Study of SPG11's splicing, in neuronal and non-neuronal cells, by long-read sequencing: implication on phenotype

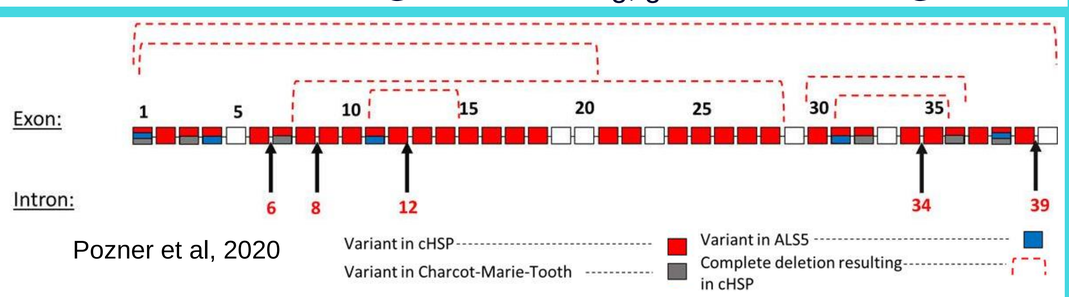
Leblanc M[1,2], Gareau T[2], Bertrand M[2], Mohammad A[2], Liu E[2], Jost Mousseau C[2], Le Goff G[2], Marie Y[2], Guégan J[2], Bohl B[2], Carlile T[3], Stevanin G[1,2,4]

[1]EPHE, PSL University, Paris, France; [2]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP, Hôpital de la Pitié Salpêtrière, Paris, France; [3]Biogen, Cambridge (MA), USA; [4]INCIA, University of Bordeaux, CNRS, EPHE, Bordeaux, France

Contacts : marion.leblanc@icm-institute.org; giovanni-b.stevanin@inserm.fr

Introduction

The SPG11 gene is one of the most frequently mutated genes in the autosomal recessive forms of hereditary spastic paraplegias (HSP) and also accounts for cases with amyotrophic lateral sclerosis (ALS5) and Charcot-Marie Tooth (CMT) neuropathy. The SPG11 gene encodes a ubiquitously expressed 7.8kb full-length transcript. This transcript codes for Spatacsin, a 2 443 amino acids protein with a suspected role in autophagy. Little is known about the regulation of the expression of this gene.

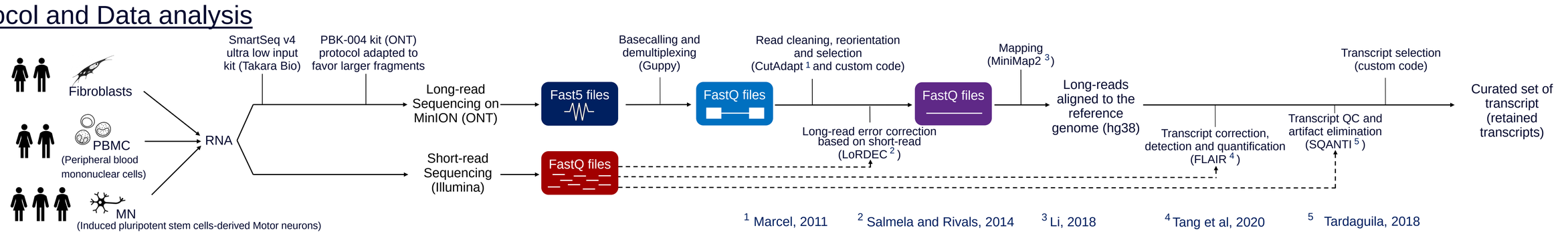


Pozner et al, 2020

Aim of the project

To identify all isoforms of SPG11 and establish their full expression profiles, first in non-pathological models of the disease, but also in pathological models to determine if the presence of remaining isoforms could explain the variability in clinical presentation.

