Emma Saarinen

GBA gene – Frequency of possibly pathogenic variants in Finnish population

Helsinki Metropolia University of Applied Sciences

Bachelor of Engineering

Biotechnology and Food Engineering

Bachelor's Thesis

9 September 2016

Metropolia

Author(s) Title Number of Pages Date	Emma Saarinen *GBA* gene – Frequency of possibly pathogenic variants in Finnish population 36 pages + 1 appendix 9 September 2016
Degree	Bachelor of Engineering
Degree programme	Biotechnology and Food Engineering
Specialisation option	Biomedicine
Instructor(s)	Kaisa Kettunen, Ph.D. Senior Reseacher Tiina Soininen, Principal Lecturer (AMK)

The aim of this thesis was to investigate the frequency of possibly pathogenic variants in the *GBA* gene in the Finnish population. The main interest was in variants causing Gaucher disease. Gaucher disease is the most prevalent lysosomal storage disorder worldwide. It is an autosomal, recessively inherited disease. The disease is caused by a deficiency of a lysosomal enzyme called glucocerebrosidase which causes the accumulation of a substrate, glucocerebroside into lymphoid organs resulting in a wide variety of phenotypes and symptoms. There are over 200 mutations in the *GBA* gene known to cause Gaucher disease. *GBA* has also a highly homologous, nearby pseudogene, *GBAP1*. The homology and close physical location of *GBAP1* also causes difficulties when studying *GBA* as the pseudogene gets easily amplified with the functional gene.

The thesis was carried out in the Institute for Molecular Medicine Finland (FIMM) and conducted as a part of a project of setting up a next-generation sequencing (NGS) panel for candidate genes behind various cytopenias. The basis of the study was existing NGS exome data from three different clinical projects, also conducted at FIMM. The NGS data was analyzed regarding the variants found in the *GBA* gene, the interest being mainly on known pathogenic variants and also possible novel variants. The findings were then validated by use of Sanger sequencing to find out if the variants were actually located in the functional *GBA* gene or its pseudogene, *GBAP1*. In addition, Sequencing Initiative Suomi (SiSu) project data was analyzed for known pathogenic variants.

The results from the validations showed that the presence of the pseudogene affects the NGS exome data and can cause false results when analyzing variants in the *GBA* gene if the analysis is made merely based on NGS data. Therefore, *GBA* specific amplification and sequencing is needed in order to verify the true variants in *GBA*.

Keywords	*GBA* gene, *GBA* pseudogene, *GBAP1*, Gaucher disease, NGS

Tekijä(t) Otsikko Sivumäärä Aika	Emma Saarinen *GBA*-geeni – Mahdollisten patogeenisten varianttien yleisyys suomalaisessa väestössä 36 sivua + 1 liite 9.9.2016
Tutkinto	Insinööri (AMK)
Koulutusohjelma	Bio- ja elintarviketekniikka
Suuntautumisvaihtoehto	Biolääketeknologia
Ohjaajat(t)	Kaisa Kettunen, vanhempi tutkija Tiina Soininen, lehtori (AMK)

Työn tarkoituksena oli tutkia *GBA*-geenin mahdollisten patogeenisten varianttien yleisyyttä suomalaisessa väestössä. Pääkiinnostuksen kohteena oli Gaucherin tautia aiheuttavat variantit. Gaucherin tauti on yleisin lysosomaalinen kertymäsairaus maailmanlaajuisesti. Tauti periytyy autosomeissa peittyvästi. Gaucherin taudin aiheuttaa lysosomaalisen entsyymin, glukoserebrosidaasin puutos, joka johtaa substraatin, glukoserebrosidin kerääntymiseen imukudoksen elimiin aiheuttaen laajan kirjon erilaisia oireita. *GBA*-geenistä on tiedossa yli 200 mutaatiota, joiden tiedetään aiheuttavan Gaucherin tautia. *GBA*-geenin lähistöllä sijaitsee erittäin homologinen pseudogeeni, *GBAP1*. Pseudogeenin samankaltaisuus sekä läheinen sijainti aiheuttavat ongelmia tutkittaessa *GBA*-geeniä, koska pseudogeeni monistuu helposti sen yhteydessä.

Tämä opinnäytetyö suoritettiin Suomen molekyylilääketieteen instituutissa (FIMM) ja tehtiin osana projektia, jonka tarkoituksena oli pystyttää uuden sukupolven sekvensointipaneeli (next-generation sequencing, NGS) erilaisia sytopenioita aiheuttaville geeneille. Työn lähtömateriaalina käytettiin olemassa olevaa NGS-eksomidataa kolmesta eri kliinisestä projektista, jotka on tuotettu FIMM:ssä. NGS-data tutkittiin *GBA*-geenin varianttien suhteen, joista kiinnostuksenkohteena olivat jo tunnetut patogeeniset sekä mahdolliset uudet variantit. Löydökset validoitiin Sanger-sekvensointimenetelmällä, jotta saatiin selville, sijaitsivatko variantit toimivassa *GBA*-geenissä vai sen pseudogeenissä, *GBAP1*:ssa. Lisäksi Sequencing Initiative Suomi (SiSu) -projektin data analysoitiin tunnettujen patogeenisten varianttien osalta.

Validoinnin tuloksista huomattiin, että pseudogeenin läsnäolo vaikuttaa NGS-exomidataan ja voi aiheuttaa vääriä tuloksia analysoitaessa *GBA*-geeniä, jos analyysi tehdään ainoastaan NGS-datan perusteella. Tämän vuoksi *GBA*-geenin varianttien varmistamiseen on käytettävä GBA-spesifistä monistusta sekä sekvensointia.

Avainsanat	*GBA*-geeni, *GBA* pseudogeeni, *GBAP1*, Gaucherin tauti, NGS

Contents

Abbreviations

1 Introduction 1

2 *GBA* gene and its pseudogene *GBAP1* 2

3 Gaucher disease 3

3.1 Clinical types of Gaucher disease 4
3.1.1 Type 1 GD 5
3.1.2 Type 2 GD 5
3.1.3 Type 3 GD 6
3.2 Diagnosis and treatment 6

4 NGS exome workflow (Illumina) 8

4.1 Library preparation 9
4.2 Cluster generation 11
4.3 Sequencing and Data analysis 12

5 Materials and Methods 13

5.1 NGS Exome Data 13
5.2 Data analysis 14
5.3 Validations 15
5.3.1 Touchdown PCR 15
5.3.2 Nested Sanger-sequencing 17

6 Results and conclusions 18

6.1 Variant selection for validation 18
6.2 SiSu project 19
6.3 Long-range Touchdown PCR 20
6.4 Nested Sanger-sequencing results 24

7 Discussion 31

References 33

Appendix 1. Creating annotation files and using grep-command in Putty

Abbreviations

LSD	Lysosomal storage disorder
GBA	Glucosylceramidase beta
GBAP1	Glucosylceramidase beta pseudogene 1
NGS	Next-Generation Sequencing
GD	Gaucher disease
SNP	Single-nucleotide polymorphism
ERT	Enzyme replacement therapy
WES	Whole exome sequencing
Exome	The protein-coding regions (exons) of a genome as a whole
dNTP	Deoxynucleoside Triphosphate, stands for all four nucleotides, dATP, dTTP, dGTP, and dCTP
HGMD	Human Genome Mutation Database
ExAc	Exome Aggregation Consortium
dbSNP	The Single Nucleotide Polymorphism Database
TD PCR	Touchdown PCR
SNV	Single-nucleotide variation
IGV	Integrative Genome Viewer
DM	Disease-causing mutation
FP	Functional polymorphism

1 Introduction

Lysosomal storage disorders (LSD) are a heterogeneous group of rare diseases caused mainly by a lack of one of the hydrolases. Reduced amount of a specific enzyme results in accumulation of the corresponding substrate into lysosomes and causes a vast repertory of various clinical phenotypes. The severity of the disorders vary significantly depending on the particular substrate in question.

Gaucher disease is the most common lysosomal storage disorder worldwide, and it occurs in between 1:50,000 and 1:100,000 live births in general population. It is estimated that GD affects roughly 30,000-100,000 people worldwide. In some populations the disease occurs more frequently, for example Ashkenazi Jew population where the frequency is as high as 1 in 500 live births. The prevalence of the disease varies between populations but seems to be rarer in European origins. In Gaucher disease, the deficiency of glucocerebrosidase causes the accumulation of glucocerebroside into lymphoid organs. The enzyme deficiency is caused by mutations in *GBA* gene. Gaucher disease is a type of sphingolipidosis which is a subgroup of LSDs. [1; 7; 11.]

This thesis was carried out in the Institute for Molecular Medicine Finland (FIMM). The topic of the thesis was received from FIMM Genomics department. The aim of the thesis was to study the prevalence of possibly disease causing mutations in *GBA* gene in the Finnish population, the interest being especially in variants causing Gaucher disease. The study was based on existing NGS exome data from three different projects, also carried out at FIMM. The aim was to search for known pathogenic and unknown, novel mutations from the existing datasets. The most interesting variants with were validated with the use of PCR and Sanger-sequencing to separate the functional GBA from its pseudogene and find out if the variants were in fact located in the functional GBA and not in the pseudogene. The study was conducted as a part of a project of setting up a next-generation sequencing (NGS) panel for candidate genes behind various cytopenias.

2 *GBA* gene and its pseudogene *GBAP1*

GBA, glucosylceramidase beta, is a protein coding gene located in chromosome 1. *GBA*s cytogenetic location is 1q22 and its genomic coordinates are chr1:155,234,447-155,244,861 (genome assembly GRCh38/hg38). The *GBA* gene has 12 exons and 11 introns [48]. *GBA* and its pseudogene was first sequenced in 1989. [2; 3.]

*GBA*s function is to produce an enzyme called glucocerebrosidase (EC 3.2.1.45), also known as glucosylceramidase or D-Glucosyl-N-acylsphingosine glucosyl hydrolase. Glucocerebrosidase is a lysosomal enzyme that plays a factor in glycolipid metabolism by hydrolyzing glucocerebroside. Glucocerebroside is a sphingolipid with a glucose head group. Glucocerebrosidase enzyme degrades this glycolipid into N-acylsphingosine (ceramide) and D-glucose with the presence of H_2O (Figure 1). Other chemical names for glucocerebroside are glucosylceramide and D-glucosyl-N-acylsphingosine. [2; 3; 4; 5; 6.]

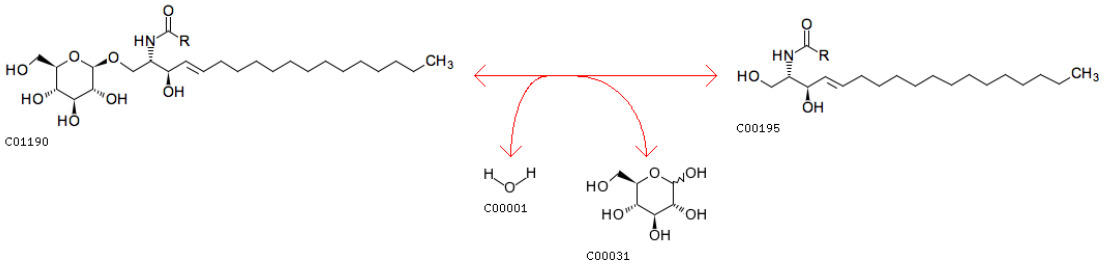

Figure 1. Hydrolysis of D-glucosyl-N-acylsphingosine (glucocerebroside) into D-glucose and N-acylsphingosine. (http://www.genome.jp/dbget-bin/www_bget?rn:R01498)

Diseases associated with the *GBA* gene are Gaucher disease and Parkinson's disease. These diseases are mainly caused by point mutations in the coding area of the gene, but also insertions, deletions and splice site mutations occur in the gene. There have also been studies about correlations between GD and Parkinson's disease. [7.]

