Aberrantly spliced genes in human diseases – a deep sequencing approach

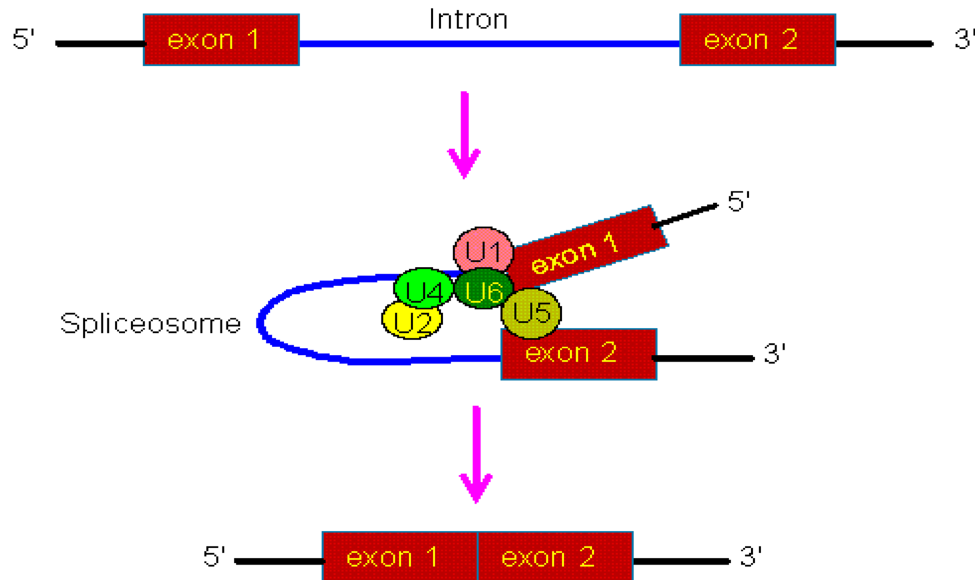
Henriette S. Andersen[1], Thomas Koed Docter[1], Jane Serup[1], Lone Sundahl[1] , Akio Masuda[2], Kinji Ohno[2] and Brage S. Andresen[1].
[1]Biochemistry and Molecular biology, University of Southern Denmark, Odense M, DK-5230, Denmark
[2]Neurogenetics, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, Japan.

Introduktion

RNA splicing is a fundamental process required for correct expression of almost all protein coding genes. Splicing is a modification of the primary RNA transcript, in which introns are removed and exons are joined. In eukaryotic introns, splicing consists of a series of reactions which are catalyzed by the spliceosome, which is a dynamic RNA-protein complex composed of more than 150 different proteins and five small nuclear ribonucleoproteins (snRNP). Specific splicing signals in the pre-mRNA (*cis*-elements) are recognized by the spliceosome, which after a complex enzymatic reaction, removes the intron and join together the two neighboring exons (Fig.1). Mutations/SNPs that affect *cis*-elements in disease genes and/or changes that affect the expression of splicing trans-factors are frequent causes of diseases due to aberrant splicing. That is the case in neuromuscular diseases such as spinal muscular atrophy (SMA) and myotonic dystrophy (DM), which are the focus of the present study. Some but not all is known on the molecular mechanism and factors involved for both diseases.

Our approach is based on deep sequencing of patient's cells and/or animal tissues from disease models in order to achieve a global identification of aberrantly spliced genes in SMA and DM. A combination of RNA-CLIP followed by sequencing is used to identify new RNA segments (*cis*-elements) bound to the known splicing trans-factors that are dysregulated in SMA and DM.

