Identification and characterization of splicing defects by Single-Molecule Real-Time sequencing technology (PacBio)

Marco Savarese, PhD[1,2]*, Talha Qureshi, MSc[1,2]*, Annalaura Torella, PhD[3,4], Pia Laine, MSc[5], Teresa Giugliano, PhD[3], Per Harald Jonson, PhD[1,2], Mridul Johari, MSc[1,2], Lars Paulin, MSc[5], Giulio Piluso, PhD[3], Petri Auvinen, PhD[5], Vincenzo Nigro, MD[3,4], Bjarne Udd, MD[1,2,6], Peter Hackman, PhD[1,2]

1 Folkhälsan Research Center, Helsinki, Finland;
2 Department of Medical Genetics, Medicum, University of Helsinki, Helsinki, Finland;
3 Dipartimento di Medicina di Precisione, Università degli Studi della Campania "Luigi Vanvitelli", Naples, Italy;
4 Telethon Institute of Genetics and Medicine, Pozzuoli, Italy;
5 Institute of Biotechnology, University of Helsinki, Helsinki, Finland;
6 Vaasa Central Hospital, Vaasa, Finland.
*These authors contributed equally to the work

Corresponding author:

Marco Savarese, Ph.D.

Folkhälsan Institute of Genetics

Dpt. of Medical Genetics

University of Helsinki

Biomedicum, Haartmaninkatu 8 -Pb 63 00014 Helsinki -Finland

E-Mail marco.savarese@helsinki.fi

Tel.:+358-294125069

Running title: SMRT-based splicing characterization

Abstract

Although DNA-sequencing is the most effective procedure to achieve a molecular diagnosis in genetic diseases, complementary RNA analyses are often required.

Reverse-Transcription polymerase chain reaction (RT-PCR) is still a valuable option when the clinical phenotype and/or available DNA-test results address the diagnosis toward a gene of interest or when the splicing effect of a single variant needs to be assessed.

We use Single-Molecule Real-Time sequencing to detect and characterize splicing defects and single nucleotide variants in well-known disease genes (*DMD*, *NF1*, *TTN*). After proper optimization, the procedure could be used in the diagnostic setting, simplifying the workflow of cDNA analysis.

Keywords: long-read sequencing, Single Molecule Real Time, PacBio, splicing

Main text

DNA sequencing is the most commonly used and effective strategy for the routine diagnostics of genetically inherited diseases (1-3). With the introduction of the next-generation sequencing platforms over ten years ago, clinical sequencing has become a routine approach (1, 2). Targeted, exome and genome sequencing have contributed to increase the diagnostic rate of a large number of rare diseases (1-4). Variants predicted as being splice-disrupting need to be characterized using RNA from the tissue of interest (5, 6). Usually, a reverse transcription-polymerase chain reaction (RT-PCR) is performed using primers amplifying the region of interest (7). In case of a splice defect caused by a heterozygous variant, RT-PCR products are often sequenced after gel extraction or after cloning them in plasmids. The procedure is even more complex when the variant occurs in regions undergoing alternative splicing or when a single variant causes multiple mis-splicing events.

On the other hand, over 50% of patients remain undiagnosed after DNA sequencing (3). Some of them carry variants that are not identified or not properly interpreted at the DNA-level (3, 8, 9). RNA sequencing or cDNA analysis is an effective second-tier test, when the tissue of interest is easily accessible (8, 10, 11). Usually, RNA sequencing is performed to globally evaluate the transcripts and search for an outlier in a gene-independent way (8, 10). Alternatively, the sole transcript of interest can be analyzed when its coding gene is strongly associated with the

observed phenotype and DNA analysis (usually focused on the coding region) has not identified any causative variant. For example, in patients with a clinical suspicion of Duchenne muscular dystrophy, when dystrophin expression is abnormal/absent and no mutation in the *DMD* gene is found with MLPA or DNA sequencing, the transcript needs to be analyzed (12, 13). Multiple RT-PCRs, followed by a direct Sanger sequencing on purified RT-PCR products, are performed to identify possible splicing defects (14).

Our study aims to verify if RT-PCR products can be pooled for library preparation and analyzed with a long-read sequencer (PacBio-RS-II) to identify and characterize single nucleotide variants and splicing defects.

We extracted RNA from the muscle biopsies of four patients with a Duchenne/Becker muscular dystrophy (DMD/BMD) using the TRIzol reagent (Invitrogen, CA, USA). Similarly, RNA from the muscle biopsies of four patients with bi-allelic causative variants in the titin (*TTN*) gene was extracted using the QIAGEN RNeasyPlus Universal Mini Kit (Qiagen, Germany). Finally, we used the PAXgene Blood RNA Kit (Qiagen) to extract RNA from the blood of two patients with a neurofibromatosis type-I, caused by heterozygous variants in the *NF1* gene.