GBA has a highly homologous pseudogene, *GBAP1*, approximately 16 kb downstream of the functional gene. *GBAP1* is a product of gene duplication event that has retained the exon-intron -structure of the functional gene. *GBAP1* has a 96% sequence homology to the exonic area of the functional *GBA* gene and over 98% homology of the regions between intron 8 and 3'-UTR (3'-untranslated region). *GBAP1* is approximately 2kb

shorter than *GBA*, missing some *Alu*-insertions, i.e. repetitive DNA sequences about the length of 300bp. [8.]

The high homology between *GBA* and *GBAP1* enables gene conversion events between the two genes. These conversions and unequal crossing over produce recombinant alleles, most of them located at the 3'-end after exon 8 of the sequence. Recombination alleles usually cause severe forms of GD and has been shown to be lethal in homozygosity in most of the cases. The most common recombination allele is *RecNcil* which has a crossover junction area from intron 9 to exon 10 and as a result a segment of *GBAP1* is incorporated into the functional *GBA*. The *RecNcil* allele includes three missense mutations L444P, A456P, and V460V (nucleotide changes [1448T>C;1484G>C;1496G>C]).

The short distance and similarity between *GBA* and *GBAP1* genes causes difficulties when studying *GBA* as the pseudogene is easily amplified with the functional gene when using PCR. Hence, *GBAP1* has to be first separated from the functional gene to prevent false results. [3; 8; 9; 10.]

3 Gaucher disease

Gaucher disease (GD) is a rare, autosomal recessively inherited metabolic disorder in which glycolipid, glucocerebroside is not degraded properly. The disorder is caused by the deficiency of an enzyme called glucocerebrosidase that is encoded by *GBA* gene. The deficiency of the enzyme causes the accumulation of glucocerebroside to lysosomes, mainly in macrophages and monocytes, affecting primary and secondary lymphoid organs such as spleen, liver, bone marrow and brain. The affected macrophages, so called Gaucher cells, cause chronic inflammatory response by activating a number of cytokines. [11.]

There are over 200 mutations in *GBA* gene that are known to cause GD. Most of them are missense point mutations, also called SNPs (single-base polymorphism) that change one amino acid to another and alter the structure of the enzyme resulting in loss of function. Null mutations cause a total lack of enzyme production, which results in more severe forms of GD. In most cases, GD is caused by homozygous variations in the *GBA*. Compound heterozygous variants where the carrier has two different pathogenic mutations

are known to cause GD as well and symptoms can vary significantly depending on the two variants. The disease is not gender specific and affects both males and females. [7; 12; 13.]

The most common pathogenic variant known to cause type 1 GD in homozygous and compound heterozygous form, is variant N370S (Figure 2). The homozygous L444P variant has been shown to cause neurological complications that manifest in GD types 2 and 3.

Variants [1]	% of Affected Individuals [2, 3]
N370S/N370S	29%
N370S/?	20%
N370S/L444P	16%
N370S/84GG	12%
L444P/L444P [4]	6%
L444P/?	3%
N370S/IVS2+1	3%

Figure 2. Four most common pathogenic variants and their distribution among affected individuals [7].

Variants 84GG and IVS2+1 are considered lethal as homozygous because none have been found in live born patients. Both variants as compound heterozygous are associated to type 3 GD. *RecNcil* recombinant allele that includes three missense mutations (L444P, A456P, and V460V) has been found to cause perinatal-lethal form of GD. [7; 9.]

3.1 Clinical types of Gaucher disease

GD has been divided into three common clinical types based on the symptoms and onset of the disease. The three types vary significantly from the symptoms that they cause. There are also two additional subtypes of the disease, perinatal-lethal and cardiovascular form that can be categorized as distinct forms of the disease.

3.1.1 Type 1 GD

Type 1 GD is called non-neuropathic or non-cerebral form of GD and in most of the cases does not affect the primary central nervous system. Type 1 is the most common type of GD and more than 90% of patients with the disease have type 1 GD. The course of the disease and symptoms may differ substantially between patients, from severe complications to asymptomatic. Type 1 is also known as the adult form of GD as the symptoms usually occur at early adulthood. Symptoms include anemia, bruising due to low amount of platelets (thrombocytopenia), decreased amount of white blood cells (leukocytopenia) and enlarged spleen and liver (hepatosplenomegaly) (Figure 3).

Figure 3. Severe hepatosplenomegaly in a patient with type 1 GD [15].

Also bone abnormalities such as osteopenia or osteoporosis and bone degeneration (avascular necrosis) due to loss of blood supply to the bone, are common among patients with type 1 GD. Patients usually have almost normal life expectancy depending on the severity of the symptoms. [7; 14.]

3.1.2 Type 2 GD

Type 2 of GD is known as acute neuropathic or infantile form of the disease. First symptoms usually occur in the first 6 months of life and include neurological complications due to the accumulation of glucocerebroside in the brain. Other symptoms are enlarged

spleen (splenomegaly) and/or liver (hepatomegaly), different types of cytopenias, involuntary spasms and seizures and respiratory distress. Some patients with type 2 GD also suffer from pulmonary lung disease and difficulty swallowing which results in reduced growth velocity and overall development. Patients with type 2 of the disease have a significantly shorter life expectancy than normal and patients usually die before 4 years of age.

3.1.3 Type 3 GD

Type 3 is called chronic or subacute neuropathic form of GD. It is also known as juvenile form as onset of the disease is usually during juvenile period. Type 3 is milder form of type 2 GD and the course of the disease is slower. Neuropathic symptoms occur also in type 3 GD. Other symptoms are anemia, bone abnormalities such as osteoporosis, seizures and difficulties with eye movement (horizontal gaze palsy). Also a significant part of patients with GD type 3 develop pulmonary lung disease. Depending on the course of the disease, patients usually live well into adulthood but may need assistance with daily life as the disease progresses.

The two other subtypes of GD are perinatal-lethal form and cardiovascular form. Perinatal-lethal type is considered a specific form of type 2 GD and differs from it with a few specific symptoms. These include non-immune hydrops fetalis in which fluid accumulates into different parts of the fetus and causes swelling due to serious anemia. Also skin abnormalities and distinct facial features are common in perinatal-lethal form. Perinatal-lethal form is a severe form of GD where the affected usually die while still in womb or within first weeks of life. This specific form of GD is caused by homozygous null mutations or recombinant alleles, such as *RecNcil* [16].

Cardiovascular form causes calcification of aortic and mitral valves. Also mild hepatomegaly may occur. Cardiovascular GD is a distinct form of type 3 GD. [7; 12; 13; 15.]

3.2 Diagnosis and treatment

The prognosis of Gaucher disease depends significantly on the course of the disease and severity of the symptoms. The diagnosis of GD cannot be made solely based on

clinical description because of overlapping symptoms in the different types of the disease. There are number of tests that are used to verify the diagnosis. The first is measuring glucocerebrosidase enzyme levels in leukocytes or fibroblasts. Reduced enzyme levels can indicate Gaucher disease, but other tests are used to confirm the diagnosis because enzyme levels can vary considerably between healthy patients as well. Complete blood count (CBC) shows the low levels of blood cells in anemia, thrombocytopenia and leukopenia. Imaging with tomography or ultrasonography is used to reveal hepatosplenomegaly.

Bone marrow biopsy is used to review the changes in macrophages. Gaucher cells (Figure 4) have distinct features such as abnormal placement of nucleus and cytoplasm that resembles “wrinkled tissue paper”. Genetic testing of patients and relatives is used to determine the specific mutations in the *GBA* gene.

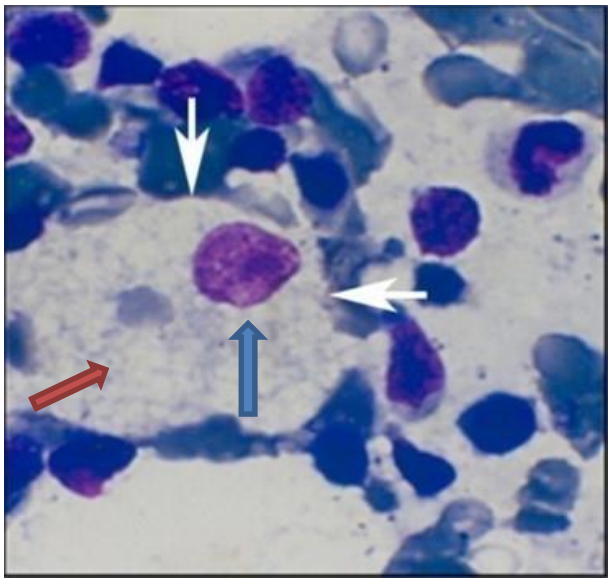

Figure 4. White arrows pointing the Gaucher cell. Blue arrow showing the placement of nucleus and red arrow pointing at wrinkled cytoplasm [15].

Treating patients with GD is often difficult due to the variability of symptoms between patients. Enzyme replacement therapy (ERT) is the most common treatment of GD and has been shown to be effective especially to patients with type 1 GD by reversing symptoms like hepatosplenomegaly and bringing blood cell counts to normal level. Also skeletal problems can be treated with ERT, but the response time has been shown to be longer. ERT is well tolerated and only a small fraction of patients produce antibodies against the enzymes given. Gaucher disease was the first LSD to be treated with ERT.

Substrate reduction therapy (SRT) can be used in addition to ERT that can further enhance the effect of ERT. In SRT the formation of substrate, glucocerebroside is reduced with inhibitors. On rare occasions, bone marrow transplantation (BMT) may be needed if ERT is not an option. Due to the morbidity and mortality of BMT, it is not used widely anymore for treating GD.

A number of patients with GD may also need surgical care. Splenectomy was once a common treatment for splenomegaly in GD, but due to ERTs success, it is now performed only in specific cases where the spleen is severely enlarged or in a risk of rapture. Patients with serious bone disease can also need hip replacements and other skeletal procedures.

The problem with treating GD patients with types 2 and 3 of the disease is that neurological complications cannot be reversed with ERT. Also, most drugs are unable to cross the blood-brain barrier (BBB) which complicates drug distribution to the brain. [6; 14; 16.]

There are several treatment methods under investigation for GD and other LSDs. One of them is pharmacological “chaperones”, chemical compounds that bind to the active site of the enzyme and stabilize the structure. As a result, more of functional enzyme is transferred to lysosomes. One of the advantages of chaperone therapy is that it enables chemical distribution to the brain by penetrating the BBB due to the small size of the compound. For that reason, chemical chaperones might be the future treatment for types 2 and 3 GD. [18; 36.]

4 NGS exome workflow (Illumina)

Genetic testing is widely used in diagnosis of genetic disorders, such as Gaucher disease. The newest method used for genetic testing is Next-generation sequencing (NGS). NGS is a high through-put sequencing system that enables fast and accurate sequencing on a larger scale than Sanger-sequencing. NGS offers many applications, for example, WGS (whole genome sequencing), exome sequencing, targeted sequencing panels and RNA-sequencing.

Targeted sequencing can be used when the interest is on a specific set of regions or genes. One of the NGS applications is Whole Exome Sequencing (WES) where the targets are the protein-coding regions of the genome. The exome covers approximately 1% of the whole human genome, but this portion of the genome is responsible for all the protein production in our system. The majority of pathogenic mutations are located in exons, which makes WES a popular application for NGS. The exome is targeted using a target enrichment method where the exons are captured with specific probes and the target is then amplified and sequenced. WES is more cost-effective compared to WGS if the particular targets of interest are known. With WES the sequencing depth can also be adjusted to the level where even the rarest variants can be detected. [22.]

Illumina Inc. (San Diego, California, United States) is the most widely known and applied NGS platform. The sample library preparation workflow of Illumina's WES and targeted gene panels follow the same guidelines and steps as any NGS application by Illumina. The following overview of targeted exome sequencing will cover library preparation and sequencing of paired-end libraries using Illumina's platform. Sequence capture is carried out by hybridization with oligonucleotide probes.