Protocol and Data analysis



[1] Marcel, 2011 [2] Salmela and Rivals, 2014 [3] Li, 2018 [4] Tang et al, 2020 [5] Tardaguila, 2018

Figure 1: Schema of data generation and processing for transcript discovery. Library preparation and data analysis were optimized to allow the discovery of transcripts on larger lowly expressed genes. Notably, short-read sequencing data were generated on the same samples as long-read data and used for long-read correction and new splice junctions' validation. Transcript selection was made based on stringent criteria to increase confidence (coding capacity, not predicted for NMD degradation, transcript start site corresponding to CAGE data, presence of a polyadenylation motif, …). Of note, data from multiple samples were pooled for transcript detection to allow better identification of the transcripts, as recommended in the SQANTI manual (https://github.com/ConesaLab/SQANTI3/wiki).

Results

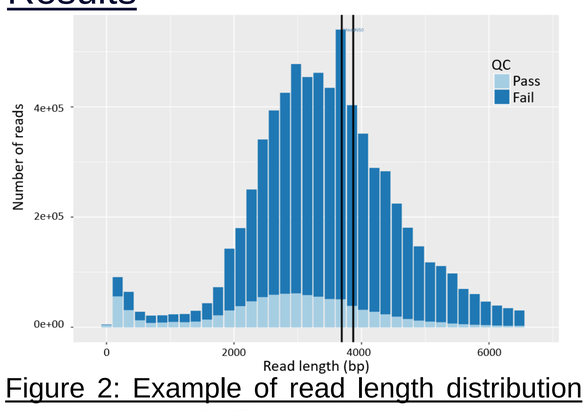


Figure 2: Example of read length distribution in one of our ONT sequencing experiments. The average sequenced read length for all our considered experiments is 3578bp+/-330bp (n=9).

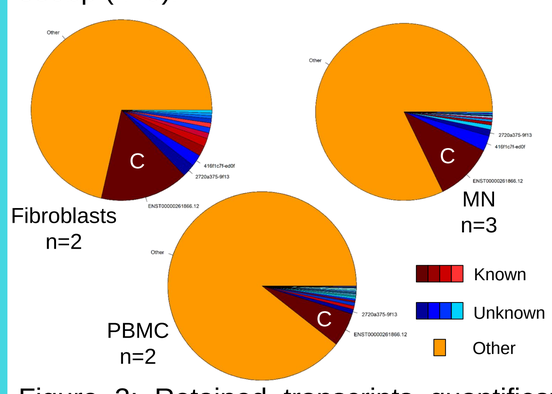


Figure 3: Retained transcripts quantification against total mapped reads on SPG11 in control samples. In orange, reads mapping on SPG11 not supporting retained transcripts. The canonical transcript is marked with a "C".

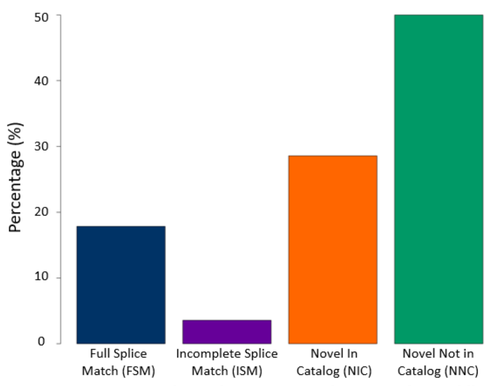


Figure 4: Retained transcripts classification (based on Tardaguila, 2018). FSM: same number of exons as a reference, each internal junction agrees. ISM: fewer exons than a reference; each internal junction agrees. NIC: novel isoform, a combination of known donor/acceptor sites. NNC: novel isoform, at least one donor or acceptor site that is not annotated. Results were obtained on a curated set of transcripts (n=7 control samples from our 3 conditions pooled for the analysis).

