The Journal of Molecular Diagnostics, Vol. 19, No. 5, September 2017

ELSEVIER

the Journal of Molecular Diagnostics

jmd.amjpathol.org

GBA Analysis in Next-Generation Era

Pitfalls, Challenges, and Possible Solutions

Stefania Zampieri, Silvia Cattarossi, Bruno Bembi, and Andrea Dardis

From the Regional Coordinator Centre for Rare Diseases, Academic Hospital Santa Maria della Misericordia, Udine, Italy

Accepted for publication May 19, 2017.

Address correspondence to Andrea Dardis, Ph.D., Regional Coordinator Centre for Rare Diseases, Academic Hospital Santa Maria della Misericordia, Piazzale Santa Maria della Misericordia 15, 33100 Udine, Italy. E-mail: andrea.dardis@asuiud.sanita.fvg.it.

Mutations in the gene encoding the lysosomal enzyme acid β-glucosidase (*GBA*) are responsible for Gaucher disease and represent the main genetic risk factor for developing Parkinson disease. In past years, next-generation sequencing (NGS) technology has been applied for the molecular analysis of the *GBA* gene, both as a single gene or as part of gene panels. However, the presence of complex gene-pseudogene rearrangements, resulting from the presence of a highly homologous pseudogene (*GBAP1*) located downstream of the *GBA* gene, makes NGS analysis of *GBA* challenging. Therefore, adequate strategies should be adopted to avoid misdetection of *GBA* recombinant mutations. Here, we validated a strategy for the identification of *GBA* mutations using parallel massive sequencing and provide an overview of the major drawbacks encountered during *GBA* analysis by NGS. We implemented a NGS workflow, using a set of 38 patients with Gaucher disease carrying different *GBA* alleles identified previously by Sanger sequencing. As expected, the presence of the pseudogene significantly affected data output. However, the combination of specific procedures for the library preparation and data analysis resulted in maximal repeatability and reproducibility, and a robust performance with 97% sensitivity and 100% specificity. In conclusion, the pipeline described here represents a useful approach to deal with *GBA* sequencing using NGS technology. *(J Mol Diagn 2017, 19: 733—741; http://dx.doi.org/10.1016/j.jmoldx.2017.05.005)*

The *GBA* gene (Online Mendelian Inheritance in Man number 606463; Human Reference Genome: GRCh37/hg19 Chromosome 1: 155,204,239 to 155,214,653), encoding the lysosomal enzyme acid β-glucosidase (GBA; Enzyme Commission number 3.2.1.45), comprises 11 exons and 10 introns spanning 7.6 kb of sequence. It is located on chromosome 1q21 within a complex locus containing seven genes and two pseudogenes, likely originating from a duplication event of this chromosomal region. Indeed, a highly homologous 5.7-kb pseudogene (*GBAP1*; Online Mendelian Inheritance in Man number 606463; Human Reference Genome: GRCh37/hg19 Chromosome 1: 155,183,616 to 155,197,325) is located approximately 16 kb downstream of the functional *GBA* gene.[1] The exonic region of the *GBAP1* shares 96% sequence homology with the coding region of the *GBA* gene, whereas the sequence homology reaches 98% in the region between intron 8 and the 3′ untranslated region. *GBAP1* is shorter than *GBA*, missing several *Alu* insertions in intronic tracts and carrying a 55-bp deletion in exon 9[2,3] (Figure 1). The high homology and the physical proximity between *GBAP1* and *GBA* enables both nonreciprocal and reciprocal homologous recombination events,[4,5] resulting in complex gene—pseudogene rearrangements. The most frequent recombinant alleles arise from nonreciprocal recombination events, leading to gene conversions in which a portion of the active gene sequence is replaced by the homologous sequence of the pseudogene, resulting in fusion alleles.[6]

Recessive mutations in the *GBA* gene cause Gaucher disease (GD), a rare lysosomal storage disorder in which the deficient activity of β-glucosidase leads to the progressive accumulation of glucosylceramide and other glycosphingolipids within the lysosomes, resulting in multiorgan dysfunction.[7]

Supported in part by Shire Human Genetics Therapies, Inc.

Disclosures: A.D. and B.B. have received travel grants and research funding from Shire, Inc., Actelion Pharmaceuticals, Orphazyme Aps, and Genzyme.

http://dx.doi.org/10.1016/j.jmoldx.2017.05.005

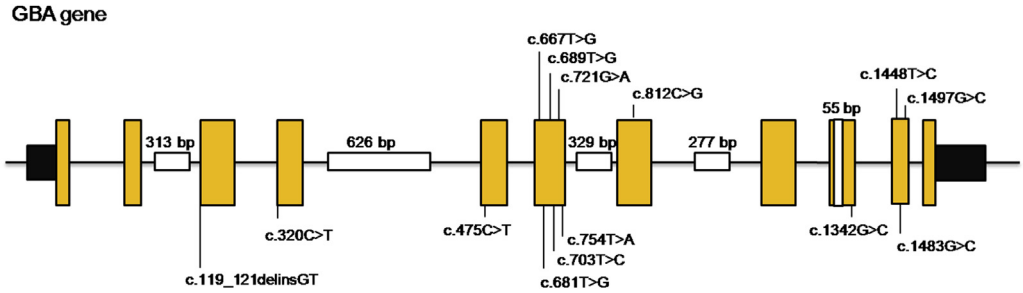

Figure 1 Schematic representation of the exonic (**yellow boxes**) and intronic (**gray lines**) structure of the *GBA* gene with the main pseudogene-derived mismatches and amino acid changes associated. Both *GBA* and *GBAP1* comprise 11 exons and the exon/intron boundaries are identical. **Bars** represent the portion of GBA genes deleted in the pseudogene.