4.1 Library preparation

Sample library preparation for NGS consist of fragmentation, end-repair, A-tailing and ligation. In fragmentation, the genomic DNA (gDNA) sample is degraded into smaller, approximately 250-350 bp long fragments. Fragmentation is done with acoustic shearing or focused sonication. End-repair enzyme T4 DNA polymerase, which has a 5'->3' polymerase and 3'->5' exonuclease activity, then removes single stranded overhangs from the double stranded DNA fragments to form blunt ends and simultaneously phosphorylating the 5'-ends of the fragments.

In A-tailing, Klenow fragment, N-terminal domain of DNA polymerase 1 from *E. coli*, adds dAMP (A-base) to the 3'-end of the fragment [28]. This enables ligation of Y-adapters with complementary dT-overhangs at the ends of the fragment. In ligation, T4 DNA ligase enzyme catalyzes a phosphodiester bond between 5'-phosphate and 3'-hydroxyl ends of the fragments and adapters. Adapters contain index tags that act as identification sequences for each sample. In paired-end sequencing the adapters are ligated to both ends of the fragment (Figure 5). Special tails, called P5 and P7 are also added to each sample fragments. These tails attach to the complementary oligos on the surface of a

flowcell during cluster generation. After ligation the sample library is then amplified. After every step of the sample library preparation (fragmentation, end-repair, A-tailing and ligation), the sample is purified, for example, with magnetic beads, to dispose of impurities such as remaining enzymes.

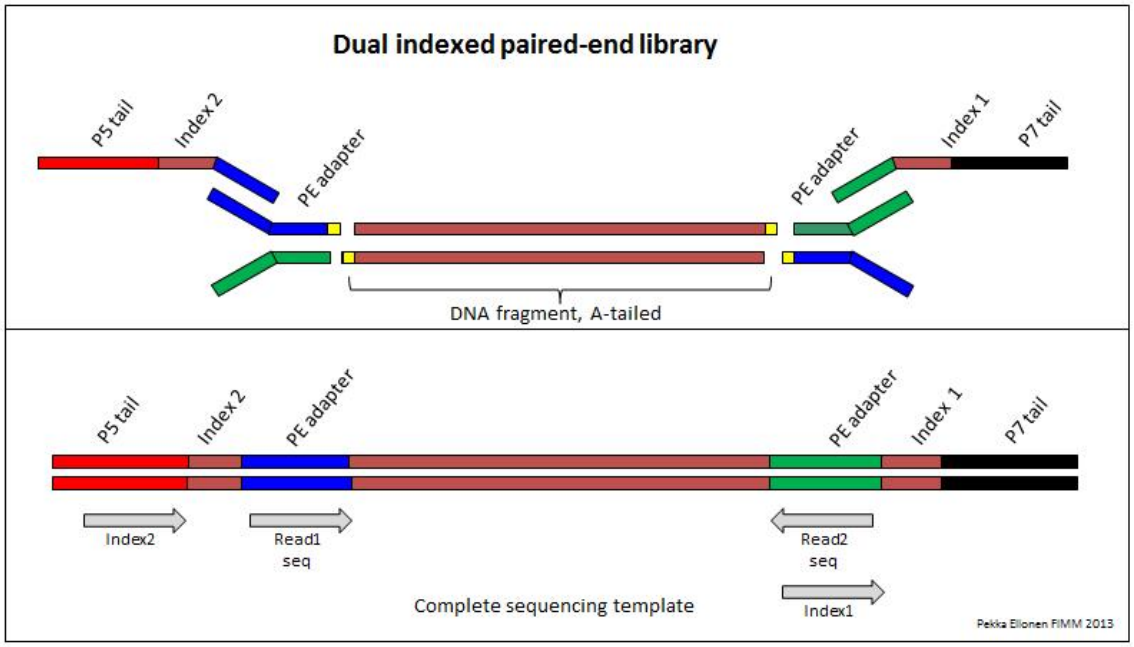

Figure 5. Complete indexed sample library [20].

An alternative method for sample library preparation is tagmentation by Nextera (Nextera DNA Sample Prep Kit, Illumina) where transposase enzyme fragments the gDNA and adds adapters to the ends of the fragment, simultaneously in one tube. Prepared library is then amplified and indexed using PCR. Nextera takes considerably less time compared to traditional library preparation. [19; 22; 23; 24.]

As mentioned above, in WES the exons of a genome are captured with specific probes that are biotinylated. This is called target enrichment and is carried out by hybridizing the probes to the target. The probes for WES are included in kits intended for exome sequencing. Depending on the manufacturer of the kit, the prior ligated adapters are usually blocked with complementary oligos to prevent adapter-dimer formation. Also known repetitive elements, such as *Alu*-sequences, are blocked using COT human DNA, for more efficient and accurate results. After probe hybridization, the target is captured with magnetic streptavidin beads that attach to biotin on the probes. By use of the magnetic beads, the target can be separated from the off-target fragments and the excessive products are washed away.

4.2 Cluster generation

Prior to sequencing, the previously captured and enriched targets are attached to the surface of a flowcell that contains oligos complementary to P5-/P7-tails that act as sequencing primers. The sequence template attaches to the P7-tail on the flowcell. Complementary sequence is then synthesized after which the original template is cleaved off, leaving only the newly formed complementary sequence attached to the flowcell. The sequence is amplified with bridge amplification where the open end of the template curves over and is attached to P5-oligo on the flowcell (Figure 6). Complementary strand is created during amplification and after denaturation the two templates can be cleaved from one end of the flowcell, creating two single stranded copies of the template. These two sequences then curve over again and bind to the complementary oligos creating more copies of the template. After amplification, the complementary reverse strands attached to the P5-oligo are denatured and removed. As a result, only the forward templates attached to P7-oligos are left on the flowcell. [26.]

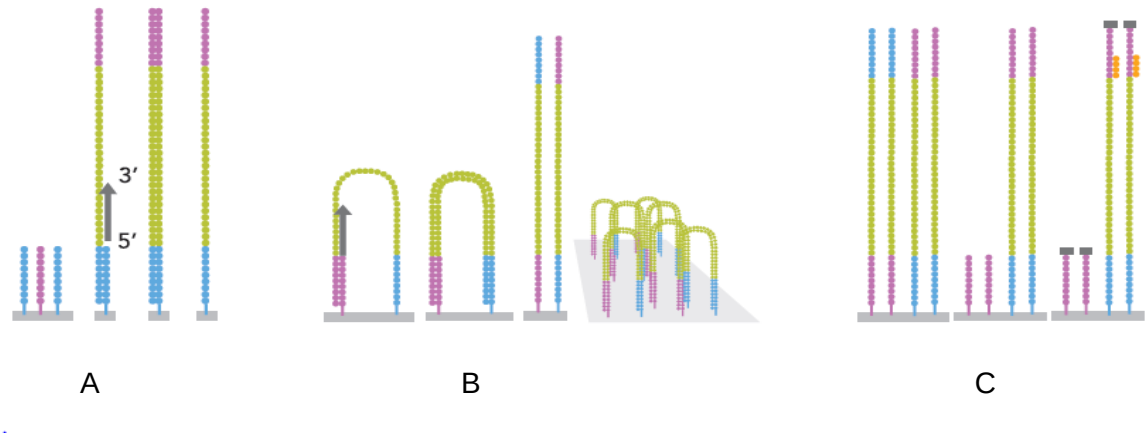

Figure 6. Cluster generation by bridge amplification. A) The template is attached to the P7-tail on the surface of the flowcell and complementary sequence is synthesized and original template removed. B) The complementary strand curves over and attaches to P5-tail and complementary sequence is created. The two strands are denatured leaving single stranded copies of the template on the flowcell. C) Complementary reverse strands are cleaved and only forward templates are left on the flowcell attached to P7-tails. The P5-oligos are blocked and sequencing primers are attached to the adapters [26].

The formed groups of template clones are called clusters. One cluster consists of roughly 1,000 copies of the fragment. Clusters are detected through imaging and can be sorted out based on their indexes. Cluster generation is done either within the sequencer (e.g. Illumina's HiSeq1500) or in a separate instrument, cBOT (Illumina) that primes the flowcell and amplifies the sequence templates.

4.3 Sequencing and Data analysis

NGS technique utilizes sequencing by synthesis-method (SBS), where fluorescently labeled dNTPs (A, T, G, C) are synthesized on to the sequence template one by one. The sequencing templates and free oligonucleotides are blocked from 3'-end to prevent unspecific binding. Sequencing primers are hybridized to the adapters attached to the templates. Nucleotides are reversible terminator-bound dNTPs which means that their 3'-OH groups have been blocked. During every sequencing cycle, one dNTP is hybridized independently to the template after which the fluorescent emission of the dNTP is detected by laser. The OH-block is then cleaved and another dNTP can be hybridized to the sequence. This technique greatly reduces sequencing bias and unspecific bonding of dNTPs. In paired-end sequencing the sequence template is read from both ends. Before the second sequence read, the cluster is re-synthesized. The template bridges over and is attached to P5-oligo. A sequence is synthesized and denatured, as before, after which the original template is cleaved and washed away. All of the synthesized templates are now attached to P5-oligo. This enables the second sequence read from the other end of the template. One read is produced per each cluster on the flowcell, paired-end sequencing produces two reads.

Sequencing depth means the average amount of reads produced per template, for example, 30x generates an average of 30 reads per template. The depth can be adjusted to accommodate different objectives. The deeper the sequencing depth, the better the specificity and sensitivity to detect variants [27].

In data analysis, the different reads are aligned to a reference genome. From the alignment, different variations, such as SNPs and indels (insertion-deletion), can be identified. Alignment and data analysis is carried out by different data analysis pipelines. [22; 25.]

Following the data analysis, the possible findings from the data need to be validated with a different method to verify the results due to the potential misleading analysis of the pipelines. A method commonly used for validations is PCR amplification followed by Sanger-sequencing.

5 Materials and Methods

5.1 NGS Exome Data

The data used in this study was produced at the Institute for Molecular Medicine Finland (FIMM). The data consisted of existing next-generation sequencing exome data from three different projects that had been sequenced at FIMM Sequencing Unit. The three projects were FHRB, ClinPIDD and Myeloma. All three projects are investigating hematological diseases or related deficiencies.

FHRB (Finnish Hematology Registry and BioBank) include a collection of samples from different fields of hematology and from patients that express similar symptoms as in Gaucher disease. ClinPIDD project is comprised of patients with primary immunodeficiencies, PIDDs, with possible overlapping symptoms with hematological disorders. The PIDDs consists of a various group of disorders that affect the immune system and usually manifest in the first year of life [29]. The myeloma project consist of patients with suspicion for myeloma which is a cancer of the plasma cells. As the cancer affects the blood cells it can manifest similar symptoms to GD, such as anemia and bone disease. The sample amount of the study was >450 samples in total, the majority consisting of ClinPIDD patients.

In addition to the three projects, data from the SiSu project (Sequencing Initiative Suomi) was included in the study. SiSu is an international collaboration project with an aim build databases and tools for research utilizing the Finnish inheritance. The project includes whole genome and exome sequencing data from >10000 samples that have been sequenced for different disease-specific and Finnish population genetic studies. The SiSu project is coordinated in FIMM [47].

NGS exome workflow was performed following Roche Nimblegen SeqCap EZ library SR capture technology using Illumina's HiSeq1500 and HiSeq2500 sequencers at FIMM Sequencing Unit. SeqCap EZ MedExome Enrichment kit (Roche Holding AG, Basel, Switzerland) or Agilent SureSelect Clinical Research Exome kit (Santa Clara, California, United States) was used to capture the targeted exons [20].