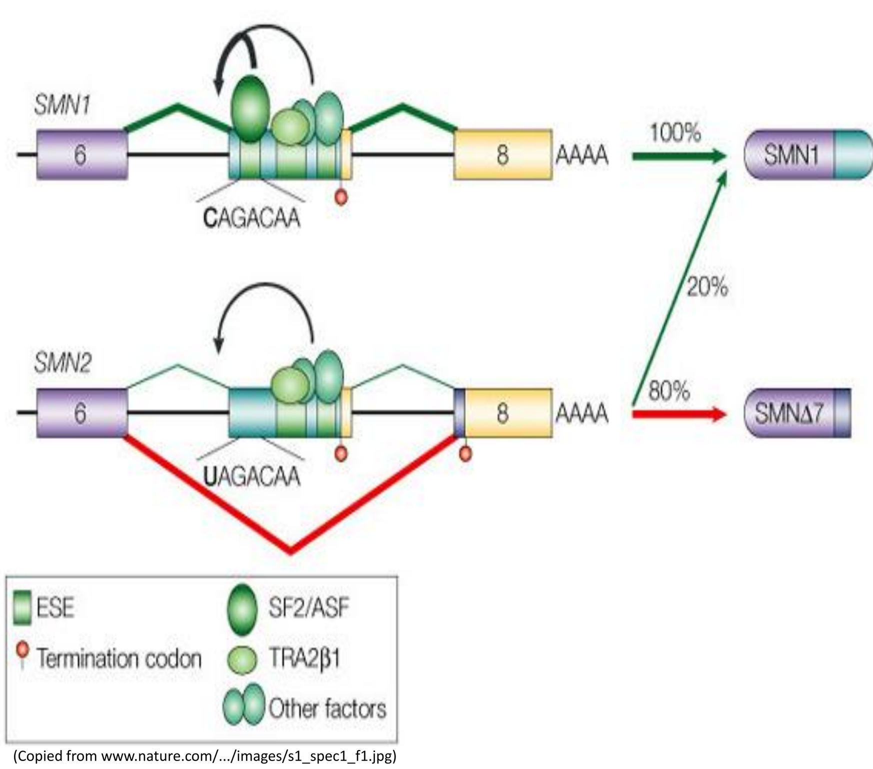
Fig. 1 General Splicing mechanism



Spinal muscular atrophy (SMA)

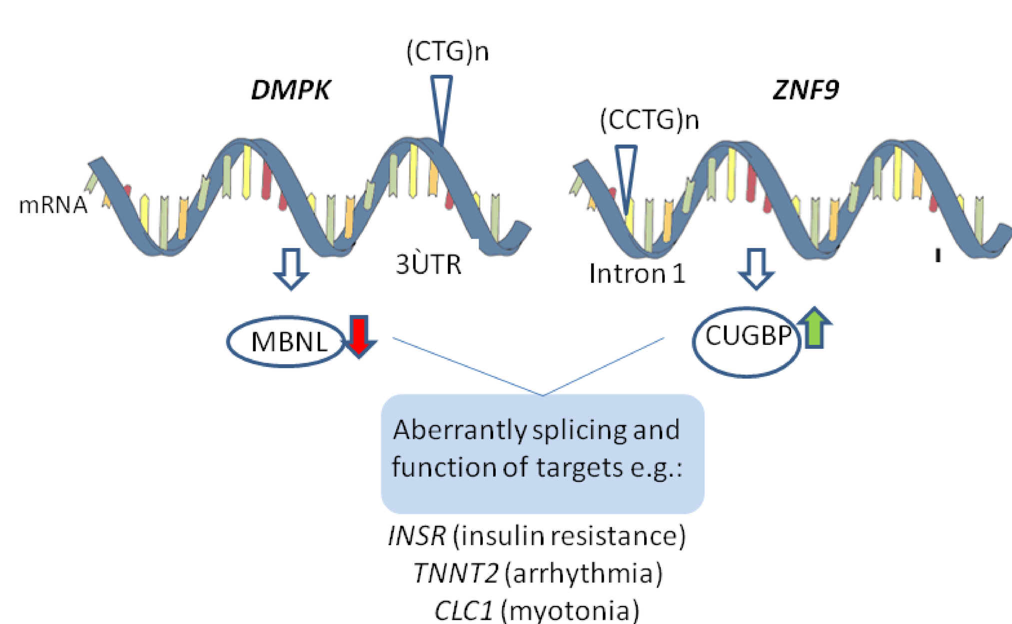
Spinal muscular atrophy (SMA) is an autosomal recessive neuromuscular disorder associated with the loss of α motor neurons in the spinal cord. SMA is caused by deletion or mutation of the survival motor neuron 1 gene (*SMN1*) significantly reducing SMN protein expression. A nearly identical copy, *SMN2* harbors a silent single nucleotide mutation +6 C >T in exon 7 that causes aberrant splicing of exon 7 (SMNΔex7)[1] . From *SMN2* only about 10% of full length SMN protein is being produced and high levels of the C-terminally truncated unstable SMN protein lacking exon 7 (SMNΔex7) (Fig.2). SMA patients with several copies of *SMN2* produce more full-length SMN and have a less severe phenotype. Why reduced levels of SMN results in degeneration of α motor neurons remains to be determined. SMN is involved in a number of cellular processes, including transcription, splicing , and neurite and axonal outgrowth. It is essential for the biogenesis of the spliceosome [2], and lower levels of SMN protein causes aberrant splicing of numerous target genes in a mouse SMA model.

