Sequencing and molecular characterization of variations in the glycine N-acyltransferase gene

C Herfurth

20254962

Dissertation submitted in partial fulfilment of the requirements for the degree *Magister Scientiae* in Biochemistry at the Potchefstroom Campus of the North-West University

Supervisor: Prof AA van Dijk

Co-supervisor: Dr FH O'Neill

May 2014

I would like to express my appreciation and gratitude to the following individuals and institutions. Without their support and contributions the completion of my M.Sc. would not have been possible.

- My supervisor, Prof. Albie van Dijk and co-supervisor, Dr. Frans O'Neill, for their support and assistance.
- Dr. Carina Mels and Mr. Lardus Erasmus for their assistance and detoxification data.
- Me Lizelle Zandberg for her assistance and SNP6 chip data.
- National Research Foundation for their financial support during this study.
- My parents and family for all their love, moral support and motivation. I would not have made it this far without them and for that I am eternally grateful.
- Willem van Dalen for his encouragement and moral support.

ABSTRACT

Humans are continuously challenged by harmful endogenous and xenobiotic substances. Detoxification is the ability to neutralise and remove these substances from the body. Glycine N-acyltransferase, EC 2.3.1.13 (GLYAT) is a key enzyme in detoxification. GLYAT catalyses an amino acid (glycine) conjugation reaction in phase II of detoxification. It is expected that, similar to what has been observed in the Cytochrome P450 enzymes, variations within the *GLYAT* gene may lead to altered enzyme activity that may affect the efficacy of detoxification. The aim of this study was to identify genetic variations within the *GLYAT* gene of a cohort of individuals whose GLYAT activity has been biochemically characterized. Biochemical profiles of phase I and II detoxification of a number of individuals was screened to select those with possible aberrant GLYAT activity. Eighteen selected individuals agreed to participate in the study. The 23.21 kb *GLYAT* gene of the participants was amplified in four fragments and sent for pyrosequencing (Roche GS FLX titanium) at Inqaba Biotec. The results were analysed with the Lasergene software package from DNAStar (Madison, Wisconsin, USA). A total of 94 variations were identified from the Next Generation Sequencing data. Of these three found in the exons were known variations and four variations located in the exons were novel. A total of 62 known and 25 novel variations were identified in the introns of the *GLYAT* gene. Sanger sequencing verified 70.29% (68 in total) of the variation, which included 12 novel variations, of which one is located in exon six. Real-time quantitative PCR (qPCR) experiments were conducted and the data analysed using CopyCaller software to identify copy number variations within the cohort. It was found that participant 17 may have multiple copies of parts of the 3-terminal end of the gene (exons five and six), which might have an effect on GLYAT activity. Variations could possibly affect GLYAT activity, but the data was inconclusive and must be confirmed. Some of the variations could possibly affect GLYAT activity, but no correlation could be made between the variations identified during this study and the cohort's detoxification ability. Further studies needs to be conducted to establish the effect of the variations in combination with one another on GLYAT activity. If some of these variations affect GLYAT activity such data might shed some light on variations observed between the glycine conjugation ability of individuals. Such information could eventually be of value in treatment of inborn errors of metabolism.

Key words: GLYAT, Detoxification, Genetic variations, SNP, Copy number variance.

OPSOMMING

Mense word voortdurend aan skadelike stowwe blootgestel. Detoksifikasie is die vermoë om hierdie stowwe te neutraliseer en uit die liggaam te verwyder. Glisien N - acyltransferase, EC 2.3.1.13 (GLYAT) is 'n belangrike ensiem in detoksifikasie. Die GLYAT ensiem kataliseer 'n aminosuur (glisien) konjugasiereaksie in fase II van detoksifikasie. Daar word verwag dat, soortgelyk aan wat in die P450 ensieme waargeneem word, variasies binne die *GLYAT* geen kan lei tot veranderde ensiem aktiwiteit wat die doeltreffendheid van detoksifikasie beïnvloed. Die doel van hierdie studie was om genetiese variasies binne die *GLYAT* geen van 'n groep individue, wie se GLYAT aktiwiteit biochemies gekaraktariseer is, te identifiseer. 'n Aantal individue, met biochemiese bebaalde fase I en fase II detoksifikasie profiele, is gefynkam om individue te identifiseer met moontlike afwykende GLYAT aktiwiteit. Agtien individue is geïdentifiseer en het ingestem om deel te neem aan hierdie studie. Die 23.21 kb *GLYAT* geen van die deelnemers is geamplifiseer in vier fragmente en gestuur vir pyrosequencing (Roche GS FLX titanium) by Inqaba Biotec. Die resultate is ontleed met die Lasergene pakket van DNAStar (Madison, Wisconsin, VSA). 'n Totaal van 94 variasies is geïdentifiseer met behulp van pyrosequencing. Drie van die variasies wat geïdentifiseer is, is bekende variasies in die eksons terwyl vier nuwe variasies geleë in die eksons waargeneem is. ‘n Totaal van 62 bekende en 25 nuwe variasies is in die introns van die *GLYAT* geen geïdentifiseer. In total is 70.29% (68 in total) van hierdie variasies geverifieer, wat 12 nuwe variasies insluit, waarvan een geleë is in ekson ses. Kwantitatiewe PCR (qPCR) eksperimente is uitgevoer en ontleed met behulp van CopyCaller sagteware om kopiegetal afwykings binne die groep te identifiseer. Daar is gevind dat deelnemer 17 moontlik veelvuldige kopieë aan die 3-teminaal kant van die geen het (ekson vyf en ses), wat moontlik 'n effek op GLYAT aktiwiteit kan hê. Variasies kan moontlik GLYAT aktiwiteit beïnvloed, maar die data was onbeslis en moet bevestig word. Verdere studies moet dus gedoen word om die effek van variasie kombinasies te bestudeer op GLYAT aktiwiteit. Indien sommige van die variasies 'n effek het op GLYAT aktiwiteit, kan dit moontlik insig lewer op verskille waargeneem tussen individue se glisien konjugasie vermoë. Hierdie navorsing sal moontlik van groot waarde kan wees in die behandeling van aangebore metaboliese siektes.

Sleutel woorde: GLYAT, Detoksifikasie, Genetiese variasies, SNP, Kopiegetal afwyking

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

Abbreviation	Meaning
AA	Amino acid
ATP	Adenosine triphosphate
CNV	Copy number variation
CoA	Coenzyme A
CPGR	The Centre of Proteomic & Genomic Research
dNTP	Deoxyribonucleotide triphosphate
ddNTP	Dideoxyribonucleotide triphosphate
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
EDTA	Ethylene diamine tetraacetic acid
FAM	6-carboxy-fluorescein
GLYAT	Glycine N-acyltransferase
GLYATL1	Glycine N-acyltransferase like 1
GLYATL2	Glycine N-acyltransferase like 2
GNAT	GCN5-related *N*-acetyltransferase
IEM	Inborn errors of metabolism
IVD	Isovaleryl coenzyme A dehydrogenase
MCAD	Medium-chain acyl-CoA dehydrogenase
MID	Molecular identifier
MGB	Minor Groove Binder
NGS	Next generation sequencing
NTC	No template control
PCR	Polymerase chain reaction
qPCR	Quantitative polymerase chain reaction
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
TAE	Tris-acetate-EDTA
TAMRA	Tetramethylrhodamine
TE	Tris-EDTA
UTR	Un-translated region

CHAPTER 1 - LITERATURE REVIEW

1.1 Human Genome

The human genome consists of 3.2 billion base pairs (nucleotides), which form the 23 chromosomes. Only 1.5 % of the three billion base pairs codes for proteins (LANDER *et al.* 2001). Nucleotide variations between individuals are expected every 300 base pairs (MANCINELLI *et al.* 2000; FEUK *et al.* 2006). Thus, between the genomes of two individuals there would exist approximately 10 million variations (MANCINELLI *et al.* 2000; FEUK *et al.* 2006).

During the replication of DNA, mutations or single nucleotide polymorphisms (SNPs) can be introduced. A SNP is defined as a single nucleotide alteration within a DNA sequence that occurs in more than 1% of the population (SACHIDANANDAM *et al.* 2001; CHORLEY *et al.* 2008), whereas a mutation is defined as damage or alterations within a DNA sequence that occurs in less than 1% of the population. Mutations are not only due to single nucleotide changes, but can also be the result of insertions and deletions within a DNA sequence. Both SNPs and mutations are permanent and can be associated with disease (MANCINELLI *et al.* 2000; CHORLEY *et al.* 2008) and drug efficacy. Personalized medicine could be applied if a correlation can be made between a DNA sequence and an individual's phenotype. When a correlation is found the effects thereof should be fully elucidated for better development of personalized medicine, this could be done by obtaining the whole genome sequence data of a vast amount of individuals (HERT *et al.* 2008). From the pharmacogenetic approach it is clear that once these variations are identified, it may be possible to customise treatment for patients based on their individual response to certain drugs (WATTS *et al.* 1990; MANCINELLI *et al.* 2000).

It is not only SNPs and mutations that can be associated with disease, but copy number variations (CNV) as well. A CNV is a segment of DNA which is either deleted or duplicated and can lead to copy number variations in a genetic sequence. According to McCarroll et al. (2007), CNVs can influence human phenotypes. Complex diseases might be more susceptible to variations (such as SNPs, CNVs, and mutations) in the introns, which could alter enzyme activity rather than

terminating the enzyme function completely (MCCARROLL and ALTSHULER 2007). Van der Sluis et al. demonstrated this effect, where enzyme activity is altered, by using a recombinant human GLYAT. They found that introducing a variation in the gene could have an effect on the enzyme kinetic properties (VAN DER SLUIS *et al.* 2013).

Each gene consists of introns and exons. The exons are spliced when DNA is transcribed to mRNA which can be translated into amino acids to form a protein. If the nucleotides are incorrectly spliced the amino acid sequence can change and thus the possibility exists that the activity of the enzyme can also change. A sequence can be incorrectly spliced if the splice site or branch point was changed by either a SNP or mutation.

1.2 Inborn errors of metabolism

In some cases a variation within a gene can cause an amino acid to change. If this amino acid is, for instance, a key amino acid in terms of folding within the affected protein, it can alter the function of the protein and increase susceptibility to disease (PRITCHARD 2001; CHORLEY *et al.* 2008).

Inborn errors of metabolism (IEM) are diseases caused by mutations in a gene coding for an enzyme involved in metabolic pathways. These mutations can cause an enzyme to be defective and results in a range of symptoms. An example of this is a mutation in the isovaleryl Coenzyme A dehydrogenase (IVD) gene which causes the enzyme, isovaleric acid-CoA dehydrogenase, to be defective (SWEETMAN and WILLIAMS ; GUAN *et al.* 2007). This causes the metabolic disorder, Isovaleric acidemia, where isovaleric acid levels will be elevated in the urine. Affected individuals will also exhibit episodic vomiting and other symptoms like dehydration and ketosis (GUAN *et al.* 2007).

Isovaleric acidemia can be treated with glycine supplementation (ITO *et al.* 1995). Isovaleric acid will be conjugated with glycine to form isovalerylglycine which will be

excreted via the urine or bile. Infants only have 5% to 40% of mature metabolizing ability of glycine conjugation. For this reason, infants treated with benzoate or salicylate could have a metabolic disadvantage, since these substances have to be removed via the glycine conjugation pathway. Because of their lower glycine conjugation ability, infants treated with glycine for isovaleric acidemia could also have a metabolic disadvantage (MAWAL *et al.* 1997).

Medium-chain acyl-CoA dehydrogenase (MCAD) deficiency is another example of an IEM. The medium-chain fatty acids, octanoic, decanoic and *cis*-4-decenoic acids, carnitine derivates, octanoylcarnitine, decanoylcarnitine and decenoylcarnitine will accumulate in tissues. The elevated levels of hexanoylglycine, phenylpropionylglycine and suberylglycine will be present in the urine of individuals with a MCAD deficiency (DE ASSIS *et al.* 2006). The glycine derivates are due to the conjugation of the toxic metabolites, resulting from MCAD deficiency. It was established by a study conducted by. de Assis and co-workers (2005) that the glycine derivates from MCAD did not affect the Na^+, K^+ ATPase activity, whereas metabolites such as decanoic acid and octanoylcarnitine inhibited the Na^+, K^+ ATPase activity. The Na^+, K^+ ATPase enzyme generate membrane potential by transportation of sodium and potassium ions within the central nervous system. The inhibition of the Na^+, K^+ ATPase enzyme can cause cerebral ischemia, epilepsy and neurodegenerative disorders (DE ASSIS *et al.* 2006). Since the glycine derivates of MCAD deficiency do not cause inhibition of the Na^+, K^+ ATPase activity, this indicates that glycine supplementation is a suitable treatment for MCAD deficiency.

Glycine supplementation is a general treatment for many IEMs. This treatment is not always 100 % effective and the response of individuals to glycine treatment varies. Supplementation with too high glycine has neurotoxic effects (NEWELL *et al.* 1997; BARTH *et al.* 2005), while too low glycine supplementation will not be efficient for detoxification (ITO *et al.* 1995). With the help of pharmacogenomics it could be possible to determine the correct dosage for optimum glycine conjugation.

CNVs, splice sites and branch point changes can also cause a disease or defective enzyme. Although CNVs have only been correlated with a small percentage of the

approximately 2000 mendelian diseases explained at a molecular level thus far, CNVs none the less can be associated with disease (MCCARROLL and ALTSHULER 2007).

Apo C-II is a cofactor for lipoprotein lipase in triglyceride metabolism. Fojo et al (1988) amplified part of the Apo C-II gene of a patient and treated the amplicon with restriction enzymes, Dde I and Hph I. They established that the patient investigated was homozygous for a G-to-C mutation. This is a donor splice site mutation and thus can lead to a defective enzyme an ultimately to Apo C-II deficiency (FOJO *et al.* 1988). From the study by Fojo et al (1988) it is clear that mutations and SNPs can have an effect on splicing and can thus cause disease.

1.3 Detoxification

Humans are continuously exposed to endogenous toxins (toxic metabolites produced by the cells) and exogenous toxins (xenobiotics) during their lifetime.

Detoxification is an individual's ability to remove toxins from the body. This process can be divided into two phases, phase I and phase II. In phase I an active group is added to fat-soluble toxins to form an active compound. These active compounds are then deactivated via conjugation in phase II to form water-soluble compounds that are easily excreted. Glycine N-acyltransferase, EC 2.3.1.13, (GLYAT) catalyses an amino acid conjugation reaction, with glycine, in phase II of detoxification. Glycine supplementation is a general treatment for a large number of inborn errors of metabolism (IEM). Glycine supplementation is not always effective. The treatment can be effective in one individual, but prove to be ineffective in another with the same symptoms or disorder (MANCINELLI *et al.* 2000). The variability between individuals can be due to genetic variations, such as single nucleotide polymorphisms, mutations and copy number variations.

In the case of an IEM an inherited trait results in an enzyme in a metabolic pathway being defective. The defective enzyme can cause toxic metabolites to accumulate in

the body. Some of these toxic metabolites are highly reactive and able to cause damage to proteins, DNA and RNA within the cells. They therefore have to be removed via alternative pathways and detoxified.

Xenobiotics originate from substances that the person is exposed to, for example, pharmaceuticals, food supplements and toxic fumes. Each individual will react differently to a consumed substance because of the impact of the environment, lifestyle and genetic variability (WILLIAMS 1978; LISKA 1998). An alcoholic's lifestyle that causes damage to the liver is an example of a lifestyle that influences detoxification ability. Metabolites such as acetaldehyde and free radicals produced by the degradation of ethanol in the consumer's liver, can cause damage to the liver and ultimately cause inflammation. Ethanol consumption will inhibit the defence of the individual against free radicals by means of antioxidants, which will lead to tissue damage. Since the liver is the primary detoxification organ, damage will cause impaired detoxification ability (PACIFICI *et al.* 1990).

Metabolites derived from IEMs and xenobiotics can be toxic to the cells and thus have to be removed by transforming it to water-soluble substrates so that it can be eliminated via the urine or bile. Well known metabolic processes do not necessarily transform xenobiotics to water-soluble substrates and thus alternative processes and detoxification are needed. Cellular detoxification comprises an enzyme system that operates in two phases, namely phase I and phase II. In phase I the xenobiotic will be activated, which can result in highly reactive products that is often more toxic than the original compound and thus phase II is needed to deactivate the compounds formed in phase I. Active compounds can enter directly into phase II, while compounds from phase I should enter phase II as soon as possible due to the high toxicity of some of the compounds formed during phase I. In Figure 1.1 the two phases of detoxification are represented schematically (BIOMATRIX 2009). The general function of the detoxification enzyme system is to remove harmful xenobiotics from the body.

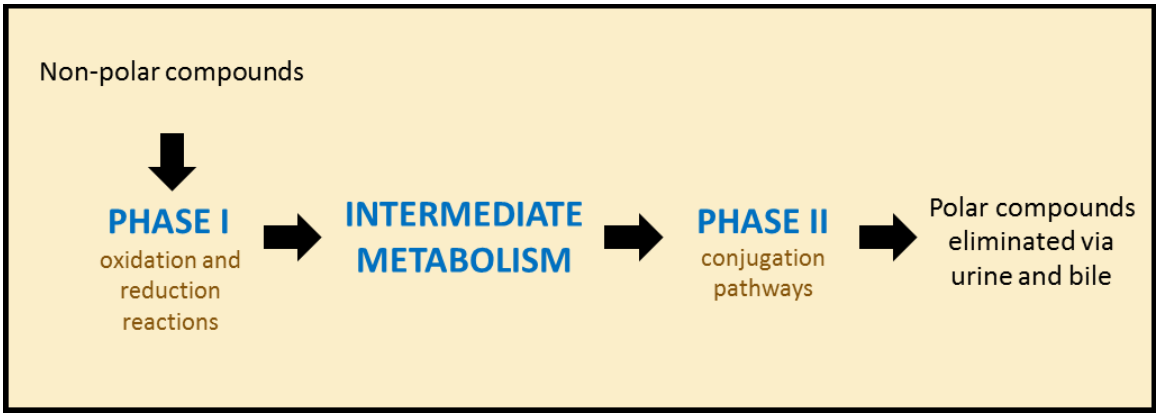

Figure 1.1: Schematic representation of phases I and II of detoxification. A reactive group is added to a non-polar substrate in phase I of detoxification by oxidation or reduction reactions to form an active compound. In phase II this active compound is conjugated to form a polar compound, which can then be eliminated via the urine or bile.

1.3.1 Phase I

In phase I of detoxification, a toxic compound is changed from a non-polar compound to a more polar compound. This is achieved, by the addition of a reactive group to the toxic compound. The reactive group enables the body to add a group to the toxic compound in the second phase of detoxification so that the compound becomes water-soluble and can be excreted. In some cases phase I is not needed, since some molecules are suited for phase II and thus these compounds can enter phase II directly. Typical phase I reactions include oxidation reactions (for example, dehalogenation, desulfuration, hydroxylation, deamination and sulfoxidation) and reduction reactions (for example, azo reduction, reductive halogenation, aromatic nitro reduction and aldehyde and ketone reduction) (LISKA 1998).

The major enzymes, of the approximately 50 enzymes needed for phase I detoxification to be functional, is the cytochrome P450 mono-oxygenase system, these include Cyp3A3, Cyp1A1, Cyp1A2, Cyp2D6 and Cyp2C (LISKA 1998; DORNE *et al.* 2004; XU *et al.* 2005). The cytochrome P450 enzymes add a reactive group (for example a hydroxyl group) to the toxic compound through the use of, amongst others, oxygen and NADH as a cofactor. This reaction often causes the compound formed to be more toxic than the original compound and can thus cause damage to DNA, RNA or proteins within the cell (LISKA 1998). The liver is the primary organ

where phase I detoxification takes place and as a result thereof the highest concentration of cytochrome P450 enzymes can be found within it. Other organs where detoxification takes place are the lungs and kidneys.

1.3.2 Phase II

In phase II of detoxification the active compound, formed during phase I, or other reactive metabolites directly entering phase II is conjugated to form a water-soluble substrate that can be excreted through the urine (compounds with low molecular weight) (LISKA 1998) and bile (compounds with high molecular weight) (HIROM *et al.* 1972).

The conjugation reactions that take place in phase II includes acetylating (where the substrate is conjugated with acetyl coenzyme A to form a mercapturic conjugate) and amino acid conjugation (a compound is conjugated with an amino acid and coenzyme A and carboxylic acid) (XU *et al.* 2005). The most common amino acid used for the amino acid conjugation reaction is glycine (ITO *et al.* 1995; KASUYA *et al.* 1996) and glycine N-acyltransferase catalyses part of this reaction. Other conjugation reactions include glucuronidation, glutathionation, methylation and sulfuridation (ITO *et al.* 1995).

1.3.3 Variability in the efficacy of detoxification

Variability in the efficacy of detoxification can occur between individuals, because the enzymes that catalyse detoxification reactions can be influenced by various factors. The factors leading to variability in detoxification between individuals includes the response of the individual to the environment, differences in lifestyle and genetic variations (NEBERT and FELTON 1976; LISKA 1998; MANCINELLI *et al.* 2000).

Variations in detoxification is not only limited to inter-individual differences, there can also be variations between species, such as the domestic cat (*Felis catus*) and pig (*Sus domestica*). The cat can conjugate phenol with sulphate whereas the pig

conjugates phenol with glucuronic acid because sulphate conjugation in the pig is defective (WILLIAMS 1978). These variations between species can also be an indication of variability within the same species.

In some cases the responses of individuals with the same IEM that receive identical treatment *e.g.* supplementation with the amino acid, glycine, will differ. This is suggestive of differences in the detoxification efficiency and more specifically the efficiency of glycine conjugation between these individuals. Variability in the efficacy of detoxification also occurs in healthy individuals since the tempo of phase I and II detoxification reactions can differ between individuals (DORNE *et al.* 2004).

The phenomenon of variability in the efficiency of detoxification between individuals raises an important question. The question being whether genetic variations in the glycine N-acyltransferase gene that leads to differences in glycine conjugation in phase II of detoxification contributes to the observed variability in detoxification efficacy.

1.4 Glycine N-acyltransferase

Glycine N-acyltransferase, EC 2.3.1.13, (GLYAT) is an enzyme which catalyses an amino acid conjugation reaction in phase II of detoxification. It uses acyl-coenzyme A to acylate its substrates, and thus it forms part of the GCN5-related *N*-acetyltransferase (GNAT) superfamily. All the members of the GNAT superfamily that have been structurally characterized have structural similarities in the form of a conserved fold (VETTING *et al.* 2005). This fold consists of an N-terminal strand followed by two helixes, three antiparallel β strands followed by a central helix, a β strand, a further α helix and a β strand at the C-terminal end of the structure (VETTING *et al.* 2005).

Three of the GLYAT family members, namely GLYAT, glycine N-acyltransferase like 1 (GLYATL1) and glycine N-acyltransferase like 2 (GLYATL2) are localized on chromosome 11 to the 11q12, Figure 1.2 (HAOXING *et al.* 2007). GLYAT and

GLYATL2 transcribes on the reverse strand, while GLYATL1 transcribes on the forward strand, as indicated by the arrows in Figure 1.2.

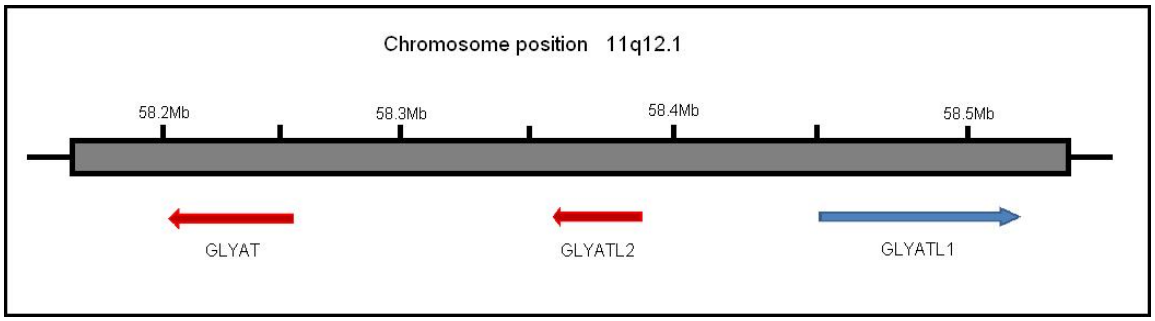

Figure 1.2: Members of the GLYAT family mapped to chromosome 11 at position 11q12.1. GLAYT as well as GLYATL2 transcribed on the reverse strand indicated by red arrows, whereas GLYATL1 is transcribed on the forward strand, indicated by the blue arrow.

The human *GLYAT* gene is located on chromosome 11 at position 11q12 (ENSEMBL 2009), and spans 23 215 base pairs. The gene consists of six exons and two isoforms of the GLYAT enzyme exist. *GLYAT* isoform A consists of all six exons of the *GLYAT* gene and has a transcript length of 2052 base pairs. *GLYAT* isoform B on the other hand consist of only five exons as can be seen in Figure 1.3 (ENSEMBL 2009). Isoform B has a transcript length of 1147 base pairs and a translation length of 163 AA residues, whereas isoform A has a translation length of 296 AA residues. Only the enzyme encoded for by Isoform A has enzyme activity and the typical GNAT fold at the C-terminal part of the protein.

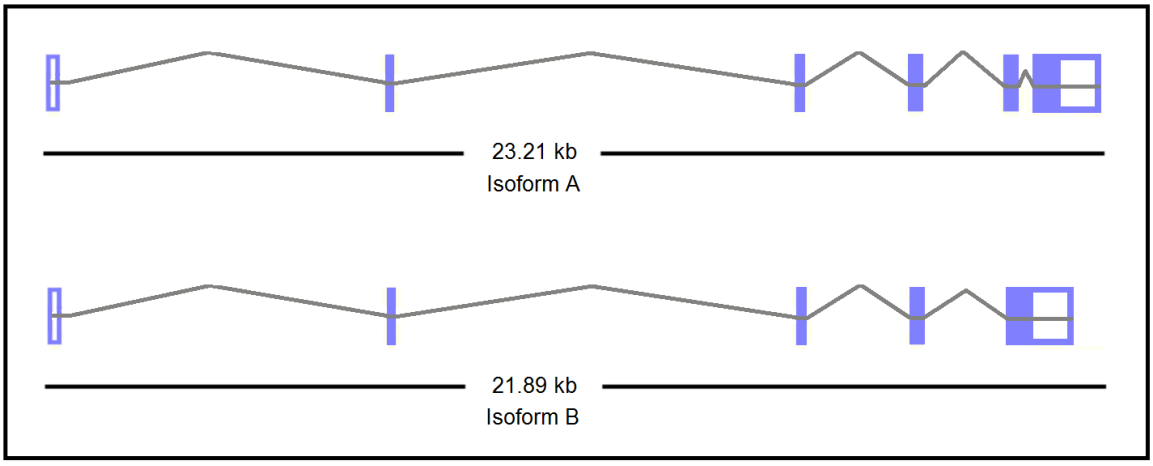

Figure 1.3: Schematic representation of *GLYAT* isoforms A and B: The blue rectangles represent the exons of the *GLYAT* gene and the grey lines represent the introns. The open blue blocks represent un-translated regions (UTR).

As of February 2013, on www.ensembl.org, 668 known single nucleotide polymorphisms (SNP) have been reported within the *GLYAT* gene (BADENHORST *et al.* 2013). To date, no link has been established between the SNPs identified within the *GLYAT* gene and the rate of conjugation with glycine. It was recently found that the SNPs identified within the *GLYAT* gene may have an effect on enzyme activity (VAN DER SLUIS *et al.* 2013). During the study conducted by van der Sluis et al., six known variations in the human *GLYAT* gene were investigated, which included K16N, S17T, R131H, N156S, F168L, and R199C. The K16N, S17T and R131H variations had similar enzyme activity as the wild type GLYAT, whereas N156S had increased and F168L had lower GLYAT activity compared to the wild type.

As previously stated the glycine *N*-acyltransferase enzyme catalyses an amino acid conjugation reaction in the second phase of detoxification. The GLYAT enzyme catalyses a reaction where an acyl-coenzyme A substrate is conjugated to glycine (LISKA 1998; VAN DER WESTHUIZEN *et al.* 2000). This reaction forms two products, coenzyme A (CoA) and N-acylglycine, which is water-soluble and can thus be excreted via the urine or bile. An example of a typical reaction catalysed by GLYAT is shown in Figure 1.4 (ITO *et al.* 1995). In this Figure the formation of isovaleryl-coenzyme A from isovaleric acid is catalysed by Coenzyme A ligase. The isovaleryl-coenzyme A acts as a substrate for GLYAT and is conjugated with glycine to form isovalerylglycine (ITO *et al.* 1995; KASUYA *et al.* 1996; BADENHORST *et al.* 2013).