5.2 Data analysis

Annotation files were created from the NGS exome data by use of various scripts in Unix environment using Putty [41]. Gene annotation is a sequence information file that enables the analysis of raw sequencing files. In creating annotation files, the sequencing data is aligned to reference genome which gives specific locations for each of the variants found. Based on the location of the variant and information from various databases, the gene annotation gives information on the variants functionality, whether it is a point mutation, deletion, insertion or splice site mutation and gives predictions of a variants pathogenicity and possible consequences for protein structure. The annotation file includes all variants found in genes that are included in the target. [30.]

The regions and variants of interest, in this case the interest being in the *GBA* gene, were extracted from the annotation files created. The extraction was done using grep-command in Putty. Grep-command enables the extraction of specific lines, in this case lines with gene name *GBA*, from large annotation files. Scripts used in creating annotation files and using grep-command can be seen in Appendix 1.

From the extracted annotation files, now consisting of only *GBA* gene variant information, all the variants and their prediction of pathogenicity was first checked from HGMD (Human Genome Mutation Database) based on their rs-number and start location in the genome [31]. Genomic start location was used as well, because a portion of the variants lack rs-numbers. The interest was particularly in exonic, nonsynonomous variants. The possibly pathogenic variants and other interesting variants were then researched from other databases, such as ClinVar, ExAc (Exome Aggregation Consortium) and dbSNP [32; 33; 34]. The unknown variants (not found in HGMD), especially exonic and splicing variants without rs-number, were also searched from databases mentioned above to find out if they were novel and possibly disease causing. The data was also analyzed in respect of both homozygous and compound heterozygous variants, as GD being a recessively inherited disorder can be caused either by homozygous or compound heterozygous mutations.

The SiSu project data was searched for known pathogenic mutations from the unrestricted data from the internet site. All of the listed variants in the *GBA* gene were first extracted from the internet site and then searched from HGMD on the basis of their rs-number or start location. The variants belonging to disease-causing mutation variant

class DM (according to HGMD) were further checked from ClinVar and ExAc to get more information on their prevalence and clinical significance [32; 33]. Original literature references from HGMD were also evaluated in order to determine the pathogenicity of the found variants.

5.3 Validations

The purpose of the validation was to find out, by use of another method, if the variants found from the patients were in fact located in the functional *GBA* gene or in its pseudogene, *GBAP1*. The NGS data analysis pipelines are unable to differentiate in which of the highly homologous genes the variants are located. For this reason, the validations have to be done manually by use of PCR and Sanger-sequencing.

5.3.1 Touchdown PCR

The primers used in the validations were designed to unique locations in *GBA* gene using Primer3 [49]. The *GBA* gene was amplified in two parts because the length of the gene did not allow amplification as one fragment (Table 1). GBA_1-segment covers exons 6-12 (3'-end) and GBA_1 exons 1-7 (5'-end). Long-range PCR and *GBA* specific primers were used in order to separate the functional *GBA* gene from its pseudogene. Phusion High-Fidelity protocol and Master Mix (New England BioLabs Inc, (NEB), Ipswich, Massachusetts, United States) was used in the amplifications according to the manufacturer's instructions [37].

Table 1. Primers for *GBA* PCR amplification in two parts.

	Primer-F	Primer-R	Product size (bp)	Coordinates
GBA_1	CACAGACCCACACAGAGACT	TTCACCGCCTATCATCTGCT	5958	chr1:155202693-155208650
GBA_2	CTTGAGTGACCCCTTCCCAT	GTGTGCTGGCGGGAAAAC	6743	chr1:155207975-155214714

Touchdown PCR (TD PCR) program was previously optimized with FIMM internal control sample, FimmX. In TD PCR, the annealing temperature is decreased progressively at each PCR cycle. This reduces primer-dimer production and optimizes the target template amplification as unspecific products form usually in lower temperatures [38]. Extension time was extended to 4 minutes due to the long product sizes of the two amplicons

(Phusion protocol: extension time 15-30s/kb). Touchdown temperature range was determined at 72 °C to 60 °C, decreasing 1 °C at each cycle (Table 2). Negative control, or no template control (NTC) and FimmX samples were used to see if there are any contaminations or other errors in any of the reactions. PCR reactions were done using GS4 multi block thermal cycler (G-Storm, Somerset, United Kingdom).

Table 2. PCR reaction mix (Phusion) and Touchdown PCR program.

Reagent	**1x**
2x Phusion MM (1x)	10 µl
Primer F (0,5 µM)	2 µl
Primer R (0,5 µM)	2 µl
DNA (20 ng/µl)	1 µl
H2O	5 µl
Total volume	**20 µl**

TD-PCR program	
Denaturation	1. 98°C, 30s
Annealing 11 cycles	2. 98°C, 10s
	3. 72-60°C, 30s
	4. 72°C, 4min
Annealing 17 cycles	1. 98°C, 10s
	2. 60°C, 30s
	3. 72°C, 4min
Elongation	4. 72°C, 10min
	5. 10°C, ∞

The results from the PCR were verified with a LabChip GX –instrument (PerkinElmer, Waltham, Massachusetts, United States) using genomic DNA -kit to see if the formed product was the right size and if there were substantial amounts of primer-dimer that needed to be purified prior to sequencing. The reaction was done by following PerkinElmer's Genomic DNA LabChip GX User Guide [39].

The PCR product was purified with NucleoMag magnetic beads (NucleoMag NGS Clean-up and size select, Macherey-Nagel, Dueren, Germany) following the manufacturer's protocol. The ratio between PCR-product and magnetic beads was 1:1 of the volume. The sample was eluted to 20 µl of PCR grade water.

All of the PCR products were concentrated from 20 µl to 10 µl to improve the quality of the Sanger-sequencing results. The concentration was performed using DNA Clean & Concentrator™-5 –kit by following the protocol (Zymo Research, Irvine, California, United States).

5.3.2 Nested Sanger-sequencing

In nested Sanger-sequencing the previously amplified PCR product was used as a template. The specific targeted variants were then covered by sequencing with nearby primers. The primers were designed using Primer3 and cover all the 12 exons of *GBA* gene. The primer locations used in nested Sanger-sequencing were visualized using Integrative Genome Viewer (IGV) that images the variants and primers to the target gene with the coordinates imported to the software [40]. The nearest primers to the variants were chosen for nested sequencing.

Sanger-sequencing was done using the BigDye Terminator v3.1 Cycle Sequencing Protocol and Kit (Thermo Fisher Scientific, Waltham, Massachusetts, United States). The dilution factor for BigDye was 8 (1:8 mix) and was determined on the basis of previous testing. The sequencing reaction mix can be seen in Table 3.

Table 3. Sanger-sequencing reaction mix using BigDye.

1x	1:8 mix
Water	4,85 µl
Seq.Buffer (0,75x)	1,5 µl
BigDye (0,25x)	1 µl
Total	7,35 µl
MIX	7,35 µl
PCR	2 µl
primer (3,25 pmol)	0,65 µl
Total volume	10 µ

Purification was done with Performa DTR v3 filter plates to remove the leftover dye terminator. The template size to be sequenced was set to >800 bp to ensure that all the variants will be covered by sequencing. The templates were sequenced with ABI3730xl DNA Analyzer (Applied Biosystems, Foster City, California, United States) [42].

6 Results and conclusions

6.1 Variant selection for validation

In the analysis of FHRB, ClinPIDD and Myeloma exome data, five SNP variants were found that had some obscurity of their pathogenicity. All the variants of interest were found to be heterozygous. Only the most prevalent intronic or intergenic variants were present in homozygous state, which was not considered to be a cause for concern as they were known to be benign. Compound heterozygosity of variants was not detected in the samples either. The selected variants for further examination, their function, nucleotide variation and frequencies from different databases can be seen in Table 4 and Table 5.

Table 4. Variants selected for validation.

SNP	Variant class (HGMD)	Ref	Alt	Function	Exonic Function
rs2230288	DM?	C	T	Exonic	Nonsynonomous SNV
rs188978150	FP	T	C	Intronic	NA
rs75548401	DM?	G	A	Exonic	Nonsynonomous SNV
rs555143723	NA	C	G	Splicing	NA
rs150466109	DM?	T	G/C	Exonic	Nonsynonomous SNV

The specific variants were selected for validation based on HGMDs variant class specifications, DM? and FP (Table 4). Variant classified as DM? implies to a likely pathogenic mutation, but some degree of doubt has been indicated on its pathogenicity. FP variant is an *in vitro* or *in vivo* functional polymorphism that has been reported to affect the structure or function of the gene and possibly affect the gene product. This variant class includes promoter variants, such as rs188978150, that have been shown to affect the promoter function by reducing its activity. This variant has not yet been associated to GD, but according to HGMD, variants such as this should always be viewed with caution [31]. A rare and not yet well-known splicing variant, rs555143723, was found in ExAc database based on its starting location (Table 4 and Table 5). This variant was selected also for validation among the other variants.

Table 5. Selected variants: their frequencies and phenotypes based on different databases.

SNP	Variant class (HGMD)	ClinVar	ExAc Freq. (homozygous)	dbSNP frequency
rs2230288	DM?	Benign	0.009792 (12)	
rs188978150	FP promoter variant	NA	NA	0.0159000
rs75548401	DM?	Uncertain significance	0.006571 (6)	
rs555143723	NA	NA	0.00005875 (0)	
rs150466109	DM?	Benign	0.007052 (23)	

Three of the selected variants were exonic, nonsynonomous SNVs (Table 4). Nonsynonomous variations are missense point mutations that change one amino acid to another and affect the structure of the encoded protein [35]. One of the two others was an intronic variant (rs188978150) and the other a splicing variant, rs555143723 (Table 4). All the validated variants were single base substitutions.

6.2 SiSu project

There were 8 variants found from the 10,000 samples of SiSu project data that were considered to be disease-causing mutations by HGMD [31]. The variants found in the data, their location, prevalence according to ExAc and clinical significance based on ClinVar can be seen in Table 6. Also the heterozygous carrier amount and type of GD caused by the specific variant are listed on table below (Table 6). The reference information for variant rs760307559 was not available.

Table 6. Disease-causing mutations found from SiSu data.

Variant	Position	Annotation	HGMD	Het. (N)	ClinVar	ExAc (freq)	Type GD
NA	155205105	SPLICE_SITE	DM	1	NA	NA	3
rs104886460	155210420	ESSENTIAL_SPLICE_SITE	DM	1	pathogenic	0.0001154	1, 3
rs364897	155208006	NON_SYNONYMOUS_CODING	DM	1	pathogenic	0.00006813	1
rs369068553	155204996	NON_SYNONYMOUS_CODING	DM	2	NA	0.0000659	1
rs381737	155207932	NON_SYNONYMOUS_CODING	DM	1	pathogenic	0.0000164	3
rs421016	155205043	NON_SYNONYMOUS_CODING	DM	34	pathogenic/likely pathogenic	0.003099	2, 3
rs760307559	155206076	NON_SYNONYMOUS_CODING	DM	1	NA	NA	-
rs76763715	155205634	NON_SYNONYMOUS_CODING	DM	18	pathogenic, risk factor	0.00221	1

Most of the listed disease-causing variants, based on HGMD, were found also from ClinVar and ExAc databases [32; 33]. According to the reference information of the variants in HGMD, all of the variants were reported as pathogenic in original articles. For example, first variant on the list (the variant lacks rs-number) was not found in either ClinVar or

ExAc, but in the reference article, the variant was defined as existing as a compound heterozygous variant in patients with GD. The two detected heterozygous variants were c.762-1GNC(IVS6-1GNC) and c.1389-3CNG(IVS9-3CNG) and were said to cause type 3 GD [43]. Another variant, rs104886460 was listed as compound heterozygous in the original article. The other variant was a common N370S mutation that is known to cause type 1 of GD [44]. Variant rs76763715 was found both as homozygous and compound heterozygous based on the reference information. The variant causes type 1 GD. [45]

Two of the listed variants, rs421016 and rs76763715 prevalence was quite high in ExAc which also translated to the carrier amount of the variants in the SiSu data. The other variants in the list were rare and only one heterozygous carrier was found for each variant. All of the variants were detected in heterozygous state in SiSu data. Due to the limited access to data, it was not possible to analyze if the carriers were compound heterozygous for the detected *GBA* variations. The projects samples were not available for further analysis; therefore, validations of the findings could not be performed from the SiSu project.