Fig. 2 Aberrant splicing of SMN



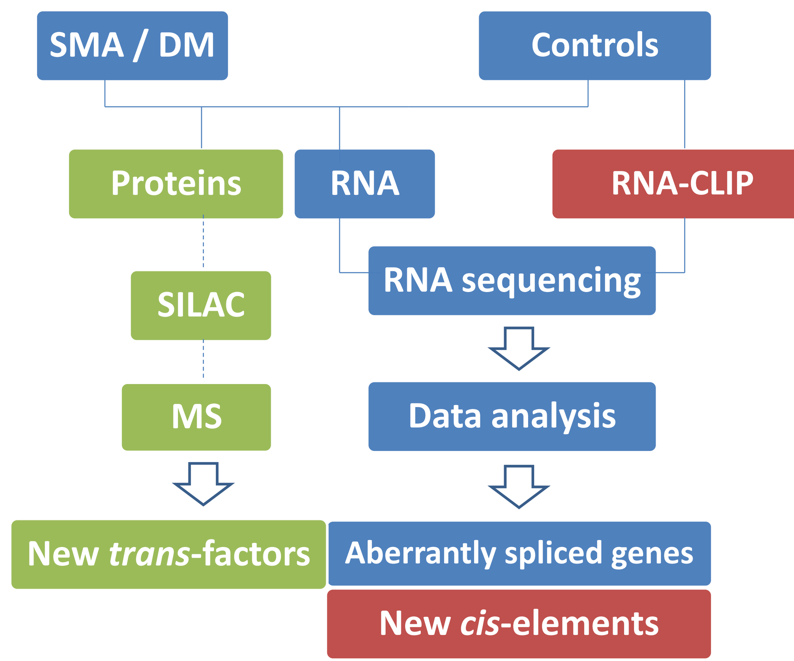
Myotonic dystrophy (DM)

Myotonic dystrophy (DM) is an autosomal dominant multisystem disorder affecting skeletal muscles, eye, hart, the endocrine system and the central nervous system. DM is caused by abnormally expanded CTG repeats in the 3´UTR region of the *dystrophia myotonica kinase gene*[3] (*DMPK*, DM type 1), or by abnormally expanded CCTG repeats in intron 1 of the *zink finger 9 gene* [4] (*ZNF9*, DM type 2)). The presence of these expanded repeats causes dysregulation of two splicing regulatory trans-factors ; muscle blind (MBNL1) and CUG-binding protein (CUGBP1) resulting in aberrant splicing of their target genes.



