioanalyzer 2100 (Agilent technologies, California). In order to run the tagmentation products on a High Sensitivity DNA kit (Agilent technologies, California), the tagmented DNA samples were diluted to 400 ρg.

The well fragmented samples were amplified via a limited-cycle PCR. This amplification was used to add index 1 and 2 as well as adapters required for cluster generation and sequencing. All indexes should uniquely identify one sample. Thus, a 5μL aliquot of each index primers was added into an empty tube. In each tube were firstly added 15μL of Nextera Master Mix and then 5μL of PCR Primer Cocktail, finally was also added 20μL of tagmented DNA. After pipetting up and down, samples were centrifuged at 280xg at 20°C for 1 minute. The thermal cycler program was set up as follows; 72°C for 3 minutes, 98°C for 30 seconds, 5 cycles of: 98°C for 10 seconds, 63°C for 30 seconds, 72°C for 3 minutes and hold at 10°C. The amplifications were clean up with AMPure XP beads (Beckman-Coulter, California). A total of 30μL of beads was added to 50μL of PCR product. These mixes were pipetted up and down 10 times and incubated 5 minutes at room temperature. All samples were placed into a

magnetic rack and after 2 minutes the supernatant was discarded. All samples were washed twice with freshly-prepared 80% ethanol. After the second wash, samples were removed from magnetic stand and air-dried for 15 minutes and resuspended with 32.5µL of Resuspension Buffer and move into the magnetic rack again. After 2 minutes the supernatant of each sample was cleared and transferred in a new tube. Each Nextera libraries was validated and quantified by a microfluidics analysis using Bioanalyzer 2100 with a High Sensitivity DNA kit (Agilent technologies, California). In order to prepare this run, the samples were diluted to 400 ρg. Each sample library was controlled in size distribution, if fragments were too small a second clean-up with beads were performed.

1.7 MiSeq platform sequencing run

The MiSeq workflow is composed by many steps, each sequencing run needs all the following:

P**repare the pre-filled reagent cartridge for use.** The MiSeq reagent cartridge is a single-use consumable consisting of foil-sealed reservoirs pre-filled with clustering and sequencing reagents sufficient for sequencing one flow cell. Before each run the cartridge was placed in a water bath containing enough room temperature deionized water to submerge the base of reagents. When the reagent cartridge was thawed, it was removed from water and air-dried. The cartridge was inverted 10 times to mix the thawed reagents and then each reagent was visually inspected to make sure that was free of precipitates and bubbles.

Denature and dilute libraries. All libraries were denatured and diluted following the most appropriate protocol for obtained libraries. The correct protocol attended to the chemistry of reagents kit and the obtained initial library concentration.

Since the v2 reagent kit was always used and library concentration always fit, the 2nM protocol was always used for all sequencing reactions of this project.

Firstly, the library was diluted to 2nM. In a new tube was dispensed a volume of 5µL of both, NaOh 0.2N and the DNA sample. The 5-minute incubation at room temperature allowed single strand denaturation of libraries. Finally, the sample was added with 990µL of Hybridization Buffer to obtain a 20pM library. Moreover, the 10nM PHix library was diluted to 12.5pM and then processed as the target library was. The Phix control should be used at 1% of total library, thus 6µL was added to 594µL of target library.

Load of library mix onto the reagent cartridge in the designated reservoir. When 1mL pipette tip pierced a hole through the foil seal over the load sample reservoir, the library mix (600µL) was pipette into the reservoir.

Start the run setup steps. From the welcome screen, the Sequence button started a series of interactive screens that guided the operator through the next setup steps.

Wash and thoroughly dry the flow cell. The flow cell was removed from its plastic case with plastic forceps. It was rinsed to remove salts. The flow cell and cartridge were thoroughly dried using a lint-free lens cleaning tissue. Furthermore, the flow cell was alcohol wiped until the glass was free of streaks, fingerprints, and lint or tissue fibers. After visually inspection of cell ports for obstructions, the flow cell stage into the sequencer was also carefully wiped. Finally the flow cell

was placed on the flow cell stage. The lower-left corner of the screen confirmed the successfully read of the flow cell. The screen invited to go to the next step.

Load the regent bottle. The cold reagent bottle was firstly inverted and then the lid removed. When the reagent compartment was open, the bottle was placed in the right of the reagent chiller and the waste bottle was empty. The lower-left corner confirmed that bottles were read correctly. The screen invited you to go further. The reagent and sample cartridge was inserted into the chiller door. The lower-left corner of the screen confirmed also the cartridge read.

Start run. After a pre-run check, the run started successfully. Before starting the run, with the change sample sheet command on the load reagents screen was inserted the appropriate sample sheet. The selected sample sheet for the *SMN2* sequencing was the chromosome 5 but the telomere.

Post-run washing. Immediately after the completing of the sequencing run, a wash was always performed. The software prompts showed how and where to load reagents.

The sequencing was performed on Illumina's MiSeq platform using a paired read run of 150 bp generating more than 1 Gb of sequence.

1.8 NGS data analysis

After the sequencing run, reads were put through the sequencing pipeline consisting of base calling using Illumina Pipeline, trimming with rNE (Vezzi F. et al. 2012), alignment on the human chr5 (NCBI build 37/UCSC hg19) truncated for the telomeric region containing a highly homologous *SMN1* gene with BWA (Li H. and Durbin R., 2009), selecting uniquely mapping reads with

proprietary script, PCR duplication removal with SAMTools (Li et al., 2009) and single nucleotide variants and indel calling with VarScan (Koboldt et al., 2009). Alignment statistics were obtained using Picard (http://picard.sourceforge.net). All the variants from reference sequence output from VarScan with consensus quality ≥30, read depth ≥10, and variant depth ≥25 were filtered versus NCBI dbSNP137 and 1000 Human Genomes Project catalog (2010 Nov) and were functionally annotated by Annovar (Wang et al., 2010). Variant calling was also performed on only properly paired reads (reads that are mapping at the expected distance and are properly oriented) and at reduced coverage of 50X.

1.9 NGS validation

Next-generation sequence variant calls were validated against Sanger sequencing.

The **Sanger sequencing** protocol set up during this project was also applied during diagnostic process in our laboratory.

A final volume of 25µL of PCR mix was added to 30ng of genomic DNA of each sample. The PCR mix was composed by 1X Kapa2GFasta HS ready mix (KAPA Biosystem, Massachussetts), 0.5µM of primers, 0.5mM of $MgCl_2$ (Applied biosystem, California), 1.25µL of dimethyl sulfoxide-DMSO ≥99.5% (Sigma Aldrich, Missouri) and water was added to fill in the required volume. The thermal protocol to amplify exon 1, 3, 4 and 6 was: 95°C for 5 minutes, 35 cycles at 95°C for 30 seconds, 63°C for 30 seconds, 72°c for 45 seconds and a final elongation step at 72°C for 7 minute. Exons 7 and 8 were amplified in the same amplicon with annealing temperature of 59°C, instead of 63°C. The exons 2a, 2b and 5 were further amplified with an annealing temperature of 60°C.

Amplicons were purified by AMPure beads (Beckman-Coulter, California), 18μL of beads (0.6X) were added to 10μL of each amplicon. Samples were incubated for 5 minutes at room temperature before they were placed into a magnetic rack.

When the supernatant was cleared, it was discarded and 2 washes with ethanol 85% were performed. After the second wash, samples were air dry for 8 minutes and then resuspended in 26μL of water. Each purified amplicon was divided in two different tubes, 5μL for each strand sequencing reaction. Each sample was added with 3.4μL of Big Dye sequencing mix (Applied Biosystem, California) and 320nM of primer. The thermal cycler was run for 35 cycles: 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes.

When the sequencing reaction ended, the precipitation of each sequencing reaction product with 25 μL of EtOH 100% and 2.5μL of EDTA 125mM was performed by centrifuge for 55 minutes at 2900 rcf. The supernatant was discarded by overturning the plate which was then centrifuged facing down for 2 minutes at room temperature at 200 rcf. The plate was left facing up on the bench for 30 minutes to let the remaining ethanol evaporate. The precipitation products were resuspended with 10 μL of Hi-Di formamide and LIZ500 (Applied Biosystems, California) and denaturated for 2 minutes at 95°C. The automated sequencer used to analyze the samples was a 96-capillary DNA Analyzer (Applied Biosystems, California). The sequencer output data was shown through the software Phred Phrap/Consed. The Phred software identified a base sequence from the fluorescence trace data registered by the automated sequencer (base calling); Phred also produced a quality score assigned to each base call.

Phrap determined highly accurate consensus sequences and used Phred quality scores to estimate their quality. Consed allowed the user to view and edit the sequences assemblies created with Phrap. Finally, all sequence obtained were aligned with BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) against human reference sequence.

In this project, Sanger sequencing was also performed to validate long range amplicons before being used as starting material for the Nextera protocol. Although the Sanger sequencing workflow was the same, the long range amplicons were too long to be purified by beads. These amplicons were indeed purified by hydrolytic enzymes as Exonuclease I and Shrimp Alkaline Phosphatase. The Exonuclease I removes all single stranded DNA, therefore any exceeding primer and PCR by product will be degraded. The Shrimp Alkaline Phosphatase hydrolyzes the dNTPs that were not used in the reaction. In order to purify 15µL of PCR product, 1U of exonuclease and 2U of shrimp alkaline phosphatase were added to each sample and then tubes were placed in a thermal cycler at 37°C for 30 minutes. The enzymes were finally inactivated by heating the samples to 80°C for 15 minutes. After this purification step, samples were processed as described above with the sequencing reaction.

The **SBE** was an additional method used for the validation of NGS variants. The templates were amplified by a PCR with specific primer that select the target region. All resulting templates were in solution, along with primers, dNTPs, enzyme and buffer components. To avoid the precipitation in the subsequent primer-extension reaction, primers and unincorporated dNTPs were removed. The enzymatic purification method was selected to clean up the templates.

In order to purify 15μL of PCR product, 1U of exonuclease and 2U of shrimp alkaline phosphatase were added to each sample and tubes were then placed in a thermal cycler at 37°C for 30 minutes. The enzymes were finally inactivated by heating samples to 80°C for 15 minutes. These purified templates were ready for the single base extension reaction. This reaction was performed during this project in two different approaches: simplex and multiplex. The first approach detected one target per reaction due to a single-primer addition in each sample mix. The second approach was obtained by pooling many primers in the same reaction and thus detecting different loci in the same sample. Both approaches required a 0.2μM final concentration of primer. Furthermore, the satisfactory amount of purified PCR product was at least 1pM per reaction. Thus, the single base extension mix was composed by 1μL of Snapshot ready reaction mix (Applied biosystem, California), up to 3μL of purified PCR, and 0.2μM primer/s (single or pooled primers). The Snapshot ready reaction mix contained both Taq polymerase and fluorescently labeled ddNTPs. The fluorescent dye was assigned to the individual ddNTPs as follows:

ddNTP	Dye Label	Color of Analyzed Data
A	dR6G	Green
C	dTAMRA™	Black
G	dR110	Blue
T (U)	dROX™	Red

Table 4 Dye assignments.

When each sample was added with single base reaction mix, samples were incubated into a thermal cycler. The protocol was repeated for 26 cycles: 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 30 seconds. When the extension

ended, each sample was added with 1U of shrimp alkaline phosphatase and incubated at 37°C for 1 hour to remove the unincorporated ddNTPs. The enzyme was then inactivated at 80°C for 15 minutes. Finally, samples were resuspended with 19,6μL of Hi-Di formamide and 0.2μL of LIZ120 (Applied Biosystem, California) and denaturated for 2 minutes at 95°C. The automated sequencer used to analyze the samples was a 96-capillary DNA Analyzer (Applied Biosystems, California). The capillary electrophoresis results were analyzed by Peak scanner software 2.0 (Applied biosystem, California).

RESULTS AND DISCUSSION

1 SMN2 Next Generation Sequencing

Despite there are a lot of method to specifically target the NGS sequencing, the first part of this project was focused on finding the best library preparation to target the entire *SMN2* (27kb) sequence. The comparison was between Haloplex (Agilent Technologies, California) and Nextera (Illumina, California) library preparation.

1.1 Haloplex

The Haloplex (Agilent Technologies, California) protocol is specifically developed for NGS, it uses single-tube target amplification and removes the need for library preparation to reduce total sample processing time and cost without the need for dedicated instrumentation or automation. The 4 steps of the Haloplex workflow (Figure 18) are: digest sample DNA, hybridize probes, purify and ligate targets and finally amplify target fragments with PCR. After the DNA sample fragmentation using restriction enzymes, the probe library is added and hybridized to the targeted fragment. Each probe is an oligonucleotide design to hybridize to both ends of a targeted DNA restriction fragment, thereby guiding the target fragments to form circular molecules. Moreover probes also contain a method-specific sequencing motif that is incorporated during circularization and a sample barcode sequence is also incorporated to the fragment sample in this step.