To date, more than 400 different disease-causing *GBA* mutations, including single base changing, splicing alterations, partial and total deletions, insertions, and gene−pseudogene rearrangements have been reported in patients affected by GD (Human Gene Mutation Database, *http://www.hgmd.org*, last accessed February 21, 2017). A prevalent class of mutant alleles encountered in patients with GD is caused by recombination events occurring between intron 2 and exon 11.[8] Indeed, more than 20 recombinant alleles have been described to date, with RecNcil and Recdelta55 being the most common, either alone or in combination with additional point mutations.[8,9] RecNcil derives from a cross-over junction area from intron 9 to exon 10, resulting in the incorporation of a segment of *GBAP1* that includes missense mutations p.L483P (L444P), p.A495P (A456P), and p.V499V (V460V) into the functional *GBA* gene. Recdelta55 encompasses a 55-bp deletion in exon 9 of the *GBA* gene, corresponding to the deleted portion of the pseudogene.[10] Recombinant alleles that include the 55-bp deletion within exon 9 either in association with the p.D448H (D409H) alone or with the p.D448H (D409H) and the RecNcil have been identified as well.[11,12] The site of cross-over of these recombinants is located within the region extending from the end of intron 8 to the beginning of exon 9.[13,14]

In addition, recombinant events involving the 3′ untranslated region, introns 2, 3, and 6 of the *GBA*, and introducing larger segments of pseudogene into the recombinant allele, have been reported.[15−20]

Although recessive mutations in the *GBA* gene cause GD, a large collaborative international multicenter study of more than 5000 patients affected by Parkinson disease (PD) and age-matched controls definitively showed the association between the presence of heterozygous *GBA* mutations and PD.[21] These data were supported further by studies performed in different clinical cohorts of PD patients from several regions, including Europe,[22−25] America,[26−28] Africa,[29] and Asia.[30,31] As expected, the highest frequency of *GBA* mutations in patients affected by PD has been found in Ashkenazi Jewish populations.[32,33] These studies provided the bases for the inclusion of *GBA* sequence analysis in the diagnostic work-up of patients affected by PD. Although DNA sequencing using the Sanger method has been the gold standard over the past 30 years for mutation identification in patients affected by GD who present with an incidence of 1:40,000 births, it is not suitable for the systematic analysis of *GBA* in patients affected by a common disease such as PD, which has an estimated incidence rate of 8 to 18 per 100,000 person-years.[34] The advent of next-generation sequencing (NGS) platforms, able to generate enormous amounts of sequence data in a short time at an affordable cost, provided a high-quality, efficient, and affordable replacement for conventional sequencing. In fact, lately NGS has been applied for the analysis of the *GBA* gene, both alone or as part of gene panels, to investigate the genetic background in large cohorts of patients affected by PD.[35−38] However, the presence of *GBA*−*GBAP1* complex rearrangements makes NGS analysis of *GBA* challenging and adequate strategies should be adopted to avoid misdetection of *GBA* recombinant mutations. For this reason, in addition to NGS analysis, some investigators have analyzed at least the 3′ region of the gene by Sanger sequencing,[35,39] or considered only the already most frequently reported mutations.[40] It is worth noting that many studies using NGS technology without the support of Sanger sequencing did not report the presence of recombinant mutations, raising the possibility that the presence of these allele was underestimated. Indeed, despite the advances in sequencing technology, Sanger sequencing remains the gold standard method to analyze the *GBA* gene, even in large patient cohorts.[28,31,41−43]

Here, we validated a strategy for *GBA* genotyping using parallel massive sequencing and provide a comprehensive review of the main pitfalls encountered during *GBA* analysis by NGS.

Materials and Methods

Samples

A total of 38 GD samples (carrying 36 different alleles detected by Sanger sequencing of all exons and intronic flanking regions of the *GBA* gene) and 209 normal controls (418 wild-type alleles) were resequenced using an Illumina MiSeq platform (Illumina, San Diego, CA).

Genomic DNA was obtained from patients affected by GD diagnosed at the Regional Coordinator Centre for Rare Diseases of Udine and normal controls, with the patients' (and/or a family member's) written informed consent. The study was approved by the Regional Ethical Committee.

The clinical diagnosis of GD was confirmed by showing reduced levels of β-glucosidase activity in peripheral blood leukocytes or fibroblasts, followed by the molecular analysis of the 11 exons and the flanking intronic sequences of the *GBA* gene by Sanger sequencing. The genotypes of the patients are summarized in Table 1.

Mutation Nomenclature

All mutations are described according to the current mutation nomenclature guidelines (Human Genome Variant Society, *http://www.hgvs.org/mutnomen*, last accessed February 21, 2017), ascribing the A of the first ATG translation initiation codon as nucleotide 1 (*https://www.ncbi.nlm.nih.gov/clinvar*; accession number NM_000157.3).[44,45] Traditional amino acid residue numbering, which excludes the first 39 amino acids of the leader sequence, nevertheless also has been provided within parentheses and designed without the p. prefix.

Library Preparation

For library preparation, the whole genomic sequence of the *GBA* gene was amplified by long-range PCR (LR-PCR) in two overlapped fragments of 2870 bp and 4492 bp, using primers designed to amplify the gene selectively and not the pseudogene.[46] The purified amplicons were processed using a Nextera XT DNA sample preparation kit (Illumina) according to the manufacturer's recommendation to generate paired-end libraries. Briefly, simultaneous PCR amplicon fragmentation and adapter sequence ligation (tagmentation) was performed using a transposase enzyme. Adapter sequences were used to amplify the obtained DNA fragment by PCR. The PCR reaction also adds dual index sequences (barcode), which are needed for sample identification. The sample library then subsequently is purified and quantified using the Quant-iT Picogreen double-stranded DNA Assay Kit (Thermo Fisher Scientific, Waltham, MA). Equal volumes of normalized libraries were sequenced on an Illumina MiSeq platform using an Illumina MiSeq Reagent kit v2 (MS-102-2002) for 300 cycles.

Data Analysis

MiSeq Control software version 2.3.0.3 (Illumina) was used for monitoring the runs and quality control, whereas secondary analysis including paired-end sequence alignment, duplicate removal, single-nucleotide variant calling, and indel detection was performed automatically using MiSeq reporter v2.4.60. Paired-end reads were mapped against the whole human reference sequence build GRCh37/hg19. A quality control check was performed using the FastQC tool software version 0.11.5 (*https://www.bioinformatics.babraham.ac.uk/projects/fastqc*),[47] and only samples with a sequencing quality control score ≥30 and with a minimal read depth of 200× were considered for variant analysis. As output, a Variant Call Format file is generated, containing the list of the identified mutations.

Variant annotation and filtering was performed using wANNOVAR (*http://wannovar.wglab.org*, last accessed February 21, 2017).[48]

In addition, to confirm the presence of a recombinant allele, reads then were forced to re-align against the *GBA* gene using a portion of human genome corresponding to the sole *GBA* gene as a reference, hereafter referred to as the *GBA* genome (Chromosome 1: 155211205 to 155204618).