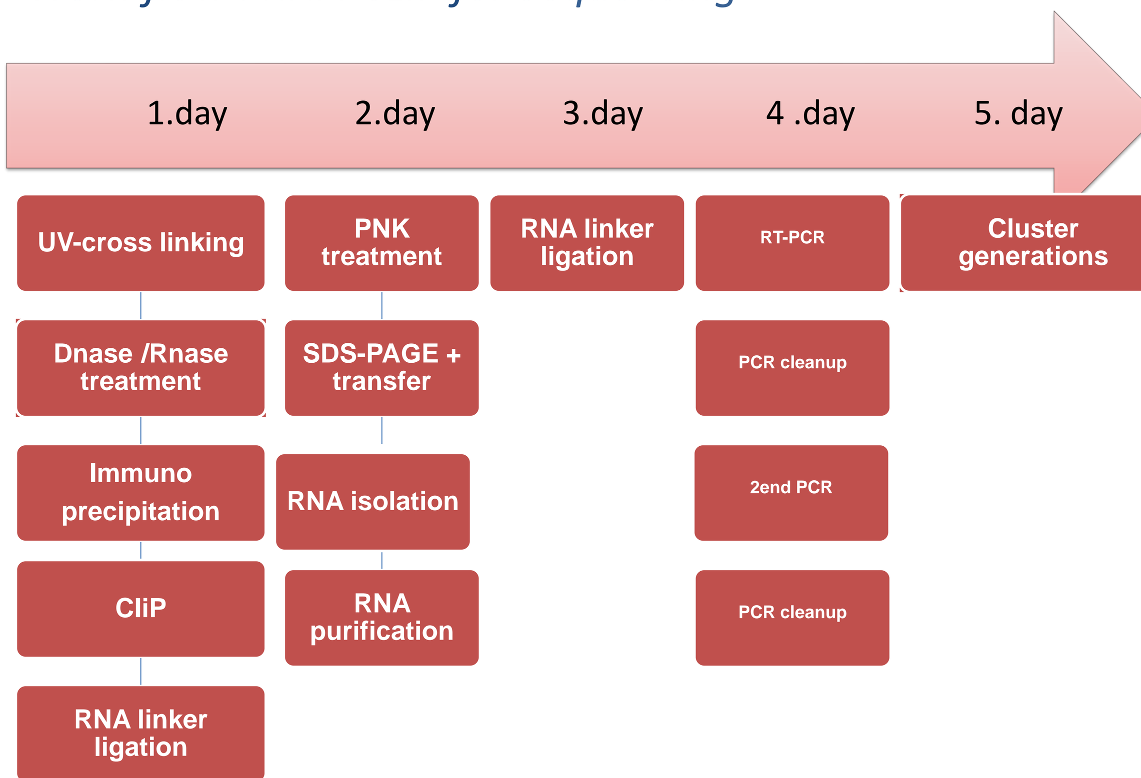
Approach

- **Aberrantly spliced genes** are identified in patients with DM and SMA by global sequencing using the SOLEXA sequencing system. Muscle and brain of DM1 patients and controls, and cultured fibroblasts isolated from SMA patients and controls will be analyzed.
- RNA-CLIP method will be used to identify **RNA *cis*-elements** recognized by known trans-factors.
- Once *cis*-elements have been identified we indentify **new trans-acting splicing factors** by RNA affinity chromatography and mass spectrometry techniques (MS).