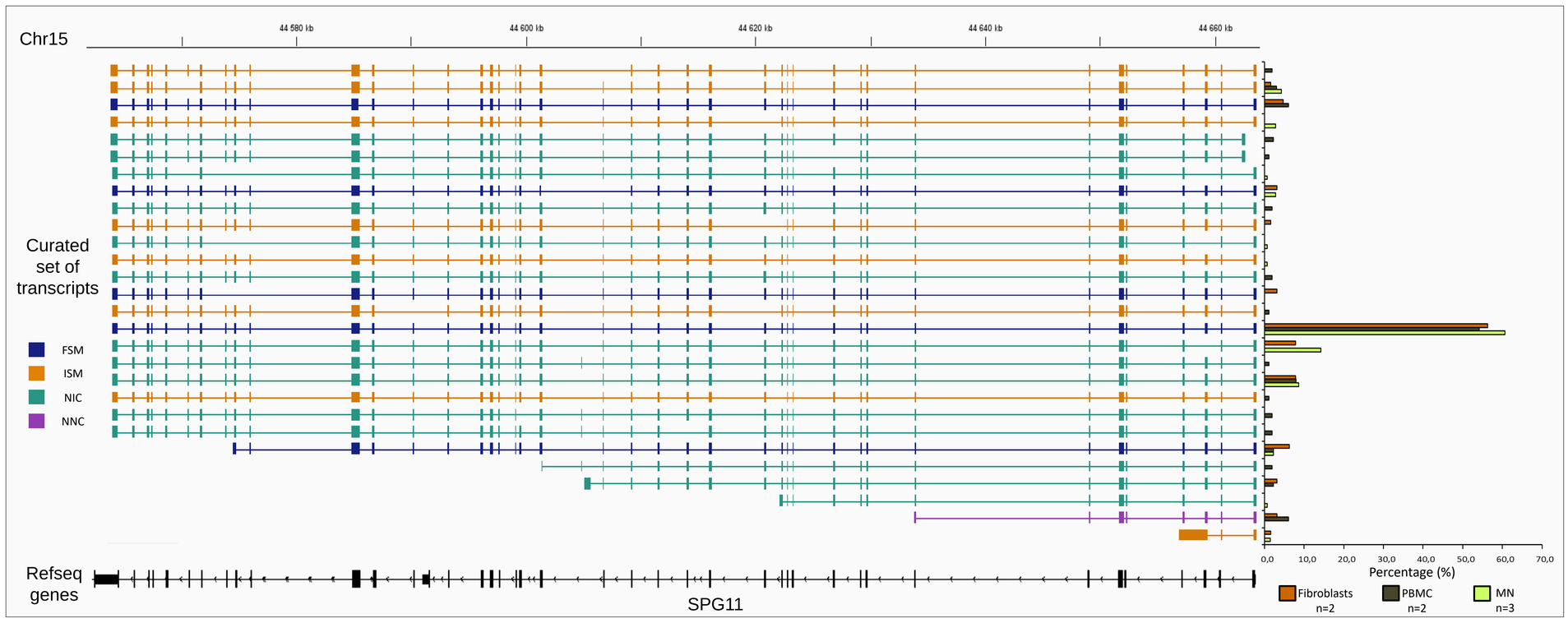


Figure 5: IGV genome browser track showing the curated set of transcripts detected in our three control conditions. Transcripts' colours are based on their classification as mentioned in Figure 4: Full-splice match (FSM) in dark blue, Incomplete spice-match (ISM) in purple, Novel in catalog (NIC) in orange and Novel not in catalog (NNC) in green. Genomic reference annotation from Refseq Genes is displayed in black at the bottom. Results displayed are the curated set of transcripts detected in control fibroblasts, PBMCs and MNs (n=7 control samples pooled for the analysis with 2 fibroblasts, 2 PBMCs and 3 MNs). On the right, grouped bar plot showing the mean percentage of expression of each retained transcript among the total of retained transcripts, by cell type (Fibroblasts n=2, PBMC n=2, MN n=3). The canonical transcript is the most expressed in all conditions, as expected. 36% of the transcripts are detected in multiple conditions, with sometimes differences in their expression level, whereas the left 64% are cell type-specific, in our experiments.