The .bam files were loaded into Integrative Genomics Viewer[49,50] (IGV software version 2.3; The Broad Institute, Cambridge, MA) for read visualization. The variants identified were compared with the variant obtained with the previous alignment against the whole genome.

All pathogenic variants were confirmed by Sanger sequencing.

Results

The GD and control samples were used to assess the quality assurance and assay reliability, assay reproducibility, and analytical sensitivity and specificity.

Quality Assurance and Assay Reliability

Quality control of NGS data implied the assessment for each sample of the total number of reads, read quality control, and the percentage of reads that were paired properly and mapped to the reference genome.

At the amplicon level, the mean coverage was assessed and, as reported in the *Materials and Methods* section, only samples with a mean coverage >200 were taken into consideration. However, the length of the target sequence (6600 bp) and the number of samples/run guaranteed a higher mean coverage (>5000 reads) across samples. In agreement with previous reports, a depth bias characteristic of transposase-based protocols[51] was observed. Indeed, the transposon strategy used to process samples implies the cleavage of DNA amplicons into fragments of approximately 300 base pairs, integrating transposon sequences onto both 5′- and 3′-ends of the fragments. Because the transposases are less favorable to bind to the very ends of the amplicons obtained by LR-PCR, a strand bias between forward and reverse strands was observed at the 3′- and 5′-termini of the *GBA* gene. However, once again, the number of samples/run guaranteed a high coverage (>500 reads) also in these regions and ensured the reliability of the data, despite the low amplicon uniformity. It is worth noting that coverage is one of the most important data quality metrics in NGS because it has a direct relationship with the sensitivity and specificity of variant detection.

Table 1 Genotypes of Patients with GD Analyzed in This Study

N	Allele 1	Allele 2
1	c.1226A>G, p.N409S (N370S)	c.203delC, p.P68fs*23 (P29fs*23)
2	c.259C>T, p.R87W (R48W)	c.1448T>C, p.L483P (L444P)
3	c.1174C>G, p.R392G (R353G)	c.1174C>G, p.R392G (R353G)
4	c.1448T>C, p.L483P (L444P)	c. 1505G>A, p.R502H (R463H)
5	c.750A>G, p.Y251_K254del (Y212_K215del)	c.1226A>G, p.N409S (N370S)
6	c.626G>C, p.R209P (R170P)	c.[1225-10delC;1225-14T>A], p.N409Sfs*20 (N370Sfs*20)
7	c.1226A>G, p.N409S (N370S)	c.1370A>G, p.Y457C (Y418C)
8	c.475C>T, p.R159W (R120W)	c.721G>A, p.G241R (G202R)
9	c.1226A>G, p.N409S (N370S)	c.187G>A, p.D63N (D24N)
10	c.1226A>G, p.N409S (N370S)	c.1319C>T, p.P440L (P401L)
11	c.[882T>G; 1342 G>C], p.[H294Q; D448H] (H255Q+D409H)	c.1448T>C, p.L483P (L444P)
12	c.592C>A, p.P198T (P159T)	g.4641_J03060.1, g.2828 (RecI)
13	c.741delC, p.W248Gfs*6 (W209Gfs*)	c.1448T>C, p.L483P (L444P)
14	c.1226A>G, p.N409S (N370S)	g.4179_5042conJ03060.1, g.2367_2910
15	c.589-13C>G, p.I197Lfs*4 (I158Lfs*4)	c.625C>T, p.R209C (R170C)
16	c.1226A>G, p.N409S (N370S)	c.[882T>G; 1342 G>C], p.[H294Q; D448H] (H255Q+D409H)
17	c.475C>, p.R159W (R120W)	c.1448T>C, p.L483P (L444P)
18	c.T703C, p.S235P (S196P)	c.1226A>G, p.N409S (N370S)
19	c.1226A>G, p.N409S (N370S)	c.1263_1317del, p.L422Pfs*4 (L383Pfs*4)
20	c.115+1G>A	c.680A>G, p.N227S (N188S)
21	c.1226A>G, p.N409S (N370S)	c.1448T>C, p.L483P (L444P)
22	c.1226A>G, p.N409S (N370S)	c.1271TC, p.L424P (L385P)
23	c.1448T>C, p.L483P (L444P)	c.1448T>C, p.L483P (L444P)
24	c.[882T>G; 1342 G>C], p.[H294Q; D448H] (H255Q+D409H)	c.[882T>G; 1342 G>C], p.[H294Q; D448H] (H255Q+D409H)
25	c.1226A>G, p.N409S (N370S)	g.7319_7368conJ03060.1:g.4856_4905 (RecNcil)
26	c.1603C>T, p.R535C (R314C)	c.[882T>G; 1342 G>C], p.[H294Q; D448H] (H255Q+D409H)
27	c.1226A>G, p.N409S (N370S)	g.7319_7368conJ03060.1:g.4856_4905 (RecNcil)
28	c.[882T>G; 1342 G>C], p.[H294Q; D448H] (H255Q+D409H)	c.[882T>G; 1342 G>C], p.[H294Q; D448H] (H255Q+D409H)
29	c.1448T>C, p.L483P (L444P)	g.7319_7368conJ03060.1:g.4856_4905 (RecNcil)
30	c.1226A>G, p.N409S (N370S)	c.84dupG, p.L29Afs*18 (84GG)
31	c.1448T>C, p.L483P (L444P)	c.1361C>G, p.P454R (P415R)
32	c.1448T>C, p.L483P (L444P)	c.1448T>C, p.L483P (L444P)
33	c.1226A>G, p.N409S (N370S)	c.1297G>T, p.V433L (V394L)
34	c.1141T>G, p.C381G (C342G)	c.1090G>A, p.G364R (G325R)
35	c.1448T>C, p.L483P (L444P)	c.[1093G>A; 1448T>C], p.[E365K; L483P] (E326K+L444P)
36	c.1226A>G, p.N409S (N370S)	c.[1263_1317del; c.1342G>C], p.L422Pfs*4 (Rec[1263del55; c.1342G>C])
37	c.1263_1317del, p.L422Pfs*4 (Recdelta55)	ND
38	c.1226A>G, p.N409S (N370S)	c.1263_1317del, p.L422Pfs*4 (Recdelta55)

For cDNA numbering, +1 corresponds to the A of the first ATG translation initiation codon. Traditional nomenclature has been included in brackets without the p prefix. RefSeq cDNA: NM_000157.3.