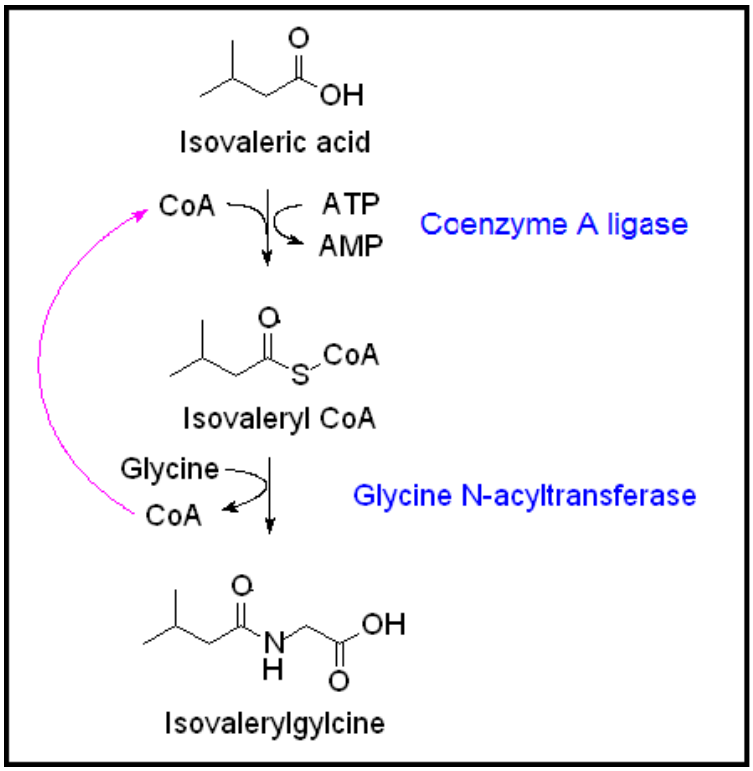

Figure 1.4: Conjugation of isovaleric acid to isovalerylglycine catalysed by glycine N-acyltransferase (EC 2.3.1.13). Isovaleryl CoA is formed from isovaleric acid. Isovaleryl CoA is a substrate for GLYAT which conjugates the substrate with glycine to form isovalerylglycine. This is a two-step reaction where the first step is catalysed by Coenzyme A ligase (EC 6.2.1.3) and the second step is catalysed by Glycine N-acyltransferase.

1.5 Sequencing

1.5.1 Sanger Sequencing

The Sanger sequencing technique is based on chain termination. The technique requires single-stranded DNA (ssDNA). The method initially required the sample to be split in to four reaction mixtures, to which dNTPs, primers and DNA polymerase is added. A single ddNTP is added to each reaction mixture respectively (ddATP to reaction one, ddTTP to reaction two, ddCTP to reaction three and ddGTP to reaction four). Once a ddNTP is built into a sequence, the DNA polymerase cannot extend or replicate the DNA further, leaving a fragment of specific size. Since the fragments have different sizes and each reaction contains a different ddNTP, the sequence can be determined (OBENRADER 2007). ddNTPs contain a hydrogen group on the 3' end,

while dNTPs contain a hydroxyl group. ddNTPs terminate the reaction because of a phosphodiester bond which cannot be formed with the next nucleotide on the amplicon (SANGER *et al.* 1977; OBENRADER 2007).

Technology has advanced since 1974 and the previous method is out of date (OBENRADER 2007). The new technology is based on the original Sanger sequencing method, but has become automated. The automated Sanger sequencing method is performed in one tube, which contains all the reagents needed for the reaction. Each of the ddNTPs is labelled with dyes which fluoresce at different wavelengths when excited by a laser. The fluorescence is measured by a camera which will ultimately determine the sequence. The results, after the fragments have been separated according to size, are illustrated in a electropherogram with peaks of different colours for the different nucleotides, starting from small to large fragments, or put another way beginning to end of the DNA to be sequenced (OBENRADER 2007).

1.5.2 Next Generation Sequencing

Next-generation sequencing involves non-electrophoretic methods to enable the generation of vast amounts of sequence information (HERT *et al.* 2008). The next generation sequencing instruments makes use of massively parallel DNA sequencing systems. The first of these instruments were designed and manufactured by 454 Life Sciences (MARDIS 2008), followed by Life Technologies (Applied Biosystems) as well as Illumina (HERT *et al.* 2008). The next generation technologies designed by these companies include the following, pyrosequencing, fluorescently labelled sequencing by synthesis and sequencing by hybridization and ligation (HERT *et al.* 2008).

Pyrosequencing is only one of the previously mentioned technologies which is a technique based on sequencing by synthesis. The sample DNA is fractionated into fragments of 300-800 bp. Adaptors, which will be used in the purification, amplification and sequencing steps, specific for the 3' and 5' ends are added to the fragments. The first four nucleotides of the adaptor serve as a calibration to enable

the instrument to estimate light emitted by a cascade of reactions (MARDIS 2008). One fragment will be bound to one bead during the library preparation by mixing the library fragments with agarose beads containing nucleotide sequences complimentary to the adaptors. The library will be emulsified with amplification reagents in an oil and water mixture (MARDIS 2008). In this mixture a microreaction will take place with each of the beaded library fragments and would result in millions of identical copies of each fragment. These fragments will be loaded on to a picotiter plate, whereupon enzyme containing beads will catalyse the reaction. Nucleotides will flow in an exact order across the wells and DNA capture beads whereupon a range of enzyme reactions will take place to produce a cheminumilescent signal, which will then be recorded by a CCD camera.

This technique encompasses sequencing by synthesis. A nucleotide is added by polymerase to extend DNA strands, this process leads to a light-generating reaction which the instrument will detect and can then determine the DNA sequence of each fragment and since the DNA fragments overlap the DNA sequence of the whole gene can be determined by aligning the fragments into contigs.

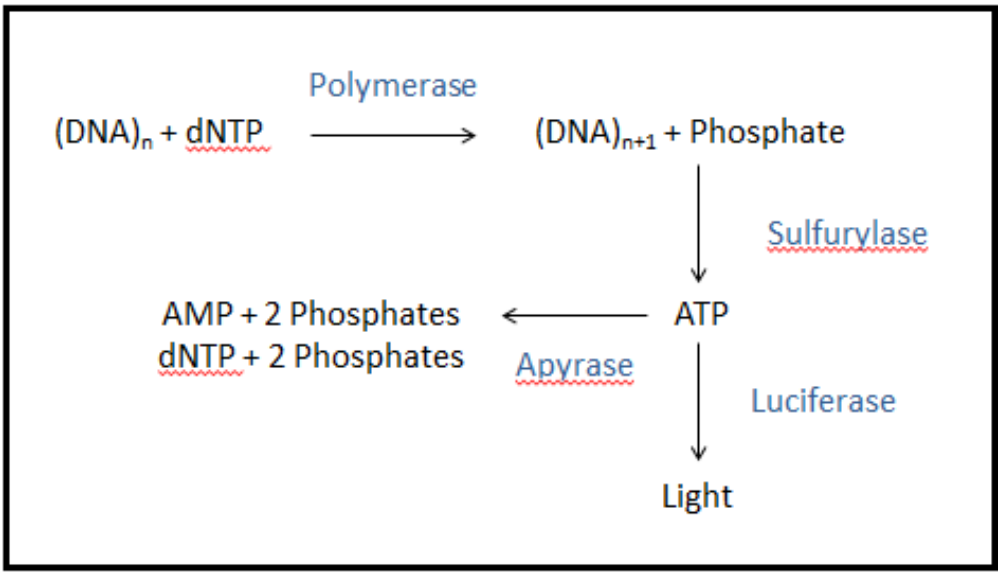

Figure 1.5: Biochemistry of pyrosequencing. The sequence by synthesis process involves incorporation of a dNTP by polymerase. A phosphate is released which is used by sulfurylase to form ATP. Luciferase then uses the ATP to produce light, which a CCD camera will detect. Apyrase breaks up the ATP into AMP and two phosphates and then forms a dNTP and two phosphates (RONAGHI *et al.* 1998).

The biochemistry of pyrosequencing, as shown in Figure 1.5 (RONAGHI *et al.* 1998), involves a PCR template that is hybridized to an oligonucleotide and incubated with DNA polymerase, ATP sulphurylase, luciferase and apyrase, whereupon a reaction will take place where dNTPs are added and will release a pyrophosphate. The pyrophosphate will be converted to ATP by ATP sulphurylase, which will drive the luciferin-mediated conversion of luciferin to oxyluciferin to generate light which is the cheminumilescent signal detected (RONAGHI *et al.* 1998; MARGULIES *et al.* 2005; WILSON and WALKER 2007). A advantage of pyrosequencing is that it does not make use of hybridization techniques and that an average of 100 million bases can be sequenced in a single 7.5h run (HERT *et al.* 2008).

The Illumina Genome Analyzer library preparation takes place on the surface of an oligo-derivatized flow cell. Bridge amplification takes place on the flow cell and uses DNA polymerase to amplify approximately a million DNA copies (QUAIL *et al.* 2012). The Illumina Genome Analyzer utilizes sequence by synthesis (QUAIL *et al.* 2012) which is similar to the pyrosequencing techniques, but still makes use of hybridization (METZKER 2010). A base-unique fluorescence label is carried by each nucleotide together with a chemical group that blocks the 3'-OH group. After a nucleotide is washed across the flow cell an image is taken of each cluster on a flow cell lane. The instrument can thus detect which nucleotide is incorporated at each step. The imaging step is followed by removing the chemical group blocking the 3'-OH group to prepare the amplicon for the next nucleotide incorporation (MARDIS 2008).

Solid Sequencer from Life Technologies is another one of the available next-generation sequencers. The Solid Sequencer, like many other next-generation sequence platforms use adapter-ligation fragmented libraries. The Solid system makes use of emulsion PCR and magnetic beads, similar to the techniques used by Roche/454 pyrosequencing, to amplify the amplicons before sequencing take place. The emulsion PCR is followed by deposition onto a flow cell slide, along with DNA ligase and fluorescent eight-mers, with a fluorescent group attached to the 4^{th} and 5^{th} nucleotides. A ligase reaction takes place and the fluorescence can be detected. The ligation step is followed by a regeneration step where the eight-mers along with the fluorescent groups are removed. Ligation and regeneration will be repeated until a sequence of 25 – 35 bp are obtained for each fragment (MARDIS 2008).

The Ion Torrent, one of the newest Next Generation Sequencing techniques, makes use of non-optical sequencing techniques, where a proton is released upon base incorporation during synthesis (QUAIL *et al.* 2012). An ion-sensitive field-effect transistor sensor will detect the proton, which is a hydrogen ion released during PCR by DNA polymerase (ROTHBERG *et al.* 2011). The DNA libraries for sequencing are prepared by fragmenting DNA and adding adaptors by a ligation reaction. This is then added to a magnetic bead. The primers used for sequencing and DNA polymerase are bound to the library and added to the ion chip. Each sensor well on the chip will contain only one bead. Nucleotides will be added to the Ion chip in an order and in the event that a nucleotide can be incorporated in to the sequence, a hydrogen ion will be released by the DNA polymerase. This will cause the pH to increase proportional to the number of nucleotides incorporated into the sequence, which can then be detected by the sensor (ROTHBERG *et al.* 2011). The Ion Torrent has a sequence yield of up to one Gb per run, with a read length of approximately 200 bases (QUAIL *et al.* 2012).

Next generation sequencing has advantages as well as limitations. These technologies can determine a vast number of DNA strands at the same time, thus millions of reads can be generated in a single experiment. Pyrosequencing enables one to sequence 100 million DNA bases within a 7.5 hour run with a 95 % base accuracy (HERT *et al.* 2008), which is the longest sequence reads obtained compared to other next generation techniques, such as Illumina Genome Analyzer and Solid Sequencer (HURD and NELSON 2009). Since the technology produces parallel sequence data a reference genome is helpful with data analysis, but a reference genome is not essential (HURD and NELSON 2009). Another advantage of pyrosequencing is that this technology does not make use of hybridization and thus do not have experimental bias due to hybridization such as the Solid Sequencer and Illumina Genome Analyzer (HERT *et al.* 2008; HURD and NELSON 2009). Pyrosequencing was thus chosen as the method of sequencing, since high throughput sequences can be obtained, but without the bias introduced due to hybridization. A huge limitation of next generation sequencing is the fact that it cannot detect more than six homopolymeric bases in length (MARDIS 2008), whereas electrophoresis-based platforms such as Sanger sequencing have the ability to accurately detect homopolymeric regions (HERT *et al.* 2008; SCHUSTER 2008; HURD

and NELSON 2009). Thus one should verify homopolymeric region data generated by next generation sequencing with Sanger sequencing.

1.6 Hypothesis

The clinical presentation of individuals with identical IEMs can differ from one another. This can be true even if these patients receive identical treatment, such as glycine supplementation. The detoxification of healthy individuals also shows differences indicating it is not just a problem limited to individuals with an IEM.

I hypothesise that sequence variations found in the *GLYAT* gene can potentially have an effect on glycine N-acyltransferase activity and might play an important role in the observed variability of detoxification efficiency. Van der Sluis et al. (2013) found that some variations have an effect on GLYAT activity (VAN DER SLUIS *et al.* 2013) and thus I believe that there may be more variations in the *GLYAT* gene that might influence GLYAT activity.

1.7 Aims

The aim of this study was to identify sequence variations in the *GLYAT* gene such as known SNPs, novel variations and CNVs by:

1. Amplification and pyrosequencing, followed by verification with Sanger sequencing, of the *GLYAT* gene in a cohort with possible low GLYAT activity based on the biochemical detoxification profile of each individual.
2. Identification of sequence variations in the *GLYAT* gene such as known SNPs and novel variations using DNAStar software (Madison, Wisconsin, USA).
3. Real time PCR based copy number assays to identify CNVs using the CopyCaller software (Life Technologies).

CHAPTER 2 – *GLYAT* GENE AMPLIFICATION OF A COHORT WITH A LOW GLYCINE DETOXIFICATION PROFILE

2.1 Introduction

Glycine N-acyltransferase (GLYAT) is an enzyme which catalyses an amino acid conjugation reaction in phase II of detoxification. The GLYAT enzyme catalyses a reaction where an acyl-coenzyme A substrate is conjugated to glycine (LISKA 1998; VAN DER WESTHUIZEN *et al.* 2000). Van der Sluis et al. (2013) found that some variations have an effect on GLYAT activity (VAN DER SLUIS *et al.* 2013). Six known variations in the human *GLYAT* gene were investigated, which included K16N, S17T, R131H, N156S, F168L, and R199C. Site-directed mutagenesis was used to create six variants of GLYAT. The six variants were expressed and purified after which they were enzymatically characterized. The K16N, S17T and R131H variations had similar enzyme activity as the wild type GLYAT, whereas N156S had increased and F168L had lower GLYAT activity compared to the wild type. Variations could be identified in the *GLYAT* gene, which could have an impact on drug response or increase susceptibility to diseases. A correlation between detoxification efficiency and variations in the *GLYAT* gene might enable personalized medicine for smaller subpopulations with the same phenotype and specific genetic profiles (MANCINELLI *et al.* 2000).

An individuals' biochemical detoxification profile can be determined by substrate loading tests. The Phase II detoxification profile is determined by the admission of paracetamol and salicylic acid. Paracetamol is used to determine the glucuronide-, sulphate- (ALBERTI *et al.* 1999; COURT *et al.* 2001) and glutathion conjugation (HADERSLEV *et al.* 1998), while salicylic acid is used to determine glycine conjugation (KUEHL *et al.* 2006).

The main aim of this study is to identify sequence variations in the *GLYAT* gene by means of sequencing and copy number analysis. In order to identify variations the *GLYAT* gene needs to be amplified by means of polymerase chain reaction (PCR). The experimental approach was to amplify and gel purify the *GLYAT* gene of each

participant in four fragments, between 2364 bp and 7829 bp in size. The purified fragments were to be used to identify possible variations in the *GLYAT* gene by means of pyrosequencing. Only a limited number of MIDs were available and thus the four fragments of the individuals were pooled into eight samples. The GLYAT gene of each of the participants were amplified in 29 smaller fragments, gel purified and Sanger sequenced for verification of the pyrosequencing data.

2.2 Materials and Methods

2.2.1 Participant selection

After ethical approval (NWU-0096-08-A1) was obtained, individuals were selected for this study. The selection of individuals for inclusion in the study was based on the biochemical detoxification profile of each participant obtained from Mr E. Erasmus and Dr. C. Mels (Table 2.1). Only individuals with detoxification profile abnormalities in phase II of detoxification were assessed for participation in this study. Abnormalities in phase I of detoxification do not serve as an indication of GLYAT variability, as GLYAT forms part of phase II of detoxification. An individuals' biochemical detoxification profile can be determined by substrate loading tests. The Phase II detoxification profile is determined by the administration of paracetamol and salicylic acid. Paracetamol is used to determine the glucuronide-, sulphate- (ALBERTI *et al.* 1999; COURT *et al.* 2001) and glutathion conjugation (HADERSLEV *et al.* 1998), while salicylic acid is used to determine glycine conjugation (KUEHL *et al.* 2006). The detoxification profile lists parameters in phase II indicative of glycine-, glucuronide-, sulphate- and glutathion conjugation. The primary selection criterion was altered glycine conjugation. If one or two of the other conjugation reactions were altered, participants were still considered, but participants exhibiting overall low second phase conjugation reactions were not considered as this is an indication of an energy-related problem, rather than an enzyme-related problem. Only eighteen participants were identified as possible candidates for this study. Unfortunately, only participants with low glycine detoxification ability were identified and none with high glycine detoxification ability. Since only participants with altered glycine detoxification ability were identified for this study, the probability of identifying variations in the *GLYAT* gene, which could possibly affect GLYAT activity, was increased.

Selected individuals were telephonically contacted and invited to participate in the study. They were given full details regarding the study and if they still agreed to participate they were asked to complete a questionnaire regarding their lifestyle, health and medication/supplements taken as well as sign an informed consent form.

Table 2.1: Glycine conjugation of participants is shown as % recovery of glycine conjugates after a salicylic acid loading test. The reference range is 30-52% for glycine conjugation. Glycine conjugation of each participant was evaluated and interpreted as low, normal or high.

Participant	Glycine conjugation (% recovery)	Glycine conjugation interpretation
1	21.8	Low
2	9.8	Low
3	12.0	Low
4	11.0	Low
5	30.7	Normal
6	31.1	Normal
7	42.8	Normal
8	22.0	Low
9	17.1	Low
10	31.6	Normal
11	14.2	Low
12	36.2	Normal
13	14.7	Low
14	31.3	Normal
15	33.3	Normal
16	N/A	N/A
17	27.7	Low
18	26.7	Low

2.2.2 DNA isolation from blood

Subsequent to informed consent being given, blood was obtained via venipuncture performed by a phlebotomist at the Potchefstroom branch of Lancet Laboratories. Three 4.5 ml ethylene diamine tetraacetic acid (K_2-EDTA) tubes of blood were obtained from each participant. The tubes with anticoagulated blood were centrifuged at 2500 x g for 10 minutes, after which the buffy coat of each participant was removed from the three tubes and pooled. The buffy coat consists primarily of white blood cells. DNA was isolated from leucocytes using a FlexiGene DNA kit (Qiagen Inc., California, USA).

The manufacturer's protocol for isolation of DNA from buffy coat was followed. In brief the protocol followed is described below:

1. Buffer FG1 was added to the leucocytes in order to breaks the cell membrane and release the cell nuclei and mitochondria.
2. FG2/QIAGEN protease was added to release the DNA from the cell nuclei and mitochondria by denaturisation of the membrane, while protease digests proteins.
3. DNA was precipitated by adding isopropanol. The isopropanol dehydrates the DNA and as a result the DNA precipitates.
4. The DNA was then washed with 70% ethanol to remove contaminants.
5. FG3 buffer, containing 10 mM Tris-HCl at pH 8.5, was added to rehydrate and resuspend the DNA.
6. The DNA concentration was determined using a Nanodrop spectrophotometer.

2.2.3 Polymerase chain reaction

To amplify the *GLYAT* gene, six primer sets were tested. We decided upon four of the six primer sets to amplify the *GLYAT* gene in four fragments. The six amplicons, primers used for generation of these amplicons and expected amplicon sizes are listed in Table 2.2. All primer used to amplify the six amplicons were obtained from previous work done within the research group, except the fragment 2 forward primer. This primer was designed using the NCBI Primer-BLAST tool. A reference sequence of the GLYAT gene was entered into the PCR template field, specifying the range of the forward primer and the reverse primer sequence. Primer sequences were generated by the NCBI Primer-BLAST tool, whereupon the best possible forward primer was chosen in combination with the existing reverse primer. The reaction mixtures for PCR of the six amplicons spanning the *GLYAT* gene have different compositions as can be seen in Table 2.3. In brief the major differences between the reaction mixtures of fragments one, two, three, and four were the concentrations of the template DNA, $MgCl_2$ and dNTP mix. The enzyme used for fragment one, three and four was TaKaRa Ex Taq DNA Polymerase (Takara Bio U.S.A., Madison, Wisconsin) and Phusion DNA polymerase (BioLabs inc., New Englang) for fragment two. The enzyme used for the 5'-1 and 3' amplicons was Kapa long range DNA polymerase (Kapa Biosystems (Pty) Ltd, Cape Town, South-Africa). The DNA polymerases used in this study were all high-fidelity enzymes. All reactions were set up in a total volume of 50 μl.

Table 2.2: Primer combinations and expected sizes of the six primer sets used in attempt to amplify the *GLYAT* gene for sequencing.

Amplicon	Primer name	Oligo-sequence	Expected size
5'-1 amplicon	L-Fwd	5'-atggtatttccatgacctgtgag-3'	12372 bp
	L-iRev	5'-tcaagagagccttgttaatctgc-3'	
3' amplicon	L-iFwd	5'-gttattgcaacaaaggctacgtg-3'	12897 bp
	L-Rev	5'-agataattaggccacaagaacgtc-3'	
Fragment 1	L-Fwd	5'-atggtatttccatgacctgtgag-3'	7829 bp
	Exon 1 Rev	5'-gtaaagagcagctaaactccactcatg-3'	
Fragment 2	Fragment 2 Fwd	5'-cgagtctcatttttctcttggatttgct-3'	9604 bp
	Exon 3 Rev	5'-gcctggctctaccatattg-3'	
Fragment 3	Exon 3 Fwd	5'-agtggttgtctgccctctgtg-3'	5110 bp
	Exon 5 Rev	5'-cattagatcccagcacacagg-3'	
Fragment 4	Exon 5 Fwd	5'-tagcaccaagcccagaacc-3'	2364 bp
	L-Rev	5'-agataattaggccacaagaacgtc-3'	

Table 2.3: Varying concentrations of the reaction mixtures used for amplification of the *GLYAT* gene in four fragments. Reaction mixtures for fragment one, two, three and four were used to amplify the *GLYAT* gene.

Reaction component	5'-1 amplicon primer combination	3' amplicon primer combination	Fragment 1 primer combination	Fragment 3 primer combination	Fragment 4 primer combination	Fragment 2 primer combination
$MgCl_2$	2.25 mM	2.25 mM	2.25 mM	2.25 mM	2.25 mM	0 mM
dNTP mix	0.3 mM	0.3 mM	0.25 mM	0.25 mM	0.25 mM	0.2 mM
DNA	1.5 ng/µl	1.5 ng/µl	1.75 ng/µl	1.5 ng/µl	1.5 ng/µl	1.5 ng/µl
Forward primer	0.5 µM	0.5 µM	0.3 µM	0.3 µM	0.3 µM	0.5 µM
Reverse primer	0.5 µM	0.5 µM	0.3 µM	0.3 µM	0.3 µM	0.5 µM
DNA polymerase	2.5 U Kapa long range DNA polymerase	2.5 U Kapa long range DNA polymerase	1.5 U TaKaRa Ex Taq DNA polymerase	1.5 U TaKaRa Ex Taq DNA polymerase	1.5 U TaKaRa Ex Taq DNA polymerase	1 UPhusion polymerase

In order to optimize the PCR the following variables were adjusted by means of several troubleshooting steps: primer concentration, DNA concentration and dNTP concentration. During the optimization the denaturation temperature was raised to 95.0˚C. This high temperature causes the hydrogen bonds of the double-stranded DNA (dsDNA) to denature and single-stranded DNA (ssDNA) to form, while also preventing secondary structures to form. The annealing temperatures were set at 60.1˚C for fragment 1, 63.1˚C for fragment two, 65.0˚C for fragment three and 59.7˚C for fragment four. These temperatures are dependent on the Tm values of the primers and were optimized for the primers to anneal to the ssDNA. The DNA polymerase will bind to the primer template hybrid whereupon the synthesis of dsDNA starts. The extension temperature used for the reactions of the *GLYAT* gene four-fragment amplification was 72.0˚C for six minutes and the extension time for the amplification of the two *GLYAT* fragments were 13 minutes at 72.0˚C. The denaturation, annealing and extension steps were repeated for 30 cycles during which the DNA fragments were amplified exponentially.

Data generated from pyrosequencing had to be verified with Sanger sequencing in order to eliminate possible artefacts generated during pyrosequencing. The *GLYAT* gene was divided into 29 smaller fragments to ensure that all the possible variations were included in 29 fragments. The NCBI Primer-BLAST tool was used to design the primer sets for the 29 short fragments. A reference sequence of the GLYAT gene was entered into the PCR template field, specifying the range of the forward and revers primers. Primer sequences were generated by the NCBI Primer-BLAST tool, whereupon the 29 primer sets were designed. The smaller fragment primers were designed so that the amplicon sizes did not exceed 850 bp in length since the average read length of Sanger sequencing is approximately 750 bp in length, although under ideal conditions read lengths in excess of 1000 bp can be obtained. Cycling conditions, primer concentration, $MgCl_2$ concentration and dNTP concentration were adjusted during optimization. DMSO was also added in concentrations of 2.5% and 5.0%, but had no effect on the reaction. The supposed effect of DMSO in a PCR is to increase the effectiveness of amplification. DMSO does this by destabilising the dsDNA, thus the bonds between base pairs, especially in GC rich regions, disrupts and ssDNA is formed (MASOUD *et al.* 1992). Reaction conditions for amplification of the smaller fragments are shown in Table 2.4. Reaction conditions were as follows, initial denaturation for three minutes at 94.0˚C,

followed by 30 cycles of denaturation for 30 seconds at 94.0˚C, annealing at primer specific temperatures as stipulated in Table 2.5 and extension at 72.0˚C for 48 seconds. A final extension for three minutes at 72.0˚C was performed after the 30 cycles. Primer sequences used for the amplification of the 29 fragments as well as the expected size of the amplicons are listed in Table 2.5.

Table 2.4: Concentrations of reaction mixtures for Sanger sequencing verification. Each PCR tube contained a reaction mixture of the concentrations listed in the Table below.

Reaction component	**Final concentration**	**Per 50 µl reaction**
PCR grade water	-	34.26 µl
10 X Taq Buffer	1X	5.00 µl
10 mM dNTP Mix	2.5 mM each dNTP	4.00 µl
Forward primer (10 µM)	0.5 µM	2.50 µl
Reverse primer (10 µM)	0.5 µM	2.50 µl
Template DNA (100ng/µl)	6ng/µl	3.00 µl
TaKaRa Ex Taq (1 U/µl)	0.25 U	0.25 µl

Table 2.5: The primer combinations, expected amplicon sizes and annealing temperatures for 29 fragments of the 23.21 kb *GLYAT* gene used for Sanger sequencing verification. Amplicon sizes range between 315 bp and 838 bp and annealing temperatures range from 49˚C to 56˚C.