6.3 Long-range Touchdown PCR

All of the selected variants were validated from two different samples. The samples used for validation and the amplified part of *GBA* can be seen in Table 7.

Table 7. Location of variants, samples for validation and PCR amplified part of *GBA*.

SNP	Samples	PCR
rs2230288	PID207, FHRB_5240	GBA_1
rs188978150	PID123, PID230	GBA_2
rs75548401	PID231, PID232	GBA_1
rs555143723	PID087, PID088	GBA_2
rs150466109	PIDD235, PIDD343	GBA_2

The program used for TD PCR can be seen in Materials and Methods section of the thesis. The success of the PCR was determined with the LabChip GX instrument using genomic DNA kit. The PCR products were diluted to 1:10 of PCR grade water. Lower marker, LM can be seen at 0.0 on horizontal scale in all of the samples.

Samples ClinPIDD207, FHRB5240, ClinPIDD231 and ClinPIDD232 were amplified using GBA_1 primers as they all were located in exon 9. The product size for GBA_1 should be around 5.9 kb. The multiple overlay electropherogram shows that approximately the right size product was amplified from all four samples (Figure 7).

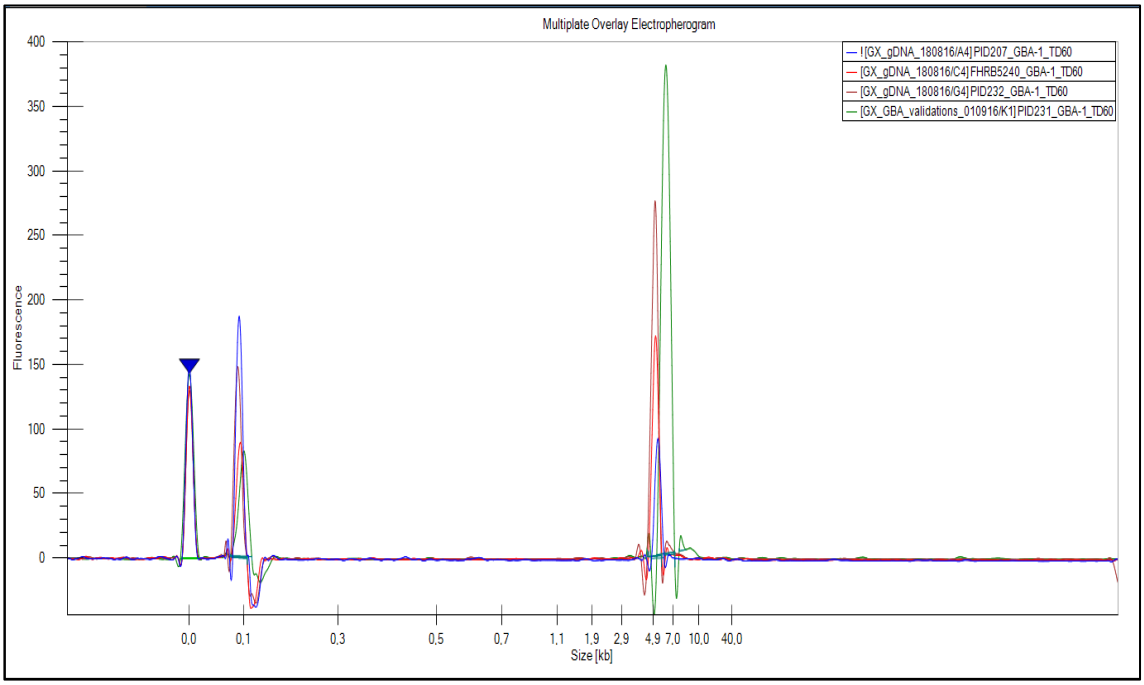

Figure 7. LabChip GX results from samples ClinPIDD207 (blue), FHRB5240 (red), ClinPIDD231 (brown) and ClinPIDD232 (green) amplified regarding amplicon GBA_1.

The product amplified from sample ClinPIDD231, shown in green, seemed to be a bit longer than the other samples but could be due to the x-axis scale not being linear. In all of the samples, primer-dimer was formed that needed to be washed away with purification as smaller products may interfere in the sequencing reactions. Primer-dimer can be seen at 0.1 peak in all of the electropherograms. The most amount of primer-dimer was formed from sample ClinPIDD207 (blue) and also the least amount of actual product was amplified from mentioned sample. The most amount of product was amplified from sample ClinPIDD231 (brown line). Samples FHRB5240 and ClinPIDD207 produced quite small amounts of product which may affect the success of nested sequencing as the PCR sample could be too diluted.

Samples ClinPIDD123, ClinPIDD230, ClinPIDD087, ClinPIDD088, ClinPIDD235 and ClinPIDD343 were amplified with GBA_2 primers (Figure 8). Also in this case, all of the amplified products were nearly the same size. The covered product size was 6.7 kb and

results show that the right size product was amplified from all six samples. As with GBA_1, primer-dimer was formed in all of the samples. The least amount of product was amplified from sample ClinPIDD87 which can be seen in blue in the overlay electropherogram. Substantial amount of primer-dimer was formed from sample ClinPIDD123 (red) and ClinPIDD230 (turquoise). The most amount of product was amplified from ClinPIDD88 (brown). As with GBA_1, some of the GBA_2 samples were not amplified efficiently, which may affect the performance of the sequencing reaction.

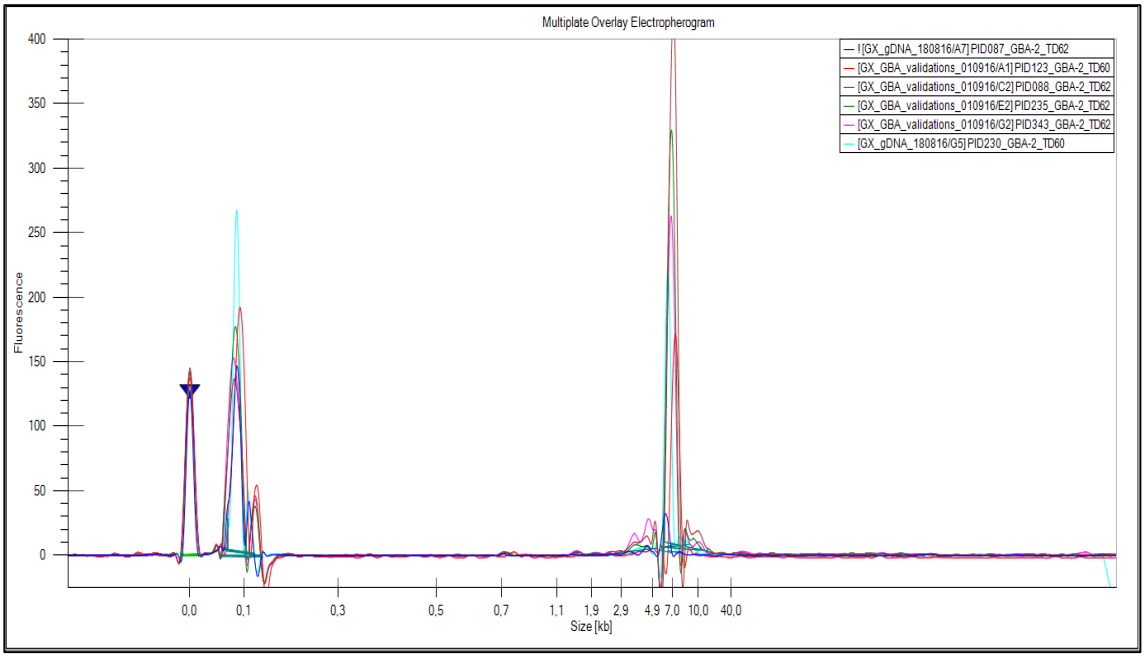

Figure 8. LabChip GX results from samples ClinPIDD087 (blue), ClinPIDD088 (brown), ClinPIDD123 (red), ClinPIDD230 (turquoise), ClinPIDD235 (green) and ClinPIDD343 (pink) amplified regarding amplicon GBA_2.

Samples ClinPIDD087, ClinPIDD088, ClinPIDD123, ClinPIDD235 and ClinPIDD343, shown in Figure 8, were amplified with a different TD PCR range, 72 °C to 62 °C, degreasing 1°C per cycle which optimized the product amplification and also decreased the amount of primer-dimer.

FimmX control sample was amplified in both fragments of the gene, GBA_1 and GBA_2 (Figure 9). Blue line shows GBA_1 and red line GBA_2. Based on the results, the right size products, 5.9 kb and 6.7 kb, were amplified from both GBA fragments, GBA_2 fragment being approximately 0.8 kb bigger than GBA_1. GBA_2 was amplified more compared to GBA_1 but also a substantially more primer-dimer was formed. GBA_2 from

FimmX was amplified using TD PCR program with a range of 72 °C to 62 °C, as well, because it produced more of the desired product and less primer-dimer.

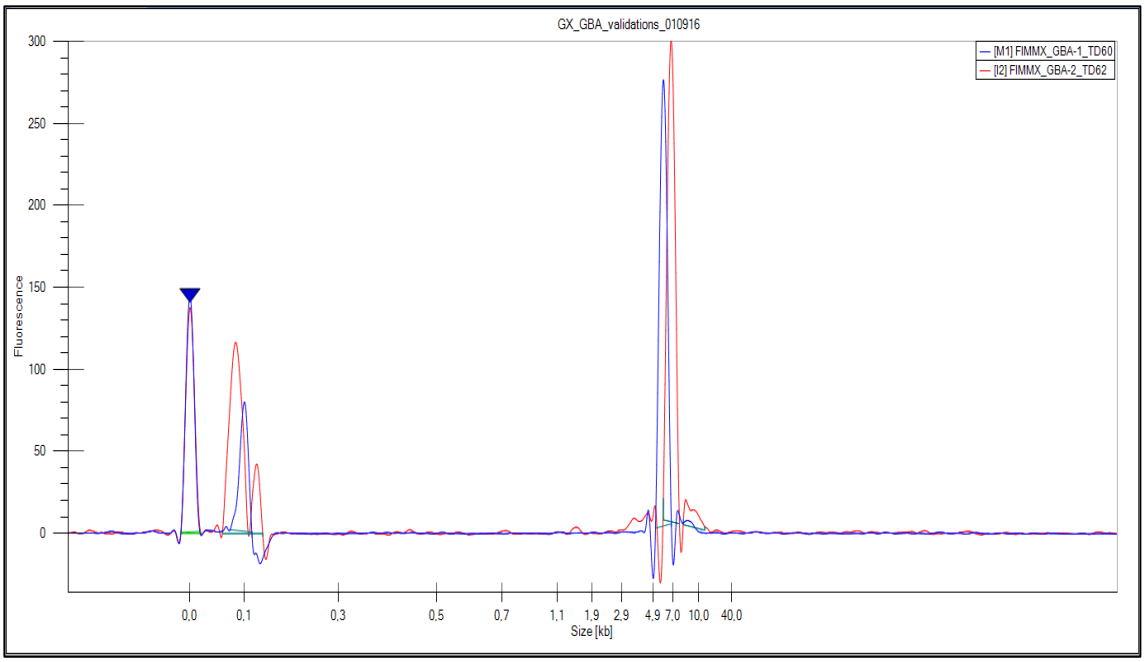

Figure 9. LabChip GX results from FimmX sample amplified regarding both amplicons, GBA_1 (blue) and GBA_2 (red).

Negative control, NTC shows that no contamination occurred in the reactions as no peaks has been formed in the electropherograms (Figure 10). The NTC sample was amplified in all reactions to find out if any of the primers may have been contaminated.