ND, none detected.

Importantly, the amplicon and bp coverage patterns remained uniform among intrasample and intersample runs, as well as the quality control performance metrics.

The high homology between *GBAP1* and *GBA* combined with alignment against the whole genome did not affect mapping quality and coverage significantly when missense mutations not derived from recombination were identified. However, the presence of recombinant mutations determined a poor mapping quality and a decrease in coverage in the part of the *GBA* gene that has recombined, as observed using IGV software after alignment against the whole genome.

Analytical Sensitivity and Specificity for *GBA* Mutation Detection

Point Mutations

All missense mutations (22), microdeletions or insertions (6), intronic mutations (1), and complex alleles (2) identified previously by Sanger sequencing were confirmed by NGS analysis.

It is worth noting that analyzing the whole *GBA* gene enabled us to identify 10 variants located in an intronic position far from the classic donor/acceptor splice site. All of

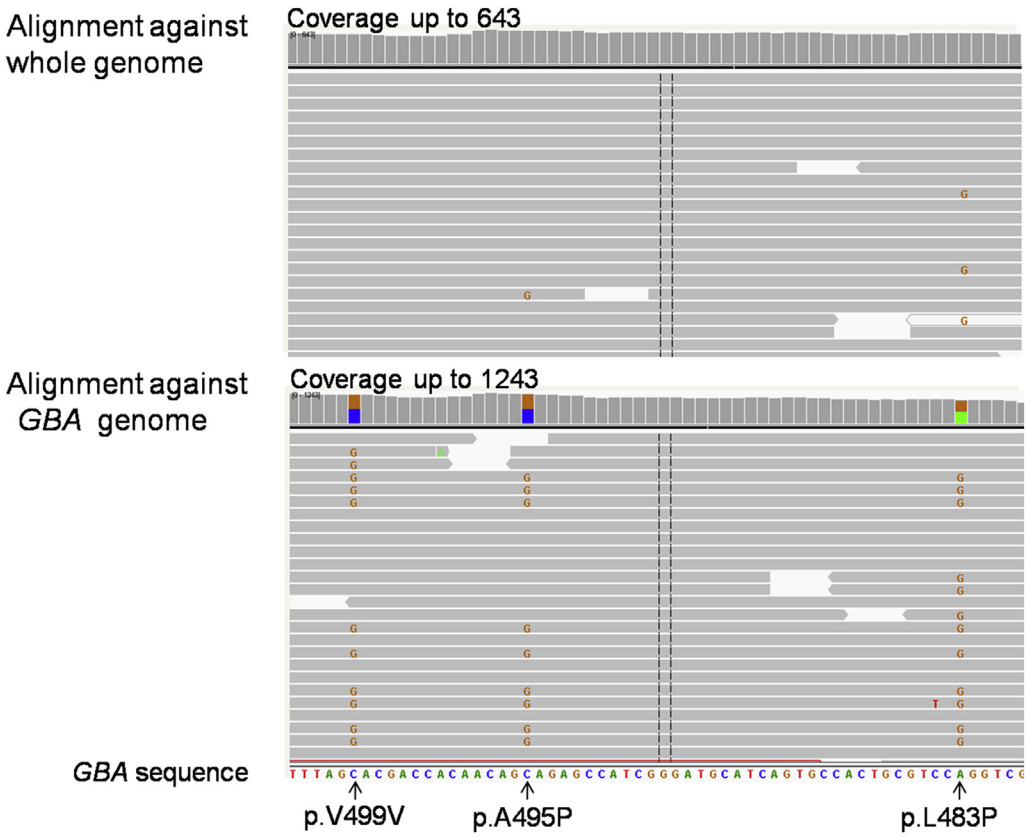

Figure 2 Analyses of *GBA* reads using IGV software. Alignment of *GBA* reads in a patient carrying the RecNcil allele, using either the whole genome or the *GBA* genome as a reference. IGV software screenshot shows the recombinant region including the missense mutations p.L483P (L444P), p.A495P (A456P), and p.V499V (V460V). Mutations have been identified as variants only after *GBA* genome alignment.

them have been reported previously as polymorphisms, according to the Database of Single Nucleotide Polymorphisms (*https://www.ncbi.nlm.nih.gov/snp*), or found both in patients with GD and healthy controls (Supplemental Table S1), except for the c.454+456_457delinsAG and the c.455-54C>T, both identified within intron 4 in patients 37 and 4, respectively. As shown in Table 1, Sanger sequencing allowed the identification of the Recdelta55 allele in patient 37; however, the mutation present in the second allele remained unknown. An *in silico* analysis using the NNsplice program (*http://www.fruitfly.org/seq_tools/splice.html*, last accessed February 21, 2017) showed that this variant might generate a new acceptor splice site (score: 0.86). Whether this sequence variant actually leads to the transcription of an aberrant spliced transcript needs to be investigated further.

Conversely, NNsplice analysis of the c.455-54C>T mutation did not affect the strengths of either natural donor or acceptor sites.

Recombinant Alleles

Five different recombinant alleles, identified by Sanger sequencing in eight patients with GD (Table 1), have been almost completely lost after alignment against the whole genome. Indeed, reads belonging to the recombinant allele aligned preferentially against the *GBAP1* gene. This was owing to poor alignment and mapping quality. Therefore, the variant calling analysis failed to recognize the mutations corresponding to the recombination, as shown in the example reported in Figure 2 for the recombinant allele carrying mutations p.L483P (L444P), p.A495P (A456P), and p.V499V (V460V). Instead, alignment against the *GBA* genome enabled us to identify and annotate all of these variants correctly. An example of the result obtained in a heterozygous patient for the RecNcil allele (homologous recombination of exons 10 to 11) is shown in Figure 2. It was possible to perform this alignment because the *GBA* gene was amplified specifically before sample processing and therefore all sequencing fragments were derived unequivocally from the *GBA* gene.

Furthermore, a manual inspection of the .bam file using IGV software showed that a given recombinant mutation represented only 11% to 13% of the reads in that specific position when the alignment was performed against the whole genome, whereas it represented the expected 48% to 52% of the reads when the alignment was performed against the *GBA* genome.