Primer set name	Forward	Reverse	Size	Annealing temperature (˚C)
1	5'-acatgagtggagtttagctgctctt-3'	5'-aagagagaacaaaggcaaacagga-3'	698 bp	52.0
2	5'-tgagactgtggtcagattatgtaggc-3'	5'-ttggaaccaccctctaaggatgta-3'	621 bp	56.0
4	5'-ggtgcaattctaggactcacctttt-3'	5'-tgcactttagttcttttccttcttgc-3'	633 bp	53.0
5	5'-aaattacttgggtgtggtgttgc-3'	5'-cagtagacagagaacctaaaactggtga-3'	595 bp	53.5
6	5'-tcctgtccctatctgttggtacatt-3'	5'-ctgggttccaagattgacttttgt-3'	576 bp	53.0
7	5'-gaaaaagtagaaatgcagggacca-3'	5'-taattttggggtcttgtgtccatt-3'	785 bp	51.0
8	5'-caagaccccaaaattatgcctaca-3'	5'-aaagggcattctgacatggtctaa-3'	691 bp	52.0
9	5'-aatcccaatttgatgcagctagtc-3'	5'-aggaatgaatacctgcaatgtgct-3'	394 bp	52.0
11	5'-tgaaatttgcccagagttctaca-3'	5'-tcatttagggaatgggtaatttga-3'	342 bp	49.0
12	5'-catcagagatgcattgagacaaca-3'	5'-ttgaataaatagaatgggaggcaag-3'	815 bp	51.5
13	5'-atggctgactttatggctcaaact-3'	5'-aaataaccacgtagcctttgttgc-3'	802 bp	52.0
14	5'-atgtgttcatgtgtgcatgtttgt-3'	5'-gtttggggctgggttcatagtta-3'	838 bp	52.0
15	5'-acaaaagtgcttaaatgcgatactaaa-3'	5'-atccaattacattaccccaattatttt-3'	674 bp	49.0
16	5'-tggataaatatttgtaggcaaattggt-3'	5'-tgcaggagttaaagctctgtgatt-3'	611 bp	51.0
17	5'-atggaggagttgggcatgtaaga-3'	5'-attcatatttggcaatgccttctg-3'	581 bp	51.5
18	5'-gaattggtgtttggttttctttgg-3'	5'-tgaaaaagattggttcagtccatgt-3'	594 bp	50.5
19	5'-catggaaaattgttgccttgtgt-3'	5'-tccatgttaaatctggggcttct-3'	649 bp	51.0
20	5'-agaagccccagatttaacatgga-3'	5'-ggtcccaaaatctcatttttagca-3'	725 bp	50.5
21	5'-tctccctttcaccctattttctct-3'	5'-ttttagtcctactggaccccaaa-3'	667 bp	52.0
22	5'-ttccccacttcaagtattcatgc-3'	5'-agtgtccttatgattttggggttg-3'	750 bp	52.0
23	5'-aacaggatacagaaagaggccaaa-3'	5'-agcttatgcatttttatcttgaaatgtg-3'	714 bp	51.0
24	5'-gccaaaccatatcaactatgaaaagc-3'	5'-gaaataagccaaacacaggcagat-3'	481 bp	52.0
25	5'-aggttcaacataggggaagcaata-3'	5'-ttgcagttccagagaaacaaacaa-3'	427 bp	51.0
26	5'-tgttgaggatgatggtggtaagaa-3'	5'-tttccagttgatgagttctggtga-3'	639 bp	52.0
27	5'-aagaatgggcttcataaagggaac-3'	5'-cacccatcaactacagctacatgagta-3'	528 bp	54.0
28	5'-tcttccactggtcttatctgggttt-3'	5'-atatgaggcaatgacccatgattt-3'	408 bp	52.0
Exon 2	5'-cgagtctcattttctcttggatttgc-3'	5'-cagtccctttccctcctctttcac-3'	315 bp	55.0
Exon 5a	5'-tagcaccaagcccagaacc-3'	5'-ggaaagccagagtgaatgcag-3'	377 bp	52.0
Exon 6	5'-gacctctatgacacttcatcagataca-3'	5'-gattctcacagacaccaaatctgctg-3'	821 bp	56.0

2.2.4 Agarose gel electrophoresis

The four different *GLYAT* amplicons, used for pyrosequencing, were all separated on 0.7% (w/v) agarose gels, 10 cm in length. The 29 smaller fragments, used for verification with Sanger sequencing, were separated on 1% (w/v) agarose gels, also 10 cm in length. This low concentration of agarose yields a less-dense matrix which improves the separation of the bigger DNA fragments during electrophoresis. All gels were made with TAE buffer and 3.0 μl of a 10 mg/ml ethidium bromide was added prior to casting. The gels were run at a constant voltage of 100 V for an hour and visualized on a Chemi genius bio-imaging system (Syngene, Cambridge, UK).

O'GeneRuler DNA ladder mix (SM1173, Fermentas Inc., Maryland, USA) was used as a size marker during gel electrophoresis. The marker serves a multi-purpose, firstly it provides an indication of how far the fragments have migrated on the agarose gel, secondly it allows one to establish the approximate size of the amplicon or amplicons produced during PCR and thirdly it contains glycerol, which is more dense than the buffer and thus the DNA and loading dye mixture will sink to the bottom of the well when loaded onto the agarose gel. As indicated in Table 2.2, the expected sizes of the four fragments were fragment one, 7829 bp, fragment two, 9430 bp, fragment three, 5110 bp and fragment four, 2364 bp. The expected sizes of the 29 amplicons range between 315 bp and 839 bp and therefore O'Gene 1 kb DNA ladder (SM1163, Fermentas Inc., Maryland, USA) was used as a marker. O'Gene 1 kb DNA ladder range in size from 250 to 10 000 bp and is thus suited for use as a marker for all the amplicons.

2.2.5 Gel extraction

All the *GLYAT* amplicons of the participants were excised from the agarose gels during visualization on a Dark Reader Transilluminator (Clare Chemical Research, Inc., Dolores, USA) and placed into labelled 1.5 ml microcentrifuge tubes. The Dark Reader was used because it does not make use of UV-light. UV-light can damage the DNA and thus reduces the reliability of the results obtained. The four excised amplicons, amplified for pyrosequencing, were gel-extracted with the Nucleospin kit (Clontech Laboratories, Inc., Cape Town, South Africa) according to the protocol supplied by the manufacturer. In brief the Nucleospin kit protocol is as follows:

1. NT buffer was added to the excised gel, followed by heating at 50.0˚C for 10 minutes in order to dissolve the agarose gel.
2. The DNA fragments were then bonded to the membrane of the spin column.
3. NT3 was used to wash the silica membrane of the spin column in order to remove contaminants.
4. The DNA fragments were eluted in NE buffer. The NE buffer was preheated and added to the spin column, which was left for one minute before centrifugation to further increase the yield.
5. Concentrations of all the samples for pyrosequencing were determined using a Nanodrop fluorometer at Inqaba Biotech.

The 29 amplicons, for Sanger sequencing, were gel-extracted with the Zymoclean gel DNA recovery kit (Zymo Research Corporation, Irvine, U.S.A) according to the protocol supplied by the manufacturer. The protocol is briefly described below:

1. Weight of excised amplicons was determined followed by adding three volumes of the ADB buffer.
2. The excised amplicon and ADB buffer were placed on a heating block at 55.0˚C for 10 minutes in order for the agarose gel to dissolve in the ABD buffer.
3. Melted agarose gel and ABD buffer were transferred to a spin column to bind the DNA the membrane of the spin column.
4. The membrane of each spin column was washed with wash buffer. The wash step was performed in order to ensure that the eluted DNA was pure.
5. DNA fragments were eluted with molecular grade water.

2.3 Results and Discussion

2.3.1 Long-Range Polymerase Chain Reaction of the 23.21 kb *GLYAT* gene

2.3.1.1 Amplification of the *GLYAT* gene in two amplicons

Initial attempts were focussed towards generating two amplicons to span the *GLYAT* gene. Three sets of primers shown in Figure 2.1 were used in this attempt. Two primer sets were used in the attempt to amplify the 5' end of the *GLYAT* gene and another set of primers were used in the attempt to amplify the 3' end of the *GLYAT* gene.

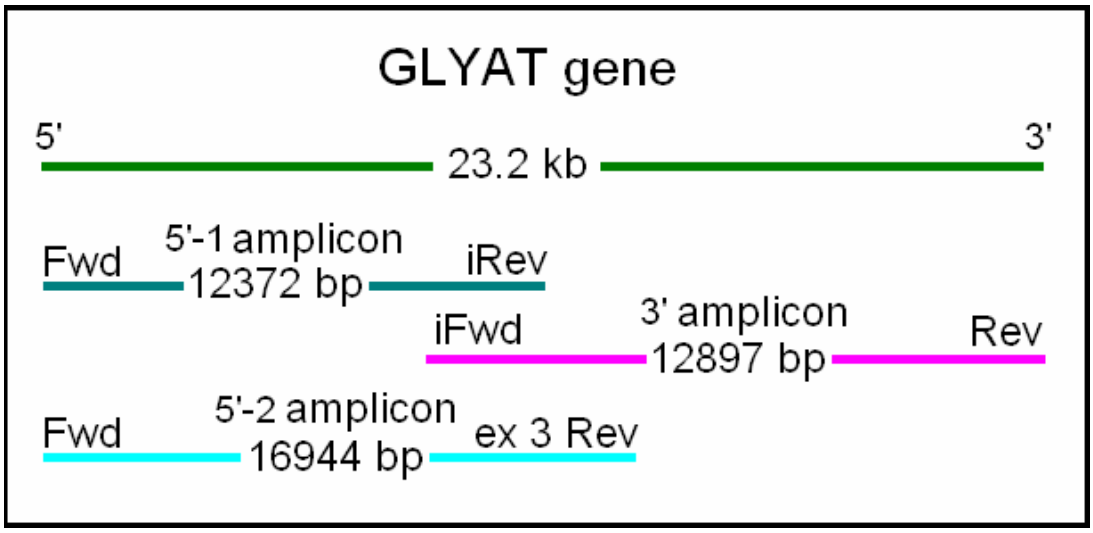

Figure 2.1: Schematic representation of the position of the three amplicons on the *GLYAT* gene. Two sets of primers were used in the attempt to amplify the 5' end of the *GLYAT* gene. The amplicons generated using the 5'-1 and 5'-2 primer sets, were expected to be 12372 bp and 16944 bp in size respectively. The amplicon generated using the 3' primer set was expected to be 12897 bp in size. The 5' amplicons overlap with the 3' amplicon in the middle of the *GLYAT* gene to ensure continuity of the obtained sequence.

Figure 2.2 illustrates the typical results that were obtained while trying the two-amplicon (12-13 kb) approach to amplify the 23.21 kb *GLYAT* gene. Kapa Long Range DNA polymerase and TaKaRa Ex Taq DNA polymerase (Takara Bio U.S.A., Madison, Wisconsin) were compared. These long-range PCR enzymes contain a polymerase and a proofreading enzyme, whereas ordinary PCR enzymes only

contain DNA polymerase. The post-PCR reaction analysis was carried out using agarose gel electrophoresis (Figure 2.2). Amplicons of 12372 bp in size were expected in lane one and three and a 12897 bp amplicons were expected in lane two and four. Relatively high molecular weight smears were observed in lane one and two (Figure 2.2). Faint smears of lower molecular weight were also observed as indicated with the arrows in Figure 2.2 lanes three and four. The smears indicated that the Kapa DNA Polymerase had amplification of high yield whereas the TaKaRa Ex Taq also had non-specific amplification, but of lower molecular weight amplicons as well as low amplification yield for these specific reaction conditions. Due to the higher yield and size of the amplicons it was decided that Kapa DNA Polymerase would be the better option for these specific reactions with optimized reaction conditions, for future application. Smears could also be indicative of DNA damage and thus the next step was to compare the previously isolated DNA to freshly isolated DNA. DNA integrity is very important in long-range PCR. Freeze-thaw cycles can cause nicks in the DNA that prevents successful amplification of longer products. A number of factors such as spin columns and excessive vortexing can lead to DNA damage during isolation. The damage incurred may also hinder successful amplification of the isolated DNA.

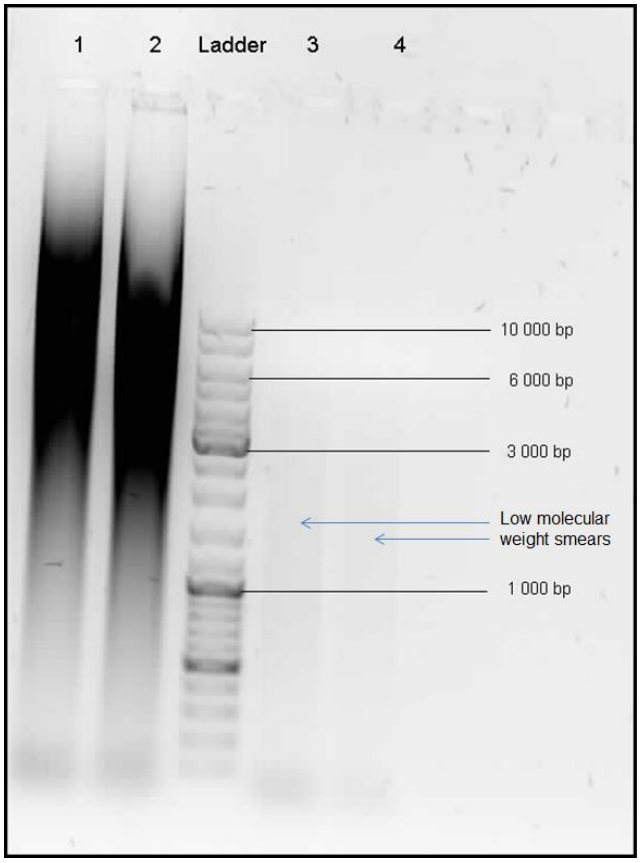

Figure 2.2: Comparison of GLYAT gene amplicons generated by Kapa Long Range and TaKaRa Ex Taq DNA Polymerases on a 0.7% agarose gel. Lanes: 1) 5'-1 amplicon synthesised, using Kapa Long Range DNA Polymerase; 2) 3' amplicon synthesised using Kapa Long Range DNA Polymerase; 3) 5'-1 amplicon synthesised using TaKaRa Ex Taq; 4) 3' amplicon synthesised using TaKaRa Ex Taq. The ladder used was O'Gene Ruler DNA Ladder Mix (Thermo Scientific) and the gel was run at a constant voltage of 100 V for an hour.

2.3.1.2 Template DNA integrity assessment

The PCR reactions of the 5'-1 amplicon and 3' amplicon were repeated using Kapa Long Range DNA polymerase, however duplicates using previously isolated DNA and freshly isolated DNA were set up. An image of the agarose gel of electrophoresed PCR products is shown in Figure 2.3. The freshly isolated DNA was successfully amplified in both instances (lanes one and two), but the previously isolated DNA gave poor amplification (lanes three and four). The freshly isolated DNA was isolated the week preceding the experiment whereas the previously isolated DNA was stored at 4˚C for more than a month. The result of this experiment clearly indicates the importance of template DNA integrity in long range PCR. The results with the freshly isolated DNA were unfortunately not reproducible and thus the

primers also had to be tested. Primers can also be damaged by freeze-thaw cycles and thus the next step was to test the integrity of the primers by using different primer combinations.

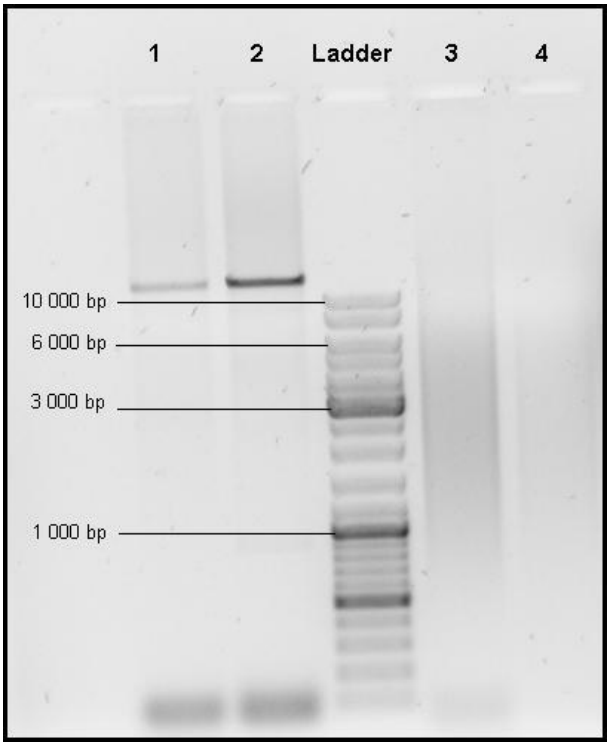

Figure 2.3: Template DNA comparison of freshly isolated DNA and DNA stored at 4˚C for >1 month on a 0.7% agarose gel. Lanes: 1) 5'-1 amplicon synthesised with freshly isolated DNA; 2) 3' amplicon synthesised with freshly isolated DNA; 3) 5'-1 amplicon synthesised with previously isolated DNA that was stored at 4˚C for >1 month; 4) 3' amplicon synthesised with previously isolated DNA that was stored at 4˚C for >1 month. The ladder used was O'Gene Ruler DNA Ladder Mix (Thermo Scientific) and the gel was run at a constant voltage of 100 V for an hour.

2.3.1.3 Primer integrity assessment

An assessment of primer integrity was performed by combining some of the available primers used for amplification of the *GLYAT* exons to yield amplicons with expected sizes of 16944 bp and 1996 bp. In lane one of the agarose gel (Figure 2.4) the 3' amplicon was included since amplification with greater apparent yield than the 5' amplicon was observed previously. The primer set used for amplification of the 3' amplicon were freshly diluted from stock to minimise possible primer degradation as

a result of freeze-thaw cycles. In lane two of the agarose gel, the exon one forward and exon three reverse primers were used in combination to generate an amplicon with expected size of 16944 bp. An amplicon of low yield, but in the correct size range was observed. In lane three, Figure 2.4, the exon two forward and exon four reverse primers were used and an amplicon of 1996 bp was expected. However as is evident from the assortment of amplicon sizes these primers annealed non-specifically and the expected product was not evident on the agarose gel. Unfortunately the amplification of the 3' amplicon and 5' amplicon were not reproducible, but the primer integrity assessment indicated that the primers used were sufficiently intact to anneal. The next step was to test whether the polymerase had sufficient activity. This was done by amplification of smaller fragments using primers for the amplification of the *GLYAT* exons with conditions previously optimized.

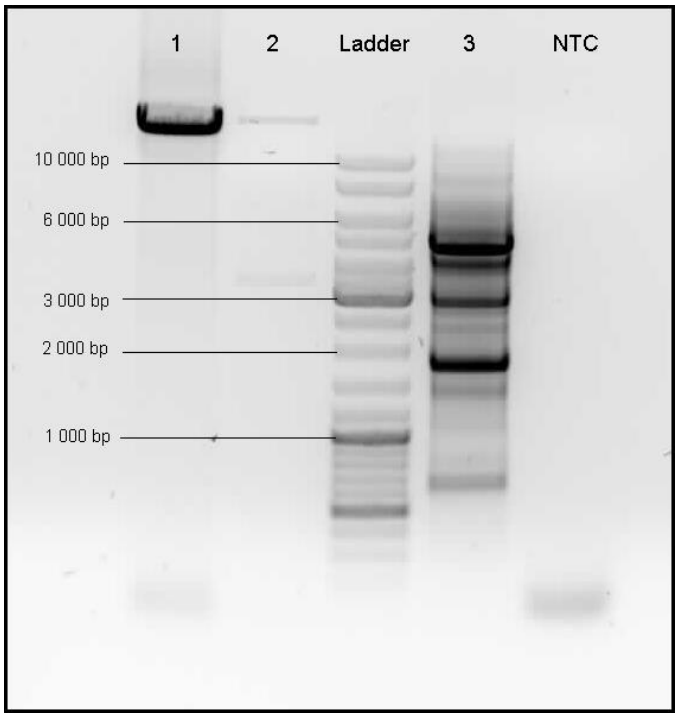

Figure 2.4: *GLYAT* exon primer combinations with Kapa Long Range DNA polymerase on a 0.7% agarose gel. Lanes: 1) 3' amplicon; 2) 16 944 bp amplicon synthesised with the exon one forward and exon three reverse primers; 3) Non-specific amplicons synthesised with the exon two forward and exon four reverse primers. The ladder used was O'Gene Ruler DNA Ladder Mix (Thermo Scientific) and the gel was run at a constant voltage of 100 V for an hour.

2.3.1.4 Testing the activity of Kapa Long Range DNA polymerase and Platinum Taq DNA polymerase

Two high-fidelity polymerases, Kapa Long Range DNA polymerase and Platinum Taq DNA polymerase (Invitrogen Corporation), and exon primer combinations yielding smaller amplicons were compared as shown in Figure 2.5. In lanes one and two of the agarose gel an amplicon of approximately 480 bp can be seen which were produced with Kapa Long Range DNA Polymerase and Platinum Taq respectively. The primers used to produce the amplicons loaded in lane one and two were the long-forward and exon one reverse primers. In lane three and four an expected amplicon of 394 bp was observed, produced by Kapa Long Range DNA polymerase and Platinum Taq DNA polymerase respectively. Primers used for the amplification of amplicons loaded in lanes three and four were exon one forward and exon one reverse. These results indicated that the DNA polymerases were still functional and that the Kapa Long Range DNA polymerase had a slightly higher apparent amplicon yield than the Platinum Taq DNA polymerase (compare Figure 2.5 lanes one and two). Since it was clear that the enzyme and primers were still working, the next step was to change the primer and dNTP concentrations, as well as to use an aliquot of the primers. The primer aliquots did not go through any freeze-thaw cycles.

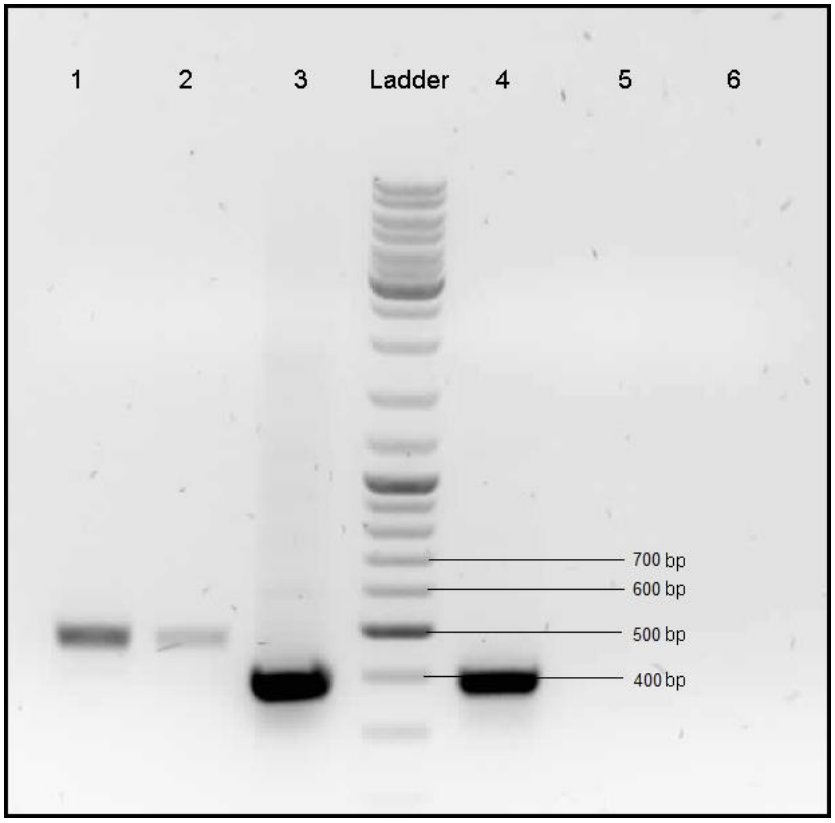

Figure 2.5: Primer combinations synthesised with Kapa Long Range DNA polymerase and Platinum Taq DNA polymerase. Lanes: 1) 480 bp synthesised with long-forward and exon one reverse primers and Kapa Long Range DNA polymerase; 2) 480 bp synthesised with long-forward and exon one reverse primers and Platinum Taq DNA polymerase; 3) 394 bp amplicon synthesised exon one forward and exon one reverse primers and Kapa Long Range DNA polymerase; 4) 394 bp amplicon synthesised exon one forward and exon one reverse primers and Platinum Taq DNA polymerase; 5) NTC using Kapa Long Range DNA polymerase; 6) NTC using Platinum Taq DNA polymerase. The gel used consisted of 1% agarose and was electrophoresed for an hour at 100 V.

2.3.1.5 Varying primer and dNTP concentrations

In previous PCRs, concentrations of 0.5 μM for primers and 0.3 μM for dNTPs were used. Primer and dNTP concentrations were changed to 0.3 μM for primers and 0.35 μM for dNTPs and used in different combinations of the primer and dNTP concentrations, in order to see whether the different reaction conditions would have an effect on amplification. Figure 2.6 portrays the analysis of the amplicons of these reactions electrophoresed on a 0.7% agarose gel. In lane one, two and three, Figure 2.6, a combination of 0.3 μM primers and 0.35 μM dNTPs of the 5'-1, 5'-2 and 3' amplicons respectively were loaded. This combination had showed no amplification for the 5' amplicons and was therefore considered not to be the optimum conditions

for these reactions. The 3' amplicons amplified with a reasonable yield. The reaction for the 5'-2 amplicon with a primer concentration of 0.3 μM and dNTP concentration of 0.3 μM was loaded in lane four, a smear was observed with a faint band indicative of an amplicon of the correct size. This indicated that these conditions were not optimal. In lane five, Figure 2.6, the reaction for 5'-2 amplicon was loaded again with only the dNTP concentration altered from the abovementioned conditions. An amplicon of the correct size can be seen in lane five. The reaction mixtures for the 3' and 5'-2 amplicons were changed to 0.3 μM primers and 0.35 μM dNTPs, since these concentrations appeared to be optimal for the reactions. Further optimization was done by combinations with altered concentrations of primers, template DNA and dNTPs.

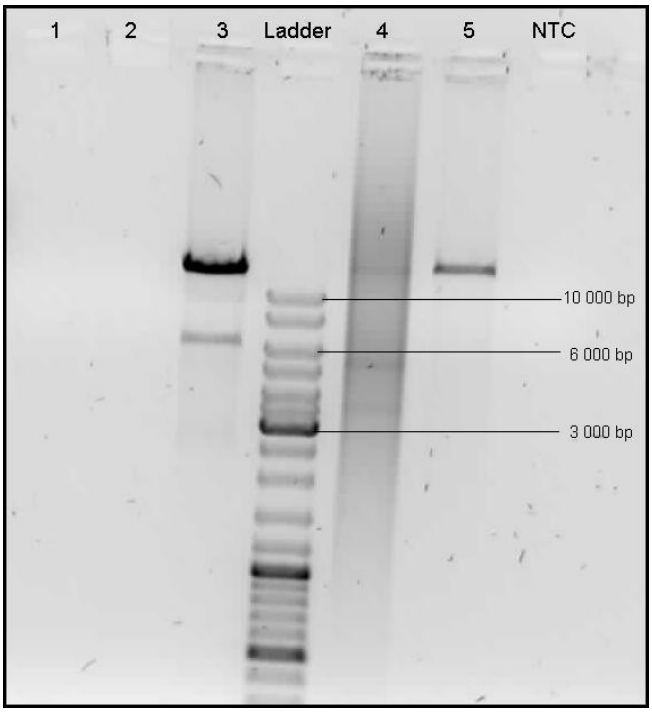

Figure 2.6: Amplification of the GLYAT gene in 2 fragments with varying primer and dNTP concentration combinations. Lanes: 1) 5'-1 amplicon synthesised with of 0.3 μM of primers and 0.35 μM of dNTPs; 2) 5'-2 amplicon synthesised with of 0.3 μM of primers and 0.35 μM of dNTPs; 3) 3' amplicon synthesised with of 0.3 μM of primers and 0.35 μM of dNTPs; 4) 5'-2 amplicon synthesised with 0.3 μM of each primer and 0.3 μM of dNTPs; 5) 5'-2 amplicon synthesised with 0.5 μM primers and 0.35 μM of dNTPs. The NTC set up with the 3' primer set was loaded the lane labelled NTC. The agarose gel (0.7%) was electrophoresed for an hour at 100 V.

2.3.1.6 Varying concentrations of primers, template DNA and dNTPs

Combinations of 1.65 ng/µl DNA, 0.35 µM primers and 0.4 µM dNTPs were made with the previous PCR conditions of 1.5 ng/µl DNA, 0.3 µM primers and 0.35 µM dNTPs. The analyses of these reactions are illustrated in Figure 2.7. The changed concentrations were not optimal for the 5'-2 amplicon, but in lane six the combination of increased DNA, primers and dNTPs gave an amplicon of very low yield. In lane eight the positive control amplicon of approximately 12897 bp in size, synthesised with the 3' primer set, can be observed. The results obtained with the 5'-1, 5'-2 and 3'- amplicons, unfortunately proved not to be consistently reproducible and thus it was decided to rather amplify the *GLYAT* gene in four smaller amplicons. The four fragment approach was done by using a combination of primers used to amplify the exons of the *GLYAT* gene and primers used for amplification of the *GLYAT* gene in two fragments.