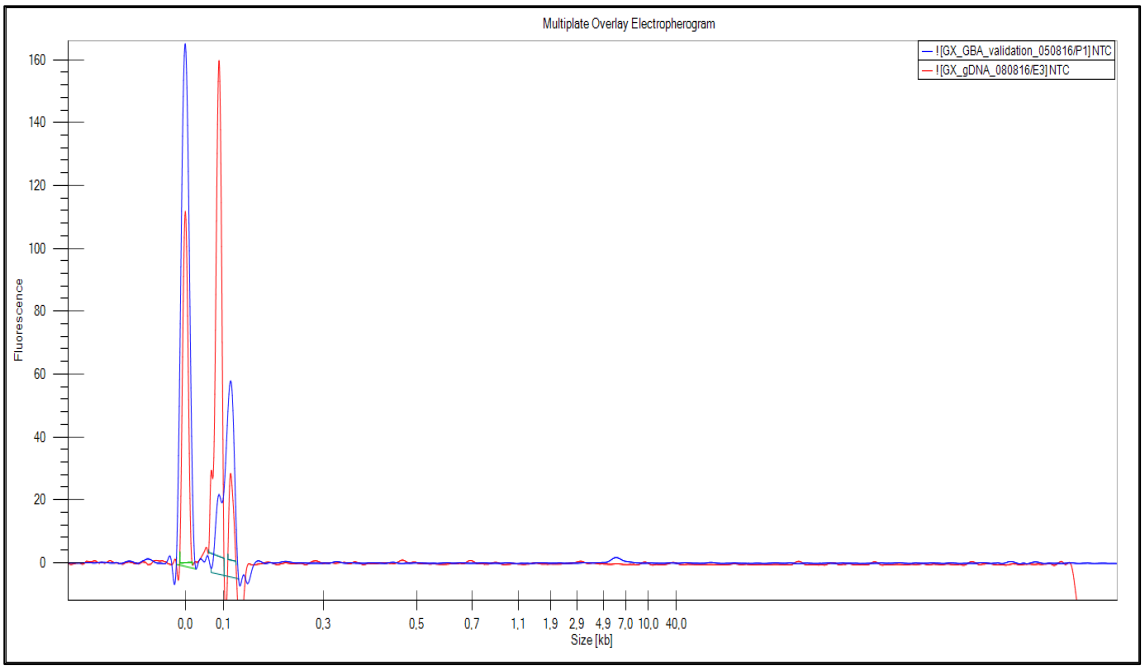

Figure 10. LabChip GX results from negative control, NTC sample amplified regarding GBA_1 and GBA_2.

6.4 Nested Sanger-sequencing results

The sequencing primers to be used were determined with IGV, as mentioned in Materials and Methods-section [40]. The imaging showed that two of the variants, rs2230288 and rs75548401 were located in exon 9 (Table 8). Two others, rs188978150 and rs555143723 could be covered with exon 2 primers and the fifth variant, rs150466109 was located in exon 3. The variants were sequenced using forward and reverse primers. The intronic variant, rs188978150 was sequenced using only exon 2 reverse primer as the designed primers target primarily only the exonic areas of the *GBA* gene and only the reverse primer could cover the targeted SNP.

Table 8. Primers used for nested sequencing determined with IGV.

SNP	Sequencing primers	Primer-F	Primer-R
rs2230288	exon_9	tcagtagttgcaaaaggggc	cagcccgagtgacagagtg
rs188978150	exon_2-R	-	tggccctggattcaaagaga
rs75548401	exon_9	tcagtagttgcaaaaggggc	cagcccgagtgacagagtg
rs555143723	exon_2	cgggaagccggaattacttg	tggccctggattcaaagaga
rs150466109	exon_3	gaggggcttgcttttcagtc	ggaggcagaggttggaatga

The results from nested Sanger-sequencing were verified by use of Sequencher 4.8-software. The sequence AB-files of the sequenced fragments, fasta-sequences of exons and sequences surrounding the variants were imported to the software which then aligns them to a single contig-file. The fasta-sequences of the variants were searched from dbSNP [34]. After the sequence alignment, the variants were searched from the chromatograms of the sequenced samples.

The reverse primers were discovered to work better and produced cleaner sequencing results than the forward primers. Chromatograms from reverse primers had also less background interrupting the analysis compared to chromatograms from forward primers. Therefore, reverse sequences were used for analysis of the Sanger-sequencing results.

The samples FHRB5240 and ClinPIDD207 were sequenced regarding the variant rs2230288. According to the NGS exome data both of the samples were heterozygous for the variant. The results from the nested Sanger-sequencing showed a Y-base, marked with a black box, on both of the sequences which indicated that the patient had a T base on one allele and an alternative base C, SNP variant, on the other (Figure 11).

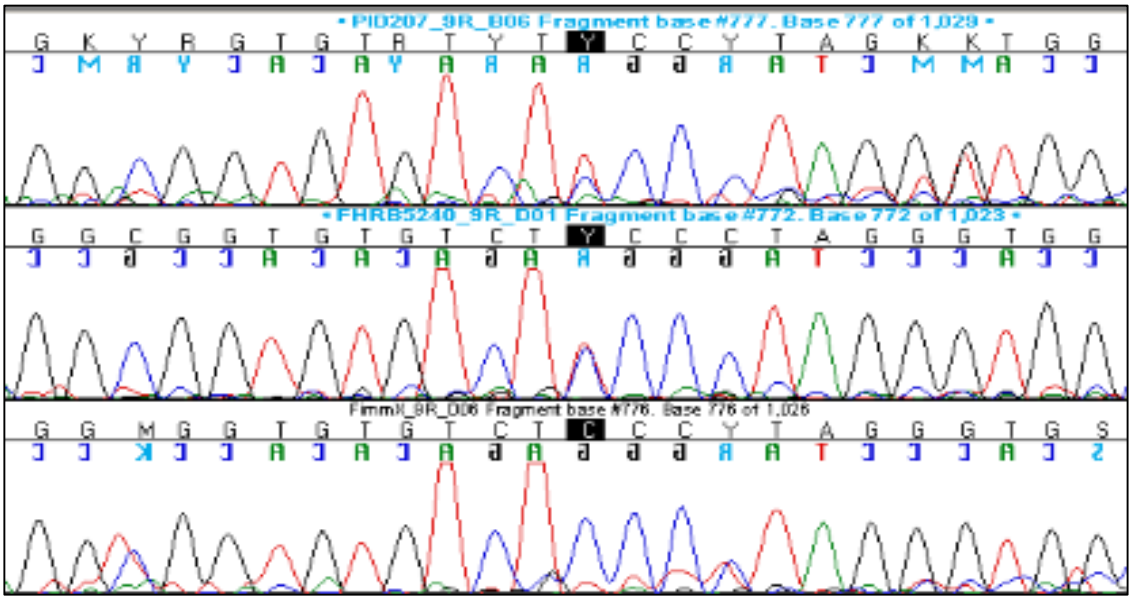

Figure 11. Sanger-sequencing results for SNP rs2230288 from samples PID207 (above), FHRB5240 (middle) and FimmX control (below).

Chromatogram results from control sample FimmX showed reference base C (below). Comparing the peaks from FimmX and the samples, the Y-bases found from the two samples were lower compared to the height of C-base from FimmX which also indicated

heterozygosity in both FHRB5240 and ClinPIDD207 samples. Both FimmX and ClinPIDD207 samples had a bit of background coming through in the chromatograms but not enough for it to interfere with the interpretation of the results. The results confirmed that both of the patients are heterozygous concerning SNP rs2230288 and that the variant is in fact located at the functional *GBA* gene on both samples. Therefore, the Sanger-sequencing results verify the exome data results.

The intronic variant rs188978150 was validated using samples from patients ClinPIDD123 and ClinPIDD230. According to NGS exome data, both of the patients were heterozygous for the specific variant, reference allele being T and alternative allele being C (Figure 12). The chromatograms from Sanger-sequencing showed a Y-base on sample ClinPIDD230 (above) which proved that the patient is heterozygous for variant rs188978150. Based on the results it can be concluded, that the variant is in the functional *GBA* and not its pseudogene. Reference allele T was detected homozygous in control sample FimmX, shown in the bottom panel of Figure 12. Also in this case the peak from Y-base was approximately half the height of the peak from reference base T in FimmX.

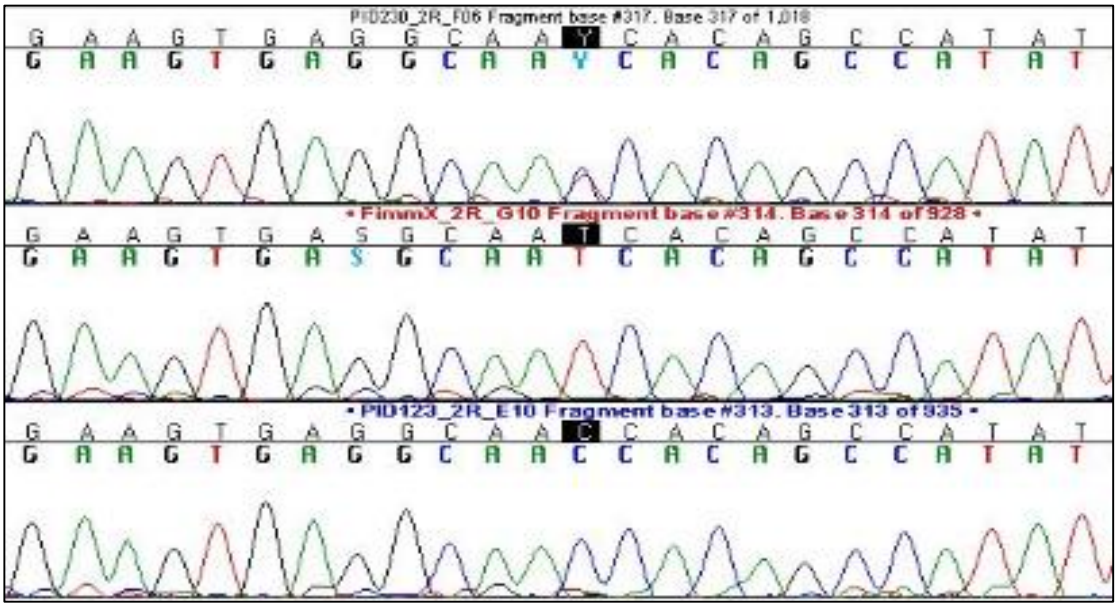

Figure 12. Sanger-sequencing results for SNP rs188978150 from samples PID230 (above), PID123 (below) and FimmX control (middle).

The results from sample ClinPIDD123 (below) showed the alternative base C instead of Y-base as in sample ClinPIDD230. The height of the peak appeared to be the same as

in FimmX, which indicated that the patient is homozygous concerning this intronic variant. The results from the validation suggest that the pseudogene is interfering with NGS results and the T-base, detected from data on sample ClinPIDD123 is actually located in the pseudogene and not in the functional *GBA*. According to the reference information of this splicing variant in HGMD, the variant reduces promoter activity by as much as 35% but is not able to cause GD by itself. Combined with a disease-causing variant, the splicing variant can cause more severe phenotype of the disease [46]. As any other possibly disease-causing variants were not found from the patient, the results from the validation did not raise a cause for concern.

The exonic variant rs75548401 was validated from samples ClinPIDD231 and ClinPIDD232. In the data analysis of the NGS exome data, the two patients were both found to be heterozygous for the specific variant, the reference allele being G. The Sanger-sequencing results from samples PID231 and PID232 both showed R-base in their chromatograms as indicated by exome data (Figure 13). R-base implies that the patient has a G base on one allele and an alternative base A in the other allele. The results shown in the figure confirmed the patients being heterozygous for variant rs75548401.