In the specific case of the Recdelta55 allele, the recombination encompassed a region lacking 55 bp. Variant calling analysis failed to identify this alteration even after alignment against the GBA genome. This allele was detected only by the analysis of .bam data by IGV software

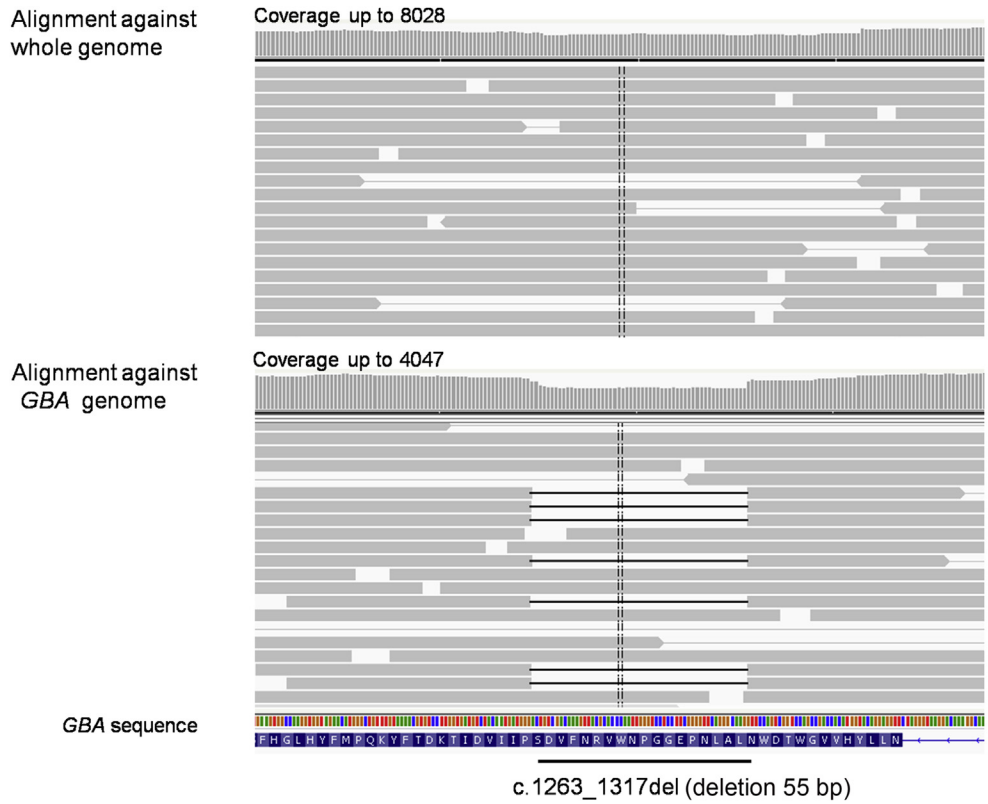

Figure 3 Alignment of *GBA* reads in a patient carrying the Recdelta55 allele, using either the whole genome or the *GBA* genome as a reference. IGV software screenshot shows a clear decrease of coverage in the recombinant region only after *GBA* genome alignment.

after *GBA* genome alignment as a specific decrease in coverage (Figure 3).

Assay Reproducibility

To assess the repeatability (inter-run) and reproducibility (intrarun), samples carrying the same mutation were used.

The mutation p.N409S (N370S), which is very common among patients with GD (17/76 alleles), always was identified among different patients in each run, indicating 100% repeatability for the detection of this pathogenic variant. The same result was obtained when comparing other variants that were present in multiple samples, such as the complex allele p.[H294Q; D448H] (H255Q + D409H) and the mutation p.L483P (L444P).

Samples 24 and 27 were selected to test the reproducibility in three different runs, which was maximal (100%).

Run-to-run comparisons were used to determine the level of multiplexing possible to ensure the optimal coverage required for data reproducibility.

Analytical Sensitivity and Specificity

Analytical sensitivity and specificity were determined considering the alignment against the whole genome and the re-alignment against the *GBA* genome according to the following formulas: sensitivity: N true positives/(N true positives + N false negatives); and specificity: N true negatives/(N true negatives + N false positives).

True positives are the alleles carrying mutations detected by Sanger sequencing, whereas true negatives are the alleles in which no mutations were detected by Sanger sequencing.

As expected, no false positives were detected using the described workflow, whereas the sensitivity was 87% when the alignment was performed against the whole genome and increased to 97% when the *GBA* genome was used. Importantly, no errors were reported in identifying the heterozygote/homozygote status of mutation.

Even though the Recdelta55 was detected manually using IGV software, it was considered as true false negative.

Discussion

NGS platforms are able to process billions of DNA fragments. Thus, the first step in NGS workflow for investigating genetic disorders involves the sample library preparation, which consists of an enrichment of the target gene sequences by multiplex PCR amplification or DNA capturing methods. If a PCR-based strategy is selected for enrichment, specific PCR primers should be designed to generate fragments of 150 to 300 bp of the target, whereas if a capture approach is used, the DNA should be fragmented first in 150- to 500-bp sequences and then specific probes should be designed to capture each fragment of the desired

target. In the case of *GBA* sequencing, during this step it is crucial to amplify/capture selectively only the active gene. In fact, because the *GBA* and *GBAP1* sequence share 96% to 98% of homology, the NGS data analysis pipelines are unable to determine whether a given sequence derives from the functional *GBA* gene or the pseudogene. As a consequence, if a specific capture of the active gene is not performed, a decrease in read depth and a decrease in mapping quality will be obtained as a result of poor alignment with the active gene. Furthermore, reads generated by the pseudogene sequencing might align with the active gene, resulting in false-positive results.

To selectively enrich the sample in *GBA* sequences and avoid *GBAP1* contamination, primers for amplicon generation or probes for capture should target sequences that are different between the gene and pseudogene. Looking at the structure of *GBA* and *GBAP1* (Figure 1), it is quite clear that it is not possible to retrieve sequences to target the *GBA* gene specifically within a size range of 150 to 500 bp.