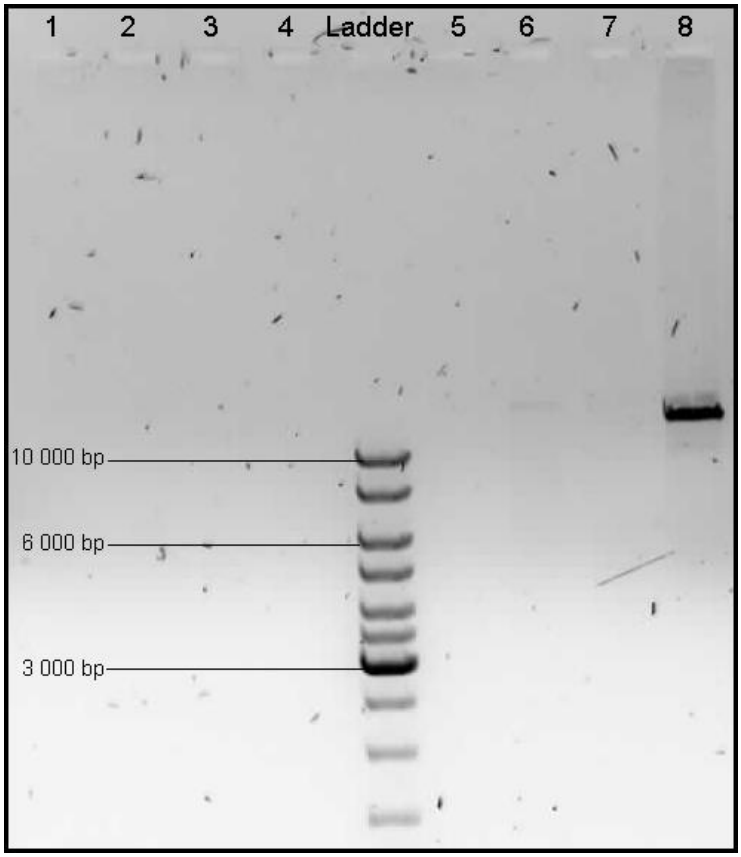

Figure 2.7: Amplification of the *GLYAT* gene with combinations of Template DNA-, primer- and dNTP concentrations. Lanes: 1) 5'-2 amplicon synthesised with altered concentration of 1.65 ng/µl template DNA; 2) 5'-2 amplicon synthesised with altered concentration of 0.35 µM primers; 3) 5'-2 amplicon synthesised with altered concentration of 0.4 µM dNTPs; 4) 5'-2 amplicon synthesised with altered concentrations of 1.65 ng/µl template DNA and 0.4 µM dNTPs; 5) 5'-2 amplicon synthesised with altered concentrations of 0.35 µM primers and 0.4 µM dNTPs; 6) 5'-2 amplicon synthesised with altered concentrations of 1.65 ng/µl template DNA, 0.35 µM primers and 0.4 µM dNTPs; 7) 5'-2 amplicon synthesised with altered concentrations of 1.5 ng/µl template DNA, 0.3 µM primers and 0.35 µM dNTPs; 8) Positive control 3' amplicon. The agarose gel (0.7%) was electrophoresed for an hour at 100 V.

2.3.1.7 Amplification of *GLYAT* in four amplicons

The four smaller amplicons were amplified using combinations of the available exon primers and forward and reverse primers of the 5'-1 and 3' amplicons respectively, as illustrated in Figure 2.8. TaKaRa Ex Taq DNA polymerase was used for fragments one, three and four and Phusion DNA polymerase was used for fragment two. Combinations of L-forward and exon two reverse (fragment one), exon two forward

and exon three reverse (fragment 2_1), newly designed fragment two forward and exon three reverse (fragment two), exon three forward and exon five reverse (fragment three) and exon five forward and L-reverse (fragment four) was used to amplify the *GLYAT* gene in four overlapping amplicons. Fragment 2_1 amplification was unsuccessful and thus a new fragment two forward primer was designed. Fragment two amplified successfully with the fragment two forward and exon three reverse primers as described in section 3.2.1.8. In lane one and two, of the analysis on a 1% agarose gel shown in Figure 2.9, no apparent amplification took place. In lane one the reaction for fragment one was loaded. The expected size of fragment one was 7829 bp. In lane two, Figure 2.9, the reaction for fragment two was run and the expected size of this fragment was 9430 bp. In lane three an amplicon of approximately 5110 bp was observed and in lane four an amplicon of approximately 2364 bp was visible. The primers used to amplify fragments one and two had Tm's that differed by more than 4.0˚C and therefore it was decided to repeat the two reactions at different annealing temperatures.

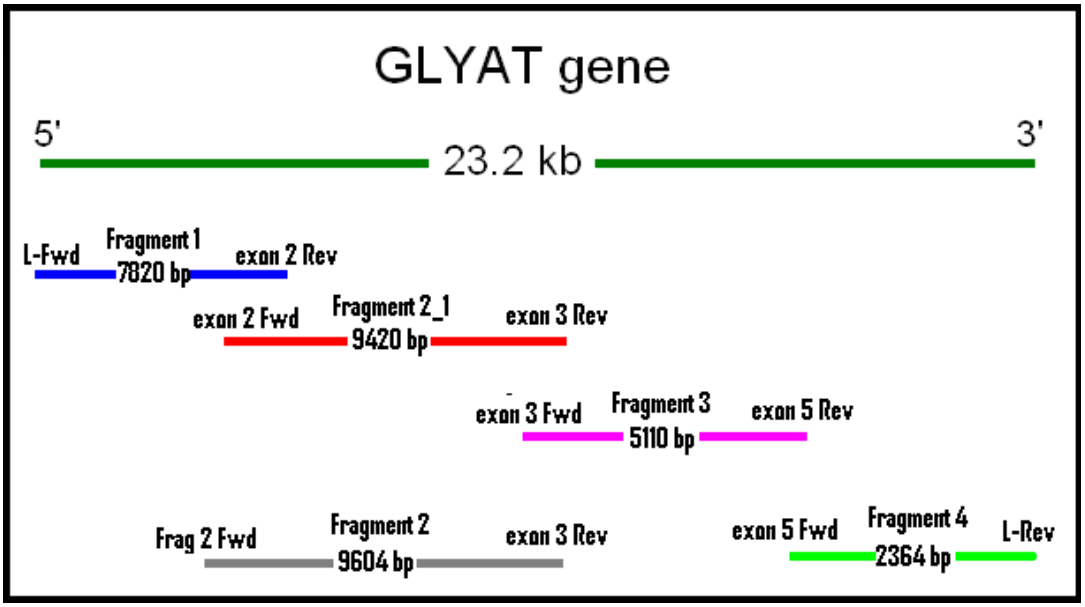

Figure 2.8: Four amplicons of *GLYAT* and five primer combinations used to amplify the gene. Primers L-Fwd and exon two reverse were used to amplify fragment one of 7828 bp in size. Exon two forward and exon three reverse were used to amplify fragment 2_1, whereas fragment two forward and exon three reverse were used to amplify fragment two. Fragment 2_1 and fragment two are 9420 bp and 9604 bp in size respectively. Fragment three was amplified using exon three forward and exon five reverse primers yielding an amplicon of 5110 bp in size. Exon five forward and L-Rev primers were used to amplify fragment four of 2364 bp in size.

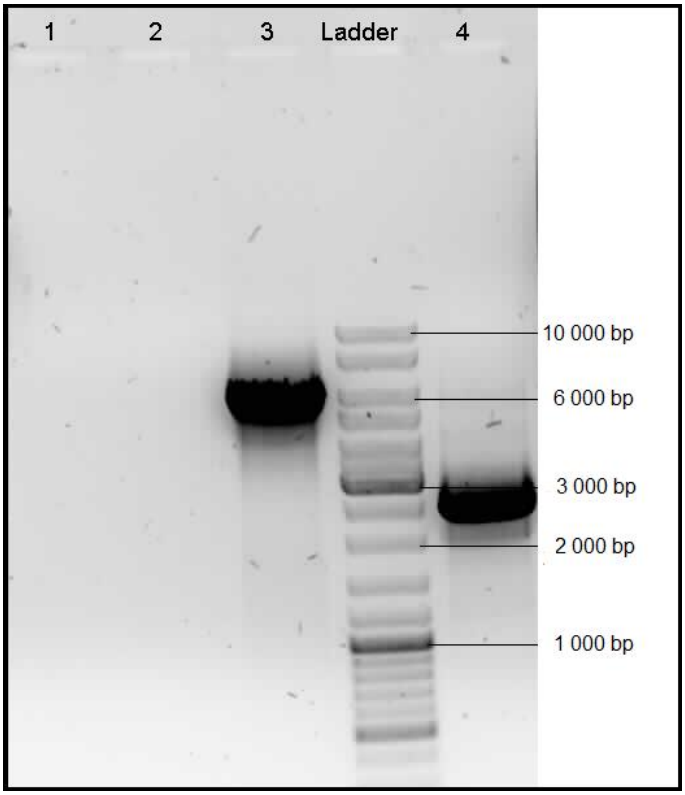

Figure 2.9: Exon primer combinations to amplify *GLYAT* in four overlapping fragments. Lanes: 1) Reaction product of fragment one with the L-Fwd and exon two reverse primers; 2) Reaction product of fragment two with the exon two forward and exon three reverse primers; 3) 5110 bp amplicon of fragment three synthesised with the exon three forward and exon five reverse primers; 4) 2364 bp amplicon of fragment four synthesised with the exon five forward and L-Rev primers. The gel used consisted of 1% agarose and was electrophoresed for an hour at 100 V.

2.3.1.8 Optimization of annealing temperatures for *GLYAT* in four fragments

Fragment one was amplified at annealing temperatures of 56.0˚C and 60.9˚C. The visualisation of these reactions was done on a 0.7% agarose gel shown in Figure 2.10. At both these temperatures fragment one amplified with approximately the same yield, as can be seen in lane one and two, Figure 2.10. Fragment 2_1 was amplified at annealing temperatures of 60.9˚C and 65.0˚C. These two reactions were loaded in lane three and four, Figure 2.10. No amplification of fragment two was observed which suggested that a new forward primer for fragment two had to be designed. The new fragment two forward primer amplified fragment two successfully (Figure 2.11 lane two). The *GLYAT* gene of each of the 18 participants in the cohort

was amplified using the four fragment approach. Primers used for the amplification of fragments one, two, three and four as previously shown in Figure 2.8.

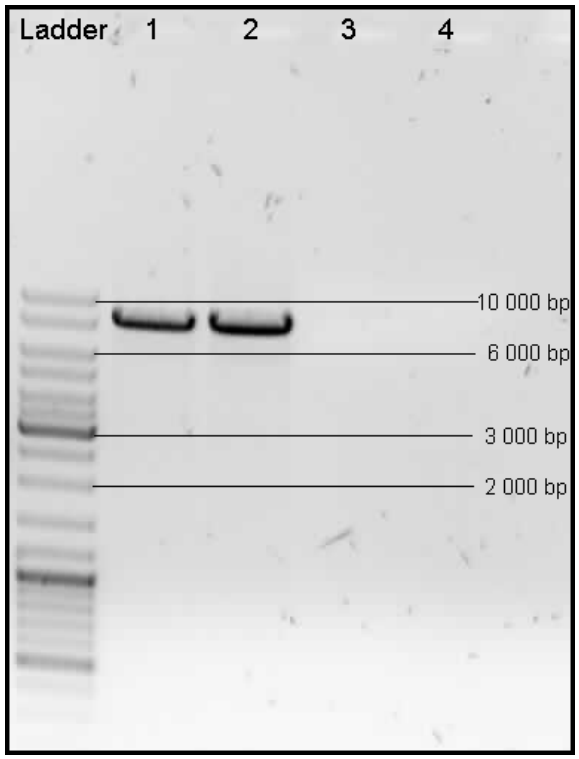

Figure 2.10: Exon primer combinations of fragment one and two at different annealing temperatures electrophoresed on a 0.7% agarose. Lanes: 1) Fragment one amplicon with annealing temperature reaction condition of 56.0˚C; 2) Fragment one amplicon with annealing temperature reaction condition of 60.9˚C; 3) Fragment 2_1 reaction product with annealing temperature reaction condition of 60.9˚C; 4) Fragment 2_1 reaction product with annealing temperature reaction condition of 65.0˚C. The 0.7% agarose gel was electrophoresed for one hour at 100 V.

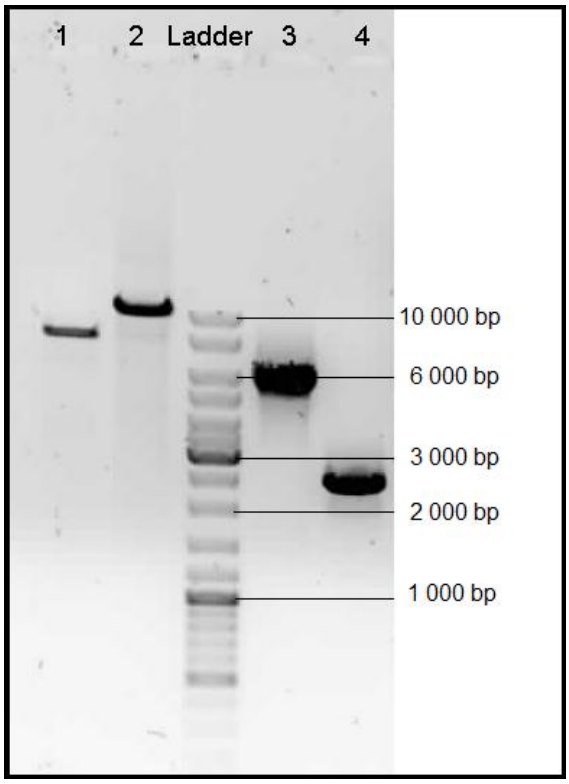

Figure 2.11: Amplicons of all four fragments spanning the full *GLYAT* gene. Lanes: 1) Fragment one amplicon; 2) Fragment two amplicon; 3) Fragment three amplicon; 4) Fragment four amplicon. The 1% agarose gel was electrophoresed at 100 V for an hour.

2.3.2. Polymerase Chain Reaction of the 29 short fragments of the *GLYAT* gene

2.3.2.1 Amplification of short *GLYAT* fragments

Data generated from pyrosequencing had to be verified with Sanger sequencing in order to eliminate possible artefacts generated during pyrosequencing. The *GLYAT* gene was divided into 29 smaller fragments to ensure that all the possible variations were included in 29 fragments. The smaller fragment primers were designed so that the amplicon sizes did not exceed 850 bp in length since the average read length of Sanger sequencing is approximately 750 bp in length, although under ideal conditions read lengths in excess of 1000 bp can be obtained.

The first attempt to amplify the short *GLYAT* fragments was done with the KAPA HiFi PCR Kit (Kapa Biosystems (Pty) Ltd, Cape Town, South-Africa). A reaction was set

up for seven of the short fragments (primer sets one, two, three, four, five, 12 and 17) and one positive control. Each reaction contained 1x KAPA HiFi buffer, 0.3 mM dNTPs, 0.3 mM forward primer, 0.3 mM reverse primer, 3.0 ng/µl template DNA and 0.5 U of KAPA HiFi DNA polymerase. The initial denaturation for the reaction was at 95.0˚C for three minutes, after which the denaturation step was done at 98.0˚C for 20 seconds and primer annealing was done at 60.0˚C for 15 seconds. The extension step was done at 72.0˚C for 30 seconds/kb and after cycling a final extension was performed at 72.0˚C for three minutes. The denaturation, annealing and extension steps were repeated for 30 cycles. None of the short fragments amplified successfully (Figure 2.12), but the positive control amplified successfully, yielding an amplicon of 394 bp (Figure 2.13). The exon one forward and exon one reverse primers were used for the positive control.

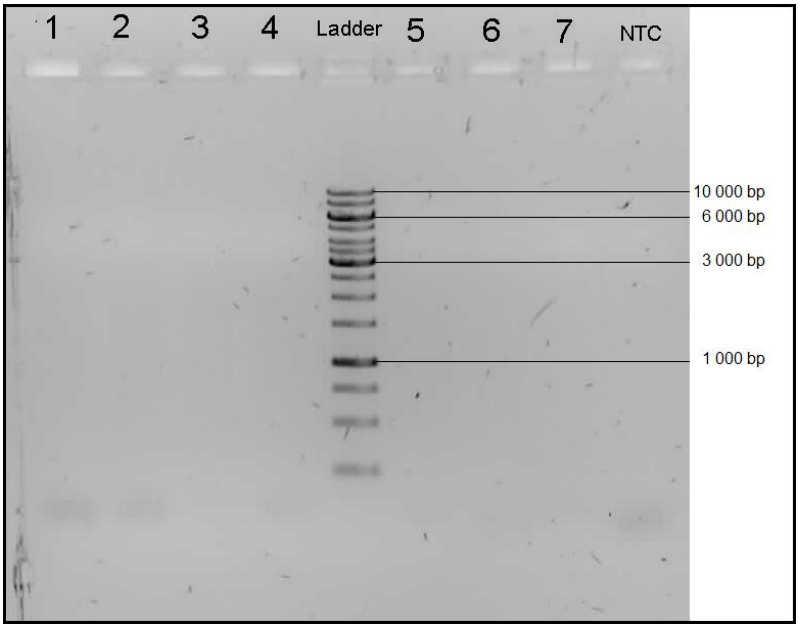

Figure 2.12: PCR products of seven short fragments of the *GLYAT* gene. Lanes: 1) Reaction product generated with primer set one; 2) Reaction product generated with primer set two; 3) Reaction product generated with primer set three; 4) Reaction product generated with primer set four; 5) Reaction product generated with primer set five; 6) Reaction product generated with primer set 12; 7) Reaction product generated with primer set 17. The NTC reaction was set up with primer set one. The 1% agarose gel was electrophoresed for one hour at 100 V.

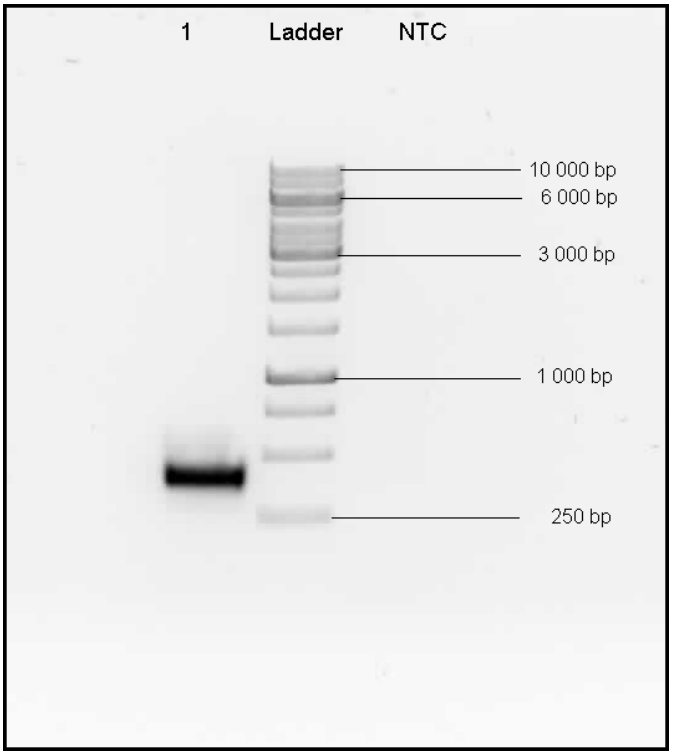

Figure 2.13: Amplicon of the positive control loaded on a 1% agarose gel. Lane 1) Amplicon generated with the exon one forward and exon one reverse. The 1% agarose gel was electrophoresed for one hour at 100 V.

No amplification of the seven short *GLYAT* fragments took place while the positive control amplified. This was an indication that the enzyme, dNTPs and template were all intact and functional. Since this was the first time these reactions were set up, it could be possible that the cycle conditions were not optimal for the seven primer sets. The results obtained from the first PCR reactions for the short *GLYAT* fragments indicated that more optimization had to be done, thus the initial denaturation time was increased from three minutes to five minutes and an annealing temperature gradient had to be done.

2.3.2.2 Optimization of annealing temperature

An annealing temperature gradient was done with primer set one for the short *GLYAT* fragments. The reaction conditions were the same as previously stated for the KAPA HiFi PCR Kit (sections 2.3.2.1) with the exception of the initial denaturation and the annealing temperature which were changed. The initial denaturation was done at 95.0˚C for five minutes, while an annealing temperature gradient was set up

for 15 seconds at 58.5˚C, 59.5˚C, 60.4˚C and 61.6˚C. The amplicons were visualized on a 1% agarose gel which was run at 100 V for an hour. An image of the agarose gel is shown in Figure 2.14.

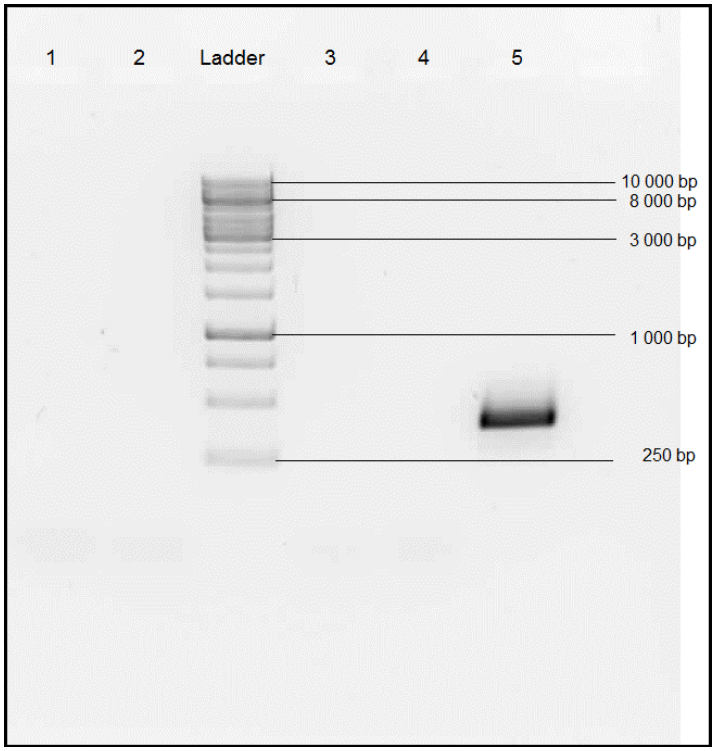

Figure 2.14: PCR products of annealing temperature gradient loaded on a 1% agarose gel. Lanes: 1) Reaction product with annealing temperature reaction condition of 58.5˚C; 2) Reaction product with annealing temperature reaction condition of 59.5˚C; 3) Reaction product with annealing temperature reaction condition of 60.4˚C; 4) Reaction product with annealing temperature reaction condition of 61.6˚C; 5) Positive control generated with primers for *GLYAT* exon one. The 1% agarose gel was electrophoresed for one hour at 100 V.

The KAPA HiFi PCR Kit has an annealing temperature recommendation of 60.0˚C and above. The calculated annealing temperature of the short *GLYAT* fragments was all below 60.0˚C and thus it was decided to use a different polymerase. Since the previous reaction did not work it was decided to use TaKaRa Ex Taq DNA polymerase (Takara Bio, Madison, Wisconsin, U.S.A.), which is a high-fidelity enzyme, which implies a lower error rate of base incorporation compared to Taq DNA polymerase, and to set up a modified Taguchi (COBB and CLARKSON 1994) reaction with the TaKaRa Ex Taq enzyme.

2.3.2.3 Taguchi reaction optimization of short *GLYAT* fragments with TaKaRa Ex Taq Polymerase

The Taguchi reaction includes the optimization of four of the reaction component parameters in one experiment (COBB and CLARKSON 1994). The components evaluated in this experiment were $MgCl_2$, dNTPs, primers and DMSO. The concentrations of the four reaction components were optimized during the Taguchi reaction. Primer set one was used in the nine Taguchi reactions. Table 2.6 summarizes the reaction setup by giving the different reaction components as well as the varying concentrations of each component.

Table 2.6: Variations of concentrations of the four reaction components in nine reaction mixtures. TaKaRa Ex Taq DNA polymerase was used to amplify the short fragment one of the *GLYAT* gene.

Reaction Component	Reaction 1	Reaction 2	Reaction 3	Reaction 4	Reaction 5	Reaction 6	Reaction 7	Reaction 8	Reaction 9
$MgCl_2$	2.00 mM	2.00 mM	2.00 mM	2.25 mM	2.25 mM	2.25 mM	2.50 mM	2.50 mM	2.50 mM
dNTP Mix	0.25 mM	0.30 mM	0.35 mM	0.25 mM	0.30 mM	0.35 mM	0.25 mM	0.30 mM	0.35 mM
Forward primer	0.25 mM	0.30 mM	0.35 mM	0.30 mM	0.35 mM	0.25 mM	0.35 mM	0.25 mM	0.30 mM
Reverse primer	0.25 mM	0.30 mM	0.35 mM	0.30 mM	0.35 mM	0.25 mM	0.35 mM	0.25 mM	0.30 mM
DMSO	0.00 %	2.50 %	5.00 %	5.00 %	0.00 %	2.50 %	2.50 %	5.00 %	0.00 %
DNA polymerase	0.25 U	0.25 U	0.25 U	0.25 U	0.25 U	0.25 U	0.25 U	0.25 U	0.25 U

In Figure 2.15 an image is shown of the agarose gel, which was used to separate the nine reactions, performed during the Taguchi reaction. No amplification was evident in any of the reactions that were set up with TaKaRa Ex Taq DNA polymerase.

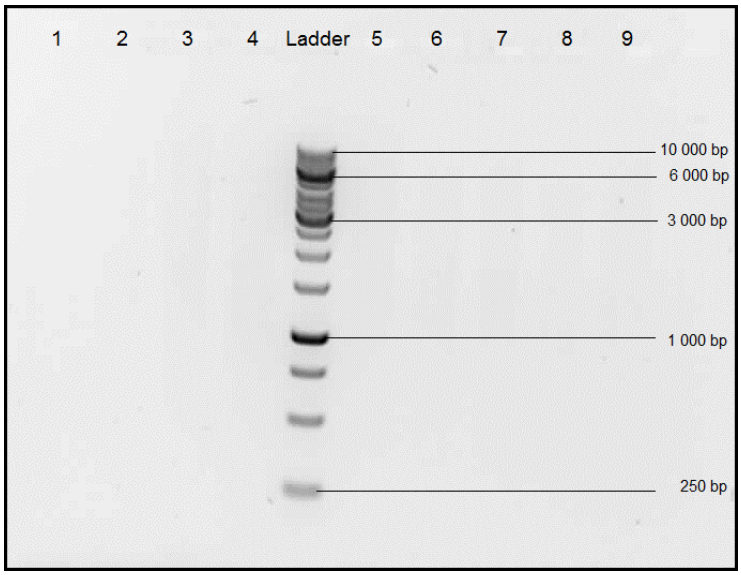

Figure 2.15: The nine Taguchi reaction mixtures separated on a 1% agarose gel. Lanes: 1) Taguchi reaction mixture one; 2) Taguchi reaction mixture two; 3) Taguchi reaction mixture three; 4) Taguchi reaction mixture four; 5) Taguchi reaction mixture five; 6) Taguchi reaction mixture six; 7) Taguchi reaction mixture seven; 8) Taguchi reaction mixture eight; 9) Taguchi reaction mixture nine. The 1% agarose gel was electrophoresed for one hour at 100 V.

Since no amplification took place with the varying concentrations of the four reaction components, the next step was to set up an annealing temperature gradient with the TaKaRa Ex Taq DNA polymerase.

2.3.2.4 Annealing temperature gradient with TaKaRa Ex Taq DNA polymerase

An annealing temperature gradient PCR was set up with the TaKaRa Ex Taq DNA polymerase. The calculated annealing temperature for primer set one was 52.0˚C but 11 annealing temperatures were tested in a range from 50.0˚C to 59.9˚C. The temperatures were 50.0˚C, 50.8˚C, 51.7˚C, 52.8˚C, 54.1˚C, 55.4˚C, 56.7˚C, 57.9˚C, 58.8˚C, 59.5˚C and 59.9˚C. No amplification took place in any of the 11 reactions at the various annealing temperatures as can be seen in Figure 2.16.