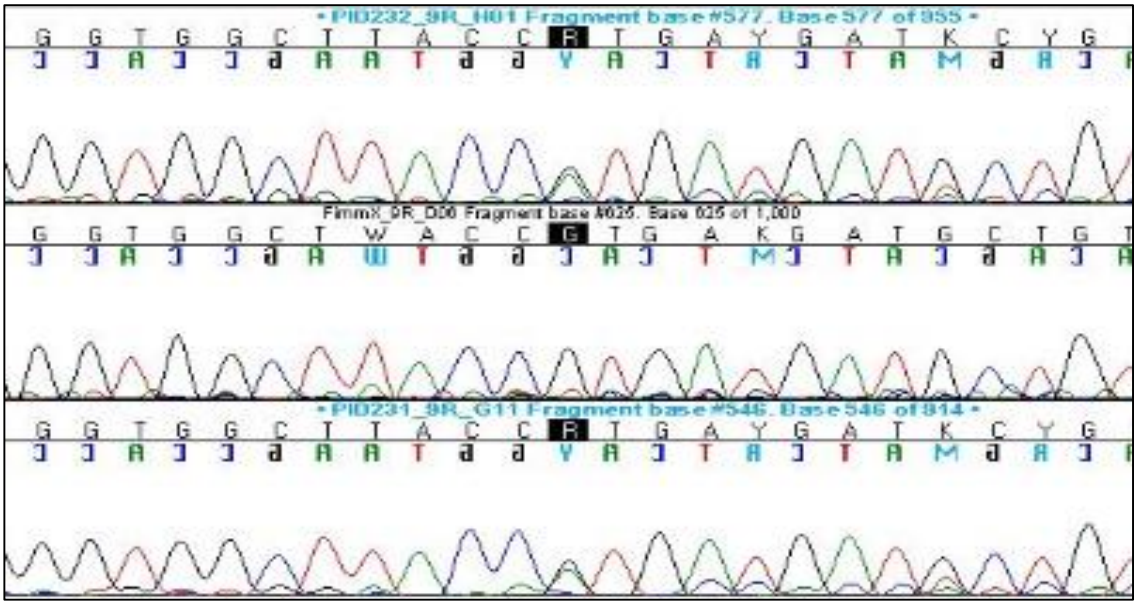

Figure 13. Sanger-sequencing results for SNP rs75548401 from samples PID232 (above), PID232 (below) and FimmX control (middle).

FimmX sample showed the reference base G in the chromatogram results (middle). The R-bases detected from the patient samples chromatograms had lower peaks compared

to the surrounding peaks, which also confirmed that both of the patients are heterozygous for the particular variant. Based on the sequencing results, it can be assumed that the heterozygous variant is located in the functional gene instead of its pseudogene, *GBAP1*, in both of the samples.

The rare and not yet well-known splicing variant rs555143723 was validated using samples ClinPIDD087 and ClinPIDD088. As with all the other variants, the two mentioned samples were both heterozygous for the variant according to the NGS exome data, reference allele being C and an alternative allele G. Figure 14 shows the chromatograms of the Sanger-sequencing from two samples. The sample ClinPIDD087, shown above, was found to be heterozygous for the specified variant, as a clear S-base was found from the chromatogram. S-base implies that patient is heterozygous for C and G alleles. By comparing the height of the peaks, the S-base was found to have a slightly lower peak then other peaks in the chromatograms. Especially when comparing to the clean C-base peak found from FimmX (middle), the S-base was approximately half of the C-base. The results from the control sample FimmX can be seen in the middle panel of Figure 14.

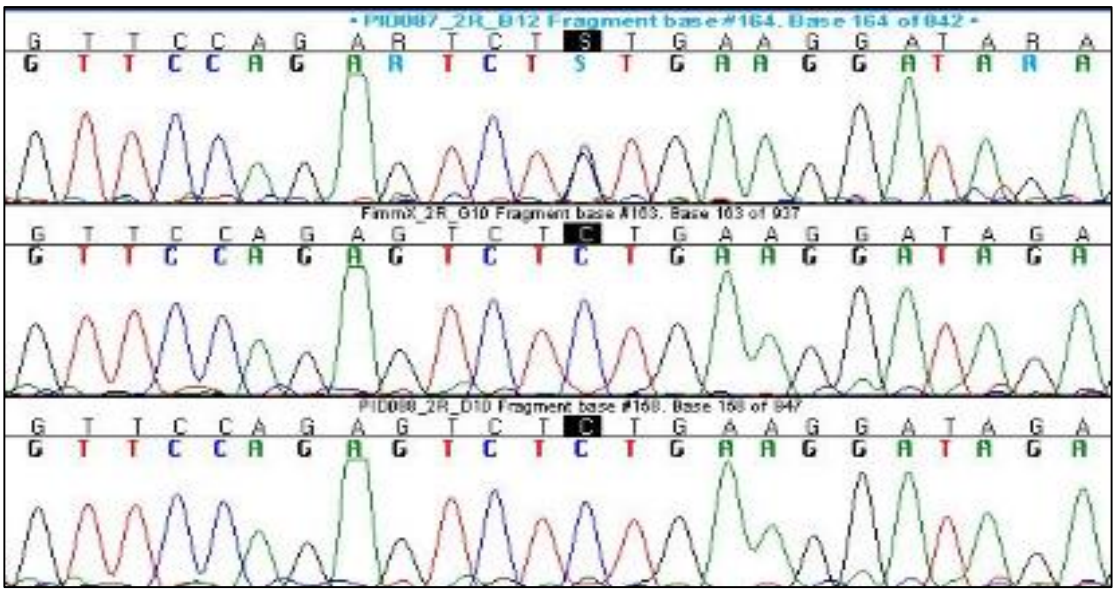

Figure 14. Sanger-sequencing results for SNP rs555143723 from samples PID87 (above), PID88 (below) and FimmX control (middle).

The expected S-base was not found from sample ClinPIDD088 (below). The results from the sequencing showed a clean C-base in the chromatogram. The peak of the base was the same height as other peaks in the chromatogram and also the C-base detected in FimmX which would suggest that the patient is in fact homozygous for the reference

allele and does not have the above-mentioned splicing variant. The results indicated that the pseudogene is, also in this case, causing false results in the exome data and that the variant is located in the carrier's pseudogene.

ClinPIDD087 was the patients sample and ClinPIDD088 was her mother's sample. The results from validation indicated that the variant was inherited from the father of the patient as the mother is homozygous for the reference allele and does not carry the variant. For the future, it might be best to validate the variant also from the fathers sample to get more reassurance of the results. It is also possible that the splicing variant occurred as *de novo*.

Exonic variant rs150466109 was validated from samples ClinPIDD343 and ClinPIDD235. According to the NGS exome data, the two samples were expected to be heterozygous for the variant. The samples had two different alternative alleles in the data, ClinPIDD235 had G-base as an alternative allele and ClinPIDD343 C-base. The reference base was T according to the data. Overall, the chromatograms had a substantial amount of background that made the interpretation of the results more challenging. However, despite the background signal, the results could still be analyzed and can be seen in Figure 15.

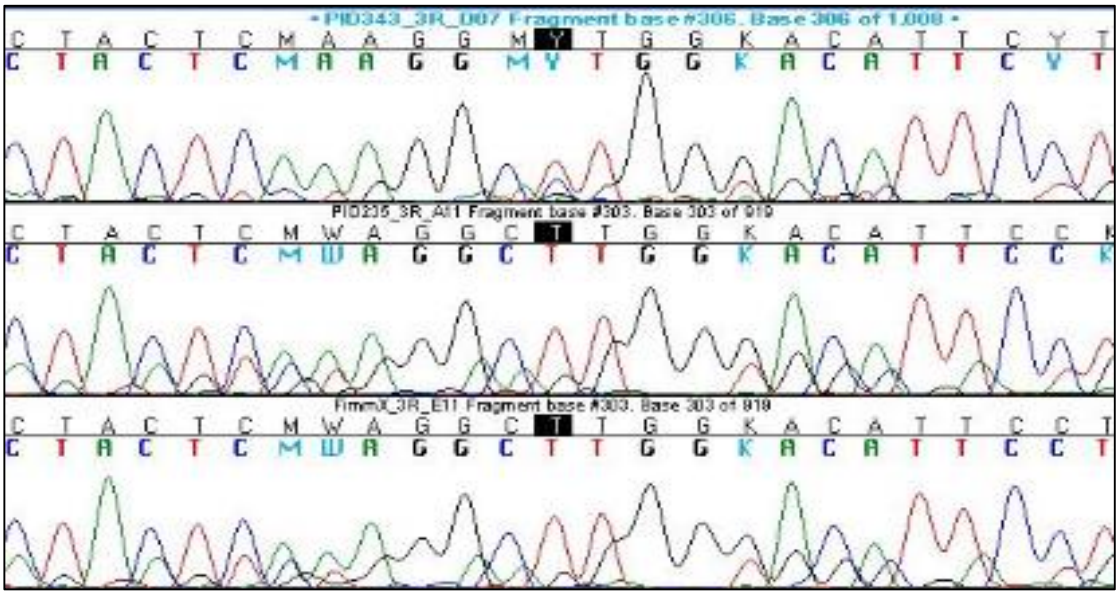

Figure 15. Sanger-sequencing results for SNP rs150466109 from samples PID343 (above), PID235 (middle) and FimmX control (below).

The chromatogram shown above from sample ClinPIDD343 was not as straightforward to interpret as the figure might seem at first. The chromatogram showed a Y-base that implied that the carrier is heterozygous regarding the variant as the exome data indicated. However, the peaks surrounding the variant were lower when compared to the other peaks in the chromatogram, for some reason. The background peak that the software had interpreted as C-base could be seen in all of the chromatograms, FimmX and ClinPID235. For that reason, the results suggested that the patient is not heterozygous but, on the contrary, homozygous regarding the reference allele, T.

The results from the sample ClinPIDD235 (middle) implied that patient was also be homozygous for the reference allele. The chromatogram of the sample showed a reference base T instead of K base that was expected. The peak of the base was compared to T-base from FimmX and they were found to be the same height. Also, when comparing the peak to other peaks in the chromatogram, the T-base was the same height as the others which indicates that the patients is actually homozygous for reference allele.

Based on the validations, it can be said that the pseudogene is distorting the exome results on both of the samples, in this case as well. The variant was not detected in the functional *GBA* from either of the samples, but the variants are, in fact, located in exon 3 of the pseudogene.

The samples, in which the validation results did not correspond with the NGS exome data where checked visually by use of IGV [40]. The purpose of the visualization was to check if there were any polymorphisms at the locations where the sequencing primers bind. Polymorphisms at the primers locations might cause allele-specific amplification that could affect the analysis of the results significantly. There were not any polymorphisms found where the primers were located; therefore, it can be concluded that the validations results are valid.

7 Discussion

There were some factors that caused difficulties with the thesis. First of all, when writing about *GBA* and its pseudogene *GBAP1,* the different databases caused some difficulties due to their differences in information about the specific genes. Databases and articles about the subject included differences in the sizes of the genes, genomic and cytogenetic locations and the amount of exons, which made it difficult to determine what information was correct and up to date. This also showed how genetics is still very much an evolving field of research. Especially *GBA* and its pseudogene, *GBAP1* and their impact on GD is still under research. It seems that the pseudogene might have larger impact on the diagnostics and interpretation of disease mutations than originally thought.

The practical part of the thesis was known to be potentially tricky from the start because of the difficult nature of the *GBA* gene. The PCR reactions had to be repeated multiple times to get enough product amplified. Also different programs had to be tested for different samples. Due to the highly diluted PCR products, there was not enough signal for the Sanger-sequencing and the first two sequencing reactions were mostly blank. The next PCR products were concentrated to improve the chances of the sequencing reaction to be successful. For some reason, the forward primers worked poorly with every sample and significant amount of background made interpretation of the results difficult or impossible. It was clear that the amplification of the gene was demanding and specific reaction conditions needed to be applied for all the amplified fragments.