For all the samples, 2 mg of total mRNA was converted into cDNA with the SuperScript III kit (Invitrogen). The cDNA of the DMD/BMD patients and of the NF1 patients was analyzed by RT-PCR using overlapping primers spanning the entire coding sequence

(12, 15). For titinopathy patients, the *TTN* cDNA was amplified with specific primers to characterize the effect of five variants in canonical splice sites, previously identified by a DNA-sequencing method (16). The RT-PCR products were then analyzed with a traditional Sanger sequencing approach (after gel extraction and/or cloning of the fragments) and a Single-Molecule Real-Time (SMRT)-Sequencing based strategy (Figure 1).

For the SMRT evaluation, all the RT-PCR products spanning the mutated regions were pooled in equimolar amounts. Library preparation and sequencing were performed according to the PacBio-RS-II protocol using P6/C4 chemistry with a prior size selection step to remove small library fragments (less than 700bp). A total of 1860 Mbp of sequence was obtained from one SMRT cell. Circular consensus sequencing (CCS) reads were generated using SMRTportal2.3 RS_ReadsOfInsert protocol at least 8 full-pass subreads.

The traditional sequencing revealed the disease-causing variants in all the patients. As summarized in Table 1, the SMRT-based approach confirmed the results obtained by the Sanger-based cDNA analysis, characterizing the splicing defects in the *DMD* transcripts, due to hemizygotic variants. Similarly, in patient NF1_1, the SMRT-analysis identified both the wild type transcript and the mutated one showing the skipping of exon 12. In patient NF1_2, the heterozygous single nucleotide variant c.5426 G>T resulting in a missense variant was detected by both approaches. The variant c.5426 G>T

p.(Arg1809Cys) has been previously identified in several families with a mild form of neurofibromatosis and it is currently listed as pathogenic in ClinVar (17, 18). In patient TTN_1, both methods identified the complete skipping of exon 67; in patient TTN_2 the insertion in the transcript of the first intronic nucleotide of intron 362; and in patient TTN_4 the partial skipping of exon 26. In patient TTN_3, the variant c.15776-1G>T caused the intron 54 retention. The second variant (c.67349-2A>C), interestingly, caused a partial skipping of exon 320 and the insertion in the transcript of the intron 319.

Long-read sequencing is increasingly used in clinical sequencing applications since it is expected to identify possible disease-causing variants usually not detected by short-read technologies (e.g. structural variants and tandem-repeats) (19-21). This preliminary study demonstrates that the SMRT sequencing is also useful for characterizing splicing defects.

A comprehensive molecular analysis, including DNA and RNA tests, is often essential for a correct diagnosis and a proper evaluation of prognosis. As seen in patient DMD_1, single nucleotide variants can be misinterpreted as missense changes although a secondary RNA analysis can reveal a deleterious effect on the splicing.

Similarly, downstream effects of splicing variants identified at the DNA-level should be carefully evaluated to finely map the consequences on the transcript(s) and, thereby, on the protein (5). However, sometimes, the characterization of splicing defects can be tedious and expensive. Our data suggests that pooling RT-PCR products in equimolar

amounts and analyzing them with a long-read sequencer, without performing any complex and time-consuming post-PCR methodology (gel extraction or cloning), is a successful method to analyze splice aberrations.

On the other hand, when there is a specific clinical suspicion and DNA tests have not identified any causative genetic variant, analysis of the cDNA of interest is still a useful second-tier test. One percent of patients with a DMD phenotype, for example, carry splicing variants only identified with a cDNA analysis (12). Similarly, cDNA analysis is a valuable second-level test for patients with a clinical diagnosis of NF1 (15), or with a suspected recessive titinopathy, when only a monoallelic truncating variant has been identified (6, 7, 22). Analyzing the full-length cDNA, especially for large genes, is complex and not easily scalable (23). Traditionally, for RNA analysis of the entire coding sequences, primer pairs that amplify partially overlapping fragments of 500–700 bp are designed (15). RT-PCR products are thereby first analyzed by agarose gel electrophoresis and subsequently by bidirectional Sanger sequencing (15).

A simplified approach could be implemented using a SMRT-based strategy. The transcript of interest could be amplified through multiple long RT-PCRs (5-10Kb), thereby, using a reduced number of primers and performing a reduced number of PCRs. Optimizing the technical aspects (primer/PCR design and equimolar pooling), RT-PCR products could be pooled and run with a PacBio sequencer, further simplifying the transcript analysis.

Acknowledgments

The authors would like to thank for the support: Magnus Ehrnrooth Foundation (MS), Päivikki ja Sakari Sohlbergin Säätiö (MS and MJ), Paulön Säätiö (MS), Biomedicum Helsinki säätiö (MJ), Jane and Aatos Erkko Foundation (PH), Medicinska Understödsföreningen Liv och Hälsa rf (PH), Folkhälsan Research Foundation (BU), Erkko Foundation (BU), Juselius Foundation (BU), Finnish Academy (BU).