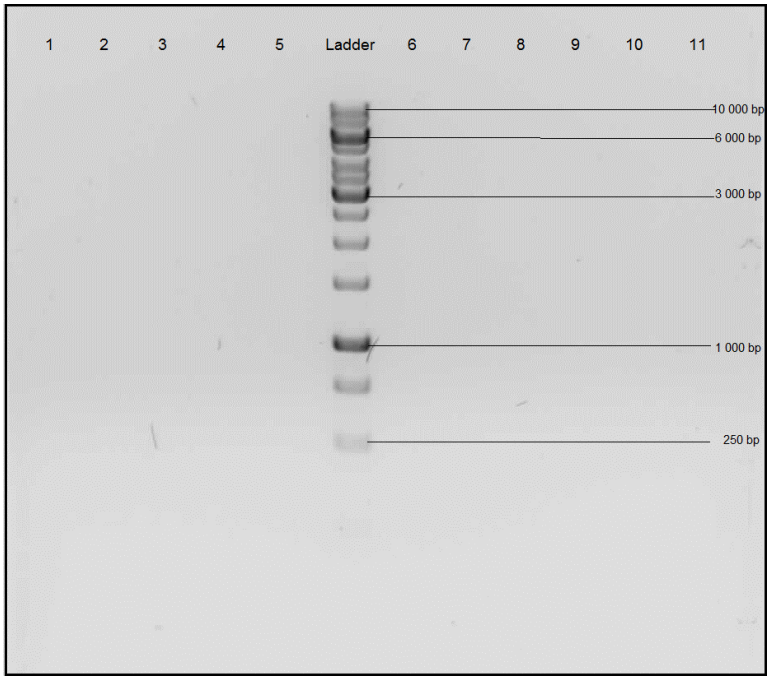

Figure 2.16: Results of annealing temperature gradient for primer set one with TaKaRa Ex Taq DNA polymerase. Lanes: 1) Reaction product with annealing temperature reaction condition of 50.0˚C; 2) Reaction product with annealing temperature reaction condition of 50.8˚C; 3) Reaction product with annealing temperature reaction condition of 51.7˚C; 4) Reaction product with annealing temperature reaction condition of 52.5˚C; 5) Reaction product with annealing temperature reaction condition of 54.1˚C; 6) Reaction product with annealing temperature reaction condition of 55.4˚C; 7) Reaction product with annealing temperature reaction condition of 56.7˚C; 8) Reaction product with annealing temperature reaction condition of 57.9˚C; 9) Reaction product with annealing temperature reaction condition of 58.5˚C; 10) Reaction product with annealing temperature reaction condition of 59.5˚C; 11) Reaction product with annealing temperature reaction condition of 59.9˚C. The 1% agarose gel was electrophoresed for one hour at 100 V.

Since no amplification took place at the different annealing temperatures, a different approach was followed. Amplification previously took place with the available exon primers for the *GLYAT* gene, thus it was decided to use these exon primers in combination with the primers for the short *GLYAT* fragments.

2.3.2.5 Combination of the *GLYAT* gene exon- and short fragment primers

The exon primers for the *GLYAT* gene were previously used as a positive control. The exon one primers amplified an amplicon of 394 bp in size. Since the reactions

with the exon one primers for the *GLYAT* gene successfully amplified an amplicon, it was decided to use these exon primers in combination with the short fragment primers. In the first attempt a combination of the forward nine and exon two reverse primers were made, with an expected amplicon size of approximately 982 bp. The combination yielded no amplification. Two exon primer sets successfully yielded an amplicon and served as positive controls. The primer sets included the internal forward primer for exon six, the reverse primer for exon six and exon two forward and reverse primers. Both the exon primers for exon six and exon two yielded amlpicons of 821 bp and 696 bp in size, respectively (Figure 2.17 lanes two and three). The amplicons were separated on a 1% agarose gel as can be seen in Figure 2.17.

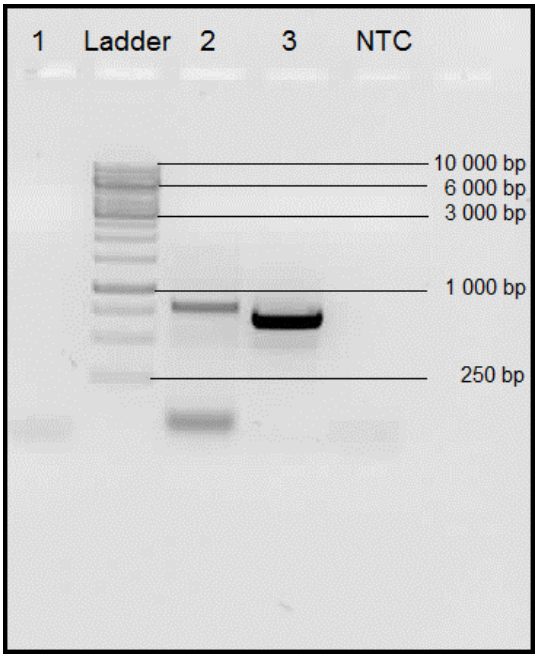

Figure 2.17: Amplicons synthesised with *GLYAT* exon primers and *GLYAT* short fragment primer combinations on a 1% agarose gel. Lanes: 1) Amplicon synthesised with the forward primer for short fragment nine and reverse primer of *GLYAT* exon six; 2) 821 bp amplicon synthesised with the exon six primer set; 3) 696 bp amplicon synthesised with the exon two primer set. The 1% agarose gel was electrophoresed for one hour at 100 V.

Since the exon primer sets amplified it was decided to start to amplify the participant DNA with these primer sets, in order to start sequencing as soon as possible. Working solutions of primers, 10 mM, were freshly prepared from primer stocks, 200 mM, by adding 10 mM TE (Tris-EDTA) buffer, pH 8. The previous reaction with the

exon primers were repeated as well as reactions for exon four and exon five. The results for the four reactions are shown in Figure 2.18. No amplification took place with the four exon primer sets.

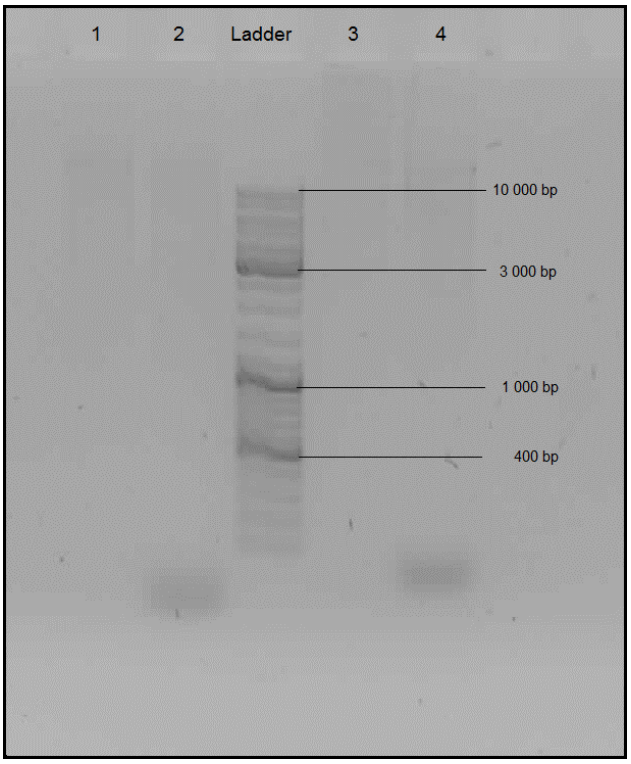

Figure 2.18: Amplification of four exons of the *GLYAT* gene. Lanes: 1) Exon two reaction product; 2) Exon four reaction product; 3) Exon five reaction product; 4) Exon six reaction product. The 1% agarose gel was electrophoresed for an hour at 100 V.

The results of exon two and six as obtained previously (Figure 2.17) could not be repeated and exon four and five did not amplify either. Further optimization was considered. The DNA concentration was not optimized in the previous reactions and therefore the next step was to consider increasing the concentration of the template DNA.

2.3.2.6 Increased template DNA concentration

Four exon primer sets were used in reactions to optimize the DNA concentration. PCRs were set up to amplify exons two, four, five and six. The expected amplicon

sizes were 315 bp (exon two), 483 bp (exon four), 377 bp (exon 5a) and 821 bp (exon six). The template DNA concentrations for these reactions were increased from 1.25ng/µl to 6.00ng/µl. None of the four exon regions amplified as can be seen in Figure 2.19.

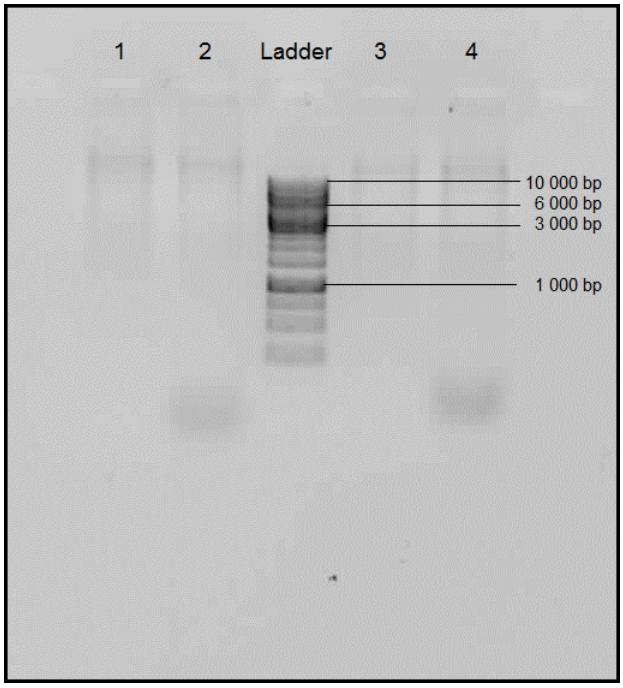

Figure 2.19: Amplification of four *GLYAT* exons with increased DNA concentration. Lanes: 1) Exon one reaction product; 2) Exon four reaction product; 3) Exon 5a reaction product; 4) Exon six reaction product. The 1% agarose gel was electrophoresed at 100 V for one hour.

At this point it was realised that a possible cause for the failure of the reactions to amplify could be because of the TE buffer in which the primers were dissolved. The TE buffer contains EDTA, which bind metal ions. Mg^{2+} is a metal ion which is a cofactor for DNA polymerases. In the case where EDTA chelates these metal ions, the concentration will be too low in the reaction and thus not enough Mg^{2+} will be present. This will cause impaired amplification. To test this theory the working primers were dissolved in molecular grade water from the stock which was dissolved in TE buffer, pH 8. The EDTA is important within the primer stock solutions as nucleases also require Mg^{2+} and therefore the EDTA will protect the primers from degradation by inhibiting possible nuclease activity. The reaction preformed with the increased DNA concentration was repeated with the freshly prepared primers.

2.3.2.7 Primers dissolved in molecular grade water

The reaction of the four exon primers (section 2.3.2.6) was repeated with the freshly prepared primers. The primers were diluted with molecular grade water rather than the TE buffer which the stock solutions were reconstituted in. All four the exon primer sets gave yield to amplicons of the expected size, namely 315 bp (exon two), 483 bp (exon four), 377bp (exon 5a) and 821 bp (exon six) (Figure 2.20).

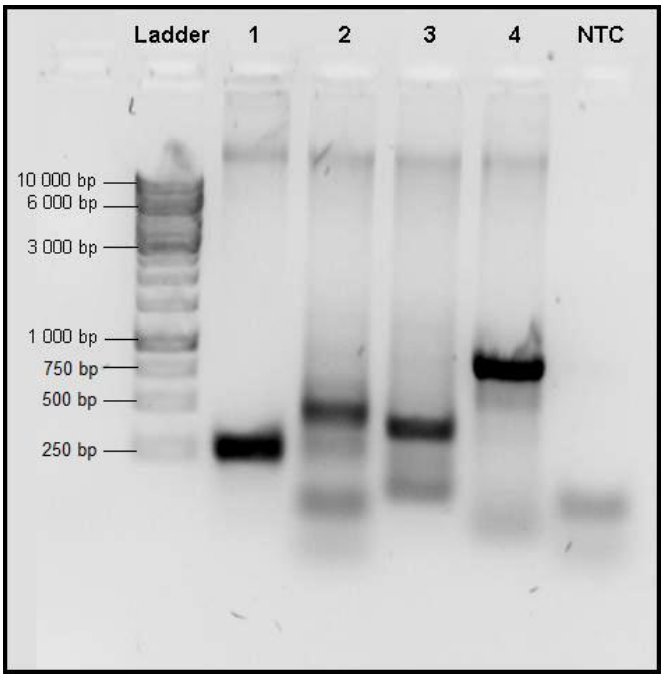

Figure 2.20: Amplicons of four *GLYAT* exons amplified with primers dissolved in molecular grade water. Lanes: 1) Exon two amplicon of 315 bp in size; 2) Exon four amplicon of 483 bp in size; 3) Exon 5a amplicon of 377 bp in size; 4) Exon six amplicon of 821 bp in size. The gel contained 1% agarose and was electrophoresed for one hour at 100 V.

All four the exon primer sets yielded amplicons of the expected size and thus it was concluded that the EDTA from the TE buffer inhibited the PCR reactions, preventing amplification from taking place. The low EDTA concentration of the TE buffer should not have influenced the PCR reactions, but the results proved the opposite. Since the reactions worked, with the primers diluted with molecular grade water, the primers for the short *GLYAT* fragments also had to be prepared with molecular grade water rather than TE buffer. The reactions for the short *GLYAT* fragments were

repeated with the freshly prepared primers with molecular grade water. All 29 short *GLYAT* fragments, which were to be used for Sanger sequencing verification of Next Generation Sequencing data, yielded amplicons of the expected size. In Figure 2.21 the results of four of the primer sets used for the verification are shown as well as one primer set which was not used for the sequencing verification since the amplicon of primer set (Figure 2.21 lane three) covered an area which did not include any variations identified during pyrosequencing. These results are representative for all of the 29 fragments.

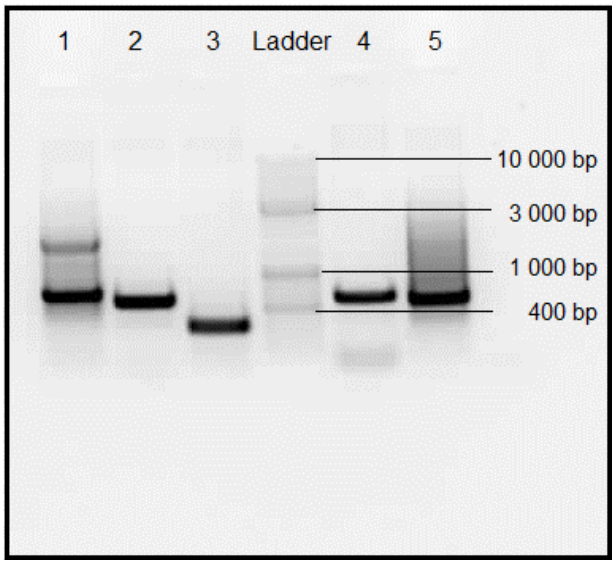

Figure 2.21: Amplification results for primer sets 1 - 5 of the short *GLYAT* gene fragments. Lanes: 1) 698 bp amplicon synthesised with primer set one; 2) 621 bp amplicon synthesised with primer set two; 3) 342 bp amplicon synthesised with primer set three; 4) 633 bp amplicon synthesised with primer set four; 5) 595 bp amplicon synthesised with primer set five. The 1% agarose gel was electrophoresed for one hour at 100 V.

After all the reactions were optimized for the amplification of the short *GLYAT* fragments to be used for Sanger sequencing verification of the pyrosequencing data, the short *GLYAT* fragments of the 18 participants had to be prepared and sequenced using Sanger sequencing.

2.3.3 Gel extraction of the *GLYAT* gene amplicons

GLYAT gene amplicons of the 18 participants had to be gel purified prior to pyrosequencing. After the amplicons of the 18 participants were gel extracted the amplicon concentrations (Table 2.7) of the four fragments were determined using a Nanodrop fluorometer at Inqaba Biotech. The samples were pooled in eight groups in preparation for pyrosequencing, where a molecular identifier (MID) was added to each group. MIDs are short nucleotide sequences added during library preparation to each sample, which allow pooled samples to be distinguished from each other in post-sequencing analysis. At the time of sequencing, the MIDs available was not sufficient in number to label each individual sample, but only pools of samples. The participants and their fragment concentration in each pool are also shown in Table 2.7. The quantity that each sample would contribute to a pool was calculated, where after it was decided which samples should be pooled together, based on the actual concentration that each participant's four fragments contribute to the pooled sample. In the case where one sample contributes a higher quantity than another sample, bias could be introduced during sequencing.

Table 2.7: The concentrations of the four *GLYAT* amplicons of the cohort. The actual concentration to the pool of each of the four fragments of the 18 participants is shown in the table below.

Pool number	Sample ID	Concentration ng/µl	Actual quantify to pool	Total quantity in pool (µg)
1	1.1	3.3816	165.7	
1	1.2	85.2496	1176.4	
1	1.3	111.7452	737.5	
1	1.4	27.6928	664.6	
1	2.1	15.1584	742.8	
1	2.2	134.5088	887.8	
1	2.3	29.44	588.8	
1	2.4	20.8256	333.2	5296.82
2	3.1	3.144	172.9	
2	3.2	143.3024	945.8	
2	3.3	58.8464	971.0	
2	3.4	26.4304	607.9	
2	9.1	1.2576	65.4	
2	9.2	75.0402	495.3	
2	9.3	79.1968	1306.7	

Pool number	Sample ID	Concentration ng/µl	Actual quantity to pool	Total quantity in pool (µg)
2	9.4	27.1664	543.3	5108.32
3	4.1	12.2088	512.8	
3	4.2	88.5216	973.7	
3	4.3	84.4496	928.9	
3	4.4	14.6848	352.4	
3	10.1	6.9568	299.1	
3	10.2	94.0144	620.5	
3	10.3	121.92	804.7	
3	10.4	51.1392	562.5	5054.73
4	15.1	4.0884	175.8	
4	15.2	29.1264	699.0	
4	15.3	42.6544	1023.7	
4	15.4	29.8016	715.2	
4	5.1	4.5828	256.6	
4	5.2	80.8512	533.6	
4	5.3	52.992	349.7	
4	5.4	28.4256	483.2	
4	6.1	9.7152	514.9	
4	6.2	147.4048	972.9	
4	6.3	25.1296	603.1	
4	6.4	20.9536	419.1	6746.98
5	7.1	4.3712	231.7	
5	7.2	82.3536	905.9	
5	7.3	119.7888	1317.7	
5	7.4	30.1328	662.9	
5	8.1	2.016	114.9	
5	8.2	135.5408	894.6	
5	8.3	108.5552	716.5	
5	8.4	12.3952	223.1	5067.22
6	11.1	4.1076	139.7	
6	11.2	127.4016	840.9	
6	11.3	61.3824	675.2	
6	11.4	45.0624	1081.5	
6	12.1	7.8656	377.5	
6	12.2	109.7408	724.3	
6	12.3	65.2768	718.0	
6	12.4	32.6176	782.8	5339.92
7	13.1	12.7356	534.9	
7	13.2	136.6096	901.6	
7	13.3	1.1376	27.3	
7	13.4	33.9472	678.9	

Pool number	Sample ID	Concentration ng/µl	Actual quantity to pool	Total quantity in pool (µg)
7	14.1	19.698	787.9	
7	14.2	112.7024	743.8	
7	14.3	102.384	675.7	
7	14.4	57.0112	940.7	5290.94
8	16.1	5.1684	87.9	
8	16.2	99.6656	657.8	
8	16.3	2.4032	58.9	
8	16.4	9.0048	171.1	
8	17.1	16.3824	360.4	
8	17.2	84.2568	926.8	
8	17.3	101.7096	1068.0	
8	17.4	8.2256	197.4	
8	18.1	7.5288	128.0	
8	18.2	122.824	810.6	
8	18.3	112.7712	744.3	
8	18.4	19.6064	470.6	5681.70

After pyrosequencing data was obtained, Sanger sequencing had to be done for verification of variations identified by pyrosequencing. 29 short *GLYAT* fragments had to be amplified and gel purified for Sanger sequencing. After the short *GLYAT* fragments were gel purified for Sanger sequencing, their respective concentrations were determined using a NanoDrop 1000 spectrophotometer. A minimum of 5 µl of purified PCR product was required for each fragment. The required concentrations for fragments smaller than 500 bp were 5ng/µl while the concentration required for the fragments of 500 – 1000 bp in size were 10ng/µl.

CHAPTER 3 – SEQUENCING

3.1 Introduction

Nucleotide variations including SNPs and mutations consisting of insertions, deletions and base pair changes, between individuals are expected every 300 base pairs (MANCINELLI *et al.* 2000; FEUK *et al.* 2006). Thus, between the genomes of 2 individuals there would exist approximately 10 million variations (MANCINELLI *et al.* 2000). These variations can be introduced during replication of DNA, are permanent and can be associated with disease (MANCINELLI *et al.* 2000; CHORLEY *et al.* 2008) and drug efficacy.

When a variation occurs in the exons, branch point or splice site, enzyme activity could be altered or the enzyme function could be terminated (MCCARROLL and ALTSHULER 2007). When a variation causes a key amino acid to change in terms of folding within the affected protein, it can alter the function of the protein and increase susceptibility to disease (PRITCHARD 2001; CHORLEY *et al.* 2008). It was found that introducing a variation in the *GLYAT* gene could have an effect on the enzyme kinetic properties (VAN DER SLUIS *et al.* 2013). In order to identify variations which could possibly alter enzyme activity, the gene in question should be sequenced. Once the particular gene is sequenced the specific nucleotide order will be known and thus variations can be identified.

Next-generation sequencing involves non-electrophoretic methods to enable the generation of vast amounts of sequence information (HERT *et al.* 2008). Pyrosequencing is one such method which is a technique based on sequencing by synthesis. In short, like all NGS technologies, a library of the sample has to be prepared. The library consists of fragmented ssDNA bound to adaptors, which can be sequenced by the various next-generation sequencing methods. The first four nucleotides of the adaptor serve as a calibration to enable the instrument to estimate light emitted by a cascade of reactions (MARDIS 2008). Once the DNA library is prepared, it is quantified usually by means of a qPCR standard curve. The library concentration is determined whereupon library amplification can take place (MARDIS

2008). The amplified library will be loaded on to a picotiter plate, whereupon enzyme containing beads will catalyse the reaction. Nucleotides will flow in an exact order across the wells and DNA capture beads whereupon a range of enzyme reactions will take place to produce a cheminumilescent signal, which will then be recorded by a CCD camera.

The Sanger sequencing technique is based on chain termination (SANGER *et al.* 1977) unlike pyrosequencing which is based on sequencing by synthesis. In this study Sanger sequencing was used to verify results obtained from pyrosequencing. During automated Sanger sequencing all the ddNTPs is labelled with dyes which fluoresce at different wavelengths when excited by a laser. The fluorescence is measured by a camera which will ultimately determine the sequence. The results are illustrated in a electropherogram with peaks of different colours for the different nucleotides, starting from small to large fragments (OBENRADER 2007).

All raw sequencing data needs to be analysed using certain software. During this study DNAStar (Madison, Wisconsin, USA) was used. The software compiles a sequence compared to a reference sequence of the gene of interest. The software will flag any variation such as base pair changes, insertions and deletions compared to the wild type gene.

3.2 Materials and Methods

3.2.1 454 Pyrosequencing

Pyrosequencing was done on the amplified *GLYAT* gene of a cohort with possible altered GLYAT activity. The massively parallel nature of 454 pyrosequencing makes it possible to identify sequence variations in the *GLYAT* gene such as known SNPs, mutations and novel variations.

3.2.1.1 Sample preparation and pyrosequencing

Sample preparation for pyrosequencing was done as follows. The amplicons (fragments one, two, three and four) spanning the *GLYAT* genes of the 18 participants were divided into eight groups, based on the DNA quantity each sample contributed to each pool, thereafter each group was labelled with a molecular identifier (MID), as illustrated in Figure 3.1. MIDs are short nucleotide sequences added during library preparation to each sample, which allow pooled samples to be distinguished from each other in post-sequencing analysis. At the time of sequencing, the MIDs available was not sufficient in number to label each individual sample, but only pools of samples. This allowed for a distinction to be made between the different groups, despite all of them being sequenced together. In our sample pool the MIDs therefore made it possible to narrow down to which participant a specific variation belonged to.

The eight pooled samples were sent to Inqaba Biotech, where the library preparation and pyrosequencing was done on a Roche GS FLX system using titanium chemistry. After the data was received from Inqaba Biotech, data assembly and analyses was done using Lasergene software from DNAStar (Madison, Wisconsin, USA).

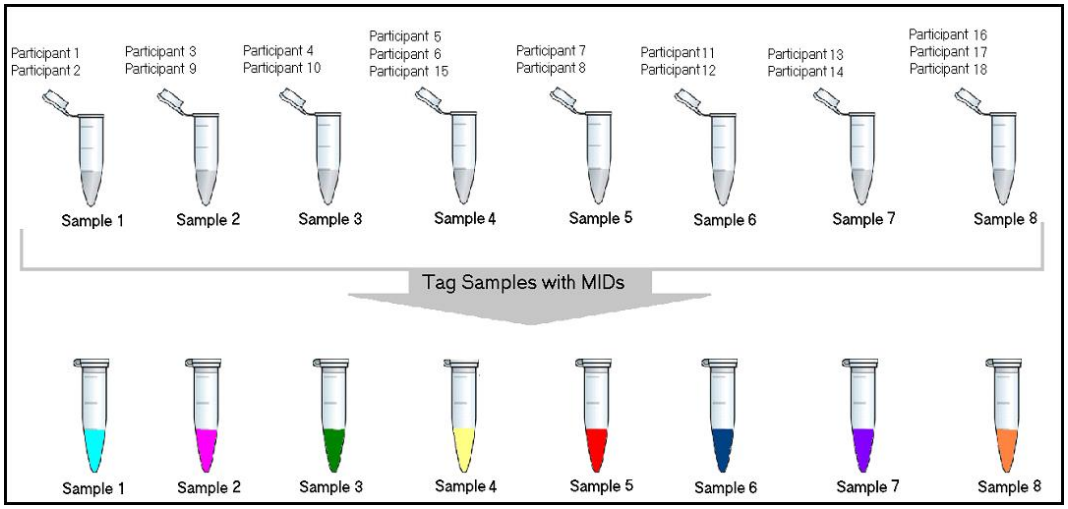

Figure 3.1: Samples group composition and tagging with MIDs: The amplicons of the 18 participants were divided into eight groups, which were tagged with MIDs. As all eight groups were sequenced together on one picotiter plate, the MIDs allowed for distinction between the eight groups after sequencing.

3.2.1.2 Pyrosequencing data analysis

The human reference genome was used to generate scaffolds from the reads obtained from pyrosequencing. Each sample group was assembled using the default settings of Lasergene Genomics Suite from DNAStar (Madison, Wisconsin, USA). After the assembly was completed possible variations were identified by the software. Each variation was assessed individually. Variations with a quality score lower than 20 and coverage lower than 20x were not considered for further analysis. Quality scores of 20 or higher are required for SNP identification while 20x coverage is also required. Quality scores of each base, indicate the probability that the base was called incorrectly. Coverage indicates the number of times a specific nucleotide of the DNA was sequenced (EWING and GREEN 1998; EWING *et al.* 1998).

3.2.2 Sanger sequencing

Variations identified after analysis of the pyrosequencing data had to be verified. This verification was done by Sanger sequencing. The *GLYAT* gene of each participant was amplified in 29 fragments whereupon the 29 fragments of each participant were sequenced using Sanger sequencing.

3.2.2.1 Sample preparation and Sanger sequencing

Potential variations identified via pyrosequencing were amplified and sent for Sanger sequencing at the Central Analytical Facilities at Stellenbosch University. This was done in order to verify the data obtained from pyrosequencing and eliminate possible artefacts. Once the Sanger sequencing data was obtained, data analyses were done using Lasergene software from DNAStar (Madison, Wisconsin, USA).

3.2.2.2 Sanger sequencing data analysis

The 29 fragments of 17 of the 18 participants were sequenced in both directions and assembled using Lasergene Genomics Suite from DNA Star, using the software default settings. As was done with pyrosequencing, the human reference genome was used to generate scaffolds of the reads obtained from Sanger sequencing. Low yield human genomic DNA was obtained from the remaining participant and thus further analyses on the remaining participant was discontinued. Variations from the reference sequence were identified by the Lasergene software and indicated.

3.2.3 Branch points and splice sites prediction

Variations identified by pyrosequencing were incorporated into the *GLYAT* reference sequence, where after the *GLYAT* reference sequence, with incorporated variation, were submitted to the Human Splicing Finder – Version 2.4.1 (DESMET *et al.* 2009). The Human Splicing Finder predicts possible splice sites and branch points. Once the list of potential splice site and branch points were generated by the Human Splicing Finder, the variations causing a possible splice site or branch point could be identified.

3.3 Results and Discussion

3.3.1 Identification of sequence variations in the *GLYAT* gene by means of pyrosequencing

Each DNA pool was assembled into a continuous sequence using the Lasergene SeqMan-suite (DNAStar, Madison, Wisconsin, USA) to align it to the human reference sequence. After the assembly was completed the software labelled the potential variations (see Figure 3.2). The potential variations are indicated in blue, while the variations rejected based on quality scores and coverage, are indicated in red. As discussed in section 1.5.2, the DNA is sheared into smaller fragments. Each one of these fragments can bind to a bead which will ultimately be sequenced. The coverage is dependent on the read length of each of these fragments as well as amount of reads.