Probes are also biotynilated and the targeted fragments can therefore be retrieved with magnetic streptavidin beads. The circular molecules are then closed by ligation, a very precise reaction that ensures that only perfectly hybridized fragments are circularized. Finally, circular DNA targets are amplified, providing an enriched and barcoded amplification product that is ready for sequencing.

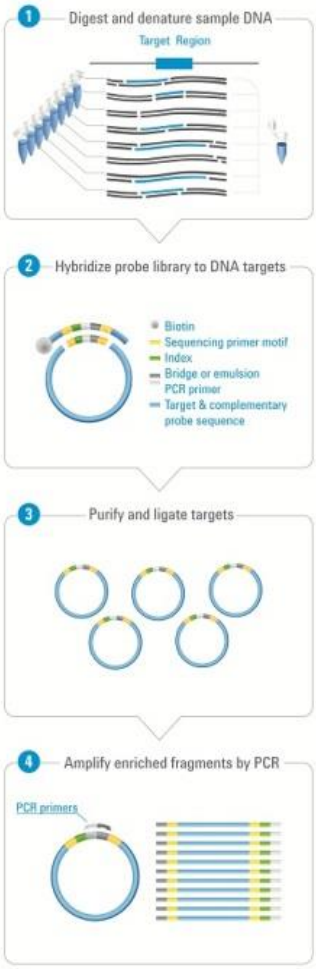


Figure 18 Haloplex workflow.

The Haloplex probes were ordered to evaluate the performance of the procedure for the enrichment of small genomic target, since the project is focused in deep sequencing of a modest region size (~30 kb). The most important step in Haloplex

protocol is the study of the target region by SureDesign. This software program identifies the best enzymes' mix to target and then enrich (as previously described) the region of interest. High sequence identity between the two genes does not allow the design of Haloplex capture probes since it is not possible to select enzymes that uniquely digest a particular genomic region that is then going to be uniquely enriched. For this reason, highly homologous *SMN1* region was masked in order to allow probe design for the SMN2 region.

Finally, the design achieved 90.6% of coverage of the target region (31394 bp out of 34659 bp). The design graphical result is represented by Figure 19. Red bars in the lower part of the figure represent the covered and uncovered target regions by fragments. Uncovered regions length range between 1-743bp and all missed fragments were located in introns or UTRs.

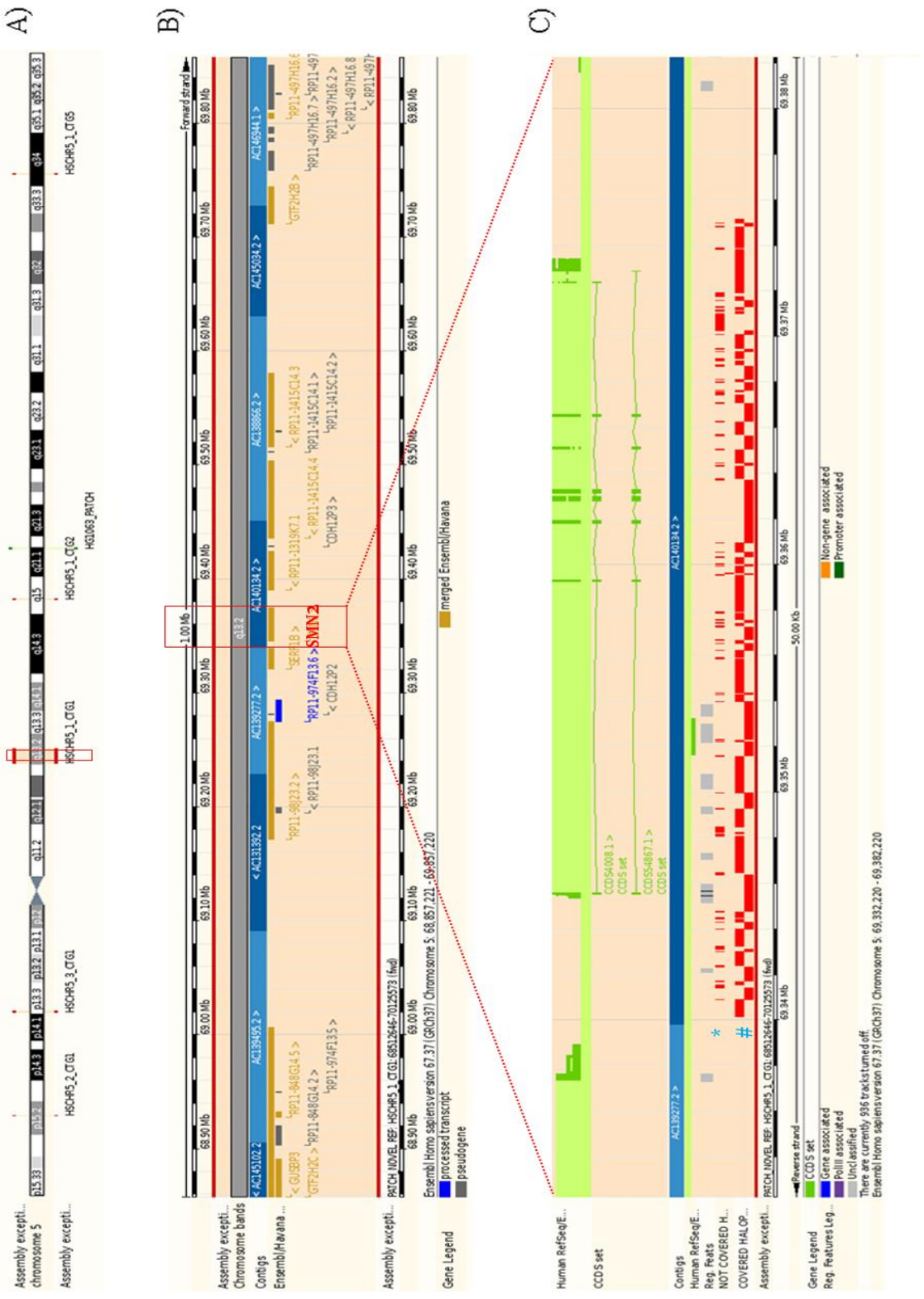


*Figure 19 Haloplex design representation. A) Chromosome 5 and SMN locus; B) Zoom into SMN2 locus; C) Zoom into SMN2 gene and graphical representation of Haloplex design. * not covered regions; # covered regions.*

Libraries were performed as described in the paragraph 1.6 of Material and Methods. Each library profile run in Bioanalyzer (Agilent Technologies, California) is shown in Figure 20.

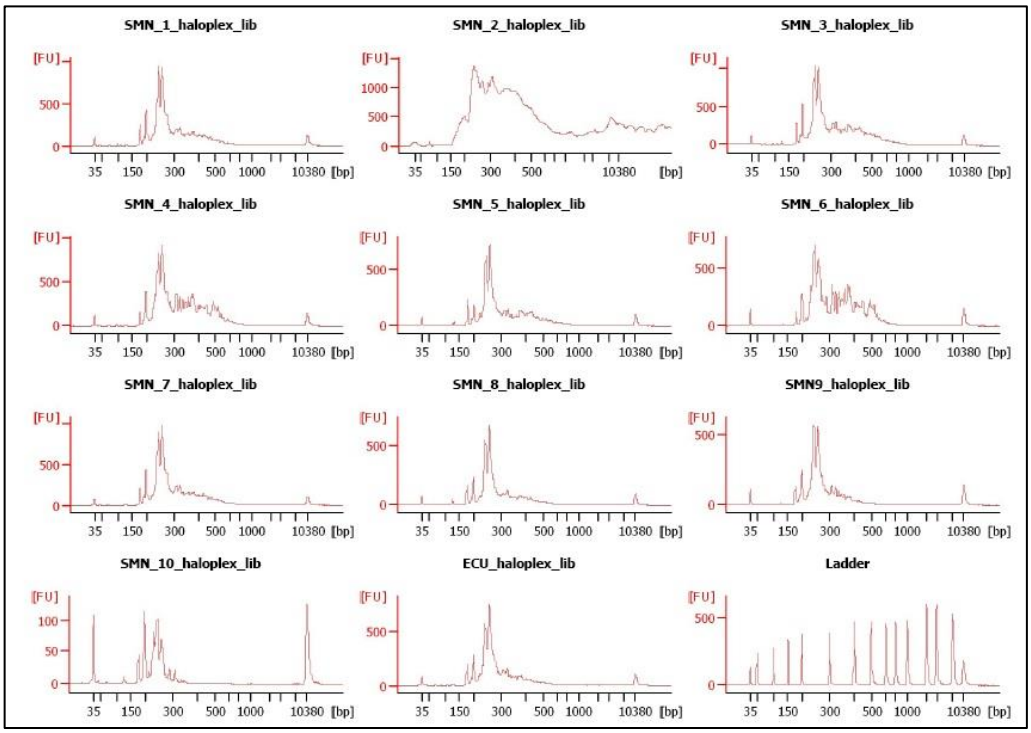


Figure 20 *Haloplex SMN2 libraries profiles.*

The profiles are quite different but all have the highest peak in the same size. Thus, the average size of the libraries is comparable, ~200bp.

Libraries were pooled in equimolar amounts. The MiSeq cartridge was charged with pooled samples and run in 50bp single run mode. The number of clusters and demultiplexing were satisfactory. However, the subsequent bioinformatics analysis gave quite disappointing results. Reads were aligned by BWA (Li H. and Durbin R., 2009) on human reference sequence with all chromosomes but chromosome 5 was truncated for the telomeric region containing *SMN1*. This deletion from the reference sequence was done to avoid aligning to a highly

homologous region that could result in discarding reads that are not mapping uniquely. The alignment was performed on whole genome scale to estimate of target enrichment by Haloplex system. A total of 2.5 Mb per sample were sequenced, but only 1% was uniquely aligned on 30Kb of target out of which 30% were covered at least 2X. These results were reproducible among all samples.

The main problem was the read length used for the sequencing run; the single read 50 bp run was not enough because the design was done for a 2x150bp. Thus, a second run was performed in paired end 2x150bp mode. Moreover during the library preparation we took advantage of a new and more efficient polymerase and increased the temperature during hybridization step. However, results were still unsatisfactory. The number of on target reads increased to only 2% with the mean coverage of the target region of 1.27X.

Surprisingly low yield could be explained by very small target size. Even though Haloplex system is recommended for small targets, these are usually in size range of Mbs of genomic sequence.

In conclusion, the Haloplex method resulted as the wrong approach for targeted resequencing of *SMN2*.

1.2 Nextera

Considering the failure of the Haloplex approach we decided to try a PCR-based approach in combination with Illumina Nextera system for the targeted resequencing of SMN2 gene.

The in vitro transposition with Nextera *Transposomes*™ simultaneously fragments and covalently labels the target DNA with 19bp transposon end tags

(orange and blue in Figure 21), producing fragments having single-stranded overhangs. The actual fragment size distribution depends on a number of factors, including the quantity and quality of starting DNA. Even by varying the concentration of *Transposome*™ complexes, the size distribution of the fragmented and tagged DNA library can be controlled. After Tagmentation reaction, platform specific adaptors and optional bar coding are introduced by 10 cycles of PCR with four primers. As shown, the sequencing adaptors (purple and pink) enable amplification with bridge PCR (bPCR) so that the amplified library can be subsequently sequenced using the appropriate primers. An optional index or barcode (triangle) is added between the downstream bPCR adaptor (pink) and the core sequencing library adaptor (orange) on both ends of the DNA, thus enabling dual-indexed sequencing of pooled libraries on any Illumina Sequencing System. The final Nextera Read 1 and Read 2 primers anneal to the 19bp transposon adapter so that the first nucleotide sequenced is target DNA.

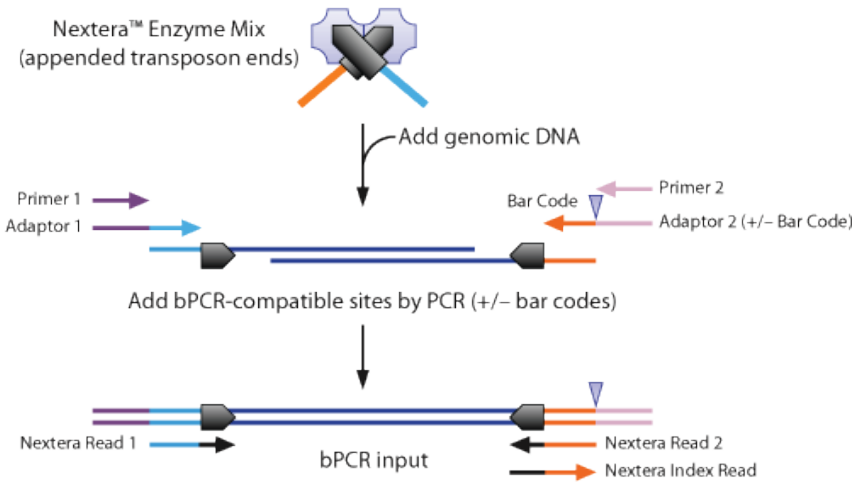


Figure 21 Schematic representation of tagmentation reaction and limited-cycle PCR with a four-primer reaction that adds bridge PCR-compatible adaptors (purple and pink). Optional barcodes (triangle) can be added.