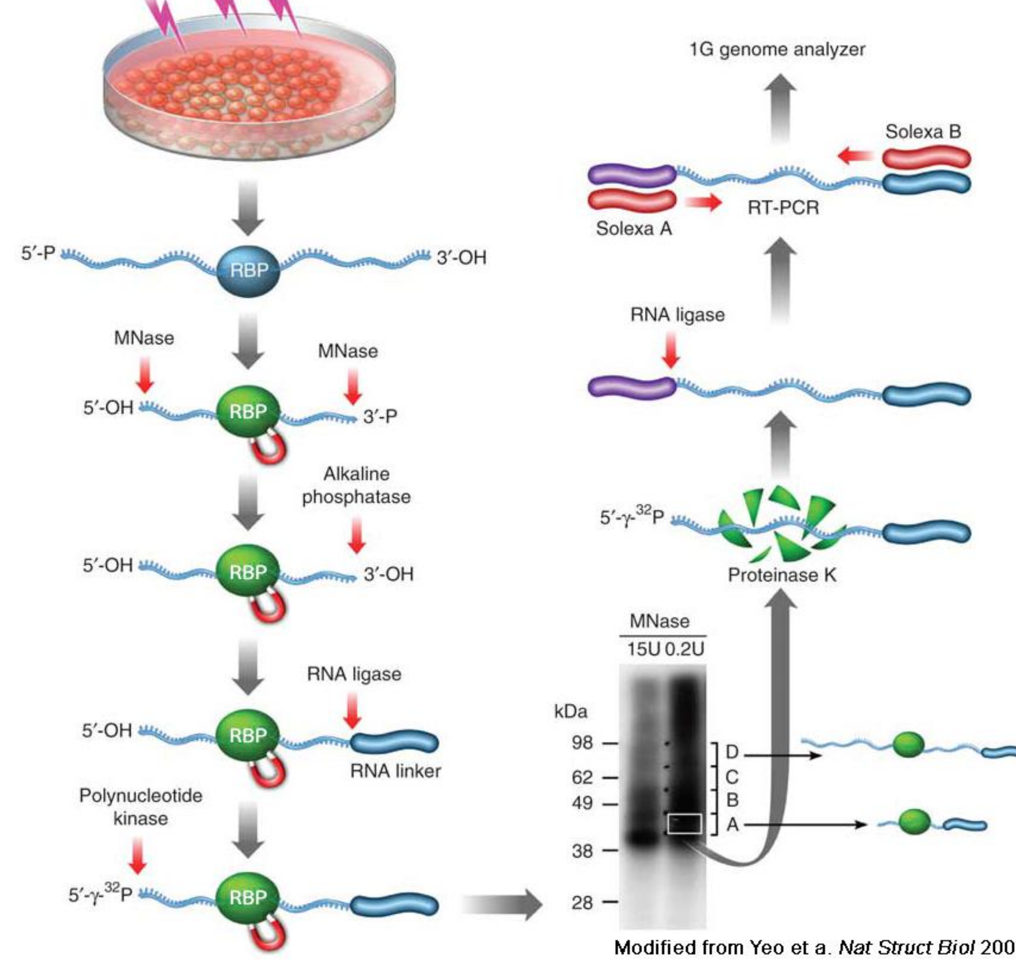
Methods

Work-flow : RNA-CLIP for Sequencing

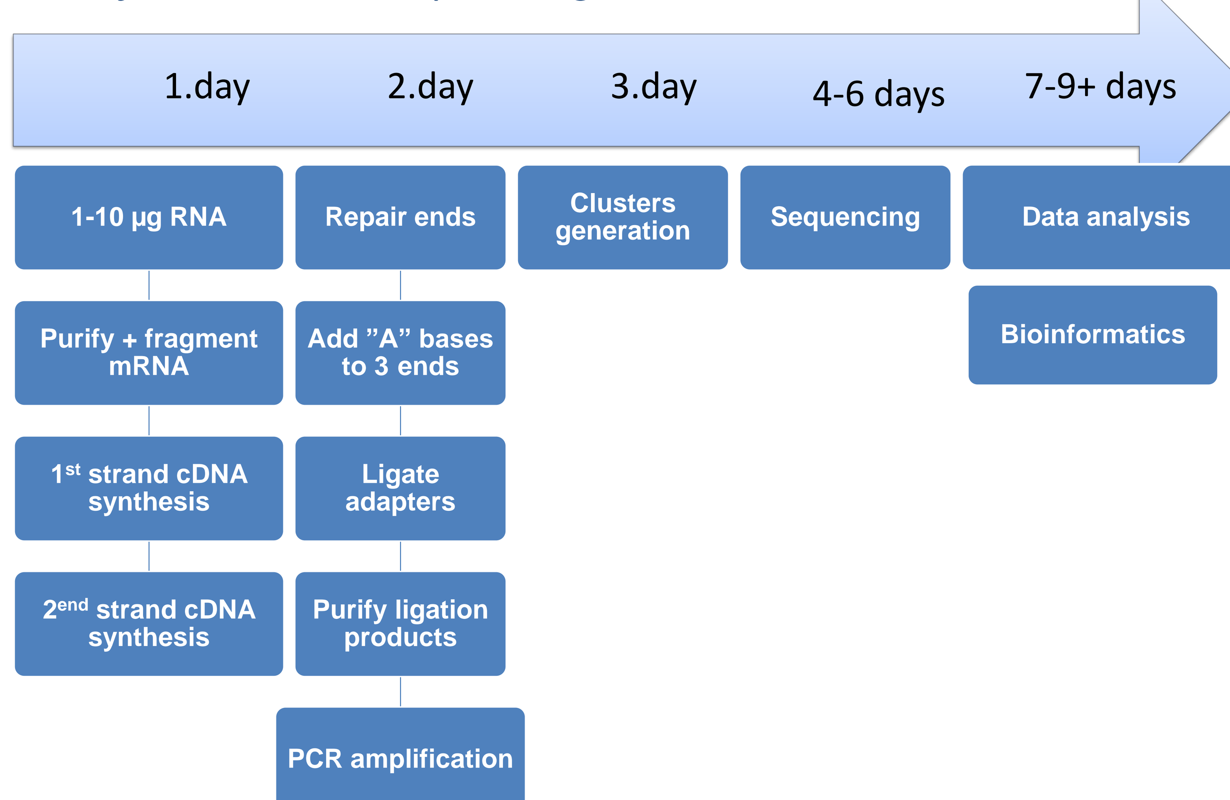


Using this RNA-cross linking chromatin immunoprecipitation (RNA-CLIP) method, living cells are exposed to UV-irradiation to induce covalent cross-links between RNA binding proteins and their *in situ* RNA targets. Prior to immunoprecipitation of a specific RNA binding protein, lysates are treated with RNAse in order to generate ~40-60 nucleotides cross-linked RNA tags. Co-purifying RNA tags can then be sequenced. Database matching of these short 40-60 nucleotide RNA CLIP "tags", which marks the native binding site of RNA binding proteins, potentially allows the entire target repertoire of an RNA binding protein to be determined.