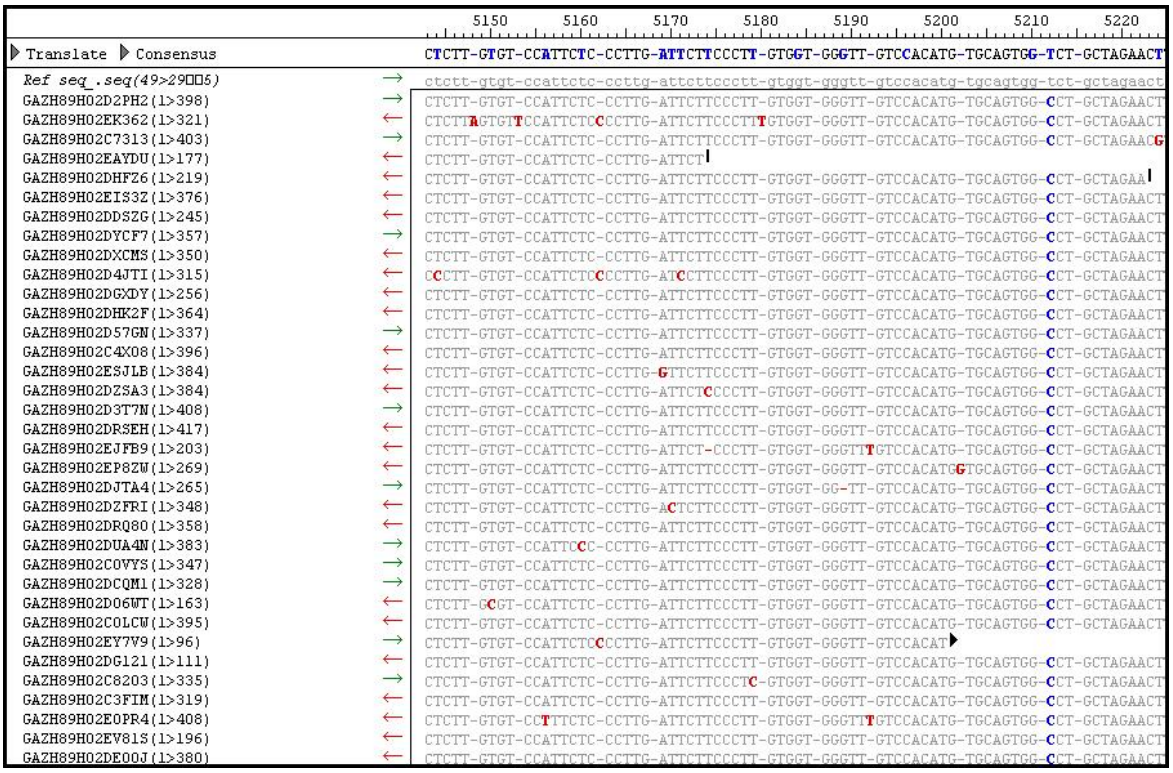

Figure 3.2: Representative sample of results from pyrosequencing data analysis. The results indicated in the Figure are a snapshot of sample pool one, which include Participants one, two and three. The bases indicated in blue are putative variations, while the bases indicated in red are variations which were rejected based on the quality scores and coverage. The red and green arrows indicate in which direction each of the fragments were sequenced.

A total of 94 variations were found within the eight sample pools. Of the 94 variations found, 65 of these variations were previously described SNPs. 29 of the 94 variations found were novel and only four of these are located in the exons of the *GLYAT* gene. Only three of the previously described variations were located in the exons. The locations of the previously identified and novel variations are summarized in Table 3.1. A total of 12 variations were due to deletions, 15 due to insertions and 67 of the variations identified were caused by base pair changes, a summary of which can be seen in Table 3.2. Table 3.3 lists all the variations found and the reference position on the *GLYAT* gene. Further to this, Table 3.3 also illustrates in which sample pool each variation was found. Table 3.1 and Table 3.2 are graphically represented in Figure 3.3. Possible branch points and splice sites identified in the *GLYAT* gene are also illustrated in Figure 3.3 and will be discussed later on.

Table 3.1: Summary of the location of variations identified in the *GLYAT* gene of the 18 participants.

Location of known variations	**n**
Exons	3
Introns	62
Location of novel variations	**n**
Exons	4
Introns	25
Total variations identified	94

Table 3.2: Summary of the types of variations identified in the *GLYAT* gene of the 18 participants.

Variation type	**n**
Deletions	12
Insertions	15
Base pair changes	67

Table 3.3: The 94 variations found in the *GLYAT* gene of the participants. The reference position (bp), variation, pools in which the variation was found as well as whether the variation is novel or previously described is indicated.

Reference position	Variation	Sample pool in which variation were found	Novel (N) / Previously described (PD)
58498835	A/T	2, 4, 6, 7	PD
58498825	C/G	1,2, 3, 5, 6, 7, 8	PD
58497569	A/T	1, 3, 6, 7, 8	PD
58495947	G/A	5	N
58495844	C/G	3	PD
58495760	A/T	6, 7	PD
58495554	G/T	1, 4, 7	PD
58495554	-/A	1	N
58495334	C/ -	1, 6, 7, 8	N
58494339	A/G	1, 2, 3, 4, 5, 6, 7	PD
58493900	A/T	1, 3, 6, 7	PD
58493900	C/T	3	N
58493889	C/G	3	N
58493888	C/T	3	N
58493839	C/G	3	N
58493668	C/T	1, 3, 5, 6, 7, 8	PD
58493420	A/G	1, 2, 3, 8	PD
58493392	C/T	1, 2, 3, 5, 6, 7, 8	PD
58493243	C/T	1, 2, 3, 5, 6, 7, 8	PD
58493212	A/G	1, 2, 3, 5, 6, 7, 8	PD
58493204	A/G	1, 2, 3, 5, 6, 7, 8	PD
58493165	C/T	1, 2, 3, 5, 6, 7, 8	PD
58493165	C/G	3, 6, 7, 8	PD
58492875	-/G	1	N
58492666	C/T	5	N
58492509	-/G	2	N
58492088	A/ -	1	N
58492086	T/ -	1	N
58492085	A/-	2, 3, 4, 6, 7, 8	N
58491921	A/ -	1, 3, 6, 7, 8	PD
58491400	A/T	1, 3, 6, 7, 8	PD
58491268	C/T	1, 3, 4, 5, 6, 7, 8	PD
58490795	A/G	1, 5	PD
58490703	C/T	1, 3, 4, 5, 6, 7	PD
58490702	-/T	3, 4, 5	PD
58490221	-/T	6, 7	PD

Reference position	Variation	Sample pool in which variation were found	Novel (N) / Previously described (PD)
58489459	C/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58488855	C/T	6	PD
58488848	-/T	4, 6, 8	PD
58488634	-/T	5, 7	PD
58488185	G/-	3, 6	N
58487994	A/C	1, 3, 4, 5, 6, 7	PD
58487393	C/T	6	PD
58486917	A/G	1, 6, 7	PD
58486850	A/C	1, 2, 4, 5, 6, 7	PD
58486737	A/C	1, 5, 6, 7	PD
58486626	A/T	6, 7	PD
58486092	C/T	6, 7	PD
58486071	-/T	7	N
58485925	-/T	4	N
58485603	C/T	1, 3, 6, 7	PD
58484971	C/T	6	PD
58484656	A/C	1, 4, 5, 6, 7	PD
58484611	A/G	1, 6, 7	PD
58484500	C/T	4	N
58484471	C/T	1, 6, 7, 8	PD
58484233	A/G	1, 6, 7	PD
58484119	A/G	1, 6, 7	PD
58484117	A/G	1, 6, 7	PD
58484086	C/T	1, 5, 6, 7	PD
58483921	G/T	1, 6, 7	PD
58483862	-/C	3, 4, 5, 6	PD
58483861	-/C	2, 7	PD
58483858	-/C	1	PD
58482423	G/T	1, 3, 6, 7, 8	PD
58482290	A/G	1, 2, 3, 4, 5, 6, 7, 8	PD
58482249	A/C	1, 3, 6, 7, 8	PD
58482187	C/G	1, 2, 3, 4, 5, 6, 7, 8	PD
58481781	C/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58481375	T/-	5, 6, 7, 8	N
58481374	A/-	2, 4	N
58481374	A/T	6, 7	N
58481370	A/ -	1	N
58481350	G/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58481196	A/T	1, 6	N
58481193	C/T	2, 4, 5, 7	PD
58481193	C/ -	1, 3, 6	N

Reference position	Variation	Sample pool in which variation were found	Novel (N) / Previously described (PD)
58480742	G/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58480738	A/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58480583	A/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58480381	T/ -	1, 2, 5, 6	N
58479908	A/G	1, 2, 3, 4, 5, 6, 7, 8	PD
58479903	A/G	5, 6	PD
58479749	C/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58479675	C/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58479305	C/G	1, 2, 3, 4, 5, 6, 7, 8	PD
58478084	G/A	1, 2, 3, 4, 5, 6, 7, 8	PD
58478039	C/T	1, 2, 3, 4, 5, 6, 7, 8	PD
58478026	A/G	1, 2, 3, 4, 5, 6, 7, 8	PD
58476844	-/GAAAGG	1, 6	PD
58476844	-/CTTTCC	1, 2, 3, 4, 5, 6, 7, 8	N
58476843	A/T	8	N
58476841	-/A	8	N
58476250	A/ -	1, 7	N

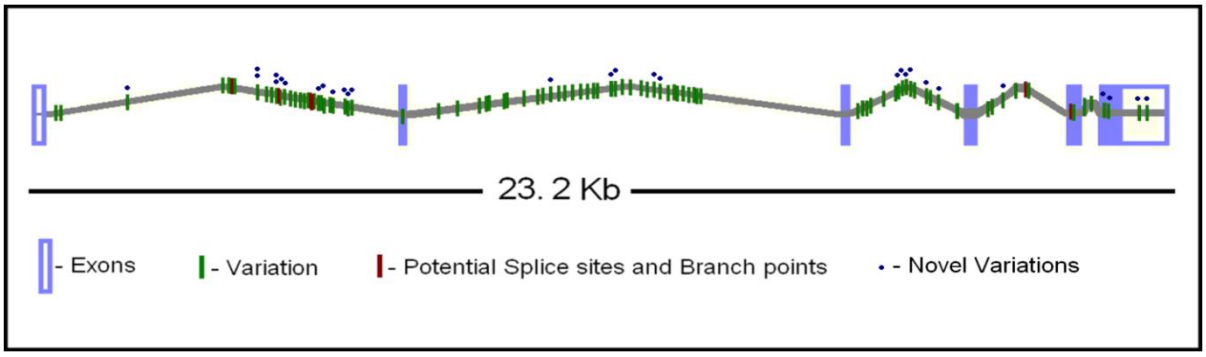

Figure 3.3: Schematic representation of the *GLYAT* gene with all variations found in the gene. Potential branch points and splice sites caused by the identified variations are indicated by red bars. The 29 novel variations found in the cohort are indicated with blue dots.

3.3.2 Pyrosequencing data verification by means of Sanger sequencing

Sanger sequencing was used to verify the variations identified from the pyrosequencing data and to distinguish which participants in each pool had the variations identified within the sample pool. Despite all the advances and

advantages of NGS, Sanger sequencing still remains the sequencing gold-standard. The results of the Sanger sequencing are summarized in Table 3.4. The Table includes the reference position of the verified variations as well as the participants in which the verified variations were identified. Only 70.21% of variations (68 variations) identified by pyrosequencing could be verified by Sanger sequencing. 12 of the variations verified are novel variations, whereof 1 is located in exon six (position 58476844). The other 29.79% could not be verified. This might be due to artefacts generated during PCR of sequencing. All though high fidelity DNA polymerases were used during amplification, incorrect nucleotide might still have been incorporated into the amplicons, since high fidelity DNA polymerases are not completely error free. A huge limitation of next generation sequencing is the fact that it cannot detect more than six homopolymeric bases in length (MARDIS 2008), which could lead to artefacts introduced into the sequence and thus these artefact had to be verified by Sanger sequencing, since Sanger sequencing have the ability to accurately detect homopolymeric regions (HERT *et al.* 2008; SCHUSTER 2008; HURD and NELSON 2009).

Table 3.4: Variations found in the *GLYAT* gene of participants illustrating which variations were verified by Sanger sequencing and in which individual each variation occurs.

Reference position	Variation	Verified by Sanger sequencing	Patients with variation
58498835	A/T	Verified	9, 10, 12, 13, 15
58498825	C/G	Verified	1, 2, 3, 4, 5, 7, 9, 11, 10, 12, 13, 14, 15, 17
58497569	A/T		
58495947*	G/A	Verified	4
58495844	C/G	Verified	1, 2, 3, 13, 14
58495760	A/T	Verified	1, 5, 10, 11, 13, 14, 15
58495554	G/T	Verified	5, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17
58495554*	-/A		
58495334*	C/ -	Verified	2, 3, 7, 8, 14, 15, 16
58494339	A/G	Verified	1, 2, 4, 12, 14
58493900	A/T		
58493900*	C/T		
58493889*	C/G		
58493888*	C/T		
58493839*	C/G	Verified	1, 2, 4, 7, 8, 10, 11, 12, 13, 14, 16, 18
58493668	C/T	Verified	1, 2, 4, 7, 8, 10, 11, 12, 13, 14, 16, 18
58493420	A/G	Verified	1, 2, 4, 7, 8, 10, 11, 12, 13, 14, 16, 18

Reference position	Variation	Verified by Sanger sequencing	Patients with variation
58493392	C/T	Verified	1, 2, 4, 7, 8, 10, 11, 12, 13, 14, 16, 18
58493243	C/T	Verified	7
58493212	A/G	Verified	1, 2, 7, 8, 9, 11, 12, 13, 14, 16, 18
58493204	A/G	Verified	1, 2, 7, 8, 9, 11, 12, 13, 14, 16, 18
58493165	C/T	Verified	1, 2, 12, 13, 14, 16, 17
58493165	C/G		
58492875*	-/G	Verified	1, 7
58492666*	C/T	Verified	1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18
58492509*	-/G		
58492088*	A/ -		
58492086*	T/ -		
58492085*	A/-	Verified	1,5,7,8,11,18
58491921	A/-	Verified	1, 2, 4, 12, 13, 14, 16, 17
58491400	A/T	Verified	1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14, 15, 16, 18
58491268	C/T	Verified	7,11,16
58490795	A/G	Verified	3, 5, 8, 9, 14, 15, 18
58490703	C/T	Verified	5, 8, 9, 10, 15, 18
58490702	-/T	Verified	1, 2, 3, 4, 11, 12, 13, 14, 16
58490221	-/T	Verified	1, 5, 7, 9, 10, 11, 12, 13, 14, 15, 17, 18
58489459	C/T	Verified	11
58488855	C/T	Verified	4, 5, 7, 8, 10, 11, 12, 13, 14, 15, 16, 18
58488848	-/T		
58488634	-/T		
58488185*	G/-	Verified	1, 2, 4, 5, 7, 8, 10, 11, 12, 13, 14
58487994	A/C	Verified	11
58487393	C/T	Verified	1, 11, 12, 13, 14
58486917	A/G	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15
58486850	A/C	Verified	1, 2, 3, 5, 7, 8, 9, 10, 11, 13, 14, 15
58486737	A/C		
58486626	A/T		
58486092	C/T		
58486071*	-/T		
58485925*	-/T	Verified	1, 2, 4, 12, 13, 14
58485603	C/T	Verified	11
58484971	C/T	Verified	1, 7, 8, 11, 12, 13
58484656	A/C	Verified	12, 16
58484611	A/G	Verified	13, 14, 16
58484500*	C/T	Verified	15
58484471	C/T	Verified	1, 16
58484233	A/G	Verified	1, 2, 4, 11, 12, 13, 16
58484119	A/G	Verified	1, 2, 4, 11, 12, 13

Reference position	Variation	Verified by Sanger sequencing	Patients with variation
58484117	A/G	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 15
58484086	C/T	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 15
58483921	G/T	Verified	1, 2, 4, 11, 12
58483862	-/C	Verified	2, 4, 11, 12, 13
58483861	-/C		
58483858	-/C	Verified	3, 7, 8, 9, 10
58482423	G/T	Verified	1, 2, 4, 12, 13, 14, 16, 17
58482290	A/G	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18
58482249	A/C	Verified	1, 2, 4, 12, 13, 14, 17
58482187	C/G	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18
58481781	C/T	Verified	1, 2, 3, 5, 7, 8, 9, 10, 11, 12, 13, 15, 16, 17
58481375*	T/-		
58481374*	A/-		
58481374*	A/T		
58481370*	A/ -		
58481350	G/T	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18
58481196*	A/T		
58481193	C/T	Verified	1, 2, 3, 4, 5, 7, 8, 9, 11, 12, 13, 14, 15, 16, 18
58481193*	C/ -		
58480742	G/T	Verified	2, 3, 4, 5, 7, 8, 10, 12, 13, 14, 15, 16, 17, 18
58480738	A/T	Verified	2, 3, 4, 5, 7, 8, 10, 12, 13, 15, 16, 17, 18
58480583	A/T	Verified	2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
58480381*	T/ -		
58479908	A/G	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
58479903	A/G	Verified	11
58479749	C/T	Verified	1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
58479675	C/T	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16
58479305	C/G	Verified	1, 2, 3, 45, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18
58478084	G/A	Verified	1, 2, 3, 4, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18
58478039	C/T	Verified	1, 2, 3, 4, 7, 8, 9, 10, 11, 13, 14, 16
58478026	A/G	Verified	3, 8, 11, 13, 14, 18
58476844	-/GAAAGG		
58476844*	-/CTTTCC	Verified	1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18
58476843*	A/T		9, 12, 13, 15
58476841*	-/A		5, 7, 9, 11, 12, 13, 14, 15, 17
58476250*	A/ -		

* - Novel variation

In addition to verifying the pyrosequencing data, Sanger sequencing also made participant-to-variant linking possible. It was also possible to establish whether the participants were homozygous or heterozygous for specific variations. Representative Sanger sequencing electropherograms of homozygous and heterozygous variations are shown in Figure 3.4. In the Figure it can be clearly seen that a homozygous variation has a single peak, while a heterozygous variation has two peaks at the same position.

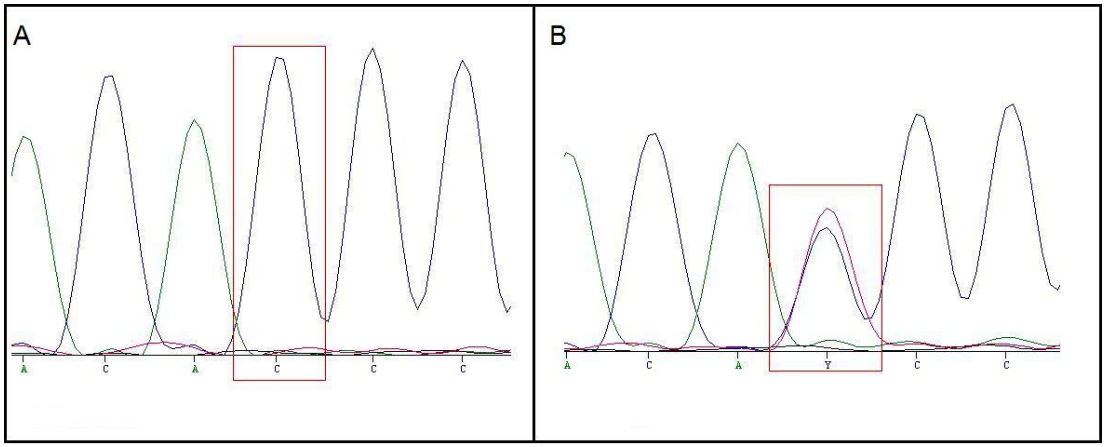

Figure 3.4: Electropherogram sections to illustrate homozygous and heterozygous variations identified with Sanger sequencing. Participant seven is homozygous for the variation found on chromosome position 58493420. This is illustrated in Figure A where only one peak is visible in the position in the red block. Participant 11 is heterozygous for the same variation as is illustrated in Figure B where two peaks are visible at the exact same position (in the red block). The two peaks (nucleotides) found in the same position is an indication of a heterozygous variation.

After the short *GLYAT* fragments of the participants were sequenced (454 pyrosequencing & Sanger sequencing) and the data was analysed, a table of the participants and their variations were drawn up. Table 3.5 lists each participant's variations as well as whether each participant is homozygous or heterozygous for each variation.

Table 3.5: The homozygous and heterozygous variations detected in the *GLYAT* gene of participants. The reference position of each of the variations identified as well as rs numbers of previously described variations are listed in the table below. Variations identified in the *GLYAT* gene of each participant are marked with an x. Blue blocks indicate a homozygous variation whereas a green block indicates a heterozygous variation.

Reference position	Rs number	Variation	Participant number																
			1*	2*	3*	4*	5	7	8*	9*	10	11*	12	13*	14	15	16	17*	18*
58498835	rs1938722	A/T								x	x		x	x		x			
58498825	rs539085	C/G	x	x	x	x	x	x		x	x	x	x	x	x	x		x	
58497569		A/T																	
58495947	rs72927718	G/A				x													
58495844	rs11229602	C/G	x	x		x								x	x				
58495760	rs2497379	A/T	x				x				x	x		x	x	x			
58495554		G/T					x	x	x	x		x	x	x	x	x	x	x	
58495554		-/A																	
58495334	rs520075	C/ -		x	x			x	x						x	x	x		
58494339	rs11229601	A/G	x	x		x							x		x				
58493900		A/T																	
58493900		C/T																	
58493889		C/G																	
58493888		C/T																	
58493839	rs621137	C/G	x	x		x		x	x		x	x	x	x	x		x		x
58493668	rs572856	C/T	x	x		x		x	x		x	x	x	x	x		x		x
58493420	rs570244	A/G	x	x		x		x	x		x	x	x	x	x		x		x
58493392	rs570146	C/T	x	x		x		x	x		x	x	x	x	x		x		x
58493243	rs547895	C/T						x											
58493212	rs591121	A/G	x	x				x	x	x		x	x	x	x		x		x
58493204	rs547178	A/G	x	x				x	x	x		x	x	x	x		x		x
58493165	rs7951340	C/T	x	x									x	x	x		x	x	
58493165		C/G																	
58492875		-/G	x					x											
58492666	rs542537	C/T	x	x	x	x		x	x	x	x	x	x	x	x		x	x	x
58492509		-/G																	
58492088		A/ -																	
58492086		T/ -																	
58492085		A/-	x				x	x	x			x							x
58491921	rs10896818	A/-	x	x		x							x	x	x		x	x	
58491400	rs1938701	A/T	x	x		x	x	x	x		x	x	x	x	x	x	x		x
58491268	rs11229598	C/T						x				x					x		
58490795	rs667445	A/G			x		x		x	x	x				x	x			x
58490703	rs11311719	C/T					x		x	x	x					x			x

Reference position	Rs number	Variation	Participant number																
			1*	2*	3*	4*	5	7	8*	9*	10	11*	12	13*	14	15	16	17*	18*
58490702	rs11357361	-/T	x	x	x	x						x	x	x	x		x		
58490221	rs499891	-/T	x				x	x		x	x	x	x	x	x	x		x	x
58489459	rs75356525	C/T										x							
58488855	rs72514070	C/T				x	x	x	x		x	x	x	x	x	x	x		x
58488848	rs34529443	-/T																	
58488634		-/T																	
58488185	rs1938700	G/-	x	x		x	x	x	x		x	x	x	x	x				
58487994	rs60207372	A/C										x							
58487393	rs1941965	C/T	x									x	x	x	x				
58486917	rs608998	A/G	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
58486850	rs483101	A/C	x	x	x		x	x	x	x	x	x		x	x	x			
58486737	rs72925789	A/C																	
58486626	rs11530801	A/T																	
58486092		C/T																	
58486071		-/T																	
58485925	rs11229597	-/T	x	x		x							x	x	x				
58485603	rs57987309	C/T										x							
58484971	rs542683	C/T	x					x	x			x	x	x					
58484656	rs11229596	A/C											x				x		
58484611		A/G												x	x		x		
58484500		C/T														x			
58484471	rs11229595	C/T	x														x		
58484233	rs11229594	A/G	x	x		x						x	x	x			x		
58484119	rs1938695	A/G	x	x		x						x	x	x					
58484117	rs673192	A/G	x	x	x	x	x	x	x	x	x	x	x	x		x			
58484086	rs572006	C/T	x	x	x	x	x	x	x	x	x	x	x	x		x			
58483921	rs1938696	G/T	x	x		x						x	x						
58483862	rs72514069	-/C		x		x						x	x	x					
58483861	rs72514069	-/C																	
58483858	rs35812997	-/C			x			x	x	x	x								
58482423	rs17152916	G/T	x	x		x							x	x	x		x	x	
58482290	rs628124	A/G	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
58482249	rs72925778	A/C	x	x		x							x	x	x			x	
58482187	rs521200	C/G	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
58481781	rs625951	C/T	x	x	x		x	x	x	x	x	x	x	x		x	x	x	
58481375		T/-																	
58481374		A/-																	
58481374		A/T																	
58481370		A/ -																	

Reference position	Rs number	Variation	Participant number																
			1*	2*	3*	4*	5	7	8*	9*	10	11*	12	13*	14	15	16	17*	18*
58481350	rs522568	G/T	x	x	x	x	x	x		x	x	x	x	x	x	x	x		x
58481196		A/T																	
58481193	rs524234	C/T	x	x	x	x	x	x		x		x	x	x	x	x	x		x
58481193		C/ -																	
58480742	rs611013	G/T		x	x	x	x	x	x		x		x	x	x	x	x	x	x
58480738	rs528136	A/T		x	x	x	x	x	x		x		x	x		x	x	x	x
58480583	rs610165	A/T		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
58480381		T/ -																	
58479908	rs596847	A/G	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
58479903	rs75697298	A/G										x							
58479749	rs558274	C/T	x	x		x	x	x	x	x	x	x	x	x	x	x	x		
58479675	rs1938698	C/T	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		
58479305	rs473287	C/G	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
58478084	rs675815	G/A	x	x	x	x		x	x	x	x	x		x	x	x	x	x	x
58478039	rs675757	C/T	x	x	x	x		x	x	x	x	x		x	x		x		
58478026	rs675423	A/G			x				x			x		x	x				x
58476844	rs5792123	-/GAAAGG																	
58476844		-/CTTTCC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
58476843		A/T																	
58476841		-/A																	
58476250		A/ -																	

* - Participant with low glycine conjugation ability

From Table 3.5 it can be seen that only one participant had a vastly different variation profile than the rest of the participants. Participant 17 has 11 homozygous variations and three heterozygous variations which is a total of 14 variations. The rest of the participants have a total of between 27 and 45 variations. No distinct pattern could be observed between the individuals. Sequence results of participants with low or normal glycine conjugation were compared, but no overall significant differences could be found between participants with low and normal glycine conjugation.

Lino Cardenas et al. (2010) reported three missense mutations in the coding regions of the *GLYAT* gene (LINO CARDENAS *et al.* 2010). Two of these missense mutations with rs numbers, rs10896818 and rs675815 were identified by sequencing. The two mutations caused Ser17Thr and Asn156Ser

substitutions respectively. Participant 12 is heterozygous for the rs10896818 variation. Participants seven, eight, nine, 11, 15 and 18 are homozygous for the rs675815 variation and participants 13, 14 and 16 are heterozygous for the rs10896818 variation as well as homozygous for the rs675815 variation. Participant 17 is homozygous for both the rs10896818 and rs675815 variations. The sequence data of the participant who has the rs10896818 and rs675815 variations was compared to the glycine conjugation of each participant, but no significant differences could be found between the participants with low and normal glycine conjugation. Both participants seven and 12 have normal glycine conjugation, but participant seven has the rs10896818 variation and participant 12 has the rs10896818 variation. Thus no correlation was found between these two variations and glycine conjugation ability based on the glycine conjugation and sequencing data.

Participant 18 was homozygous for all variations found. Participant 18 was originally diagnosed with low glycine conjugation based on low hippuric acid excretion. Benzoic acid is conjugated with glycine to form hippuric acid (LISKA 1998), thus hippuric acid excretion is often used as an indication of glycine conjugation. The data, indicating abnormal hippuric acid excretion, was obtained from Mr E Erasmus. Participant 18 was given sodium-benzoate and the urinary excretion of hippuric acid was measured, but little hippuric acid was detected. Participant 18 started treatment with carnitine where after the detoxification profile of participant 18 was biochemically determined. According to the biochemical detoxification profile, participant 18 has 26.7% glycine conjugation, which is slightly below the reference range of 30-52%. The results of participant 18 may be an indication of a problem preceding GLYAT in the metabolic pathways. As levels of carnitine increase, so does the levels of free CoA. Carnitine regulates intra-mitochondrial free CoA, thus in the case of carnitine deficiency acyl-CoA esters will accumulate, and in effect will cause inhibition of other metabolic pathways which require CoA for the reactions to take place. As previously stated, GLYAT forms part of a two-step reaction, in the first step CoA is added to a substrate to form an acyl-CoA, which is then conjugated with glycine. CoA is needed in the first step of the process and thus a CoA deficiency will impair the reaction. This impaired reaction will, in effect, cause fewer toxins to be removed via glycine conjugation and may be an explanation for what is seen in the hippuric acid excretion and biochemically determined detoxification data of participant 18.