Nextera method offers many advantages over current library preparation methods. For instance, it combines fragmentation, repair and ligation steps resulting in significant time and cost-saving. It is a scalable method that requires as little as 50 ng of starting template, compared to 5-10 µg for other current procedures. It must though be considered that distal 50-100 bp of linear fragments are less covered.

All Nextera sequencing primers have already been fully validated and are completely compatible with the standard Illumina sequencing primers. Therefore, Nextera technology represents an efficient and high-throughput method for generating bar-coded libraries compatible with multiple NGS platforms. Moreover, the Nextera DNA Sample Prep Kit can also make libraries from amplicons.

As depicted in the flowchart, for each individual the whole *SMN2* sequence was amplified in three independent LR-PCR reactions producing three partially overlapping fragments. Primers and amplification protocols are previously described (paragraph 1.3 and 1.5 of material and methods). Two set of primers were alternatively used. The principal one produced three amplicons, each amlicon length was 16kb (LRSMN_F_003/LRSMN_R_007), 5kb (LRSMN_F_008/LRSMN_R_011) and 11kb (LRSMN_F_012/LRSMN_R_004). The second set instead amplified the same total target region with amplicon lengths of 12kb (LRSMN_F_003/LRSMN_9R_12kb), 9kb (LRSMN_F_010/LRSMN_R_011) and 11kb (LRSMN_F_012/LRSMN_R_004). This second set was specifically used for samples that showed amplification problems with the first set of primers.

We tested the correctness of those amplification reactions with Sanger sequencing exons and exon-intron junction included in each fragment. The resulting Sanger sequence was aligned by BLAST to whole reference sequence and thus we demonstrated that long rage PCRs specifically amplified *SMN2* region.

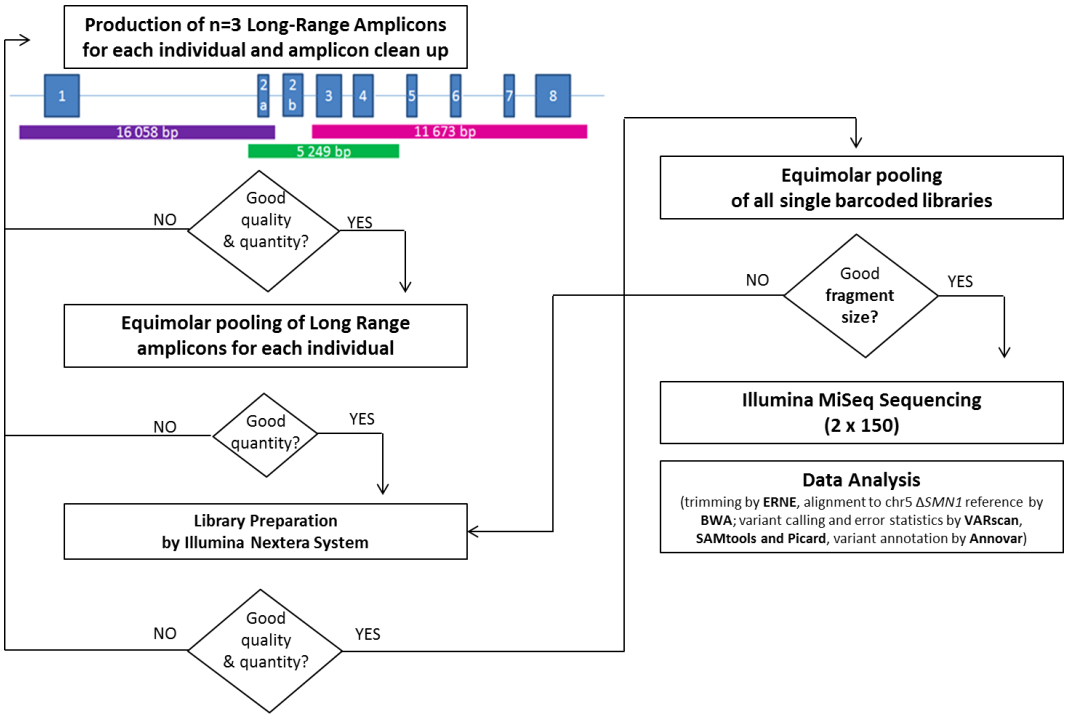


Figure 22 Workflow representation of SMN2 sequencing by Nextera library preparation. Blue squares represent exons while blue lines stand for introns; diamond shape are checking point for quality control.

As shown in Figure 22, the three amplicons for each individual were quantified and pooled in equimolar amounts. The pooled fragments for each individual were processed by the Nextera DNA sample preparation kit (Illumina). Each individual library was uniquely barcoded allowing for concurrent multisample studies. The obtained libraries were run into Bioanalyzer (Agilent Technologies, California).

Usually, traces of successfully sequenced library had a broad size distribution from 250bp to 1000bp. The SMN2 library fragment met size requirements for sequencing, as shown in Figure 23.

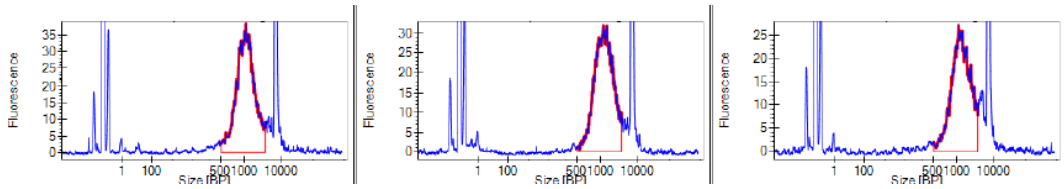


Figure 23 Distribution of fragments of three independent libraries.

Samples pooling was calculated from molarity obtained by Bioanalyzer.

The pooled libraries were then run on Bioanalyzer (Figure 24) to control distribution of fragments and the resulting molarity, before being sequenced.

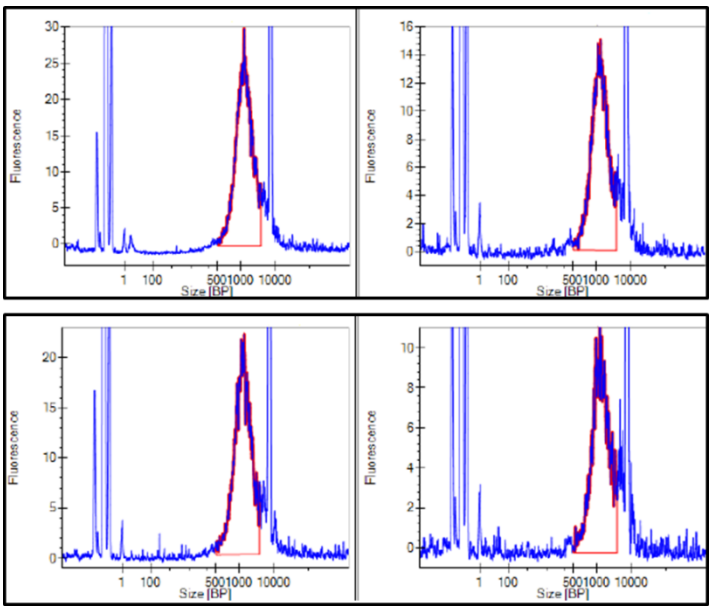


Figure 24 Serial dilutions of pooled SMN2 libraries.

Finally, sequencing was performed on Illumina's MiSeq platform (as described in paragraph 1.7 of Material and Methods) using a paired read run of 150 bp

generating more than 1 Gb of sequence. Obtained reads went through the pipeline described in paragraph 1.8 of material and methods.

Nextera system produced libraries of homogeneous insert sizes (average of 348±63, min 216±103 and max 450±93).

Individual coverage was uniform across entire target region with the expected doubling of coverage at amplicon overlaps (Figure 25). Thus this trend of coverage was a probable outcome of quantitative characteristic of this approach to library preparation.

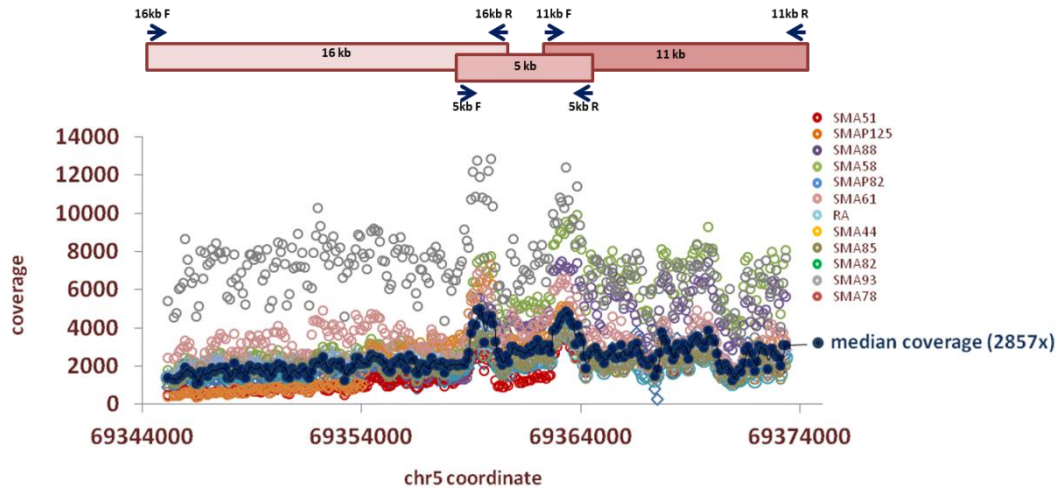


Figure 25 Representation of coverage distribution through SMN2 region.

The coverage of a subgroup of patients was represented into the graph for a better understanding of the figure. The dark blue dots stand for the median coverage of 2857X. We called variants on uniquely mapping reads in 3 different datasets: i) total mapping reads; ii) properly paired reads and iii) at sequencing coverage of 50X. The efficiency of overall variant detection was not affected at lower read counts.

A reduced representation of the variants found for SMA individuals SMA51, SMAP82 and SMAP125 with 3, 4 and 5 SMN2 copies, respectively, is shown in figure 26. The variant frequency is the percentage of reads containing the variant over the total number of reads at that particular position (Figure 26).

Obtained variant sequencies

		SMN1/SMN2 0/3 A		SMN1/SMN2 0/4 B		SMN1/SMN2 0/5 C	
POS	REF	GEN	VAR Freq	GEN	VAR Freq	GEN	VAR Freq
69345130	T					K	40.83
69345626	G					K	18.44
69348440	T	Y	33.77			Y	21.42
69365216	G			S	50.75		
69367010	C	Y	67.65			Y	60.49
69367063	G	K	30.32	K	52.63		
69368084	A	R	27.96	G	72.87		
69368329	G	R	34.91	R	72.36		
69369621	C					Y	39.23
69370574	C	S	33.20				
69371300	C					S	59.77
69371981	C	M	30.23	A	100		
69372616	G			R	49.95		
69373081	A			R	24.55		
69373667	A	R	30.29	G	99.72		
69373682	C			S	49.87		

POS= posizione; **REF**=reference; **VAR**=variante; **VARFreq**=frequenza della variante.
Y= C o T; **K**= GoT; **R**= AoG; **S**= CoG; **M**= AoC.

Attended variant sequencies

3 *SMN2* copies
1/3 = 33%
2/3 = 66%
3/3 = 100%

4 *SMN2* copies
1/4 = 25%
2/4 = 50%
3/4 = 75%
4/4 = 100%

5 *SMN2* copies
1/5 = 20%
2/5 = 40%
3/5 = 60%
4/5 = 80%
5/5 = 100%

Figure 26 Partial example of variant calling output.
POS=coordinate position; REF=reference base; GEN= individual genotype in IUPAC nomenclature; VAR Freq=variant frequency; Variants are reported with IUPAC nomenclature.

In total, we identified 115 SNPs of which 60 were never described before. Moreover we identified 11 indels and only one was reported previously.

All indels and most of SNPs were intronic, but we identified also two exonic variants from SMA patients.

We noted few single nucleotide variants with very low frequencies (<15%). When variants were called only on reads that mapped at the expected distance in proper orientation or at the reduced coverage, those low frequency variants disappeared, indicating them as erroneous calls most probably caused by misalignments.

Moreover variant mapping showed low-frequency variants as predominantly located in intronic homopolymers or placed near indels or otherwise showed ‘strand bias’. The last phenomenon arose when genotype inferred from the positive strand and negative strand were significantly different. In addition, such variants were rarely private, ubiquitously present across individuals, sustaining inaccuracies rather than a presence of a true somatic variation. So we decided to eliminate all low frequency variants from our results.

Most indels were located in intronic homopolymers. The alignment in homopolymer region is highly error prone and for this reason those indels were removed decreasing the number of identified indels from 11 to 5.