To overcome this issue, it is highly recommended to use LR-PCR to amplify the active gene selectively, followed by a fragmentation of the PCR product. Indeed, LR-PCR using specific primers, enabling generation of one or two overlapping fragments covering the whole *GBA* gene in a specific manner, has been validated and used extensively in clinical settings for the molecular diagnosis of GD.[46,52–55]

However, we have shown that even though a specific PCR amplification of the *GBA* gene was performed during the library preparation, data analysis might be compromised by poor alignment. In fact, aligner software generally aligns reads against the whole genome; thus, the reads originally belonging to the *GBA* gene might align to the pseudogene region with the consequent reduction of read depth and mapping quality. This is particularly critical in the case of recombinant alleles in which a region of the gene becomes identical to the pseudogene. Therefore, most of the generated reads aligned preferentially against the *GBAP1* gene, with a consequent decrease in read depth and loss of the recombinant allele. Indeed, using specific LR-PCR for library preparation and aligning against the whole genome resulted in an error of variant calling causing a false-negative result. However, because the *GBA* gene specifically was amplified before sample processing and therefore all sequencing fragments were derived unequivocally from the *GBA* gene, it was possible to align the reads against the *GBA* genome. This strategy, which already has been used to deal with the analysis of sequencing data of genes and highly homologous pseudogenes,[56,57] enabled us to identify and annotate correctly all recombinant variants, except the Recdelta55 allele. Indeed, variant calling analysis failed to identify this alteration. However, an inspection of the .bam data by IGV after *GBA* genome alignment showed a specific decrease in coverage of this sequence in patients carrying this mutation, suggesting that it might be possible to detect this specific kind of mutation by improving the sensitivity of the variant calling software. In the absence of such a bioinformatic tool this mutation needs to be identified by Sanger sequencing.

As expected, the method described here led to the identification of 10 deep intronic variants not identified during Sanger sequencing of exons and intronic flanking sequences. Among them, the c.454+456_457delinsAG and the c.455-54C>T. have not been described before. Therefore, this approach might lead to an increase in the mutation detection rate because deep intronic pathogenic variants have been identified in several lysosomal storage disorders.[58,59] However, results should be interpreted with caution because intronic regions are highly polymorphic and functional studies should be performed to determine the pathogenic nature of these variants.

Issues related with the application of NGS pipelines for analysis of genes that share a high degree of homology with inactive pseudogenes already have been raised and different strategies have been proposed for specific enrichment. Among them, microdroplet-based PCR has been used for enrichment of a desired genomic region[60] and copy number variant detection has been used as an indicator of structural variants, such as large rearrangements.[38] In the specific case of the *GBA* gene, the presence of mutations originated from recombination events between the gene and the pseudogene, further complicating the analysis and leading to false-negative results.

In conclusion, based on the data analysis and the validation of the findings it seems clear that NGS pipelines for *GBA* must be designed carefully. The presence of pseudogenes significantly affect data output and must be taken into account in the data analysis. Therefore, specific procedures for the library preparation and data analysis should be adopted. The pipeline described here represents a useful approach to deal with *GBA* sequencing using NGS technology.

Supplemental Data

Supplemental material for this article can be found at *http://dx.doi.org/10.1016/j.jmoldx.2017.05.005*.

References

1. Horowitz M, Wilder S, Horowitz Z, Reiner O, Gelbart T, Beutler E: The human glucocerebrosidase gene and pseudogene: structure and evolution. Genomics 1989, 4:87–96
2. Reiner O, Wigderson M, Horowitz M: Structural analysis of the human glucocerebrosidase genes. DNA 1988, 7:107–116
3. Martínez-Arias R, Calafell F, Mateu E, Comas D, Andrés A, Bertranpetit J: Sequence variability of a human pseudogene. Genome Res 2001, 11:1071–1085
4. Hong CM, Ohashi T, Yu XJ, Weiler S, Barranger JA: Sequence of two alleles responsible for Gaucher disease. DNA Cell Biol 1990, 9: 233–241
5. Latham TE, Theophilus BD, Grabowski GA, Smith FI: Heterogeneity of mutations in the acid beta-glucosidase gene of Gaucher disease patients. DNA Cell Biol 1991, 10:15–21