Subsequent to the initial test, Participant 18 was given sodium-benzoate and p-amino benzoate on two separate occasions. The excretion of metabolites of both these substrates was measured over time. Urine samples were collected over a period of time, where after the hippuric acid and p-aminohippuric acid were measured by means of tandem mass spectrometry. Excretion of hippuric acid, a metabolite of sodium-benzoate, did not fit the normal profile (Figure 3.5). Normally, approximately 90% of the initial dose would be excreted as hippuric acid after three hours (WOOD *et al.* 1978), but the excretion of hippuric acid of participant 18 showed peaks at two, five and eight hours after given a sodium-benzoate challenge. Excretion of p-aminohippuric acid derived from p-aminobenzoate for participant 18 was normal, as can be seen in Figure 3.6. Excretion of p-aminohippuric acid peaked at two hours and returned to base levels after a period of four hours. The results of abnormal excretion of hippuric acid, but normal excretion of p-aminohippuric acid might be an indication of a difference in GLYAT substrate specificity of participant 18.

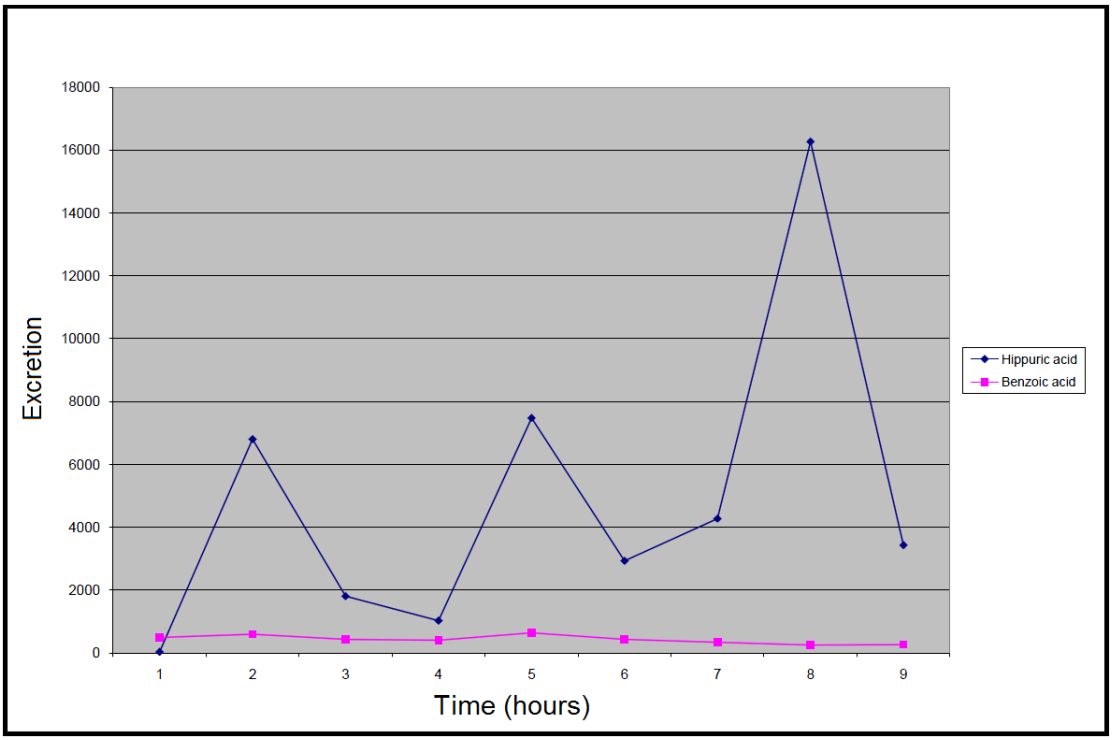

Figure 3.5: Hippuric acid excretion subsequent to a sodium-benzoate challenge. Excretion of hippuric acid was measured as absorbance over time. Hippuric acid excretion of participant 18 derived from sodium-benzoate was abnormal. The experiment was done over nine hours, after which the excretion of hippuric acid in the urine was no longer measured. Hippuric acid excretion peaked at two, five and eight hours. (Data from E. Erasmus)

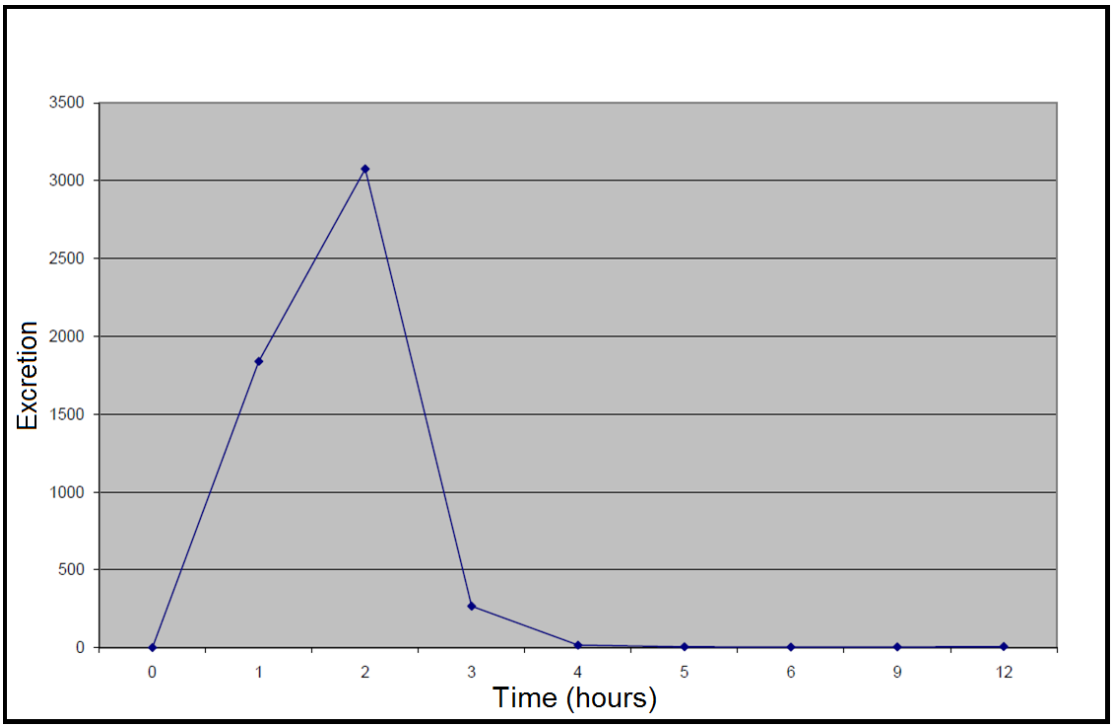

Figure 3.6: p-aminohippuric acid excretion after p-aminobenzoate challenge. Excretion of p-aminohippuric acid was measured as absorbance over time. Excretion of p-aminohippuric acid was normal for participant 18. Excretion peaked at two hours and p-aminohippuric acid was considered to be fully excreted after four hours as the measured values had returned to baseline. (Data from E. Erasmus)

3.3.3 Potential branch points and splice sites

The potential branch points and splice sites were predicted by the Human Splicing Finder (DESMET *et al.* 2009). Potential splice sites deleted, potential splice sites formed, potential branch points deleted as well as potential branch points formed were identified. The results obtained are only a prediction and have to be investigated by further studies. The results obtained from the Human Splicing Finder are shown in Tables 3.6, 3.7, 3.8 and 3.9. In Table 3.6 the results of the potential splice sites deleted and in Table 3.7 the results for the potential splice sites formed are shown. The results for the potential branch points deleted and potential branch points formed are shown in Table 3.8 and Table 3.9 respectively. All variations which could possible influence potential branch points and splice sites were verified by means of Sanger sequencing.

Table 3.6: Potential deleted splice sites, caused by variations identified in the cohort. The upper case characters represent potential exonic regions and the lower case characters represent potential intronic regions.

Splice site chromosome position	Variation chromosome position	Potential splice site deleted
58493386	58493392	TAAgttagc
58493157	58493165	ttttccctcctgCT
58492662	58492666	AAAgtacaa
58489455	58489459	CTGgtgtta
58487390	58487393	TCGgtgtaa
58486839	58486850	ggcttaaggaagTT
58482185	58482187	tggtggcaggagAA

Table 3.7: Potential splice sites that were formed as a result of variations identified in the cohort. The upper case characters represent potential exonic regions and the lower case characters represent potential intronic regions.

Splice site chromosome position	Variation chromosome position	Potential splice site formed
58495753	58495760	CAAgtgttt
58493827	58493839	gcaagaggccagAT
58493414	58493420	GACgtttga
58492652	58492666	aggcttggtgagAG
58491389	58491400	tttctattttagTG
58487382	58487393	atgaacagtcagTG
58486912	58486917	TGTgtcaga
58484458	58484471	tcagaaactcagAA
58482420	58482423	ggacattctaagAA
58482177	58482187	ggcaggacaaagAA
58478072	58478084	tttatctcccagTG

Table 3.8: Potential branch points deleted as a result of identified variations in the cohort. The upper case characters represent potential exonic regions.

Branch point chromosome position	Variation chromosome position	Potential branch point deleted
58479372	58479675	TGCTTAG
58478078	58478084	TCCCAAT

Table 3.9: Potential branch points formed as a result of variations identified in the cohort. The upper case characters represent potential exonic regions and the lower case characters represent potential intronic regions.

Branch point chromosome position	Variation chromosome position	Potential branch point formed
58493834	58493839	GCCAgAT
58490214	58490221	GCCTCAC
58489453	58489459	ACCTGAT
58486915	58486917	GtCAGAT

The participants were compared to one another regarding the variations which could potentially influence branch points and splice sites. Each of the variations which could potentially influence branch points and splice sites was assessed individually. Participants five, 11, 13, 14 and 15 have the variation found on chromosome position 58495760. Participant five and 15 are homozygous for the variation, while participants 11, 13 and 14 are heterozygous. Both participants homozygous for the variation have normal glycine conjugation, while two of the participants with the heterozygous variation have low glycine conjugation and another one has normal glycine conjugation. 17 variations identified in the cohort might have an effect on splice sites or branch points, but none of these 17 variations could be correlated with glycine conjugation as discussed using the variation found at chromosome position 58495760 as an example.

The rs675815 missense mutation identified by Lino Cardenas et al (2010) was identified by the Human Splicing Finder (DESMET *et al.* 2009) as a variation which could potentially cause a branch point to be deleted or a splice site to be formed. This variation was found in the cohort but no correlation could be found between glycine conjugation and the rs675815 mutation as previously stated.

A total of 94 variations were identified from the Next Generation Sequencing data. Four of these found in the exons were known variations and three variations located in the exons were novel. Of the 94 variations identified, 62 known and 25 novel variations were identified in the introns of the *GLYAT* gene. Sanger sequencing verified 68 in total of the variation, which included 12 novel variations, of which one is located in exon six. A total of 17 of the variations identified could possibly cause branch points and splice sites to be formed or deleted.

CHAPTER 4 – COPY NUMBER ASSAYS

4.1 Introduction

A CNV is a segment of DNA which is either deleted or duplicated and can lead to copy number variations in a genetic sequence. According to McCarroll and Altshuler (2007), CNVs can influence human phenotypes. These CNVs can be caused by new structural alterations and inherited variations. Complex diseases might be more susceptible to variations in the exons, which could alter enzyme activity rather than terminating the enzyme function completely (MCCARROLL and ALTSHULER 2007).

Real time PCR can be used as a quantitative nucleic acid analysis, such as gene expression, gene and genome quantitation (HEID *et al.* 1996). Unlike end-point PCR, real time PCR does not need post PCR manipulations to be done in order to analyse the results, which will shorten analysis time and eliminate possible causes of contamination (HEID *et al.* 1996). Real time PCR amplify target DNA exponentially and is measured in real time by means of a 5' fluorogenic exonuclease detection system or a fluorescence dye which bond to dsDNA, such as SYBR green (WILSON and WALKER 2007).

The aim of this study was to identify sequence variations in the *GLYAT* gene such as known SNPs, novel variations and CNVs. Real-time PCR based copy number assays were to be used to identify CNVs, by means of the 5' fluorogenic exonuclease detection system.

The 5' fluorogenic exonuclease detection system make it possible to detect a specific target region, while a fluorescence dye which binds to dsDNA is non-specific (WILSON and WALKER 2007). The 5' fluorogenic exonuclease detection system makes use of a probe binding to a specific target region, consisting of a fluorescence reporter dye and a quencher. The quencher quenches the reporter dye signal when the quencher and reporter dye are in close proximity on the probe. The DNA polymerase will start to amplify the DNA. Once the polymerase reaches the target

region the probe is cleaved. The quencher and reporter dye will be separated and this will increase the fluorescence of the reporter. With each cycle the fluorescence will increase as more probes are cleaved (HEID *et al.* 1996; WILSON and WALKER 2007).

4.2 Materials and Methods

4.2.1 Real-Time quantitative polymerase chain reaction

Real-time PCR (qPCR) was performed using TaqMan® chemistry. The reaction mixture for each reaction is shown in Table 4.1. The TaqMan copy number assays detect the target gDNA. The TaqMan copy number assay consists of two primers for amplification of the target region as well as a Minor Groove Binder (MGB) probe. The MGB probe includes a FAM reporter dye, a non-fluorescent quencher and a Minor Groove Binder. The RNaseP reference assay detects targets which exist in two copies in a normal diploid genome, namely the Ribonuclease P RNA component H1 (H1RNA) gene (RPPH1) on chromosome 14. The RNaseP reference assay makes use of a tetramethylrhodamine (TAMRA) probe. The TAMRA probe includes a VIC reporter dye attached at the 5' end and a TAMRA quencher attached at the 3' end. The quencher quenches the reporter dye signal when the quencher and reporter dye are in close proximity on the probe. The assay locations and their probes are listed in Table 4.2. The copy number assay and reference assay were amplified in duplex during the qPCR in a single reaction. The probe of each assay bound to a specific target on the gDNA. The polymerase started to amplify a piece of DNA using the primers. Once the polymerase reached the target region the probe was cleaved. The quencher and reporter dye was then separated, which increased the fluorescence of the reporter. With each cycle the fluorescence increased as more probes were cleaved. By measuring the fluorescence, in real-time, a plot can be generated.

Table 4.1: Reaction mixture for qPCR (TaqMan) to determine copy number.

Reaction mixture components	Volume per well
2 x TaqMan Universal Mastermix, no UNG	10.0 µl
TaqMan copy number assay, 20 x	1.0 µl
TaqMan Reference assay, 20 x	1.0 µl
Nuclease-free water	4.0 µl
Template gDNA	4.0 µl

Table 4.2: Details of the TaqMan Copy number assay. The assay ID and respective assay location and assay probe is listed in the table. All assays listed were predesigned except for GLYATe6_CCD1s4R, which was a custom designed assay.

Assay ID	Assay Location	Probe
Hs02540133_cn	Chr11:58499417	MGB
Hs01714809_cn	Chr11:58491960	MGB
Hs01519924_cn	Chr11:58482843	MGB
Hs01018714_cn	Chr11:58480251	MGB
Hs01843803_cn	Chr11:58478206	MGB
Hs00776659_cn	Chr11:58477438	MGB
Hs00160286_cn	Chr11:58476986	MGB
Hs02958972_cn	Chr11:58476803	MGB
Hs02818120_cn	Chr11:58476634	MGB
Hs00401731_cn	Chr11:58476384	MGB
GLYATe6_CCD1s4R	Chr11:58476941	MGB
TaqMan Copy Number Reference Assay RNase P	Cytoband 14q11.2	TAMRA

The reaction parameters for the assays were as follows. Each reaction was held at 95˚C for 10 minutes, followed by 40 cycles of 95˚C for 15 seconds and 60˚C for 60 seconds. A quantitation for the copy number assays was performed, which is required to capture the cycle threshold data. When the amplification plot crosses the threshold line, cycle threshold data is generated. The cycle threshold data was then used by the CopyCaller software (Life Technologies) to calculate copy numbers.

4.2.2 Copy number assay data analysis

The copy number assays were analysed using CopyCaller software version 1.0 from Life Technologies (http://www.lifetechnologies.com/za/en/home/technical-resources/software-downloads/copycaller-software.html). Before the data was analysed, the threshold and baseline had to be adjusted. The threshold was changed to 0.2. This was done for all the assays since cycle threshold (Ct) values of the experiments were compared to each other and it was therefore important to be defined in the same way. The baselines for the assays were automatically determined by the on-board software of the 7500 Real-Time PCR System (Life Technologies) thus the baseline eliminated the background noise without overlapping with the area of amplification and therefore the Ct values could be accurately determined. Once the threshold and baseline were adjusted, the data generated

during the qPCR could be exported as text files and opened for data analysis using the CopyCaller software.

CopyCaller uses relative quantitation to calculate the number of copies, with the use of the comparative Ct method. The software measures the Ct values of the assay as well as the reference assay and compares these values. The software not only calculates the copy number, but also the confidence estimate and deviation estimate also known as z-score. The confidence estimate is an indication of the probability that the copy numbers are correctly calculated, while the z-score is a value of the standard deviations of each sample. Confidence estimates and z-scores were taken into consideration when copy numbers calculated by the software were evaluated. The higher the copy number calculated the lower the confidence will be. Copy number calls were accepted or rejected based on the z-scores. Copy number calls were accepted with a z-score lower than 1.75 and copy numbers with a z-score between 1.75 and 2.65 were accepted with caution. Copy number calls with a z-score of higher than 2.65 were rejected.

Genome-Wide Human SNP Array 6.0 data of two participants (participant 16 and 17) was available within our research group. Genome-Wide Human SNP Array 6.0 includes 1.8 million genetic markers, which includes over 906 600 SNPs and 946 000 probes for CNV detection. The data was analysed with Genotyping Console version 4.1.1.384 from Affymetrix. The data from the qPCR analysis was to be compared to the Genome-Wide Human SNP Array 6.0 data in order to verify that the data from the qPCR was accurate.

4.3 Results and Discussion

4.3.1 Real-time polymerase chain reaction

Real-time PCR copy number assays were used to determine copy number variations of 11 target regions in the *GLYAT* gene within the cohort. No optimization was needed for the qPCR copy number assays as the conditions suggested within the protocol yielded results of sufficient quality. The protocol of the supplier, Life Technologies, was followed for the 11 assay reactions. The 11 assays are located on positions chr58499417, chr58491960, chr58482843, chr58480251, chr58478206, chr58477438, chr58476986, chr58476803, chr58476634, chr58476384 and chr58476941 (Table 4.2). All reactions amplified with a sigmoid curve, typical illustrations of which can be seen in Figures 4.1 and 4.2. In Figure 4.1 the amplification plots of the custom-designed, GLYATe6-CCD154R assay of 12 participants are shown and in Figure 4.2 an amplification plot of assay Hs01519924 of participant five is shown. An amplification plot is a graph of the fluorescence signal against the cycle number, the curve is sigmoid if the reaction ran normally. In the initial cycles of qPCR fluorescence levels are below the detection limit. This stage is called the baseline. After the initial stage, amplification can be detected and an increase of fluorescence can be measured and used for further analysis.

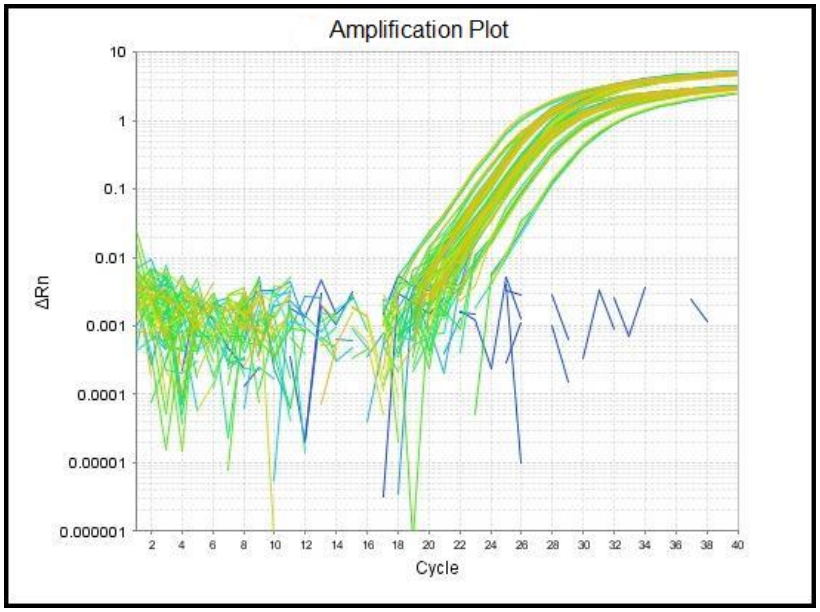

Figure 4.1: Amplification plots of assay GLYATe6-CCD154R of participants five, seven, eight, nine, 11, 12, 13, 14, 15, 16, 17 and 18. The reactions were run for 40 cycles. The yellow and green curves are the GLYATe6-CCD154R assay and reference assay plots respectively. The blue line is the no template control.

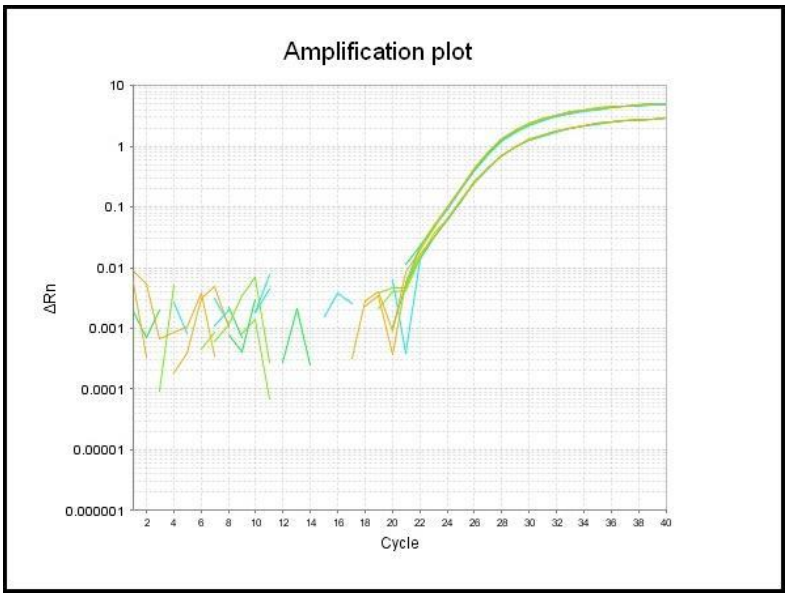

Figure 4.2: Amplification plot of assay Hs01519924 and RNaseP reference assay of participant five. Four replica reactions of the assay of participant five were run. The upper group of curves is that of assay Hs01519924, while the lower group of curves is that of the RNaseP reference assay.

4.3.2 Copy number identification with CopyCaller software

qPCR data of the 11 copy number assays, located on positions chr58499417, chr58491960, chr58482843, chr58480251, chr58478206, chr58477438, chr58476986, chr58476803, chr58476634, chr58476384 and chr58476941, was exported to CopyCaller software from Life Technologies and analysed. The location of the 11 assays on the *GLYAT* gene is represented in Figure 4.3. Analysis was done without a calibrated sample, thus without a sample with a known copy number. The CopyCaller software uses a theoretical model to predict the copy number within a specific target. When a calibrated sample is present, the software compares the Ct values of the test and calibrator samples to predict copy numbers. When a calibrator sample is not present, the software will use a theoretical model to calculate and predict copy numbers. This theoretical model may cause the predicted copy number to be higher or lower than expected. For example, if a copy number of two is expected for the specific target, and the software calculates it as 2.4, the copy number will then be predicted as three. Results for all participants, except participant six, as predicted by the CopyCaller software, are listed in Table 4.3. Unfortunately there was not enough sample material to complete the copy number assays for participant six. Confidence values for each copy number predicted will drop as the copy number increase, thus a participant with two copies will have a high prediction confidence value whereas a participant with three copies will have a lower prediction confidence. The CopyCaller software also calculates z-scores, which is the standard deviation from the mean copy number. As per the CopyCaller software copy numbers with z-scores <1.75 can be accepted, copy numbers with z-scores between 1.75 and 2.65 can be accepted with caution and copy numbers with z scores >2.65 cannot be accepted. In a diploid cell, a number of two copies of each gene is present, but this may vary for some genes, for example AMY1 may have between two and fifteen copies in a diploid cell (WAIN *et al.* 2009). A variation in copy numbers may cause structural changes, which could possible increase susceptibility to disease (WAIN *et al.* 2009). Copy number variations found in the GLYAT gene could possibly alter GLYAT activity by causing structural changes within the GLYAT enzyme.

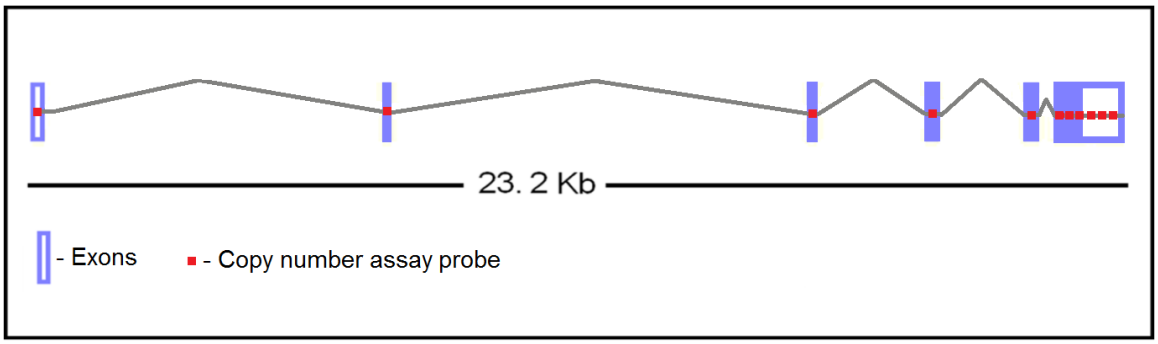

Figure 4.3: Representation of the location of the 11 copy number assay probes on the *GLYAT* gene.

Table 4.3: Summary of copy number variation in each participant listed according to numbers assigned to the participants.

Assay	Participant number																
	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16	17	18
Assay Hs02540133_cn	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2
Assay Hs01714809_cn	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2
Assay Hs01519924_cn	2	2	2	2	2	2	2	2	2	2	2	3	2	2	2	2	2
Assay Hs01018714_cn	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Assay Hs01843803_cn	2	2	2	2	2	2	2	2	2	2	2	3	3	2	2	6	2
Assay Hs00776659_cn	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	6	2
Assay Hs00160286_cn	2	2	2	2	2	2	2	2	2	2	2	3	3	2	2	6	2
Assay Hs02958972_cn	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	6	2
Assay Hs02818120_cn	2	2	2	2	2	2	2	2	2	2	2	3	3	2	2	6	2
Assay Hs00401731_cn	2	2	2	2	2	2	2	2	2	2	2	2	3	2	2	5	2
Assay GLYATe6-CCD154R	2	2	2	2	2	2	2	2	2	2	2	3	3	2	2	6	2

For assay Hs02540133_cn the CopyCaller software predicted two copies for all participants except for participant 13, for which the software predicted three copies. The z-scores for all participants, except participant 16, were below 1.75 and thus these copy number variations can be accepted. The z-score for participant 16 for assay Hs02540133_cn was 2.41, which is above 1.75, but below 2.65 and thus the copy number predicted was accepted with caution. Table 4.4 summarizes calculated and predicted copy numbers as well as confidence values and z-scores for assay Hs02540133_cn of the cohort.

Table 4.4: Copy number information for assay Hs02540133_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	2.08	2	> 0.990	0.95
2	2.05	2	> 0.990	0.36
3	1.97	2	> 0.990	0.03
4	1.97	2	> 0.990	0.03
5	2.01	2	> 0.995	0.17
7	1.95	2	> 0.995	0.01
8	1.94	2	> 0.995	0.01
9	2.03	2	> 0.995	0.30
10	2.04	2	> 0.990	0.07
11	2.02	2	> 0.995	0.26
12	2.04	2	> 0.995	0.50
13	2.53	3	0.730	1.62
14	2.02	2	> 0.995	0.20
15	2.00	2	> 0.995	0.09
16	1.62	2	> 0.995	2.41*
17	1.97	2	> 0.995	0.02
18	2.06	2	> 0.995	0.55

* - accepted with caution z-score between 1.75 and 2.65

The calculated copy numbers for assay Hs01714809_cn are shown in Table 4.5. The CopyCaller software predicted two copies for the target region for all participants except for participant 16 for whom three copies were predicted. Table 4.5 show the different confidence values, z-score, calculated copy numbers and predicted copy numbers for the cohort. All the participants had z-scores below 1.75 except for participant 17, which had a z-score of 1.83. Thus the copy numbers for all the participants except for participant 17 could be accepted. The copy number for assay Hs01714809_cn of participant 17 was accepted, but with caution.