2 Validation SMN2 sequencing

2.1 Variants validation

The NGS-sequencing results were validated by Sanger sequencing and single base extension (SBE). The SBE was performed as described in paragraph 1.9 of materials and methods. It relied on a highly specific primer which complementary binds with the 3’ end adjacent to the target SNP. The former oligonucleotide was enzymatically extended to its 3’ end by only a single base using a fluorescently labeled ddNTP which did not allow any further extension for the absence of 3'-hydroxyl group. This method was used to increase validation rate because most variants were surrounded by repetitive regions that fail to be sequenced by Sanger sequencing. Instead SBE allowed variant targeting even when repetitive regions were present in the analyzed amplicon. As shown in Figure 27, we chose to validate a total of 38 SNPs (one-third of found SNPs) and all five indels.

SNPs			
coordinate	validation	coordinate	validation
69345130	V	69367063	V
69348033	V	69368270	V
69348440	V	69368571	V
69355622	V	69368717	V
69356085	V	69368815	V
69356114	V	69368911	V
69357190	V	69370451	/
69357245	V	69370574	V
69357509	V	69370591	V
69358605	V	69370594	V
69359017	V	69370731	/
69359244	V	69370742	V
69359824	V	69370895	V
69360743	V	69371063	V
69362410	V	69371300	V
69363717	V	69371368	V
69364605	V	69371981	V
69365934	V	69372353	V
69367010	V	69373081	V

INDELS	
coordinate	validation
69344454	/
69344902	V
69345023	/
69353437	V
69357016	/
69365975	/
69371439	/

Figure 27 List of validated variants. Green check mark stands for true positive variant; red bar stand for not validated variants.

We were able to validate and confirm 36 out of 38 SNPs and only 2 out of 5 indels. Two SNPs and 3 indels were not confirmed by Sanger due to the failure of PCR amplification of variant surrounding regions.

Five of the validated variants were not present in publically available databases. The greatest observation was that the variant frequencies were indicative of the number of variant gene copies per subject, when related to the total *SMN2* copy numbers of the patient. The samples' genotype was determinate by three independent methods, real-time PCR with Taqman MGB technology (Passon et al. 2009), MLPA (Passon et al., 2010) and single base extension (unpublished data of our laboratory). All techniques quantified the number of *SMN2* by taking advantage of the single C to T substitution in exon 7 and

A to G in exon 8. The observed variant frequencies were in accordance with expected frequency considering the individuals SMN2 copy numbers.

The possibility to gather quantitative information from this approach was an important point, especially for diagnosis or prognosis purposes. The possibility of obtaining quantitative information throughout the entire gene in SMA is important in order to recognize samples with a partial deletion of target gene. While the diagnosis is focused in detecting the exon 7 deletion, the NGS sequencing identified variants through the entire gene. We found a patient with different variant frequencies between the 5' and the 3' of the gene. As clearly represented in Figure 28, the red line divides variants with frequencies of ~50% and ~33% indicating a change from four to three SMN2 copy numbers, i.e. a partial deletion. Due to the lack of informative variants the breaking point was localized between intron 4 and 7. To clearly validate the presence of a partial SMN copy, we will need to enlarge the SBE (set up for exon 7 and 8 only) to all *SMN2* exons.

69359244	exon 2a	rs26788	2611	2436	48,24%
69359824	intron 2	rs26789	3486	3683	51,33%
69360500	intron 2				
69360743	intron 2	rs26790	3598	3767	51,10%
69361437	intron 3				
69361500	intron 3				
69362410	intron 3	rs26791	3083	2966	48,98%
69362949	exon 3	rs4915	3726	4210	52,98%
69363365					
69363623	intron 6				
69363717	intron 6	rs150422			
69364605	intron 6	rs111871	5	5828	99,81%
69364893					
69365216	intron 7	rs152110	2169	4073	65,08%
69365934	intron 7				
69366414	intron 7	rs62374808			
69366625	intron 7				
69367010	intron 7	rs150419			
69367063	intron 7	rs152159			
69367348	intron 7				
69367578	intron 7				
69367742	intron 7				
69367840	intron 7	rs460163	4005	1982	33,07%
69368084	intron 7	rs212228	4012	1705	29,75%
69368206	intron 7				
69368270	intron 7				
69368329	intron 7	rs212227	3511	1698	32,59%

Figure 28 Representation of variant frequencies of one patient; red line indicates the breaking point between frequencies.

To determine if the variant (A>G at position g.69356085) with 100% frequency common to all patients have any link with SMA phenotype we analyzed by SBE in fifty individuals with no correlations with SMA. All tested individuals were homozygous for the variant indicating that that particular position is a private mutation present in the reference sequence.

2.2 Variant distribution

Validated variants are distributed along SMN2 gene as follows:

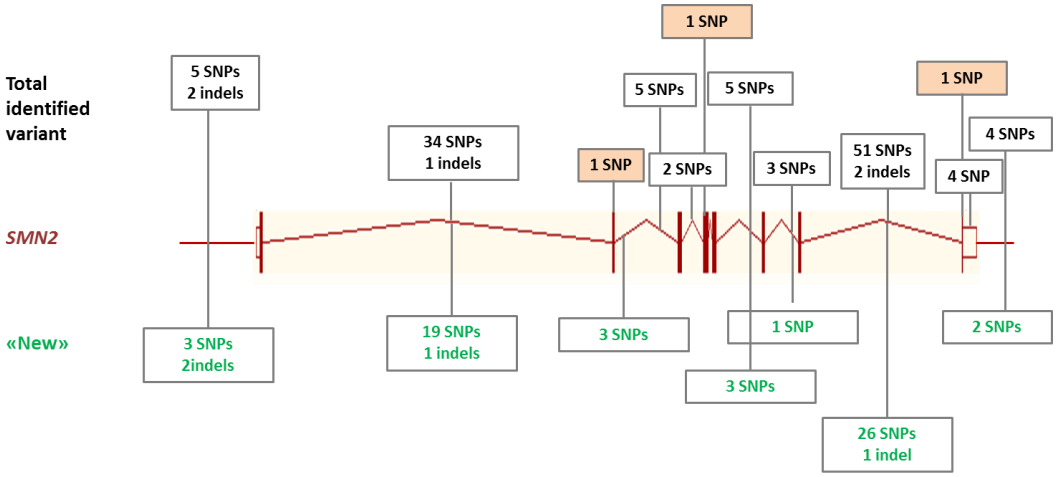


Figure 29 Repreentation of variants distribution along SMN2. Above all found variants, below "new" variant, not present in databases.

As shown in Figure 29, the variant distribution was not homogenous but rather clustered along the SMN2 gene. Particularly in intron 7 we found 51 SNPs, which is 50% more in respect to those found in intron 1. Moreover, the intron 7 was less than a half in length compared to intron 1. We therefore calculated the variant rate for each intron by dividing the total number of variants per intron by intron length. Intron 7 had the highest number of variants, with a rate of one polymorphism every 112bp whereas intron 3 was characterized by the lowest number of variants, having 9 times less variants than intron 7. The second region with the highest variant rate (1:152bp) was intron 8 and 3'UTR. Thus the two regions with the highest rate of variant were those flanking exon7. It is interesting to note that this exon, which is actually subjected to alternative splicing and directly involved in the modulation of the SMA phenotype, is flanked by regions characterized by increased polymorphisms rate as hotspot region.

After considering variant distribution, we looked if our samples present well know variants already reported in literature. Prior and colleagues in 2009

reported a c.859G>C substitution that originates a new exonic splicing enhancer element. It increases the amount of full-length transcripts resulting in less severe phenotype of analyzed patient. Thus c.859G>C substitution is a positive modifier of the SMA phenotype and indirectly demonstrates that SMN2 genes are not equivalent. Prior et al. were the first in introducing the concept of the effect of SMN2 sequence variations on the disease severity. We had found that None of 41 sequenced individual in our cohort retained the c.859G>C.

As represented in figure 29, variant distribution showed also the presence of two exonic variants that were validated by Sanger sequencing: c.C84T in exon 2a and c.A462G in exon 3, both causing synonymous substitutions, p.S28S and p.Q154Q, respectively. Exon 2a variant falls exactly in the third base of the first codon. It was present in only two individuals of the same family. This mutation introduced a potential but weak splicing element as suggested by Human Splicing Finder 2.4.1 prediction (Desmet et al., 2009). Exon 3 variant falls in the twelfth base before the end of exon3 and was characterized by a prediction tool as an EIE (Exon-Identity Element) site broken mutation and also a great new acceptor splicing element. Brahe and colleagues first-described this substitution as more frequently found in affected individual (Brahe et al., 1996). In our cohort 33 SMA patients had the variant. Many efforts were spent during this project to relate previously described mutation with our NGS variants. We could clearly associate only the c.A462G (p.Q154Q) mutation due to upgrade of coordinates and problems with renaming of variants in databases.

We found several intronic mutations that mapped in the 100bo regions flanking an exon and that are usually considered as splicing modulators. In particular intron 6

and 7 bearded two of those mutations indicating that splicing of those exons was finely modulated in *SMN2*. However, further experiments should be done in order to validate the *in silico* predictions and observations of functional activity of these mutations. And finally, experiments with mini-genes or splicing in vitro assay should be done to definitively prove the alteration of the normal splicing.

Since the *SMN2* re-sequencing revealed new indels, but their validation was not possible due to amplification problems in both Sanger sequencing and SBE, we decided to search in literature if those indels were previously described. Harahap and colleagues (Harahap et al., 2014) recently described an insertion c.320_321+GCC. They found this insertion in only one in 50 SMA analyzed patients and described it as having a "slight but significative negative effect on transcript efficiency". They finally concluded that there was a correlation between insertion and the mild SMA phenotype of the subject with only 2 *SMN2* copies. Our results showed +GCC insertion in 12 SMA affected individuals with different phenotypes (from type I to III). We did not find any other information about indels that were usually published when located in coding regions.

3 *SMN2* variants and functional transcripts

Our interest in *SMN2* re-sequencing arises from the fact that *SMN2* is a genetic modifier of SMA disease severity and an attractive molecular target in therapeutic strategies. The *SMN2* copy number was inversely correlated with the severity of the SMA phenotype (McAndrew et al., 1997; Mailman et al., 2002). However, there are SMA cases for which the severity of phenotype does not follow that correlation. Hence phenotype modulators acquired importance. The SMN2 sequencing was a great way to search for variants or haplotypes with certain function that can modulate SMN2 transcript production.

Our unpublished data of overall and full-length (FL) S*MN1* and *SMN2* steady state mRNA transcript levels in peripheral blood cells obtained by a SBE genotyping assay, which took advantage of SNPs in exon7 and 8 to distinguish between transcripts deriving from the two *SMN* genes, showed that proportions of overall *SMN1* and *SMN2* mRNA transcripts were in line with the respective allelic copy numbers across different genotypes, indicating that two genes are transcriptionally equivalent with very similar activities of their promoters. On the other hand, proportions of S*MN1* and *SMN2* -FL transcripts were in discordance with the respective allelic copy numbers, indicating that the two genes are not equivalent in FL-RNA production. The latter finding was in agreement with the literature data.

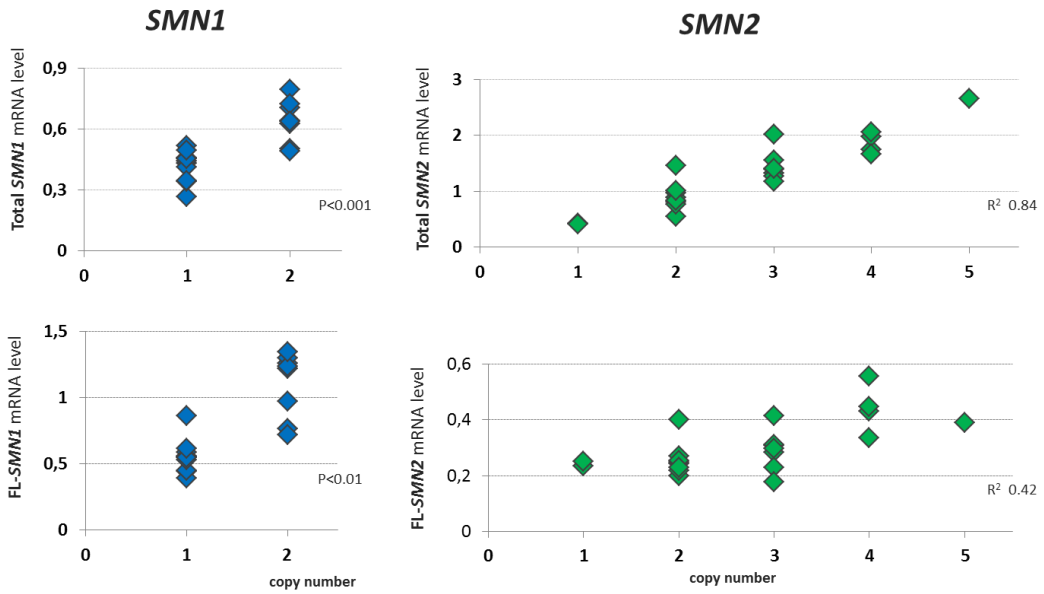


Figure 30 Total and full length mRNA expression. A SMN1 and SMN2 comparison.