Figure 1 was created with BioRender (www.biorender.com).

References

1. Stark Z, Tan TY, Chong B, Brett GR, Yap P, Walsh M, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet Med. 2016;18(11):1090-6.
2. Tan TY, Dillon OJ, Stark Z, Schofield D, Alam K, Shrestha R, et al. Diagnostic Impact and Cost-effectiveness of Whole-Exome Sequencing for Ambulant Children With Suspected Monogenic Conditions. JAMA Pediatr. 2017;171(9):855-62.
3. Nigro V, Savarese M. Next-generation sequencing approaches for the diagnosis of skeletal muscle disorders. Curr Opin Neurol. 2016;29(5):621-7.
4. Savarese M, Di Fruscio G, Torella A, Fiorillo C, Magri F, Fanin M, et al. The genetic basis of undiagnosed muscular dystrophies and myopathies: Results from 504 patients. Neurology. 2016;87(1):71-6.
5. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24.
6. Savarese M, Johari M, Johnson K, Arumilli M, Torella A, Topf A, et al. Improved Criteria for the Classification of Titin Variants in Inherited Skeletal Myopathies. J Neuromuscul Dis. 2020;7(2):153-66.
7. Bryen SJ, Ewans LJ, Pinner J, MacLennan SC, Donkervoort S, Castro D, et al. Recurrent TTN metatranscript-only c.39974-11T>G splice variant associated with autosomal recessive arthrogryposis multiplex congenita and myopathy. Hum Mutat. 2019.
8. Gonorazky HD, Naumenko S, Ramani AK, Nelakuditi V, Mashouri P, Wang P, et al. Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease. Am J Hum Genet. 2019;104(3):466-83.
9. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-48 e24.
10. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci Transl Med. 2017;9(386).
11. Scotti MM, Swanson MS. RNA mis-splicing in disease. Nat Rev Genet. 2016;17(1):19-32.
12. Aartsma-Rus A, Ginjaar IB, Bushby K. The importance of genetic diagnosis for Duchenne muscular dystrophy. J Med Genet. 2016;53(3):145-51.
13. Fratter C, Dalgleish R, Allen SK, Santos R, Abbs S, Tuffery-Giraud S, et al. EMQN best practice guidelines for genetic testing in dystrophinopathies. Eur J Hum Genet. 2020.

14. Valero MC, Martin Y, Hernandez-Imaz E, Marina Hernandez A, Melean G, Valero AM, et al. A highly sensitive genetic protocol to detect NF1 mutations. J Mol Diagn. 2011;13(2):113-22.
15. Giugliano T, Santoro C, Torella A, Del Vecchio Blanco F, Grandone A, Onore ME, et al. Clinical and Genetic Findings in Children with Neurofibromatosis Type 1, Legius Syndrome, and Other Related Neurocutaneous Disorders. Genes (Basel). 2019;10(8).
16. Evila A, Arumilli M, Udd B, Hackman P. Targeted next-generation sequencing assay for detection of mutations in primary myopathies. Neuromuscul Disord. 2016;26(1):7-15.
17. Pinna V, Lanari V, Daniele P, Consoli F, Agolini E, Margiotti K, et al. p.Arg1809Cys substitution in neurofibromin is associated with a distinctive NF1 phenotype without neurofibromas. Eur J Hum Genet. 2015;23(8):1068-71.
18. Santoro C, Maietta A, Giugliano T, Melis D, Perrotta S, Nigro V, et al. Arg(1809) substitution in neurofibromin: further evidence of a genotype-phenotype correlation in neurofibromatosis type 1. Eur J Hum Genet. 2015;23(11):1460-1.
19. Mitsuhashi S, Matsumoto N. Long-read sequencing for rare human genetic diseases. J Hum Genet. 2020;65(1):11-9.
20. Asogawa M, Ohno A, Nakagawa S, Ochiai E, Katahira Y, Sudo M, et al. Human short tandem repeat identification using a nanopore-based DNA sequencer: a pilot study. J Hum Genet. 2020;65(1):21-4.
21. Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. J Hum Genet. 2019;64(5):359-68.
22. Savarese M, Maggi L, Vihola A, Jonson PH, Tasca G, Ruggiero L, et al. Interpreting Genetic Variants in Titin in Patients With Muscle Disorders. JAMA Neurol. 2018;75(5):557-65.
23. Savarese M, Valipakka S, Johari M, Hackman P, Udd B. Is Gene-Size an Issue for the Diagnosis of Skeletal Muscle Disorders? J Neuromuscul Dis. 2020.