RNA-CLIP:

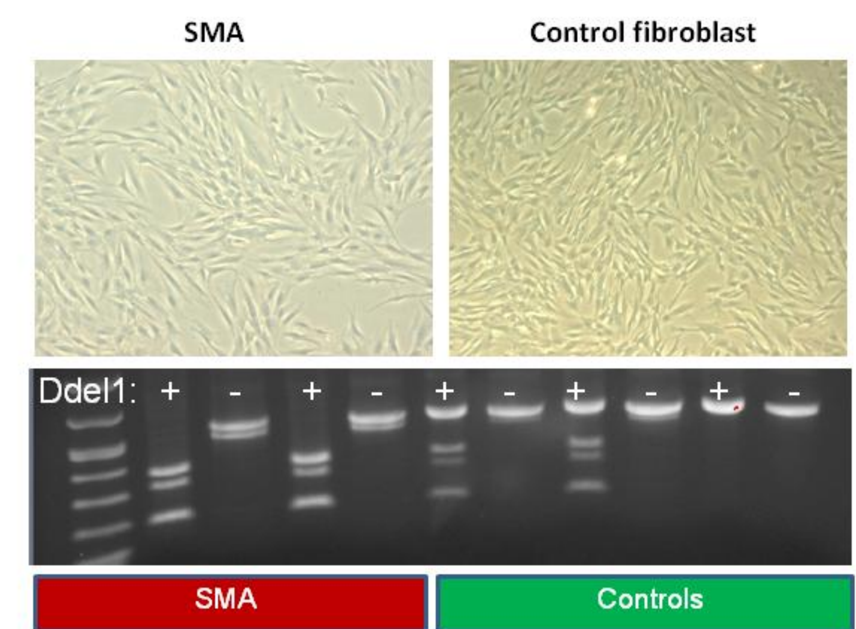


Work-flow : mRNA-Sequencing



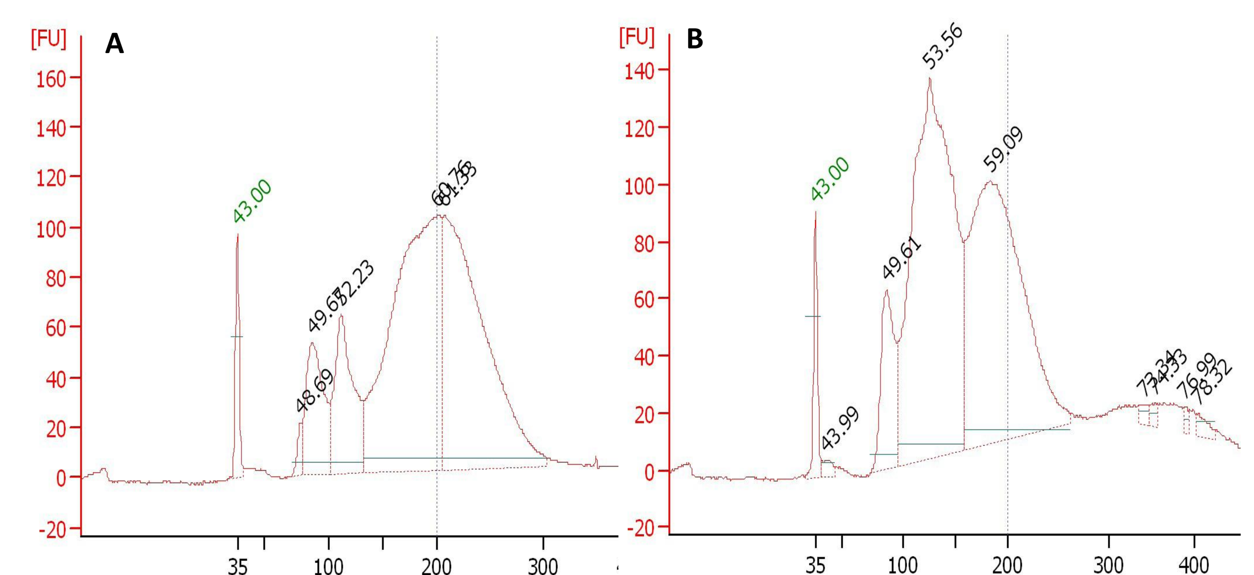
Results

SMA cells and control fibroblasts



Fibroblast from SMA patients and healthy controls are cultured for isolation of RNA for sequencing. RNA isolated from SMA and control cells are tested for absence or presents of exon 7 by PCR amplification and enzymatic digest with Ddel1 restriction enzyme, which only cuts *SMN2*. A clear lack of *SMN1* transcript in SMA patients is observed.