Both expression values were found doubled between one and two *SMN1* copies in peripheral blood cells (blue graphs in Figure 30). Similarly, there is a tight linear correlation of *SMN2* total mRNA levels with *SMN2* copy number (green graph, Figure 30 above). Nevertheless, the same correlation was not observed when measuring *SMN2* FL-mRNA levels (green graph, Figure 30 below), suggesting that *SMN2* alleles are not functionally equivalent since they produce FL-mRNA with different efficiencies.

Further experiments should be done to deepen our knowledge on correlation between a single, or a set of variants, and different levels of FL-SMN2 transcript production. In order to find out association between variants and SMN2 functionality, the SMA patient should be at least doubled to reach statistical significance.

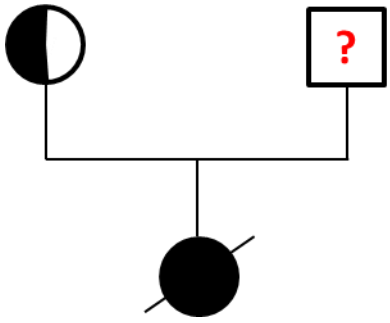
Our study, starting from the experimental design to the knowledge gained by NGS approach is a great starting point for further research. Finally, our last aim was to gather all SMN2 variant together for an easier patient management and their enrolment in clinical trials.

4 *Case reports*

We took advantage of SMN2 sequencing by NGS to include also two peculiar family cases. A deep knowledge of the SMN sequence would be helpful to genetic counseling of these families.

4.1 Family 1

According to the previous clinical diagnosis of type I SMA, the child had homozygous deletion of *SMN1*.



Subject	# SMN1	# SMN2
Mother	1	2
Father	2	2
Daughter	0	2

Figure 31 Family tree and real-time PCR resulting genotype.

While the mother was a healthy deletion carrier presenting only one *SMN1*, the father had two *SMN1* copies. The only *SMN1* dosage for these three family members was not sufficient to immediately explain the inheritance of the two deleted *SMN1* alleles in the child.

Therefore, three different hypotheses were taken into account:

- the father was a "**2/0 genotype**" indeed presenting two *SMN1* copies on the same chromosome 5 [about 2-5% of SMA carriers] and a deletion of the gene on the other, which was the one inherited by the daughter;

- the child inherited two maternal chromosome 5 carrying the *SMN1* deletion because of a **uniparental maternal isodisomy (isoUPD)** mechanism;
- ***de novo SMN1* deletion** on the paternal chromosome 5 [2% of SMA patients] occurred either during gametogenesis or in the earliest stages of embryonic development.

In order to assess whether the father was a silent deletion carrier ("2/0 genotype") the analysis was expanded to other key family members. In most cases, individuals having the "2/0 genotype" took origin from one parent with three SMN1 gene copies and the other having one *SMN1* copy, as represented in Figure 32.

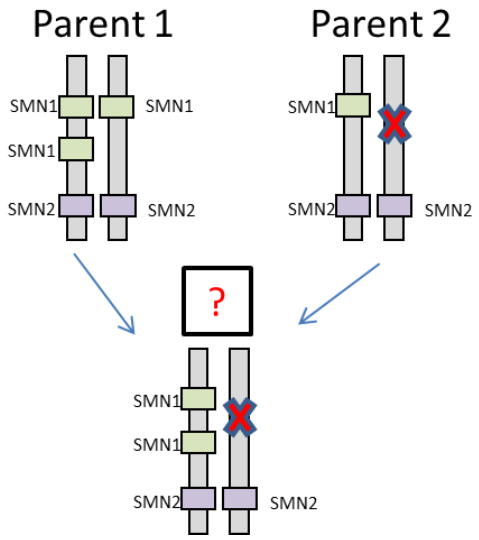


Figure 32 Most probable alleles configuration resulting in a 2-0 genotype.

Unfortunately, the grandfather of the proband died years ago so DNA sample was not available. Instead all available blood-paternal relatives were however genotyped by qPCR as previously made for the proband and her parents, looking

for at least one individual presenting either one or three SMN1 alleles. None of the analyzed relatives presented three or one SMN1 gene copies. This did not exclude, but made the "2/0 genotype" in the father really unlikely. In conclusion, the father has probably one SMN1 gene copy on each allele.

The second hypothesis was the maternal uniparental isodisomy (isoUPD), as representing the inheritance of two copies of the same parental homologue due to non-disjunction in meiosis II. Homozygosity for a recessive mutation could indeed be caused by this mechanism. If this was the case, chromosome 5 of the child should present exclusively maternal contribution. Thus the *SMN2* sequencing by NGS was useful to look through the entire gene. Many SNPs were found in the SMN genes of each of the three family members, most of all mapping across intron 6.

As represented in the Figure 33, only one common SNP between mother and newborn was found.

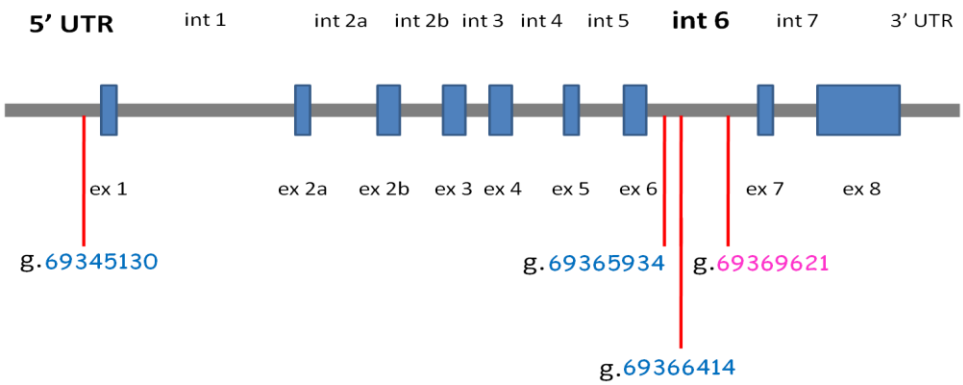


Figure 33 Schematic representation of the proband's SMN2 genes and genomic position of the unique maternally (pink) or paternally inherited variants (genomic coordinates).

On the other hand, three common SNPs with father were found suggesting that the proband inherited her SMN2 genes from both the parents. Sanger sequencing and

SBE validated all found variants confirming SNPs in common between father and child. The hypothesis of maternal isoUPD failed due to the presence of a biparental contribution as regards the SMN2 genes of the child. It must, though, be considered that the four significant variants found by our NGS approach, did not allow inferring the inheritance pattern of the region downstream the SMN2 gene of the proband, where *SMN1* is located (Figure 34).

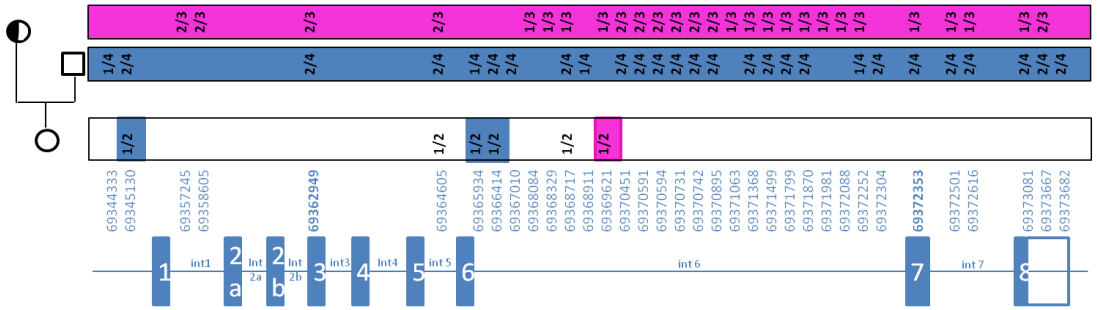


Figure 34 Representation of detected variant positions through SMN genes of parent and proband.

Thus we further performed a molecular analysis using ten polymorphic markers (VNTRs) spanning a region of about 5 Mb across the SMA locus on chromosome 5q13. As Figure 35 shows, only four of the above represented VNTRs were informative (D5S435, D5S557, D5S610, D5S351, represented in orange), one mapping about 1.2 Mb from the 5' of the SMN2 gene and three spanning about 1.3 Mb at the 3' of *SMN1*.

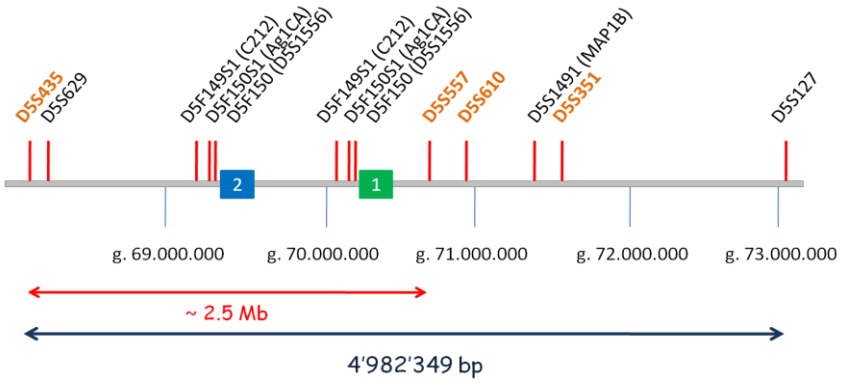


Figure 35 Schematic representation of VNTRs spanning about 5Mb across the SMA locus on chromosome 5q11.2-q13.3. Orange highlights the informative one.

The informative polymorphisms effectively demonstrated a biparental inheritance of the chromosome 5 because the proband is heterozygous at these loci, with one allele belonging to the father and one to the mother (Figure 36).

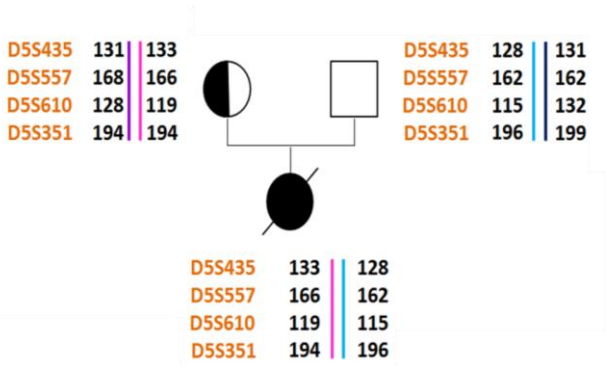


Figure 36 Four polymorphic markers demonstrating a biparental inheritance of chromosome 5.

Both NGS approach and VNTR analysis support a biparental inheritance of the SMA locus for the child, suggesting the occurrence of a *de novo* deletion of the SMN1 gene on a paternal chromosome 5. The mutation probably interested only a single paternal sperm cell. Indeed also a certain percentage was interested due to

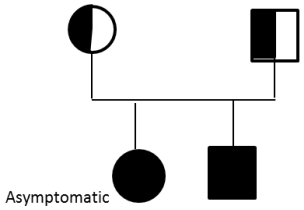
mutation of precursor sperm cell. In this latter case, the father represented a condition of *gonadal mosaicism* in which a post-fertilization mutation was confined to the gamete precursors and was not detected in somatic tissues (Anazi *et al.*, 2014). Alternatively, a *de novo* mitotic deletion of *SMN1* could have occurred spontaneously after the fertilization.

We did not identify the moment at which the *de novo* mutation occurred but the identification of the method was important for the genetic counselor.

The frequency of occurrence was rather different between the method supposed above.

4.2 Family 2

We decided to include into the *SMN2* sequencing also a second family, composed by male and female siblings and their healthy SMA carrier parents (Figure 37). Both siblings showed homozygous deletions of *SMN1* and were expected to develop SMA. However, the female was fully asymptomatic despite carrying four *SMN2* copies and a paternal-inherited gene conversion as her affected relative.



Subject	# SMN1	# SMN2
Mother	1	3
Father	1	2
Daughter	0	4
Son	0	4

Gene conversion

Figure 37 Family tree and real-time resulting genotypes.

Although siblings had the same *SMN2* copy number, the identification of differential variant between them could explain their phenotype. As shown in Figure 38, both siblings inherited the same variants.

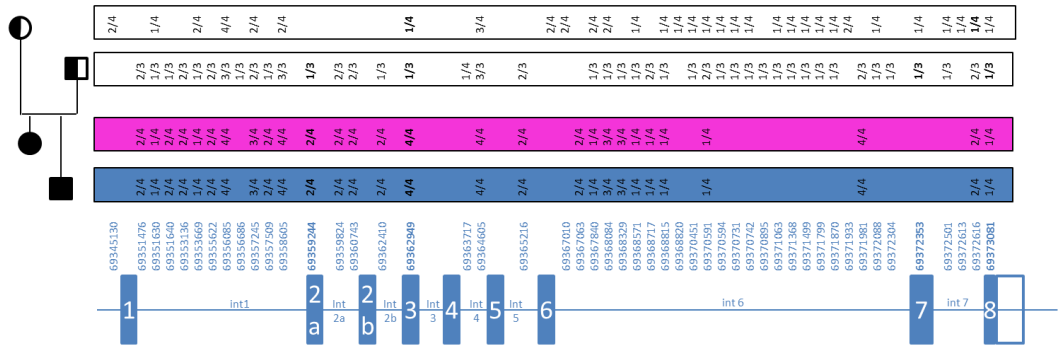


Figure 38 Representation of detected variant positions through SMN genes of parent and probands.