Figure 1: Traditional and SMRT-based cDNA analysis for splicing characterization

a) Splicing variants are traditionally identified and characterized using a reverse transcriptase-polymerase chain reaction (RT-PCR), followed by a direct Sanger sequencing on purified products. In case of multiple RT-PCR products, this requires a gel extraction or cloning of PCR products into a plasmid for sequencing.

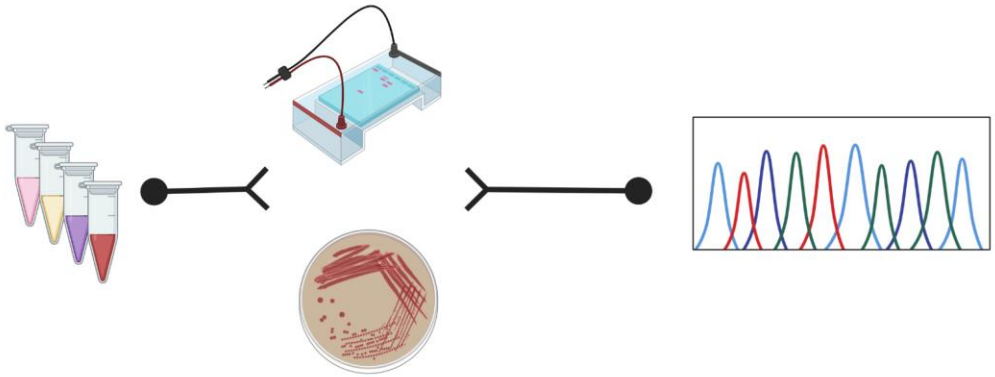
b) In the SMRT-based approach, RT-PCR products spanning the mutated regions were pooled in equimolar amounts in a single tube and the pool was used for library preparation and sequencing following the PacBio-RS-II protocol.

Table 1 – Variants identified and characterized

Patient ID	DNA analysis	Transcript analysis
DMD_1	chrX: 32632551 C>A (c.1351 G>T in exon 12)	Sequence from nucleotide r.1332 to r.1359 is deleted from the transcript (Partial skipping exon 12: r.1332_1359del)
DMD_2	chrX:32489317 C>T (c.2949+964 G>A in intron 22)	New exon is created from intron 22 and its sequence from positions 2949+889 to 2949+1147 is inserted in the transcript (r.2949_2950ins2949+889_2949+1147)
DMD_3	chrX:32591688_32591689ins(353) (c.1771_1772ins353 in exon 15)	As a consequence of an insertion of a transposon of 353 nucleotides in exon 15, the sequence from nucleotide r.1705 to r.1812 is deleted from the transcript (Complete skipping exon 15:r.1705_1812del)
DMD_4	chrX:31279418 T>C (c.9225-285 A>G in intron 62)	New exon is created from intron 62 and its sequence from nucleotide r.9225-347 to r.9225-290 is inserted in the transcript (r.9925_9926ins9225-347_9225-290)
NF1_1	chr17: 29533256 A>C (c.1261-2A>C in intron 11)	Sequence from nucleotide r.1261 to r.1284 is deleted from the transcript (Partial skipping exon 12:r.1261_1284del)
NF1_2	chr17:29654737 G>T (c.5426 G>T)	Single nucleotide variant: r.5426 G>T
TTN_1	chr2:179593225 A>T (c.19426+2T>A in intron 67)	Sequence from nucleotide r.19148 to r.19426 is deleted from the transcript (Complete skipping ex67:r.19148_19426del)
TTN_2	chr2:179392999dup (c.107377+2dup in intron 362)	First nucleotide of intron 362 is inserted in the transcript (r.107377_107378ins107377+1)
TTN_3	chr2:179598245 C>A (c.15776-1 G>T in intron 54)	Intron 54 sequence is inserted in the transcript. NOTE: nucleotide 15776-1 changed from g to t (Intronic retention: r.15775_15776ins[15775+1_15776-1])
TTN_3	chr2:179444577 T>G (c.67349-2 A>C in intron 319)	a) the sequence from nucleotide r. 67349 to r.67363 is deleted from the transcript (Partial skipping ex320:r. 67349_67363del) b) the intron 319 sequence is inserted in the transcript. NOTE: nucleotide 67349-2 changed from a to c (Intronic retention: r.67438_67439ins[67438+1_67349-1])
TTN_4	chr2:179642045 C>T (c.4646-1G>A in intron 26)	Sequence from nucleotide r.4646 to r.4659 is deleted from the transcript (Partial skipping exon 26:r.4646_4659del)

Genomic coordinates are in genome assembly GRCh37 (hg19). Reference transcripts: *DMD* (NM_004006.1), *NF1* (NM_000267.3), *TTN* (NM_001267550.1)

a)



b)