Analysis of cDNA for sequencing using the 2100 Bioanalyser



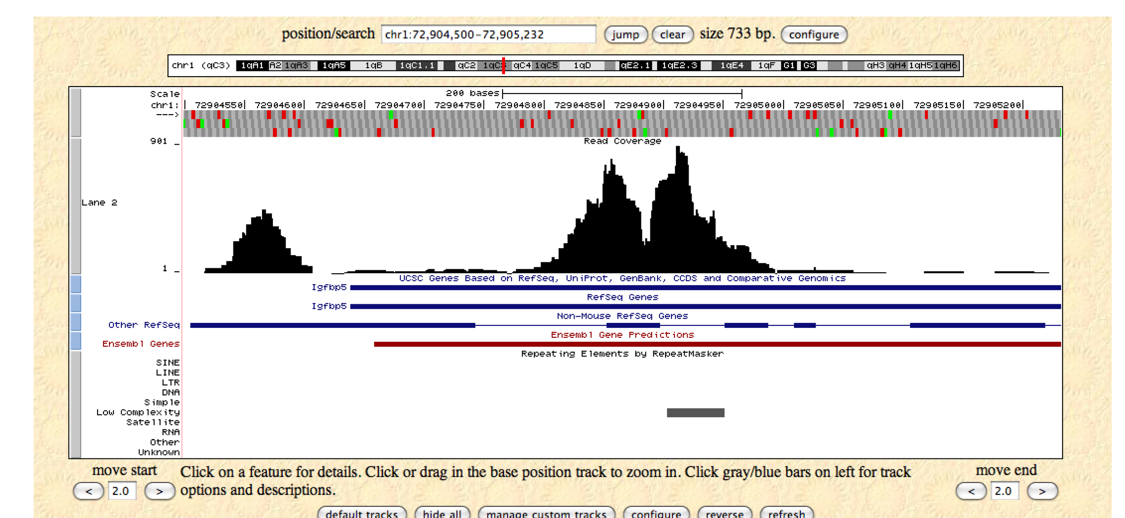
Test results from the *2100 Bioanalyzer* showing **A)** CUGB1 cDNA with major fragments between 75-200 bp. and **B)**MBNL1 cDNA with major fragments between 75-300 bp.

Summary report on RNA-CLIP cDNA sequencing

Lane Info		Tile Mean +/- SD for Lane				
Sample	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	First Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters
S2	188289	145666 +/- 7883	52302 +/- 8656	84 +/- 7	103.70 +/- 11.47	36.16 +/- 7.12
Control	19039	7778 +/- 339	5288 +/- 320	111 +/- 8	84.03 +/- 2.72	68.05 +/- 3.99
S4	162996	145009 +/- 8412	45276 +/- 5795	177 +/- 6	70.55 +/- 5.94	31.42 +/- 5.05

This table shows the output from single-read sequencing of fragments generated from CLIP experiments using two different antibodies, S2 (CUGBP1) and S4 (MBNL1). Between 45 to 52 x 10^3 clusters were generated. Only a low % of clusters passed filtering. Further optimization will increase the outcome.

Results of data-analysis



Sequencing reads were aligned to the mouse genome. From our preliminary analysis we identified hits to some of the well known *trans*-factors including *TNNT2* and genes involved in insulin regulation. Among these, *Insulin-like growth factor-binding protein 5* as demonstrated above.

Summary

Research strategy and preliminary data have been presented here. Further optimizing and analysis of the SOLEXA sequencing data are required. However preliminary analysis has identified a couple of the know target genes strengthening our confidence in the data set. Further sequencing is likely required and should include untreated control cells for comparison.

We are currently validating mRNA isolated from SMA and control cells and expect to start sample preparation for paired-end sequencing within the next few months.

We are confident that this high-throughput sequencing study will bring us closer towards elucidating the molecular mechanism and factors involved for both diseases.

References: 1) Cartegni L, et al. Nat Rev Genet 2002 3:285-298; 2) Fischer,U; et al, Cell .1997 90: 1023-1029; 3)Aslandis C, et al. Nature 1992 355:548-551; 4) Liquori CL, et al. Sciences 2001 293:864-867