Moreover our unpublished data reported also a comparable expression of PLS3 between siblings in blood samples. Indeed for this family case, none of the two previously reported modulators genes is the causative. Therefore the explanation of the asymptomatic state of the female is laid elsewhere.

4.3 *SMN* variant distribution

With the inclusion of relatives of SMA patients, we sequenced their *SMN2* as well as *SMN1*. Due to 5 nucleotide changes g.69372304A>G, g.69372353C>T, g.69372501 G>A, g.69372616G>A, g.69373081A>G located from intron 7 to exon 8 (as shown in Figure 13 of Introduction) that differ between the two genes (Bürglen et al., 1996), we were able to clearly quantify both genes for all sequenced relatives. Moreover taking advantage of variant frequencies of those

base changes, we clearly identify gene conversions in both SMA patients and relatives. This was a further validation of the quantitative aspect of this method.

The patient relative variant analysis shows four positions in intron 7 that are present uniquely in these individuals with variant frequency correlated to their SMN1 copy number and thus are probably linked to their SMN1 gene.

5 *NGS, a new approach to diagnostic genetic testing*

Disease target gene panels are the most suitable level of analysis for clinical requirements. After development of the laboratory test, it must then be fully validated (Gargis et al., 2012; Rhem et al, 2013). The validation process is composed of three interconnected components, as shown in Figure 39.

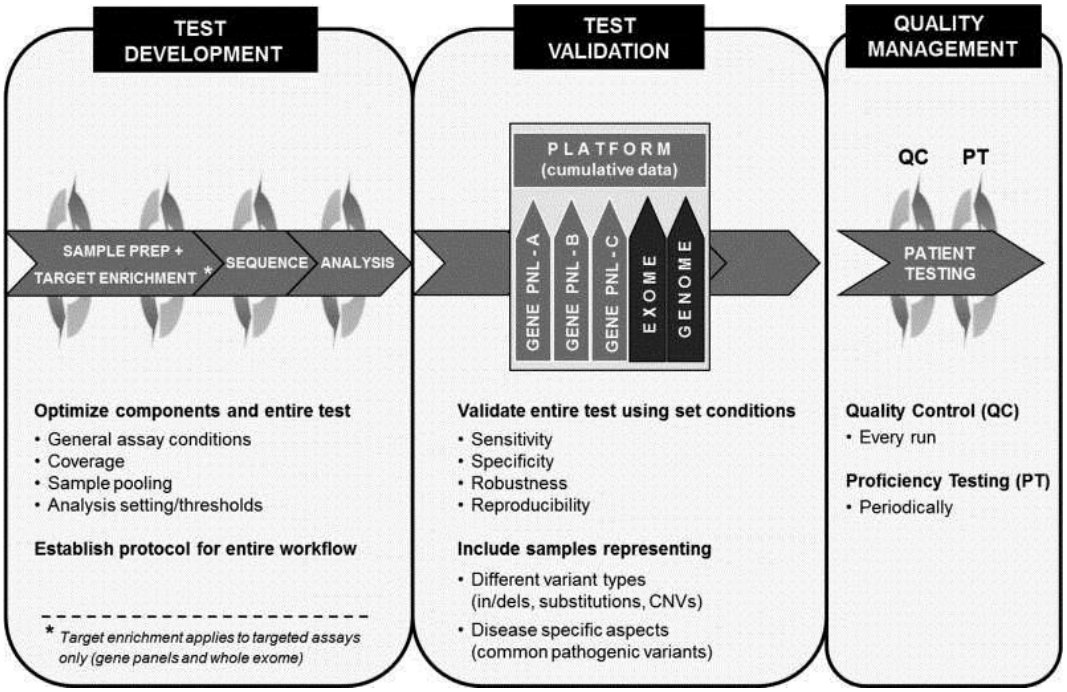


Figure 39 NGS test development and validation process.

We set up an NGS workflow that completely fulfills American College of Medical Genetics and Genomics guidelines. First of all, we successfully sequenced and analyzed uniquely the target region. Secondly, we performed the test system validation process, in which a group of samples with known variants were sequenced, in order to evaluate the ability of the workflow to identify target variants. Moreover we tested the approach precision by assessing reproducibility

and repeatability. We fulfilled those requirements since we obtained the same variant calling by the same sample that was run multiple times under either same or different conditions (i.e. different technicians and instruments). The analytical sensitivity and specificity were also obtained because there was complete concordance between NGS variant calling and Sanger sequencing of known variants. This approach was accurate due to the average depth of coverage and its uniformity. Finally, allelic read percentage was in agreement with the multigene feature of the pathology. The third validation step will be tested when the workflow will enter the clinical routine.

Finally, a further requirement for introducing NGS in clinical diagnostic is the evaluation of the turnaround time. Our approach required a month to obtain a medical report from the blood samples, highly reduced in comparison with the Sanger sequencing of the same number of individuals.

CONCLUSIONS

The combination of LR-PCR amplification and NGS of SMN2 gene allows the genetic testing in parallel of several SMA patients. We successfully identified SNPs and indels from 41 individual but in order to identify a genotype to phenotype correlation with statistical significance an increased number of individual should be employed for sequencing.

The entire workflow was set up to suit ACMG guidelines and thus our approach may prove effective in hospital laboratories since the validation process was successfully obtained.

BIBLIOGRAPHY

Abul-Husn NS, Owusu Obeng A, Sanderson SC, Gottesman O, Scott SA. *Implementation and utilization of genetic testing in personalized medicine.* Pharmgenomics Pers Med. 2014 Aug 13;7:227-40.Review.

Ackermann B, Kröber S, Torres-Benito L, Borgmann A, Peters M, Hosseini Barkooie SM, Tejero R, Jakubik M, Schreml J, Milbradt J, Wunderlich TF, Riessland M, Tabares L, Wirth B. *Plastin 3 ameliorates spinal muscular atrophy via delayed axon pruning and improves neuromuscular junction functionality*. Hum Mol Genet. 2013 Apr 1;22(7):1328-47.

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, Shendure J. *Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition*. Genome Biol. 2010;11(12):R119.

Alías L, Barceló MJ, Bernal S, Martínez-Hernández R, Also-Rallo E, Vázquez C, Santana A, Millán JM, Baiget M, Tizzano EF. *Improving detection and genetic counseling in carriers of spinal muscular atrophy with two copies of the SMN1 gene*. Clin Genet. 2014 May;85(5):470-5.

Anazi S, Al-Sabban E, Alkuraya FS. *Gonadal mosaicism as a rare cause of autosomal recessive inheritance*. Clin Genet. 2014 Mar;85(3):278-81.

Andreassi C, Angelozzi C, Tiziano FD, Vitali T, De Vincenzi E, Boninsegna A, Villanova M, Bertini E, Pini A, Neri G, Brahe C. Phenylbutyrate increases SMN expression in vitro: relevance for treatment of spinal muscular atrophy. Eur J Hum Genet. 2004 Jan;12(1):59-65.

Ansorge WJ. *Next-generation DNA sequencing techniques*. N Biotechnol. 2009 Apr;25(4):195-203. Epub 2009 Feb 3.

Azzouz M, Le T, Ralph GS, Walmsley L, Monani UR, Lee DC, Wilkes F, Mitrophanous KA, Kingsman SM, Burghes AH, Mazarakis ND. *Lentivector-mediated SMN replacement in a mouse model of spinal muscular atrophy*. J Clin Invest. 2004 Dec;114(12):1726-31.

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. 2002 Aug 30;277(35):31957-62.

Bebee TW , Dominguez CE, Chandler DS. Mouse models of SMA: tools for disease characterization and therapeutic development. Hum Genet. 2012 Aug;131(8):1277-93.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ,Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A,

Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E, Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP,Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature. 2008 Nov 6;456(7218):53-9.

Bergin A, Kim G, Price DL, Sisodia SS, Lee MK, Rabin BA. *Identification and characterization of a mouse homologue of the spinal muscular atrophy-determining gene, survival motor neuron*. Gene. 1997 Dec 19;204(1-2):47-53.

Bernal S, Alías L, Barceló MJ, Also-Rallo E, Martínez-Hernández R, Gámez J, Guillén-Navarro E, Rosell J, Hernando I, Rodríguez-Alvarez FJ, Borrego S, Millán JM, Hernández-Chico C, Baiget M, Fuentes-Prior P, Tizzano EF. *The c.859G>C variant in the SMN2 gene is associated with types II and III SMA and originates from a common ancestor*. J Med Genet. 2010 Sep;47(9):640-2.

Bernal S, Also-Rallo E, Martínez-Hernández R, Alías L, Rodríguez-Alvarez FJ, Millán JM, Hernández-Chico C, Baiget M, Tizzano EF. *Plastin 3 expression in discordant spinal muscular atrophy (SMA) siblings*. Neuromuscul Disord. 2011 Jun;21(6):413-9.

Brahe C, Servidei S, Zappata S, Ricci E, Tonali P, Neri G. *Genetic homogeneity between childhood-onset and adult-onset autosomal recessive spinal muscular atrophy*. Lancet. 1995;346(8977):741–2.

Brahe C, Clermont O, Zappata S, Tiziano F, Melki J, Neri G. *Frameshift mutation in the survival motor neuron gene in a severe case of SMA type I*. Hum Mol Genet. 1996 Dec;5(12):1971-6.

Brzustowicz LM, Lehner T, Castilla LH, Penchaszadeh GK, Wilhelmsen KC, Daniels R, Davies KE, Leppert M, Ziter F, Wood D, et al. *Genetic mapping of chronic childhood-onset spinal muscular atrophy to chromosome 5q11.2-13.3*. Nature. 1990 Apr 5;344(6266):540-1.

Brzustowicz LM, Allitto BA, Matseoane D, Theve R, Michaud L, Chatkupt S, Sugarman E, Penchaszadeh GK, Suslak L, Koenigsberger MR, et al. *Paternal isodisomy for chromosome 5 in a child with spinal muscular atrophy*. Am J Hum Genet. 1994 Mar;54(3):482-8.

Burghes AH. When is a deletion not a deletion? When it is converted. Am J Hum Genet. 1997 Jul;61(1):9-15.

Bürglen L, Lefebvre S, Clermont O, Burlet P, Viollet L, Cruaud C, Munnich A, Melki J. *Structure and organization of the human survival motor neurone (SMN) gene*. Genomics. 1996 Mar 15;32(3):479-82.

Bussaglia E, Clermont O, Tizzano E, Lefebvre S, Bürglen L, Cruaud C, Urtizberea JA, Colomer J, Munnich A, Baiget M, et al. *A frame-shift deletion in the survival motor neuron gene in Spanish spinal muscular atrophy patients*. Nat Genet. 1995 Nov;11(3):335-7.

Campbell L , Potter A, Ignatius J, Dubowitz V, Davies K. *Genomic variation and gene conversion in spinal muscular atrophy: implications for disease process and clinical phenotype*. Am J Hum Genet. 1997 Jul;61(1):40-50.

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA. *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing*. Nat Genet. 2008 Jun;40(6):722-9.

Cartegni L, Krainer AR. *Disruption of an SF2/ASF-dependent exonic splicing enhancer in SMN2 causes spinal muscular atrophy in the absence of SMN1*. Nat Genet. 2002 Apr;30(4):377-84.

Caruccio N. *Preparation of next-generation sequencing libraries using Nextera™ technology: simultaneous DNA fragmentation and adaptor tagging by in vitro transposition.* Methods Mol Biol. 2011;733:241-55.

Casals F, Idaghdour Y, Hussin J, Awadalla P. *Next-generation sequencing approaches for genetic mapping of complex diseases.* J Neuroimmunol. 2012 Jul 15;248(1-2):10-22.

Chiche JD, Cariou A, Mira JP. *Bench-to-bedside review: fulfilling promises of the Human Genome Project.* Crit Care. 2002 Jun;6(3):212-5. Epub 2002 Mar 20. Review.

Clermont O, Burlet P, Bürglen L, Lefebvre S, Pascal F, McPherson J, Wasmuth JJ, Cohen D, Le Paslier D, Weissenbach J, 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. 1994 Apr;54(4):687-94.

Clermont O, Burlet P, Lefebvre S, Bürglen L, Munnich A, Melki J. *SMN gene deletions in adult-onset spinal muscular atrophy*. Lancet. 1995 Dec 23-30;346(8991-8992):1712-3.

Cobben JM, van der Steege G, Grootscholten P, de Visser M, Scheffer H, Buys CH. *Deletions of the survival motor neuron gene in unaffected siblings of patients with spinal muscular atrophy*. Am J Hum Genet. 1995 Oct;57(4):805-8.