6. Tayebi N, Stubblefield BK, Park JK, Orvisky E, Walker JM, LaMarca ME, Sidransky E: Reciprocal and nonreciprocal recombination at the glucocerebrosidase gene region: implications for complexity in Gaucher disease. Am J Hum Genet 2003, 72:519−534
7. Beutler E, Grabowski G: Gaucher disease. Edited by Scriver CR, Beaudet AL, Valle D, Sly WS. In The metabolic and molecular basis of inherited disease. New York: McGraw-Hill, 2001. pp. 3635−3668
8. Hruska KS, LaMarca ME, Scott CR, Sidransky E: Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). Hum Mutat 2008, 29:567−583
9. Torralba MA, Alfonso P, Pérez-Calvo JI, Cenarro A, Pastores GM, Giraldo P, Civeira F, Pocoví M: High prevalence of the 55-bp deletion (c.1263del55) in exon 9 of the glucocerebrosidase gene causing misdiagnosis (for homozygous N370S (c.1226A > G) mutation) in Spanish Gaucher disease patients. Blood Cells Mol Dis 2002, 29: 35−40
10. Beutler E, Gelbart T, West C: Identification of six new Gaucher disease mutations. Genomics 1993, 15:203−205
11. Amaral O, Pinto E, Fortuna M, Lacerda L, Sá Miranda MC: Type 1 Gaucher disease: identification of N396T and prevalence of glucocerebrosidase mutations in the Portuguese. Hum Mutat 1996, 8: 280−281
12. Sarria AJ, Giraldo P, Perez-Calvo JI, Pocoví M: Detection of three rare (G377S, T134P and 1451delAC), and two novel mutations (G195W and Rec[1263del55;1342G>C]] in Spanish Gaucher disease patients. Hum Mutat 1999, 14:88
13. Hatton CE, Cooper A, Whitehouse C, Wraith JE: Mutation analysis in 46 British and Irish patients with Gaucher's disease. Arch Dis Child 1997, 77:17−22
14. Tayebi N, Reissner KJ, Lau EK, Stubblefield BK, Klineburgess AC, Martin BM, Sidransky E: Genotypic heterogeneity and phenotypic variation among patients with type 2 Gaucher's disease. Pediatr Res 1998, 43:571−578
15. Reissner K, Tayebi N, Stubblefield BK, Koprivica V, Blitzer M, Holleran W, Cowan T, Almashanu S, Maddalena A, Karson EM, Sidransky E: Type 2 Gaucher disease with hydrops fetalis in an Ashkenazi Jewish family resulting from a novel recombinant allele and a rare splice junction mutation in the glucocerebrosidase locus. Mol Genet Metab 1998, 63:281−288
16. Cormand B, Díaz A, Grinberg D, Chabás A, Vilageliu L: A new gene-pseudogene fusion allele due to a recombination in intron 2 of the glucocerebrosidase gene causes Gaucher disease. Blood Cells Mol Dis 2000, 26:409−416
17. Filocamo M, Bonuccelli G, Mazzotti R, Giona F, Gatti R: Identification of a novel recombinant allele in three unrelated Italian Gaucher patients: implications for prognosis and genetic counseling. Blood Cells Mol Dis 2000, 26:307−311
18. Stone DL, Tayebi N, Orvisky E, Stubblefield B, Madike V, Sidransky E: Glucocerebrosidase gene mutations in patients with type 2 Gaucher disease. Hum Mutat 2000, 15:181−188
19. Tayebi N, Park J, Madike V, Sidransky E: Gene rearranagement on 1q21 introducing a duplication of the glucocerebrosidase pseudogene and a metaxin fusion gene. Hum Genet 2000 Oct, 107:400−403
20. Tayebi N, Callahan M, Madike V, Stubblefield BK, Orvisky E, Krasnewich D, Fillano JJ, Sidransky E: Gaucher disease and parkinsonism: a phenotypic and genotypic characterization. Mol Genet Metab 2001, 73:313−321
21. Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, et al: Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. N Engl J Med 2009, 361: 1651−1661
22. Kalinderi K, Bostantjopoulou S, Paisan-Ruiz C, Katsarou Z, Hardy J, Fidani L: Complete screening for glucocerebrosidase mutations in Parkinson's disease patients from Greece. Neurosci Lett 2009, 452: 87−89
23. Neumann J, Bras J, Deas E, O'Sullivan SS, Parkkinen L, Lachmann RH, Li A, Holton J, Guerreiro R, Paudel R, Segarane B, Singleton A, Lees A, Hardy J, Houlden H, Revesz T, Wood NW: Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease. Brain 2009, 132(Pt 7):1783−1794
24. Lesage S, Anheim M, Condroyer C, Pollak P, Durif F, Dupuits C, Viallet F, Lohmann E, Corvol J-C, Honoré A, Rivaud S, Vidailhet M, Dürr A, Brice A: Large-scale screening of the Gaucher's disease–related glucocerebrosidase gene in Europeans with Parkinson's disease. Hum Mol Genet 2011, 20:202−210
25. Ran C, Brodin L, Forsgren L, Westerlund M, Ramezani M, Gellhaar S, Xiang F, Fardell C, Nissbrandt H, Söderkvist P, Puschmann A, Ygland E, Olson L, Willows T, Johansson A, Sydow O, Wirdefeldt K, Galter D, Svenningsson P, Belin AC: Strong association between glucocerebrosidase mutations and Parkinson's disease in Sweden. Neurobiol Aging 2016, 45:212.e5−212.e11
26. Spitz M, Rozenberg R, Pereira Lda V, Reis Barbosa E: Association between Parkinson's disease and glucocerebrosidase mutations in Brazil. Parkinsonism Relat Disord 2008, 14:58−62
27. Alcalay RN, Levy OA, Waters CC, Fahn S, Ford B, Kuo SH, Mazzoni P, Pauciulo MW, Nichols WC, Gan-Or Z, Rouleau GA, Chung WK, Wolf P, Oliva P, Keutzer J, Marder K, Zhang X: Glucocerebrosidase activity in Parkinson's disease with and without GBA mutations. Brain 2015, 138(Pt 9):2648−2658
28. Han F, Grimes DA, Li F, Wang T, Yu Z, Song N, Wu S, Racacho L, Bulman DE: Mutations in the glucocerebrosidase gene are common in patients with Parkinson's disease from Eastern Canada. Int J Neurosci 2016, 126:415−421
29. Lesage S, Condroyer C, Hecham N, Anheim M, Belarbi S, Lohman E, Viallet F, Pollak P, Abada M, Dürr A, Tazir M, Brice A: Mutations in the glucocerebrosidase gene confer a risk for Parkinson disease in North Africa. Neurology 2011, 76:301−303
30. Huang CL, Wu-Chou YH, Lai SC, Chang HC, Yeh TH, Weng YH, Chen RS, Huang YZ, Lu CS: Contribution of glucocerebrosidase mutation in a large cohort of sporadic Parkinson's disease in Taiwan. Eur J Neurol 2011, 18:1227−1232
31. Yu Z, Wang T, Xu J, Wang W, Wang G, Chen C, Zheng L, Pan L, Gong D, Li X, Qu H, Li F, Zhang B, Le W, Han F: Mutations in the glucocerebrosidase gene are responsible for Chinese patients with Parkinson's disease. J Hum Genet 2015, 60:85−90
32. Aharon-Peretz J, Badarny S, Rosenbaum H, Gershoni-Baruch R: Mutations in the glucocerebrosidase gene and Parkinson disease: phenotype-genotype correlation. Neurology 2005, 65:1460−1461
33. Vacic V, Ozelius LJ, Clark LN, Bar-Shira A, Gana-Weisz M, Gurevich T, Gusev A, Kedmi M, Kenny EE, Liu X, Mejia-Santana H, Mirelman A, Raymond D, Saunders-Pullman R, Desnick RJ, Atzmon G, Burns ER, Ostrer H, Hakonarson H, Bergman A, Barzilai N, Darvasi A, Peter I, Guha S, Lencz T, Giladi N, Marder K, Pe'er I, Bressman SB, Orr-Urtreger A: Genome-wide mapping of IBD segments in an Ashkenazi PD cohort identifies associated haplotypes. Hum Mol Genet 2014, 23:4693−4702
34. Lee A, Gilbert RM: Epidemiology of Parkinson disease. Neurol Clin 2016, 34:955−965
35. Bandrés-Ciga S, Mencacci NE, Durán R, Barrero FJ, Escamilla-Sevilla F, Morgan S, Hehir J, Vives F, Hardy J, Pittman AM: Analysis of the genetic variability in Parkinson's disease from Southern Spain. Neurobiol Aging 2016, 37:210.e1−210.e5
36. Benitez BA, Davis AA, Jin SC, Ibanez L, Ortega-Cubero S, Pastor P, Choi J, Cooper B, Perlmutter JS, Cruchaga C: Resequencing analysis of five Mendelian genes and the top genes from genome-wide association studies in Parkinson's disease. Mol Neurodegener 2016, 11:29
37. Gorostidi A, Martí-Massó JF, Bergareche A, Rodríguez-Oroz MC, López de Munain A, Ruiz-Martínez J: Genetic mutation analysis of Parkinson's disease patients using multigene next-generation sequencing panels. Mol Diagn Ther 2016, 20:481−491
38. Spataro N, Roca-Umbert A, Cervera-Carles L, Vallès M, Anglada R, Pagonabarraga J, Pascual-Sedano B, Campolongo A, Kulisevsky J, Casals F, Clarimón J, Bosch E: Detection of genomic rearrangements