The SiSu data was analyzed only *in silico* from the WEB-site due to the fact that the samples were not available for analysis. Therefore, validations of the variants could not be performed for the thesis. The data analysis seems to confirm that GDs prevalence is extremely rare in Finnish population as no homozygous variants were detected. Also, all of the heterozygous variants were quite rare and most of them had only one carrier found from the 10000 samples sequenced for the project. The drawback of the *in silico* analysis is that the detected variants could not be validated with another method. Some of these variants might actually exist in homozygous state, but the NGS data detects signals from both the *GBA* gene and its pseudogene which may influence the analysis of the sequencing results giving the impression of *GBA* allele heterozygosity.

The analysis of the SiSu data grew the study substantially wider with over 10000 Finnish samples. The SiSu project is currently the most extensive database of Finnish inheritance and with the addition of its data to the thesis, it can be said that the study gives quite broad picture of the frequency of possibly disease-causing mutations in the *GBA* gene in the Finnish population.

The results from the validations clearly show that the pseudogene affects highly the NGS data analysis results as the pipeline does not recognize if detected the variant is located in the functional gene or the pseudogene. In almost half of the samples used for validations, one of the alleles could be tracked down to the pseudogene and not the functional *GBA* gene. In that aspect, the results can be considered to be fairly substantial and helpful for future research of the gene. The results from the validations do not cause a concern regarding the patients, as all but one was found to be either heterozygous for the specific variant or homozygous for the reference allele. One of the patient was found to be homozygous for the promoter variant, but that is not concerning as the variant cannot cause the disease by itself, as mentioned in the results.

Finally, based on the data analysis and validations of the findings it seems clear that traditional NGS is not the best method for analyzing the *GBA* gene as the results are not valid due to the pseudogenes presence. All of the heterozygous and homozygous variants found should be validated separately since even the heterozygous variant might actually be homozygous in the functional gene. This could happen if the pseudogene presents with another nucleotide than the *GBA* gene resulting the variant to be interpreted as heterozygous by the data analysis pipeline. This affects significantly the results if the pseudogene is not taken into account in the analysis of the data.

References

1 Lysosomal storage disease, Michael C Kruer; Amy Kao, 2015, [Internet document], http://emedicine.medscape.com/article/1182830-overview, Read: 27.7.2016

2 *GBA*, https://ghr.nlm.nih.gov/gene/GBA, Read: 11.7.2016

3 *GBA*, http://omim.org/entry/606463, Read: 11.7.2016

4 Glucocerebrosidase, https://www.nlm.nih.gov/cgi/mesh/2011/MB_cgi?mode=&term=Glucocerebrosidase, Read: 8.7.2017

5 http://enzyme.expasy.org/EC/3.2.1.45, Read: 8.7.2016

6 KEGG Compound, http://www.genome.jp/dbget-bin/www_bget?rn:R01498, Read: 8.7.2016

7 Gaucher disease, GeneReviews® [Internet document], February 26, 2015, http://www.ncbi.nlm.nih.gov/books/NBK1269/, Read: 5.7.2016

8 Sequence variability of a human pseudogene, Rosa Martínez-Arias, Francesc Calafell, Eva Mateu, David Comas, Aida Andrés, and Jaume Bertranpetit, 2001, [Internet document], http://www.ncbi.nlm.nih.gov/pmc/articles/PMC311081/, Read: 18.7.2016

9 Aggarwal S, Jain SJMN, Das Bhowmik A,Tandon A, Dalal A. 2015. Molecular studies on parents after autopsy identify recombinant GBA gene in a case of Gaucher disease with ichthyosis phenotype. Am J Med Genet Part A 167A:2858–2860., [Internet document], http://onlinelibrary.wiley.com/doi/10.1002/ajmg.a.37251/epdf, Read: 18.7.2016

10 A New Gene–Pseudogene Fusion Allele Due to a Recombination in Intron 2 of the Glucocerebrosidase Gene Causes Gaucher Disease, Bru Cormand, Anna diaz, Daniel Grinberg, Amparo Chabas, Lluisa Vilageliu, 2000, [Internet document], http://www.ub.edu/geneticaclasses/brucormand/pdfs/17.pdf, Read: 18.7.2016

11 Gaucher disease, Genetics Home Reference, July 12, 2016, [Internet document], https://ghr.nlm.nih.gov/condition/gaucher-disease#statistics, Read: 5.7.2016

12 Gaucher disease, NORD National Organization of Rare Disorders, [Internet document], http://rarediseases.org/rare-diseases/gaucher-disease/, Read: 11.7.2016

13 Gaucher disease [Internet document], https://www.symptoma.com/en/info/gaucher-disease#workup, Read: 11.7.2016

14 Abrams CS. Thrombocytopenia. In: Goldman L, Schafer AI, eds. Goldman's Cecil Medicine. 24th ed. Philadelphia, Pa: Elsevier Saunders; 2011: chap 175. Read: 12.7.2016

15 Gaucher disease, Aabha Nagral, 2014, [Internet document], https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017182/, Read: 12.7.2016

16 Hydrops fetalis, Medscape, Jan 09, 2014, http://emedicine.medscape.com/article/974571-overview, Read: 14.7.2016

17 Gaucher Disease Treatment & Management, Medscape, Nov 24, 2014 [Internet document] http://emedicine.medscape.com/article/944157-treatment, Read: 14.7.2016

18 Chaperone Activity of Bicyclic Nojirimycin Analogues for Gaucher Mutations in Comparison with N-(n-nonyl)-Deoxynojirimycin [Internet document], 2009 Nov 23, http://www.ncbi.nlm.nih.gov/pubmed/19830760, Read: 18.7.2016

19 Library construction for next-generation sequencing: Overviews and challenges, Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian, 2014 Feb 1 [Internet document], http://www.ncbi.nlm.nih.gov/pubmed/24502796, Read: 20.7.2016

20 Next-Generation Sequencing, https://www.fimm.fi/en/services/technology-centre/sequencing/next-generation-sequencing/dna-library-preparation, Read: 20.7.2016

21 Roche NimbleGen SeqCap EZ SR User's Guide, Version 5.1, https://lifescience.roche.com/wcsstore/RASCatalogAssetStore/Articles/SeqCapEZLibraryUserGuide.pdf, Read: 20.7.2016

22 An Introduction to Next-Generation Sequencing Technology, Illumina, 2016 [Internet document], http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf, Read: 20.7.2016

23 DNA Library preparation, NEB, 2010, [Internet document], https://www.neb.com/applications/library-preparation-for-next-generation-sequencing/dna-library-preparation, Read: 21.7.2016

24 T4 DNA ligase, NEB, [Internet], https://www.neb.com/products/m0202-t4-dna-ligase, Read: 21.7.2016

25 Illumina Sequencing Technology, Illumina, [Internet document], http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf, Read: 21.7.2016

26 cBOT™ 2 Cluster Generation System, Illumina, 2016, [Internet document], http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/cbot-2-data-sheet-770-2015-029.pdf, Read: 28.7.2016

27 Comparison of solution-based exome capture methods for next generation sequencing, Sulonen AM, Ellonen P, Almusa H, Lepistö M, Eldfors S, Hannula S, Miettinen T, Tyynismaa H, Salo P, Heckman C, Joensuu H, Raivio T, Suomalainen A, Saarela J, 2011, [Internet document], http://www.ncbi.nlm.nih.gov/pubmed/21955854, Read: 27.7.2016

28 Klenow fragment, NEB, [Internet] https://www.neb.com/products/m0210-dna-polymerase-i-large-klenow-fragment, Read: 20.7.2016

29 The Molecular Pathology of Primary Immunodeficiencies, [Internet document], https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1867474/, Read: 27.7.2016

30 Genome Annotation: From sequence to biology, [Internet document], http://www.nature.com/nrg/journal/v2/n7/full/nrg0701_493a.html, Read: 25.7.2016

31 HGMD Human Genome Mutation Database, https://portal.biobase-international.com/cgi-bin/portal/login.cgi?redirect_url=/hgmd/pro/start.php?, Read: 20.5.2016

32 Exome Aggregation Consortium, http://exac.broadinstitute.org/, Read: 20.5.2016

33 ClinVar, National Center of Biotechnology Information, http://www.ncbi.nlm.nih.gov/clinvar/, Read: 20.5.2016

34 The Single Nucleotide Polymorphism Database (dbSNP), http://www.ncbi.nlm.nih.gov/SNP/, Read: 20.5.2016

35 An Introduction to Genetic Analysis, 7th edition, Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, and William M Gelbart., 2000, [Internet document], http://www.ncbi.nlm.nih.gov/books/NBK21955/, Read: 26.7.2016

36 Pharmacological Chaperones: Design and Development of New Therapeutic Strategies for the Treatment of Conformational Diseases, Marino Convertino, Jhuma Das, and Nikolay V. Dokholyan, ACS Chem. Biol. 2016, 11, 1471−1489 [Internet document], http://pubs.acs.org/doi/pdf/10.1021/acschembio.6b00195, Read: 28.7.2016

37 PCR Protocol for Phusion® High-Fidelity DNA Polymerase (M0530), NEB, https://www.neb.com/protocols/1/01/01/pcr-protocol-m0530, Read: 3.8.2016

38 High and Low Annealing Temperatures Increase Both Specificity and Yield in Touchdown and Stepdown PCR, Karl H. Hecker, Kenneth H. Roux 1996, [Internet document], http://www.biotechniques.com/multimedia/archive/00053/19962003478_53026a.pdf, Read: 3.8.2016

39 LabChip GX User Guide, PerkinElmer, 2013, http://www.bioneer.co.kr/literatures/manual/instrument/LabChip%20GX%20HT%20DNA%20High%20Sensitivity%20LabChip%20Kit.pdf, Read: 3.8.2016

40 Integrative Genomics Viewer, Version 2.3, http://software.broadinstitute.org/software/igv/, Read: 4.8.2016

41 PuTTY, www.putty.org, Read: 9.8.2016

42 Full service sequencing, FIMM, https://www.fimm.fi/en/services/technology-centre/sequencing/capillary-sequencing-services/full-service-sequencing, Read: 12.8.2016

43 Molecular characterization of type 3 neuronopathic Gaucher disease in Thai patients, P. Suwannarat, 2007, [Internet document] https://www.researchgate.net/publication/6149646_Molecular_characterization_of_type_3_neuronopathic_Gaucher_disease_in_Thai_patients, Read: 31.8.2016

44 Exhaustive screening of the acid beta-glucosidase gene, by fluorescence-assisted mismatch analysis using universal primers: mutation profile and genotype/phenotype correlations in Gaucher disease, D. P. Germain, 1998, [Internet document],http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1377310/, Read: 31.8.2016

45 Genetic heterogeneity in type 1 Gaucher disease: multiple genotypes in Ashkenazic and non-Ashkenazic individuals, S. Tsuji, 1988, [Internet document], http://www.ncbi.nlm.nih.gov/pmc/articles/PMC279989/, Read: 31.8.2016

46 Characterization of the c.(-203)A>G variant in the glucocerebrosidase gene and its association with phenotype in Gaucher disease, Alfonso P, 2011, [Internet document], http://www.ncbi.nlm.nih.gov/pubmed/21087600, Read: 6.9.2016

47 Sequencing Initiative Suomi, http://sisuproject.fi/, Read: 12.8.2016

48 Ensembl, http://www.ensembl.org/index.html, Read: 11.7.2016

49 Primer3, http://primer3.ut.ee/, Read: 1.3.2016

Creating annotation files and using grep-command in Putty

cd /projects/fimm_ga2_heckman/fhrb_germline

grun.py -n fhrb_batch -c './run_samples.bash fhrb_list.txt'

grep -w "GBA" /projects/fimm_ga2_heckman/fhrb_germline/*/*annotated.txt > /fs/projects/ms_saarela/Immunodeficiency/Clinical-exomes/GBA/FHRB_annotation