Cobben JM, Scheffer H, De Visser M, Van der Steege G, Verhey JB, Osinga J, Burton M, Mensink RG, Grootscholten PM, Ten Kate LP, Buys CH. Prenatal prediction of spinal muscular atrophy. Experience with linkage studies and consequences of present SMN deletion analysis. Eur J Hum Genet. 1996;4(4):231-6.

Coovert DD, Le TT, McAndrew PE, Strasswimmer J, Crawford TO, Mendell JR, Coulson SE, Androphy EJ, Prior TW, Burghes AH. *The survival motor neuron protein in spinal muscular atrophy*. Hum Mol Genet. 1997 Aug;6(8):1205-14.

Corti S, Nizzardo M, Nardini M, Donadoni C, Salani S, Del Bo R, Papadimitriou D, Locatelli F, Mezzina N, Gianni F, Bresolin N, Comi GP. Motoneuron transplantation rescues the phenotype of SMARD1 (spinal muscular atrophy with respiratory distress type 1). J Neurosci. 2009 Sep 23;29(38):11761-71.

Cuscó I, Barceló MJ, Baiget M, Tizzano EF. *Implementation of SMA carrier testing in genetic laboratories: comparison of two methods for quantifying the SMN1 gene*. Hum Mutat. 2002 Dec;20(6):452-9.

Cuscó I, López E, Soler-Botija C, Jesús Barceló M, Baiget M, Tizzano EF. *A genetic and phenotypic analysis in Spanish spinal muscular atrophy patients with c.399_402del AGAG, the most frequently found subtle mutation in the SMN1 gene*. Hum Mutat. 2003 Aug;22(2):136-43.

Cuscó I, Barceló MJ, Rojas-García R, Illa I, Gámez J, Cervera C, Pou A, Izquierdo G, Baiget M, Tizzano EF. *SMN2 copy number predicts acute or chronic spinal muscular atrophy but does not account for intrafamilial variability in siblings*. J Neurol. 2006 Jan;253(1):21-5.

Desmet FO, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. *Human Splicing Finder: an online bioinformatics tool to predict splicing signals*. Nucleic Acid Research, 2009, April

DiDonato CJ, Chen XN, et al. *Cloning, characterization, and copy number of the murine survival motor neuron gene: homolog of the spinal muscular atrophy-determining gene*. Genome Res. 1997;7(4):339–52.

Emilien G, Ponchon M, Caldas C, Isacson O, Maloteaux JM. *Impact of genomics on drug discovery and clinical medicine.* QJM. 2000 Jul;93(7):391-423. Review.

Engel E. *A new genetic concept: uniparental disomy and its potential effect, isodisomy*. Am J Med Genet. 1980;6(2):137-43.

Engel E. *A fascination with chromosome rescue in uniparental disomy: Mendelian recessive outlaws and imprinting copyrights infringements*. Eur J Hum Genet. 2006 Nov;14(11):1158-69.

Feldkötter M, Schwarzer V, Wirth R, Wienker TF, Wirth B. *Quantitative analyses of SMN1 and SMN2 based on real-time lightCycler PCR: fast and highly reliable carrier testing and prediction of severity of spinal muscular atrophy*. Am J Hum Genet. 2002 Feb;70(2):358-68.

Fischbeck KH, Souders D, La Spada A. *A candidate gene for X-linked spinal muscular atrophy*. Adv Neurol. 1991;56:209-13.

Fischer U, Liu Q, Dreyfuss G. The SMN-SIP1 complex has an essential role in spliceosomal snRNP biogenesis. Cell. 1997 Sep 19;90(6):1023-9.

França LT, Carrilho E, Kist TB. *A review of DNA sequencing techniques.* Q Rev Biophys. 2002 May;35(2):169-200.

Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, Lu F, Lyon E, Voelkerding KV, Zehnbauer BA, Agarwala R, Bennett SF, Chen B, Chin EL, Compton JG, Das S, Farkas DH, Ferber MJ, Funke BH, Furtado MR, Ganova-Raeva LM, Geigenmüller U, Gunselman SJ, Hegde MR, Johnson PL, Kasarskis A, Kulkarni S, Lenk T, Liu CS, Manion M, Manolio TA, Mardis ER, Merker JD, Rajeevan MS, Reese MG, Rehm HL, Simen BB, Yeakley JM, Zook JM, Lubin IM. *Assuring the quality of next-generation sequencing in clinical laboratory practice.* Nat Biotechnol. 2012 Nov;30(11):1033-6.

Gómez-Curet I, Robinson KG, Funanage VL, Crawford TO, Scavina M, Wang W. *Robust quantification of the SMN gene copy number by real-time TaqMan PCR*. Neurogenetics. 2007 Nov;8(4):271-8.

Grohmann K, Schuelke M, Diers A, Hoffmann K, Lucke B, Adams C, Bertini E, Leonhardt-Horti H, Muntoni F, Ouvrier R, Pfeufer A, Rossi R, Van Maldergem L,Wilmshurst JM, Wienker TF, Sendtner M, Rudnik-Schöneborn S, Zerres K, Hübner C. *Mutations in the gene encoding immunoglobulin mu-binding protein 2 cause spinal muscular atrophy with respiratory distress type 1*. Nat Genet. 2001 Sep;29(1):75-7.

Gubitz AK, Feng W, Dreyfuss G. *The SMN complex*. Exp Cell Res. 2004 May 15;296(1):51-6

Haas J, Katus HA, Meder B. *Next-generation sequencing entering the clinical arena*. Mol Cell Probes. 2011 Oct-Dec;25(5-6):206-11. Epub 2011 Sep 8.

Hahnen E, Forkert R, Marke C, Rudnik-Schöneborn S, Schönling J, Zerres K, Wirth B. *Molecular analysis of candidate genes on chromosome 5q13 in autosomal recessive spinal muscular atrophy: evidence of homozygous deletions of the SMN gene in unaffected individuals*. Hum Mol Genet. 1995 Oct;4(10):1927-33.

Hahnen E, Schönling J, Rudnik-Schöneborn S, Zerres K, Wirth B. *Hybrid survival motor neuron genes in patients with autosomal recessive spinal muscular atrophy: new insights into molecular mechanisms responsible for the disease*. Am J Hum Genet. 1996 Nov;59(5):1057-65.

Hahnen E, Eyüpoglu IY, Brichta L, Haastert K, Tränkle C, Siebzehnrübl FA, Riessland M, Hölker I, Claus P, Romstöck J, Buslei R, Wirth B, Blümcke I. *In vitro and ex vivo evaluation of second-generation histone deacetilase inhibitors for treatment of spinal muscular atrophy*. J Neurochem. 2006 Jul;98(1):193-202.

Hao le T, Wolman M, Granato M, Beattie CE. *Survival motor neuron affects plastin 3 protein levels leading to motor defects*. J Neurosci. 2012 Apr 11;32(15):5074-84.

Harahap NI, Takeuchi A, Yusoff S, Tominaga K, Okinaga T, Kitai Y, Takarada T, Kubo Y, Saito K, Sa'adah N, Nurputra DK, Nishimura N, Saito T, Nishio H. *Trinucleotide insertion in the SMN2 promoter may not be related to the clinical phenotype of SMA*. Brain Dev. 2014 Oct 31.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P. *Library construction for next-generation sequencing: overviews and challenges*. Biotechniques. 2014 Feb 1;56(2):61-4, 66, 68, passim.

Ingman M, Gyllensten U. *SNP frequency estimation using massively parallel sequencing of pooled DNA.* Eur J Hum Genet. 2009 Mar;17(3):383-6.

International Human Genome Sequencing Consortium.Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; *Initial sequencing and analysis of the human genome.*

Nature 2001 Feb 15;409(6822):860-921. Erratum in: Nature 2001 Aug 2;412(6846):565. Nature 2001 Jun 7;411(6838):720.

Kerr DA, Nery JP, Traystman RJ, Chau BN, Hardwick JM. Survival motor neuron protein modulates neuron-specific apoptosis. Proc Natl Acad Sci U S A. 2000 Nov 21;97(24):13312-7.

Kleyn PW , Wang CH, Lien LL, Vitale E, Pan J, Ross BM, Grunn A, Palmer DA, Warburton D, Brzustowicz LM, et al. *Construction of a yeast artificial chromosome contig spanning the spinal muscular atrophy disease gene region*. Proc Natl Acad Sci U S A. 1993 Jul 15;90(14):6801-5.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. *VarScan: variant detection in massively parallel sequencing of individual and pooled samples*. Bioinformatics. 2009 Sep 1;25(17):2283-5.

Le TT, Pham LT, Butchbach ME, Zhang HL, Monani UR, Coovert DD, Gavrilina TO, Xing L, Bassell GJ, Burghes AH. *SMNDelta7, the major product of the centromeric survival motor neuron (SMN2) gene, extends survival in mice with spinal muscular atrophy and associates with full-length SMN*. Hum Mol Genet. 2005 Mar 15;14(6):845-57.

Lefebvre S, Bürglen L, Reboullet S, Clermont O, Burlet P, Viollet L, Benichou B, Cruaud C, Millasseau P, Zeviani M, Le Paslier D, Frézal J, Cohen D, Weissenbach J, Munnich A, Melki J. *Identification and characterization of a spinal muscular atrophy-determining gene*. Cell. 1995;80(1):155–65.

Lefebvre S, Burlet P, Liu Q, Bertrandy S, Clermont O, Munnich A, Dreyfuss G, Melki J. *Correlation between severity and SMN protein level in spinal muscular atrophy*. Nat Genet. 1997 Jul;16(3):265-9.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC. *The diploid genome sequence of an individual human.* PLoS Biol. 2007 Sep 4;5(10):e254.

Li H, Durbin R. *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics. 2009 Jul 15;25(14):1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. *The Sequence Alignment/Map format and SAMtools.* Bioinformatics. 2009 Aug 15;25(16):2078-9.

Liu Q, Dreyfuss G. *A novel nuclear structure containing the survival of motor neurons protein*. EMBO J. 1996 Jul 15;15(14):3555-65.

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. 1997 Sep 19;90(6):1013-21.

López-González R, Velasco I. *Therapeutic potential of motor neurons differentiated from embryonic stem cells and induced pluripotent stem cells*. Arch Med Res. 2012 Jan;43(1):1-10.

Lorson CL, Hahnen E, Androphy EJ, 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. 1999 May 25;96(11):6307-11.

Luo M, Liu L, Peter I, Zhu J, Scott SA, Zhao G, Eversley C, Kornreich R, Desnick RJ, Edelmann L. *An Ashkenazi Jewish SMN1 haplotype specific to duplication alleles improves pan-ethnic carrier screening for spinal muscular atrophy*. Genet Med. 2014 Feb;16(2):149-56.

Mailman MD , Heinz JW, Papp AC, Snyder PJ, Sedra MS, Wirth B, Burghes AH, Prior TW. *Molecular analysis of spinal muscular atrophy and modification of the phenotype by SMN2*. Genet Med. 2002 Jan-Feb;4(1):20-6.

Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. *Target-enrichment strategies for next-generation sequencing.* Nat Methods. 2010 Feb;7(2):111-8.

McAndrew PE, Parsons DW, Simard LR, Rochette C, Ray PN, Mendell JR, Prior TW, Burghes AH. *Identification of proximal spinal muscular atrophy carriers and patients by analysis of SMNT and SMNC gene copy number*. Am J Hum Genet. 1997 Jun;60(6):1411-22.

Melki J, Abdelhak S, Sheth P, Bachelot MF, Burlet P, Marcadet A, Aicardi J, Barois A, Carriere JP, Fardeau M, et al. *Gene for chronic proximal spinal muscular atrophies maps to chromosome 5q*. Nature. 1990 Apr 19;344(6268):767-8.

Melki J, Lefebvre S, Bürglen L, Burlet P, Clermont O, Millasseau P, Reboullet S, Bénichou B, Zeviani M, Le Paslier D, et al. *De novo and inherited deletions of the 5q13 region in spinal muscular atrophies*. Science. 1994 Jun 3;264(5164):1474-7.

Monani UR, Lorson CL, Parsons DW, Prior TW, Androphy EJ, Burghes AH, McPherson JD. *A single nucleotide difference that alters splicing patterns distinguishes the SMA gene SMN1 from the copy gene SMN2*. Hum Mol Genet. 1999 Jul;8(7):1177-83.

Monani UR, Sendtner M, Coovert DD, Parsons DW, Andreassi C, Le TT, Jablonka S, Schrank B, Rossoll W, Prior TW, Morris GE, Burghes AH. *The human centromeric survival motor neuron gene (SMN2) rescues embryonic lethality in Smn(-/-) mice and results in a mouse with spinal muscular atrophy*. Hum Mol Genet. 2000 Feb 12;9(3):333-9.

Monani UR. *Spinal muscular atrophy: a deficiency in a ubiquitous protein; a motor neuron-specific disease*. Neuron. 2005 Dec 22;48(6):885-96.

Motulsky AG, Qi M. *Pharmacogenetics, pharmacogenomics and ecogenetics.* J Zhejiang Univ Sci B. 2006 Feb;7(2):169-70.