from targeted resequencing data in Parkinson's disease patients. Mov Disord 2017, 32:165−169
39. Asselta R, Rimoldi V, Siri C, Cilia R, Guella I, Tesei S, Soldà G, Pezzoli G, Duga S, Goldwurm S: Glucocerebrosidase mutations in primary parkinsonism. Parkinsonism Relat Disord 2014, 20: 1215−1220
40. Török R, Zádori D, Török N, Csility É, Vécsei L, Klivényi P: An assessment of the frequency of mutations in the GBA and VPS35 genes in Hungarian patients with sporadic Parkinson's disease. Neurosci Lett 2016, 610:135−138
41. Pulkes T, Choubtum L, Chitphuk S, Thakkinstian A, Pongpakdee S, Kulkantrakorn K, Hanchaiphiboolkul S, Tiamkao S, Boonkongchuen P: Glucocerebrosidase mutations in Thai patients with Parkinson's disease. Parkinsonism Relat Disord 2014, 20: 986−991
42. Mitsui J, Mizuta I, Toyoda A, Ashida R, Takahashi Y, Goto J, Fukuda Y, Date H, Iwata A, Yamamoto M, Hattori N, Murata M, Toda T, Tsuji S: Mutations for Gaucher disease confer high susceptibility to Parkinson disease. Arch Neurol 2009, 66:571−576
43. Mata IF, Leverenz JB, Weintraub D, Trojanowski JQ, Chen-Plotkin A, Van Deerlin VM, Ritz B, Rausch R, Factor SA, Wood-Siverio C, Quinn JF, Chung KA, Peterson-Hiller AL, Goldman JG, Stebbins GT, Bernard B, Espay AJ, Revilla FJ, Devoto J, Rosenthal LS, Dawson TM, Albert MS, Tsuang D, Huston H, Yearout D, Hu SC, Cholerton BA, Montine TJ, Edwards KL, Zabetian CP: GBA variants are associated with a distinct pattern of cognitive deficits in Parkinson's disease. Mov Disord 2016, 31: 95−102
44. den Dunnen JT, Antonarakis SE: Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 2000, 15:7−12
45. den Dunnen JT, Paalman MH: Standardizing mutation nomenclature: why bother? Hum Mutat 2003, 22:181−182
46. Miocić S, Filocamo M, Dominissini S, Montalvo AL, Vlahovicek K, Deganuto M, Mazzotti R, Cariati R, Bembi B, Pittis MG: Identification and functional characterization of five novel mutant alleles in 58 Italian patients with Gaucher disease type 1. Hum Mutat 2005, 25:100
47. Andrews S: FastQC: a quality control tool for high throughput sequence data. PLoS One 2012, 7:e30619
48. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010, 38:e164
49. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. Nat Biotechnol 2011, 29:24−26
50. Thorvaldsdóttir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013, 14:178−192
51. Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q: Impact of three Illumina library construction methods on GC bias and HLA genotype calling. Hum Immunol 2015, 76:166−175
52. Koprivica V, Stone DL, Park JK, Callahan M, Frisch A, Cohen IJ, Tayebi N, Sidransky E: Analysis and classification of 304 mutant alleles in patients with type 1 and type 3 Gaucher disease. Am J Hum Genet 2000, 66:1777−1786
53. Jeong SY, Kim SJ, Yang JA, Hong JH, Lee SJ, Kim HJ: Identification of a novel recombinant mutation in Korean patients with Gaucher disease using a long-range PCR approach. J Hum Genet 2011, 56:469−471
54. Siebert M, Bock H, Michelin-Tirelli K, Coelho JC, Giugliani R, Saraiva-Pereira ML: Novel mutations in the glucocerebrosidase gene of brazilian patients with Gaucher disease. JIMD Rep 2013, 9:7−16
55. Yoshida S, Kido J, Matsumoto S, Momosaki K, Mitsubuchi H, Shimazu T, Sugawara K, Endo F, Nakamura K: Prenatal diagnosis of Gaucher disease using next-generation sequencing. Pediatr Int 2016, 58:946−949
56. Li J, Dai H, Feng Y, Tang J, Chen S, Tian X, Gorman E, Schmitt ES, Hansen TA, Wang J, Plon SE, Zhang VW, Wong LJ: A comprehensive strategy for accurate mutation detection of the highly homologous PMS2. J Mol Diagn 2015, 17:545−553
57. Judkins T, Leclair B, Bowles K, Gutin N, Trost J, McCulloch J, Bhatnagar S, Murray A, Craft J, Wardell B, Bastian M, Mitchell J, Chen J, Tran T, Williams D, Potter J, Jammulapati S, Perry M, Morris B, Roa B, Timms K: Development and analytical validation of a 25-gene next generation sequencing panel that includes the BRCA1 and BRCA2 genes to assess hereditary cancer risk. BMC Cancer 2015, 2:215−225
58. Macías-Vidal J, Rodríguez-Pascau L, Sánchez-Ollé G, Lluch M, Vilageliu L, Grinberg D, Coll MJ; Spanish NPC Working Group: Molecular analysis of 30 Niemann-Pick type C patients from Spain. Clin Genet 2011, 80:39−49
59. Palhais B, Dembic M, Sabaratnam R, Nielsen KS, Doktor TK, Bruun GH, Andresen BS: The prevalent deep intronic c. 639+919 G>A GLA mutation causes pseudoexon activation and Fabry disease by abolishing the binding of hnRNPA1 and hnRNP A2/B1 to a splicing silencer. Mol Genet Metab 2016, 119:258−269
60. Valencia CA, Rhodenizer D, Bhide S, Chin E, Littlejohn MR, Keong LM, Rutkowski A, Bonnemann C, Hegde M: Assessment of target enrichment platforms using massively parallel sequencing for the mutation detection for congenital muscular dystrophy. J Mol Diagn 2012, 14:233−246