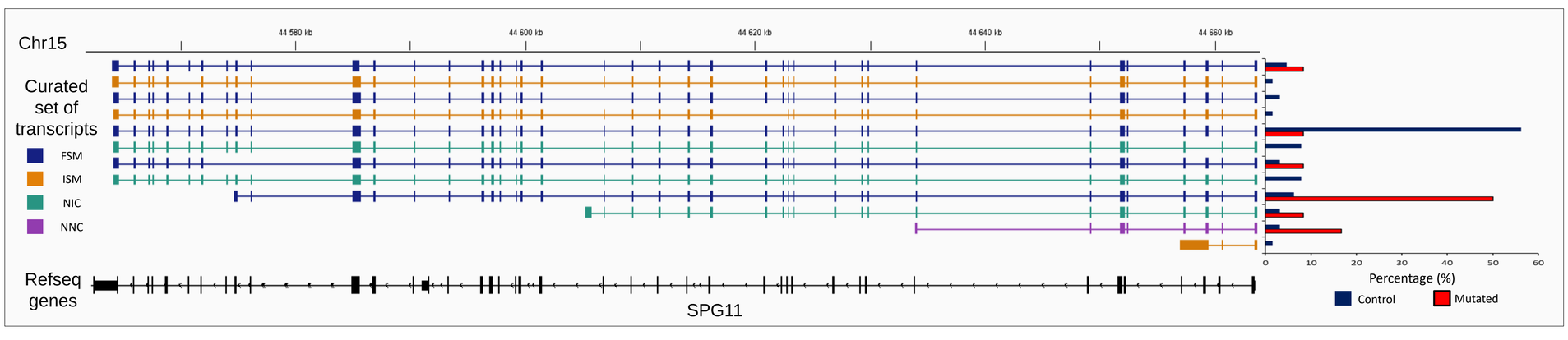


Figure 6: IGV genome browser track showing the curated set of transcripts detected in fibroblasts from control individuals and SPG11 mutated patients. Transcripts' colours are based on their classification as mentioned in figures 4 and 5: Full-splice match (FSM) in dark blue, Incomplete spice-match (ISM) in purple, Novel in catalog (NIC) in orange and Novel not in catalog (NNC) in green. Genomic reference annotation from Refseq Genes is displayed in black at the bottom. Results displayed are the curated set of transcripts detected in fibroblasts (n=4 with 2 control fibroblasts and 2 SPG11 mutated fibroblasts pooled for the analysis). On the right, grouped bar plot showing the mean percentage of expression of each retained transcript among the total of retained transcripts. SPG11 mutated fibroblasts were obtained from two patients with a homozygous nonsense mutation in exon 32 and presenting with complex form of HSP-SPG11 (early onset, peripheral neuropathy, corpus callosum and cortical atrophy). In the patients' samples, a decrease in SPG11 expression is observed, especially the canonical transcript predicted to be degraded by the NMD. However, despite the decrease, some other transcripts are still present.

Conclusions

- Average sequenced read length of 3,6kb
- Detection of 28 transcripts in our three cell types with 78% of these unknown
- Most transcripts are cell type-specific in our experiments
- Among the 28 retained transcripts, the most represented alternative splicing event is exon skipping (58,8%)
- SPG11 mutated fibroblasts: high decrease of the canonical transcript but persistence of other transcripts

Perspectives

- Sequencing of RNA extracts from fibroblasts, iPSC-derived MN and PBMC of other control individuals, and SPG11 and SPG11-mutated ALS5 patients
- Study of transcripts stability in a cellular context
- Study of protein expression in vitro -> localization and function
- Phenotype-genotype correlation (ALS5 versus SPG11 patients)

Acknowledgements

The research leading to these results has received funding in the context of the PSL-Biogen doctoral partnership (CIFRE convention n°2019/1067).

We gratefully acknowledge the iGenSeq core and Data and Analysis core facilities at ICM.

Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.