Munsat TL, Davies KE. *International SMA consortium meeting*. (26-28 June 1992, Bonn, Germany). Neuromuscul Disord. 1992;2(5-6):423-8.

Ogino S, Gao S, Leonard DG, Paessler M, Wilson RB. *Inverse correlation between SMN1 and SMN2 copy numbers: evidence for gene conversion from SMN2 to SMN1*. Eur J Hum Genet. 2003 Mar;11(3):275-7.

Oprea GE, Kröber S, McWhorter ML, Rossoll W, Müller S, Krawczak M, Bassell GJ, Beattie CE, Wirth B. *Plastin 3 is a protective modifier of autosomal recessive spinal muscular atrophy*. Science. 2008 Apr 25;320(5875):524-7.

Passini MA, Bu J, Richards AM, Treleaven CM, Sullivan JA, O'Riordan CR, Scaria A, Kells AP, Samaranch L, San Sebastian W, Federici T, Fiandaca

MS, Boulis NM, Bankiewicz KS, Shihabuddin LS, Cheng SH. *Translational Fidelity of Intrathecal Delivery of Self-Complementary AAV9-Survival Motor Neuron 1 for Spinal Muscular Atrophy*. Hum Gene Ther. 2014 Apr 28. Ahead of print.

Pearn, J., *Incidence, prevalence, and gene frequency studies of chronic childhood spinal muscular atrophy*. J Med Genet., 1978. 15(6): p. 409-413.

Pearn J. *Classification of spinal muscular atrophies*. Lancet. 1980 Apr 26;1(8174):919-22.

Pellizzoni L, Kataoka N, Charroux B, Dreyfuss G. *A novel function for SMN, the spinal muscular atrophy disease gene product, in pre-mRNA splicing*. Cell. 1998 Nov 25;95(5):615-24.

Pellizzoni L, Charroux B, Dreyfuss G. *SMN mutants of spinal muscular atrophy patients are defective in binding to snRNP proteins*. Proc Natl Acad Sci U S A. 1999 Sep 28;96(20):11167-72.

Pellizzoni L, Yong J, Dreyfuss G. *Essential role for the SMN complex in the specificity of snRNP assembly*. Science. 2002 Nov 29;298(5599):1775-9.

Poke G, Doody M, Prado J, Gattas M. *Segmental Maternal UPD6 with Prenatal Growth Restriction*. Mol Syndromol. 2013 Jan;3(6):270-3.

Prior TW, Russman BS. *Spinal Muscular Atrophy*. 2000 Feb 24 [updated 2013 Nov 14]. GeneReviews® [Internet]. Pagon RA, Adam MP, Ardinger HH, *et al.*, editors. Seattle (WA): 1993-2014.

Prior TW, Krainer AR, Hua Y, Swoboda KJ, Snyder PC, Bridgeman SJ, Burghes AH, Kissel JT. *A positive modifier of spinal muscular atrophy in the SMN2 gene*. Am J Hum Genet. 2009 Sep;85(3):408-13.

Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E; Working Group of the American College of Medical Genetics and GenomicsLaboratory Quality Assurance Commitee. *ACMG clinical laboratory standards for next-generation sequencing.* Genet Med. 2013 Sep;15(9):733-47.

Robinson WP. *Mechanisms leading to uniparental disomy and their clinical consequences*. Bioessays. 2000 May;22(5):452-9.

Ropers HH. *On the future of genetic risk assessment.* J Community Genet. 2012 Jul;3(3):229-36.

Russman BS. *Spinal muscular atrophy: clinical classification and disease heterogeneity*. J Child Neurol. 2007 Aug;22(8):946-51.

Sambuughin N, Sivakumar K, Selenge B, Lee HS, Friedlich D, Baasanjav D, Dalakas MC, Goldfarb LG. *Autosomal dominant distal spinal muscular atrophy type V (dSMA-V) and Charcot-Marie-Tooth disease type 2D (CMT2D) segregate within a single large kindred and map to a refined region on chromosome 7p15*. J Neurol Sci. 1998 Nov 26;161(1):23-8

Sanger F, Nicklen S, Coulson AR. *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7.

Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. *The real cost of sequencing: higher than you think!* Genome Biol. 2011 Aug 25;12(8):125.

Scarciolla O, Stuppia L, De Angelis MV, Murru S, Palka C, Giuliani R, Pace M, Di Muzio A, Torrente I, Morella A, Grammatico P, Giacanelli M, Rosatelli MC, Uncini A, Dallapiccola B. *Spinal muscular atrophy genotyping by gene*

dosage using multiple ligation-dependent probe amplification. Neurogenetics. 2006 Nov;7(4):269-76.

Scharf JM, Endrizzi MG, Wetter A, Huang S, Thompson TG, Zerres K, Dietrich WF, Wirth B, Kunkel LM. *Identification of a candidate modifying gene for spinal muscular atrophy by comparative genomics*. Nat Genet. 1998 Sep;20(1):83-6.

See K , Yadav P, Giegerich M, Cheong PS, Graf M, Vyas H, Lee SG, Mathavan S, Fischer U, Sendtner M, Winkler C. *SMN deficiency alters Nrxn2 expression and splicing in zebrafish and mouse models of spinal muscular atrophy*. Hum Mol Genet. 2014 Apr 1;23(7):1754-70.

Smith M, Calabro V, Chong B, Gardiner N, Cowie S, du Sart D. *Population screening and cascade testing for carriers of SMA*. Eur J Hum Genet. 2007 Jul;15(7):759-66.

Somerville MJ, Hunter AG, Aubry HL, Korneluk RG, MacKenzie AE, Surh LC. *Clinical application of the molecular diagnosis of spinal muscular atrophy: deletions of neuronal apoptosis inhibitor protein and survival motor neuron genes*. Am J Med Genet. 1997 Mar 17;69(2):159-65.

Sterky F., Lundeberg J. *Sequence analysis of genes and genomes.* J Biotechnol. 2000 Jan 7;76(1):1-31. Review.

Strasswimmer J, Lorson CL, Breiding DE, Chen JJ, Le T, Burghes AH, Androphy EJ. *Identification of survival motor neuron as a transcriptional activator-binding protein*. Hum Mol Genet. 1999 Jul;8(7):1219-26.

Stratigopoulos G, Lanzano P, Deng L, Guo J, Kaufmann P, Darras B, Finkel R, Tawil R, McDermott MP, Martens W, Devivo DC, Chung WK. *Association of plastin 3 expression with disease severity in spinal muscular atrophy only in postpubertal females*. Arch Neurol. 2010 Oct;67(10):1252-6.

Talbot K, Ponting CP, Theodosiou AM, Rodrigues NR, Surtees R, Mountford R, Davies KE. *Missense mutation clustering in the survival motor neuron gene: a role for a conserved tyrosine and glycine rich region of the protein in RNA metabolism?* Hum Mol Genet. 1997 Mar;6(3):497-500.

Thomas NH, Dubowitz V. *The natural history of type I (severe) spinal muscular atrophy*. Neuromuscul Disord. 1994 Sep-Nov;4(5-6):497-502.

van der Steege G, Grootscholten PM, van der Vlies P, Draaijers TG, Osinga J, Cobben JM, Scheffer H, Buys CH. *PCR-based DNA test to confirm clinical diagnosis of autosomal recessive spinal muscular atrophy*. Lancet. 1995 Apr 15;345(8955):985-6.

van der Vleuten AJ, van Ravenswaaij-Arts CM, Frijns CJ, Smits AP, Hageman G, Padberg GW, Kremer H. *Localisation of the gene for a dominant congenital spinal muscular atrophy predominantly affecting the lower limbs to chromosome 12q23-q24*. Eur J Hum Genet. 1998 Jul-Aug;6(4):376-82.

Van Meerbeke JP, Sumner CJ. *Progress and promise: the current status of spinal muscular atrophy therapeutics*. Discov Med. 2011 Oct;12(65):291-305.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher

A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. *The sequence of the human genome.* Science. 2001 Feb 16;291(5507):1304-51. Erratum in: Science 2001 Jun 5;292(5523):1838.

Viollet L, Barois A, Rebeiz JG, Rifai Z, Burlet P, Zarhrate M, Vial E, Dessainte M, Estournet B, Kleinknecht B, Pearn J, Adams RD, Urtizberea JA, Cros DP, Bushby K, Munnich A, Lefebvre S. *Mapping of autosomal recessive chronic distal spinal muscular atrophy to chromosome 11q13*. Ann Neurol. 2002 May;51(5):585-92.

Vitali T, Sossi V, Tiziano F, Zappata S, Giuli A, Paravatou-Petsotas M, Neri G, Brahe C. Detection of the survival motor neuron (SMN) genes by FISH: further evidence for a role for SMN2 in the modulation of disease severity in SMA patients. Hum Mol Genet. 1999 Dec;8(13):2525-32.

Vyas S, Béchade C, Riveau B, Downward J, Triller A. *Involvement of survival motor neuron (SMN) protein in cell death*. Hum Mol Genet. 2002 Oct 15;11(22):2751-64.

Wang CH, Xu J, Carter TA, Ross BM, Dominski MK, Bellcross CA, Penchaszadeh GK, Munsat TL, Gilliam TC. *Characterization of survival motor neuron (SMNT) gene deletions in asymptomatic carriers of spinal muscular atrophy*. Hum Mol Genet. 1996 Mar;5(3):359-65.

Wang CH, Papendick BD, Bruinsma P, Day JK. *Identification of a novel missense mutation of the SMN(T) gene in two siblings with spinal muscular atrophy. Neurogenetics*. 1998 Aug;1(4):273-6.

Wang K, Li M, Hakonarson H. *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res. 2010 Sep;38(16):e164.

Weiss MM, Van der Zwaag B, Jongbloed JD, Vogel MJ, Brüggenwirth HT, Lekanne Deprez RH, Mook O, Ruivenkamp CA, van Slegtenhorst MA, van den Wijngaard A, Waisfisz Q, Nelen MR, van der Stoep N. *Best practice guidelines for the use of next-generation sequencing applications in genomediagnostics: a national collaborative study of Dutch genome diagnostic laboratories.* Hum Mutat. 2013 Oct;34(10):1313-21.

Wirth B, Hahnen E, Morgan K, DiDonato CJ, Dadze A, Rudnik-Schöneborn S, Simard LR, Zerres K, Burghes AH. *Allelic association and deletions in autosomal recessive proximal spinal muscular atrophy: association of marker genotype with disease severity and candidate cDNAs*. Hum Mol Genet. 1995 Aug;4(8):1273-84.

Wirth B, Schmidt T, Hahnen E, Rudnik-Schöneborn S, Krawczak M, Müller-Myhsok B, Schönling J, Zerres K. *De novo rearrangements found in 2% of index patients with spinal muscular atrophy: mutational mechanisms, parental origin, mutation rate, and implications for genetic counseling*. Am J Hum Genet. 1997 Nov;61(5):1102-11.

Wirth B. *An update of the mutation spectrum of the survival motor neuron gene (SMN1) in autosomal recessive spinalmuscular atrophy (SMA)*. Hum Mutat. 2000;15(3):228-37.

Wirth B. *Spinal muscular atrophy: state-of-the-art and therapeutic perspectives*. Amyotroph Lateral Scler Other Motor Neuron Disord. 2002 Jun;3(2):87-95.

Wirth B, Brichta L, Hahnen E. *Spinal muscular atrophy: from gene to therapy*. Semin Pediatr Neurol. 2006 Jun;13(2):121-31.

Yamazawa K, Ogata T, Ferguson-Smith AC. *Uniparental disomy and human disease: an overview*. Am J Med Genet C Semin Med Genet. 2010 Aug 15;154C(3):329-34.

Zanetta C, Nizzardo M, Simone C, Monguzzi E, Bresolin N, Comi GP, Corti S. *Molecular therapeutic strategies for spinal muscular atrophies: current and future clinical trials*. Clin Ther. 2014 Jan 1;36(1):128-40.

Zapletalová E, Hedvicáková P, Kozák L, Vondrácek P, Gaillyová R, Maríková T, Kalina Z, Jüttnerová V, Fajkus J, Fajkusová L. *Analysis of point mutations in the SMN1 gene in SMA patients bearing a single SMN1 copy*. Neuromuscul Disord. 2007 Jun;17(6):476-81.

Zerres K, Rudnik-Schöneborn S. *Natural history in proximal spinal muscular atrophy. Clinical analysis of 445 patients and suggestions for a modification of existing classifications.* Arch Neurol. 1995 May;52(5):518-23.

Zerres K, Rudnik-Schoneborn S, Forrest E, Lusakowska A, Borkowska J, Hausmanowa-Petrusewicz I. *A collaborative study on the natural history of childhood and juvenile onset proximal spinal muscular atrophy (type II and III SMA): 569 patients*. J Neurol Sci. 1997;146(1):67–72.

SITOGRAPHY

http://www.dnamica.it
http://www.ncbi.nlm.nih.gov/pubmed
http://picard.sourceforge.net
http://blast.ncbi.nlm.nih.gov/Blast.cgi