Table 4.5: Copy number information for assay Hs01714809_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	1.94	2	> 0.990	0.34
2	1.94	2	> 0.990	0.23
3	2.05	2	> 0.990	0.51
4	1.95	2	> 0.990	0.18
5	1.95	2	> 0.995	0.06
7	1.85	2	> 0.995	0.67
8	1.85	2	> 0.995	0.70
9	1.96	2	> 0.995	0.06
10	1.99	2	> 0.990	0.04
11	1.99	2	> 0.995	0.03
12	1.98	2	> 0.995	0.05
13	2.44	2	0.980	1.30
14	1.92	2	> 0.995	0.24
15	2.03	2	> 0.995	0.08
16	3.35	3	> 0.995	0.00
17	2.17	2	> 0.995	1.83*
18	2.00	2	> 0.995	0.04

* - accepted with caution z-score between 1.75 and 2.65

For the assay Hs01519924_cn target region, a copy number of two were predicted for all the participants except for participant 13 which had three copies predicted for the target region based on the calculated copy number as shown in Table 4.6. The confidence values, z-scores, predicted copy numbers as well as calculated copy numbers for assay Hs01519924_cn are summarized in Table 4.6. The z-score for all the participants were below 1.75 and thus the predicted copy numbers could be accepted.

Table 4.6: Copy number information for assay Hs01519924_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	2.03	2	> 0.990	0.65
2	2.00	2	> 0.990	0.12
3	1.99	2	> 0.990	0.09
4	1.97	2	> 0.990	0.07
5	1.98	2	> 0.995	0.11
7	2.03	2	> 0.995	0.44
8	1.88	2	> 0.995	0.92
9	2.03	2	> 0.995	0.15
10	2.03	2	> 0.990	0.31
11	2.00	2	> 0.995	0.09
12	2.02	2	> 0.995	0.42
13	2.57	3	0.680	1.28
14	1.99	2	> 0.995	0.09
15	2.00	2	> 0.995	0.07
16	1.85	2	> 0.995	1.16
17	2.02	2	> 0.995	0.29
18	2.01	2	> 0.995	0.16

For the Hs01018714_cn target region, the software predicted that all participants had two copies of the specific target region. The z-scores, confidence values and copy numbers, both calculated and predicted, by the CopyCaller software, are listed in Table 4.7. The z-scores for all participants, except participant 13, were below 1.75, which indicated that the copy numbers predicted could be accepted. The z-score for participant 13 was 2.16, which was above 1.75, but still below 2.65, indicating that the copy number predicted could be accepted, but with caution.

Table 4.7: Copy number information for assay Hs01018714_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	2.10	2	> 0.990	0.87
2	2.03	2	> 0.990	0.34
3	1.99	2	> 0.990	0.05
4	1.97	2	> 0.990	0.02
5	2.01	2	> 0.995	0.01
7	1.92	2	> 0.995	0.26
8	1.91	2	> 0.995	0.30
9	1.98	2	> 0.995	0.03
10	2.00	2	> 0.990	0.04
11	1.97	2	> 0.995	0.03
12	2.00	2	> 0.995	0.02
13	2.45	2	0.910	2.16*
14	2.01	2	> 0.995	0.01
15	1.97	2	> 0.995	0.06
16	1.76	2	> 0.995	1.35
17	2.03	2	> 0.995	0.02
18	2.09	2	> 0.995	0.25

* - accepted with caution z-score between 1.75 and 2.65

Two copies were calculated and predicted, for assay Hs01843803_cn for all participants, except for participants 13, 14 and 17. Three copies were predicted for participants 13 and 14, while six copies were predicted for participant 17 for the target region. The confidence values and z-score as well as the copy numbers, calculated and predicted, are given in Table 4.8. Confidence estimate and z-scores could not be calculated by the software for participants one, two, three, four and 10. Possible causes include too low number of samples tested in the same qPCR run and zero copies for the target region for the samples. The qPCR run for participants one, two, three, four and 10 included seven samples, which is a sufficient number for the software to calculate the confidence and z-scores. The results for the participants without confidence values and z-scores could not be accepted, since the software calculated two copies in each of them, but the lack of confidence values suggest that the participants had zero copies. These results were not confirmed and will have to be done in future studies. All of the copy numbers predicted, for the

remaining participants, for the target region for all the participants had a z-score below 1.75 and thus all these copy numbers could be accepted.

Table 4.8: Copy number information for assay Hs01843803_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	2.04	2	N/A*	N/A*
2	2.02	2	N/A*	N/A*
3	1.98	2	N/A*	N/A*
4	1.99	2	N/A*	N/A*
5	1.93	2	> 0.995	0.02
7	1.55	2	> 0.995	1.61
8	2.18	2	0.950	0.79
9	1.85	2	> 0.995	0.12
10	1.98	2	N/A*	N/A*
11	1.84	2	> 0.995	0.12
12	1.88	2	> 0.995	0.03
13	2.46	3	< 0.455	1.15
14	2.92	3	0.970	0.00
15	1.99	2	0.990	0.02
16	2.19	2	0.930	0.9
17	6.37	6	0.680	0.01
18	2.14	2	0.970	0.55

* - not accepted, unable to calculate confidence and z-scores

Two copies were predicted from calculated copy numbers for target Hs00776659_cn in all participants, except participant 14 and 17, where three and six copies were calculated respectively. In Table 4.9 the confidence values, z-score, predicted copy numbers and calculated copy numbers are listed for assay Hs00776659_cn. Z-scores for the copy numbers for participants five, seven, eight, nine, 11, 12, 13, 14, 15, 16, 17 and 18 listed in Table 4.9 were all below 1.75 and thus all copy numbers predicted for the participants for assay Hs00776659_cn could be accepted. Confidence estimate and z-scores could not be calculated by the software for participants one, two, three, four and 10. The results for the participants without confidence values and z-scores could not be accepted, since the software calculated two copies in each of them, but the lack of confidence values suggest that the

participants had zero copies. These results were not confirmed and will have to be done in future studies.

Table 4.9: Copy number information for assay Hs00776659_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	1.95	2	N/A*	N/A*
2	1.95	2	N/A*	N/A*
3	2.05	2	N/A*	N/A*
4	1.98	2	N/A*	N/A*
5	1.93	2	> 0.995	0.03
7	1.57	2	> 0.995	1.62
8	2.15	2	0.990	0.53
9	1.82	2	> 0.995	0.39
10	2.03	2	N/A*	N/A*
11	1.87	2	> 0.995	0.13
12	1.80	2	> 0.995	0.46
13	2.33	2	0.890	1.22
14	2.83	3	0.830	0.03
15	2.01	2	> 0.995	0.02
16	2.18	2	0.980	0.63
17	5.60	6	0.800	0.13
18	2.08	2	0.990	0.20

* - not accepted, unable to calculate confidence and z-scores

Copy numbers are predicted for each participant based on the calculated copy numbers. All participants, except participants 13, 14 and 17, had a copy number of two. Participants 13 and 14 had three copies, while participant 17 had a copy number of six in the target region for assay Hs00160286_cn. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for participants for assay Hs00160286_cn are summarized in Table 4.10. The z-scores for participants five, seven, eight, nine, 11, 12, 13, 14, 15, 16, 17 and 18 were below 1.75, except for participant seven, which indicated that that copy numbers predicted were true and could be accepted. The results for participant 7 for the target region had a z-score of 1.78, which was above 1.75 but still below 2.65. The z-score calculated indicated that the copy number predicted had to be accepted with caution,

but could still be accepted. Confidence estimate and z-scores could not be calculated by the software for participants one, two, three, four and 10. The results for the participants without confidence values and z-scores could not be accepted, since the software calculated two copies in each of them, but the lack of confidence values suggest that the participants had zero copies. These results were not confirmed and will have to be done in future studies.

Table 4.10: Copy number information for assay Hs00160286_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	1.98	2	N/A**	N/A**
2	1.98	2	N/A**	N/A**
3	1.97	2	N/A**	N/A**
4	2.00	2	N/A**	N/A**
5	1.92	2	> 0.995	0.01
7	1.51	2	> 0.995	1.78*
8	2.12	2	0.970	0.52
9	1.89	2	> 0.995	0.02
10	2.07	2	N/A**	N/A**
11	1.95	2	> 0.995	0.02
12	1.91	2	> 0.995	0.02
13	2.51	3	< 0.455	0.91
14	3.12	3	0.940	0.00
15	2.05	2	0.980	0.26
16	2.03	2	0.990	0.09
17	6.33	6	0.660	0.09
18	2.15	2	0.940	0.73

* - accepted with caution z-score between 1.75 and 2.65

** - not accepted, unable to calculate confidence and z-scores

In the target region for assay Hs02958972_cn two copies were predicted for all participants, except participant 14 where three copies were predicted for the target region and participant 17 where six copies were predicted. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for the participants for assay Hs02958972_cn are listed in Table 4.11. The CopyCaller

software calculated the z-scores for all the participants to be below 1.75 and thus all the predicted copy numbers could be accepted for assay Hs02958972_cn. It can be seen in Table 4.11 that the confidence values for participant 17 is the lowest, since the most copy numbers were predicted for participant 17 within the specific target region. Confidence estimate and z-scores could not be calculated by the software for participants one, two, three, four and 10. The results for the participants without confidence values and z-scores could not be accepted, since the software calculated two copies in each of them, but the lack of confidence values suggest that the participants had zero copies. These results were not confirmed and will have to be done in future studies.

Table 4.11: Copy number information for assay Hs02958972_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	2.04	2	N/A*	N/A*
2	1.98	2	N/A*	N/A*
3	2.00	2	N/A*	N/A*
4	1.99	2	N/A*	N/A*
5	1.90	2	> 0.995	0.02
7	1.69	2	0.990	0.35
8	2.16	2	0.960	0.78
9	2.00	2	0.980	0.01
10	1.94	2	N/A*	N/A*
11	1.93	2	> 0.995	0.02
12	1.89	2	> 0.995	0.01
13	2.09	2	0.970	0.10
14	2.97	3	0.970	0.00
15	2.05	2	0.990	0.23
16	2.14	2	0.960	0.66
17	5.93	6	0.840	0.01

* - not accepted, unable to calculate confidence and z-scores

In the target region for assay Hs02818120_cn two copies were predicted for all participants, except participants 13, 14 and 17. Three copy numbers were predicted for the target region in participants 13 and 14 and six copies were predicted for

participant 17. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for the participants are listed in Table 4.12. From the table it is clear that all copy numbers predicted for participants, five, seven, eight, nine, 11, 12, 13, 14, 15, 16, 17 and 18 can be considered as correct, since the z-scores were all below 1.75. Confidence estimate and z-scores could not be calculated by the software for participants one, two, three, four and 10. The results for the participants without confidence values and z-scores could not be accepted, since the software calculated two copies in each of them, but the lack of confidence values suggest that the participants had zero copies. These results were not confirmed and will have to be done in future studies.

Table 4.12: Copy number information for assay Hs02958972_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	1.95	2	N/A*	N/A*
2	1.94	2	N/A*	N/A*
3	1.93	2	N/A*	N/A*
4	1.96	2	N/A*	N/A*
5	1.92	2	> 0.995	0.01
7	1.53	2	> 0.995	1.71
8	2.18	2	0.930	0.85
9	1.84	2	> 0.995	0.10
10	2.05	2	N/A*	N/A*
11	1.90	2	> 0.995	0.01
12	1.84	2	> 0.995	0.12
13	2.45	3	< 0.455	1.12
14	3.02	3	0.960	0.00
15	2.03	2	0.990	0.12
16	2.17	2	0.920	0.73
17	5.79	6	0.690	0.01
18	2.05	2	0.990	0.09

* - not accepted, unable to calculate confidence and z-scores

Two copies were predicted for the target region for assay Hs00401731_cn in all participants, except participant 14 and 17. Three copies were predicted for participant 14 and six copies were predicted for participant 17 for the target region.

Table 4.13 summarizes the calculated copy numbers, predicted copy numbers, confidence values and z-scores data generated by the CopyCaller software for assay Hs00401731_cn. The z-scores for the target region for participants five, seven, eight, nine, 11, 12, 13, 14, 15, 16, 17 and 18 were below 1.75 as can be seen in Table 4.13. This indicated that all the predicted copy numbers could be accepted. Confidence estimate and z-scores could not be calculated by the software for participants one, two, three, four and 10. The results for the participants without confidence values and z-scores could not be accepted, since the software calculated two copies in each of them, but the lack of confidence values suggest that the participants had zero copies. These results were not confirmed and will have to be done in future studies.

Table 4.13: Copy number information for assay Hs00401731_cn for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	1.97	2	N/A*	N/A*
2	1.95	2	N/A*	N/A*
3	1.98	2	N/A*	N/A*
4	2.02	2	N/A*	N/A*
5	1.46	2	> 0.995	1.42
7	1.44	2	0.980	1.57
8	2.07	2	0.995	0.06
9	1.84	2	> 0.995	0.29
10	2.08	2	N/A*	N/A*
11	1.92	2	> 0.995	0.06
12	1.87	2	> 0.995	0.10
13	2.38	2	0.870	1.44
14	2.98	3	0.950	0.00
15	2.04	2	> 0.995	0.06
16	2.12	2	0.990	0.26
17	5.61	6	0.630	0.12
18	2.06	2	> 0.995	0.09

* - not accepted, unable to calculate confidence and z-scores

Assay GLYATe6-CCD154R is a custom-designed assay within exon six of the *GLYAT* gene with a target at chromosome position 58476941. Two copies were predicted for the target region in all participants, except participants 13, 14 and 17. Three copies were predicted for the target region for participants 13 and 14. A copy number of six was predicted in participant 17 for the target region of assay GLYATe6-CCD154R. Calculated copy numbers, predicted copy numbers, confidence values and z-scores obtained from the CopyCaller software, are listed in Table 4.14. The z-scores of all participants were below 1.75, except for participant seven, where the z-score was 1.86 as well as participants one, two, three, four and 10, where the confidence estimate and z-scores could not be calculated. The z-scores indicated that the predicted copy number for all participants could be accepted, with the exception of participant seven, which could be accepted with caution as it has a z-score above 1.75 but still below 2.65. The results for the participants without confidence values and z-scores could not be accepted, since the software calculated two copies in each of them, but the lack of confidence values suggest that the participants had zero copies. These results were not confirmed and will have to be done in future studies.

Table 4.14: Copy number information for assay GLYATe6-CCD154R for the cohort. The calculated copy numbers, predicted copy numbers, confidence values and z-scores for each participant are presented in the table.

Participant	Calculated copy number	Predicted copy number	Confidence	z-score
1	2.06	2	N/A**	N/A**
2	2.01	2	N/A**	N/A**
3	1.95	2	N/A**	N/A**
4	1.95	2	N/A**	N/A**
5	1.96	2	> 0.995	0.01
7	1.54	2	> 0.995	1.86*
8	2.16	2	0.960	0.64
9	1.82	2	> 0.995	0.33
10	2.14	2	N/A**	N/A**
11	1.94	2	> 0.995	0.02
12	1.91	2	> 0.995	0.02
13	2.56	3	< 0.455	0.90
14	2.93	3	0.960	0.00
15	2.02	2	> 0.995	0.04
16	2.20	2	0.950	0.75
17	6.09	6	0.810	0.01
18	2.15	2	0.970	0.64

* - accepted with caution z-score between 1.75 and 2.65

** - not accepted, unable to calculate confidence and z-scores

CNV data of participants 13, 14 and 17 differed from the rest of the participants. Data for assays Hs02540133_cn, Hs01519924_cn, Hs01843803_cn, Hs00160286_cn, Hs02818120_cn and GLYATe6-CCD154R indicated that participant 13 had three copies in the target regions of the aforementioned assays. Participant 14 had three copies in target regions of assays Hs01843803_cn, Hs00776659_cn, Hs00160286_cn, Hs02958972_cn, Hs02818120_cn, Hs00401731_cn and GLYATe6-CCD154R, while participant 17 had six copies in the same target regions with the exception of assay Hs00401731_cn where five copies were predicted. Participant 13 had a calculated glycine conjugation of 14.7% and participant 17 had 27.7% glycine conjugation, which is below the reference range of 30% – 53%. Participant 14 had a calculated glycine conjugation ability of 31.3%, which is at the bottom of the reference range. The increased copies towards the end of the *GLYAT* gene (exons

five and six) could lead to reduced conjugation efficiency since the exon six codes for the active site of GLYAT.

4.3.3 SNP6 chip copy number results

SNP6 chip data was generated by Me Lizelle Zandberg and the Centre for Proteomic & Genomic Research (CPGR). Two participants from the cohort, participants 16 and 17, were included in the SNP6 chip analysis. The SNP6 chip included the following CNV target regions, chr58477685, chr58478023, chr58478039, chr58478141, chr58479252, chr58479305, chr58479675, chr58483288, chr58484471, chr58484706, chr58488664, chr58490221, chr58493392, chr584943239 and chr584950891. The SNP6 chip data indicated that participants 16 and 17 had two copies in each of the target regions of the SNP6 chip in the *GLYAT* gene. This is not the case with the qPCR data, which indicated that participant 17 has as much as six copies in some target regions. The target regions of the SNP6 chip and qPCR copy number assays were compared. It was found that the target regions of the SNP6 chip and qPCR copy number assays, for the *GLYAT* gene, differ. Three target regions of the SNP6 chip, chr58478023, chr58478039 and chr58478141 are in close proximity to one of the target regions of the qPCR copy number assays, chr58478206. Target region chr58478023 is located 183bp, chr58478039 is located 167bp and chr58478141 is located 65bp from the target region (chr58478206) of the qPCR copy number assay. Probes are usually 100-200bp in length, which indicated that there was a possibility that the three target regions of the SNP6 chip might overlap with the target region mentioned above from the qPCR copy number assay. Life Technologies (Applied Biosystems) does not disclose details surrounding the copy number assays, thus we were unable to determine whether the target regions for the SNP6 chip and qPCR copy number assays overlapped. The *GLYAT* data from the SNP6 chip and the data generated from the qPCR copy number assays could therefore not be used to verify each other.

CHAPTER 5 – CONCLUDING SUMMERY

In some cases of inborn errors of metabolism (IEM) toxic metabolites can accumulate because of a defective enzyme, as is the case in Isovaleric acidemia. The metabolites before the block, caused by the defective enzyme, will accumulate and can cause damage to the DNA, RNA and other cellular structures (SWEETMAN and WILLIAMS ; GUAN *et al.* 2007). These metabolites must be removed from the body via alternative metabolic pathways and detoxification. The clinical presentation of individuals with identical IEMs can differ from one another. This can be true even if these patients receive identical treatment, such as glycine supplementation (MANCINELLI *et al.* 2000). The detoxification of healthy individuals also shows differences indicating it is not just a problem limited to individuals with an IEM. Variations found in the *GLYAT* gene could have an effect on glycine N-acyltransferase activity and might play an important role in the observed variability of detoxification efficiency. The main aim of this study was to identify sequence variations in the *GLYAT* gene such as known SNPs, novel variations and CNVs.

A total of 68 variations were identified in the cohort and 12 of these variations were novel, whereof one was located in exon six on position chr58476844. Sequence results of participants with low or normal glycine conjugation were compared, but no overall significant differences could be found between participants with low and normal glycine conjugation. Data on variations in the *GLYAT* gene could not be used on its own to explain lower GLYAT activity in the cohort. In conjunction with GLYAT activity assays, a correlation could possibly be made between the variations identified within the cohort and their glycine conjugation ability. The two mutations causing Ser17Thr and Asn156Ser substitutions (positions chr58491921 and chr58478084) were identified in 11 of the participants. When these variations were expressed and enzymatically characterized, it was found that Ser17Thr had similar enzyme activity to the wild type GLYAT and Asn156Ser had increased activity compared to the wild type GLYAT (VAN DER SLUIS *et al.* 2013). This does not conform to the observed glycine conjugation activity of the participants. These mutations were individually expressed and not in combination. The effect of these variations in combination of the two variations on GLYAT enzyme activity should be investigated, since some participants have a combination of the two variations. This leaves room for further studies based on recombinant GLYAT enzymes.

The results of abnormal excretion of hippuric acid, but normal excretion of p-aminohippuric acid of participant 18, as described in section 3.3.2, might be an indication of a difference in GLYAT substrate specificity. Further studies will have to be done to investigate the possibility of substrate specificity. A recombinant enzyme of combinations of variations, as found in the individuals with impaired glycine conjugation, will have to be expressed. For the investigation of substrate specificity a wild type enzyme as well as recombinant enzyme of variation which might influence enzyme activity of participant 18 should be engineered. The wild type enzyme will serve as a positive control of GLYAT enzyme activity of different substrates, while the recombinant enzyme of participant 18 would enable an investigation of different substrates and thus it could be determined whether the recombinant enzyme of participant 18 is specific for any substrates. Further a recombinant enzyme should be engineered to include the novel variation found in exon 6 on position chr58476844 (-/CTTTCC).

Copy number variations as determined by the qPCR copy number assays could not be verified by the data obtained from the SNP6 chip. Six copies were calculated for participant 17 for the target region, chr58478206, which does not correspond with three target regions, chr58478023, chr58478039 and chr58478141, in close proximity to target region chr58478206. A possible cause could be that the sample gDNA used was too concentrated and thus more copies of the gene of interest could have been added to the reaction. This high lights a possible short coming of the qPCR copy number assay. This could be overcome by quantifying the gDNA using qPCR, which is more accurate than the conventional methods. A standard curve will have to be set up using gDNA and the TaqMan copy number reference assay RNaseP. This will allow one to accurately quantify sample gDNA and add the correct amount of gDNA to the reaction.

Variations could possibly affect GLYAT activity, but no correlation could be made between variations identified during this study and the cohort's detoxification ability. Further studies need to be conducted to establish the effect of the variations in combination with one another on GLYAT activity. The effect of these variations on

GLYAT activity might shed some light on variability observed between individuals' glycine conjugation ability. Such research would be of great value in treatment of inborn errors of metabolism (IEM).

The aims of the study in short, included amplification of the *GLYAT* gene in a cohort, identification of variations in the *GLYAT* gene and if possible to correlate variations found in the *GLYAT* gene to GLYAT enzyme activity. The *GLYAT* gene was amplified and sequenced in a cohort with possible altered GLYAT activity. A total of 68 variations were identified within the cohort, 12 were novel variations, whereof one was located in exon six. The variations found were investigated, but attempts to correlate the variations to GLYAT activity were unsuccessful. Further investigation is needed to establish whether combinations of variations identified will affect GLYAT activity.

REFERENCES

Alberti, A., P. Pirrone, M. Elia, R. H. Waring and C. Romano, 1999 Sulphation deficit in "low-functioning" autistic children: a pilot study. Biol Psychiatry 46**:** 420-424.

Badenhorst, C. P., R. van der Sluis, E. Erasmus and A. A. van Dijk, 2013 Glycine conjugation: importance in metabolism, the role of glycine N-acyltransferase, and factors that influence interindividual variation. Expert Opin Drug Metab Toxicol 9**:** 1139-1153.

Barth, A., L. B. Nguyen, L. Barth and D. W. Newell, 2005 Glycine-induced neurotoxicity in organotypic hippocampal slice cultures. Exp Brain Res 161**:** 351-357.

Biomatrix, 2009 A Critical Organ Demands an Advanced Formula, pp. BioMatrix Nutraceuticals

Chorley, B. N., X. Wang, M. R. Campbell, G. S. Pittman, M. A. Noureddine *et al.*, 2008 Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659**:** 147-157.

Cobb, B. D., and J. M. Clarkson, 1994 A simple procedure for optimising the polymerase chain reaction (PCR) using modified Taguchi methods. Nucleic Acids Res 22**:** 3801-3805.

Court, M. H., S. X. Duan, L. L. von Moltke, D. J. Greenblatt, C. J. Patten *et al.*, 2001 Interindividual variability in acetaminophen glucuronidation by human liver microsomes: identification of relevant acetaminophen UDP-glucuronosyltransferase isoforms. J Pharmacol Exp Ther 299**:** 998-1006.

de Assis, D. R., R. C. Maria, G. C. Ferreira, P. F. Schuck, A. Latini *et al.*, 2006 Na+, K+ ATPase activity is markedly reduced by cis-4-decenoic acid in synaptic plasma membranes from cerebral cortex of rats. Exp Neurol 197**:** 143-149.

Desmet, F. O., D. Hamroun, M. Lalande, G. Collod-Beroud, M. Claustres *et al.*, 2009 Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res 37**:** e67.

Dorne, J. L., K. Walton and A. G. Renwick, 2004 Human variability for metabolic pathways with limited data (CYP2A6, CYP2C9, CYP2E1, ADH, esterases, glycine and sulphate conjugation). Food Chem Toxicol 42**:** 397-421.

Ensembl, 2009 GLYAT, pp. Ensembl

Ewing, B., and P. Green, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8**:** 186-194.

Ewing, B., L. Hillier, M. C. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8**:** 175-185.

Feuk, L., A. R. Carson and S. W. Scherer, 2006 Structural variation in the human genome. Nat Rev Genet 7**:** 85-97.

Fojo, S. S., U. Beisiegel, U. Beil, K. Higuchi, M. Bojanovski *et al.*, 1988 Donor splice site mutation in the apolipoprotein (Apo) C-II gene (Apo C-IIHamburg) of a patient with Apo C-II deficiency. J Clin Invest 82**:** 1489-1494.

Guan, X. J., W. X. Zhang, C. C. Li, Y. M. Zheng, L. Lin *et al.*, 2007 The role of external signal regulated kinase and transforming growth factor beta(1) in asthma airway remodeling and regulation of glucocorticoids. Zhonghua Yi Xue Za Zhi 87**:** 1767-1772.

Haderslev, K. V., J. Sonne, H. E. Poulsen and S. Loft, 1998 Paracetamol metabolism in patients with ulcerative colitis. Br J Clin Pharmacol 46**:** 513-516.

Haoxing, Z., L. Qingyu, L. Jie, Z. Zhaomin, X. Fang *et al.*, 2007 Molecular Cloning and Characterization of a Novel Human Glycine-N-acyltransferase Gene GLYATL1, Which Activates Transcriptional Activity of HSE Pathway. INT J MOL SCI 8**:** 433-444.

Heid, C. A., J. Stevens, K. J. Livak and P. M. Williams, 1996 Real time quantitative PCR. Genome Res 6**:** 986-994.

Hert, D. G., C. P. Fredlake and A. E. Barron, 2008 Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. Electrophoresis 29**:** 4618-4626.

Hirom, P. C., P. Millburn, R. L. Smith and R. T. Williams, 1972 Species variations in the threshold molecular-weight factor for the biliary excretion of organic anions. Biochem J 129**:** 1071-1077.

Hurd, P. J., and C. J. Nelson, 2009 Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic 8**:** 174-183.

Ito, T., K. Kidouchi, N. Sugiyama, M. Kajita, T. Chiba *et al.*, 1995 Liquid chromatographic-atmospheric pressure chemical ionization mass spectrometric analysis of glycine conjugates and urinary isovalerylglycine in isovaleric acidemia. J Chromatogr B Biomed Appl 670**:** 317-322.

Kasuya, F., K. Igarashi and M. Fukui, 1996 Participation of a medium chain acyl-CoA synthetase in glycine conjugation of the benzoic acid derivatives with the electron-donating groups. Biochem Pharmacol 51**:** 805-809.

Kuehl, G. E., J. Bigler, J. D. Potter and J. W. Lampe, 2006 Glucuronidation of the aspirin metabolite salicylic acid by expressed UDP-glucuronosyltransferases and human liver microsomes. Drug Metab Dispos 34**:** 199-202.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409**:** 860-921.

Lino Cardenas, C. L., J. Bourgine, C. Cauffiez, D. Allorge, J. M. Lo-Guidice *et al.*, 2010 Genetic polymorphisms of glycine N-acyltransferase (GLYAT) in a French Caucasian population. Xenobiotica 40**:** 853-861.

Liska, D. J., 1998 The detoxification enzyme systems. Altern Med Rev 3**:** 187-198.

Mancinelli, L., M. Cronin and W. Sadee, 2000 Pharmacogenomics: the promise of personalized medicine. AAPS PharmSci 2**:** E4.

Mardis, E. R., 2008 Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9**:** 387-402.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader *et al.*, 2005 Genome Sequencing in Open Microfabricated High Density Picoliter Reactors. Nature 437**:** 376-390.

Masoud, S. A., L. B. Johnson and F. F. White, 1992 The sequence within two primers influences the optimum concentration of dimethyl sulfoxide in the PCR. PCR Methods Appl 2**:** 89-90.

Mawal, Y., K. Paradis and I. A. Qureshi, 1997 Developmental profile of mitochondrial glycine N-acyltransferase in human liver. J Pediatr 130**:** 1003-1007.

McCarroll, S. A., and D. M. Altshuler, 2007 Copy-number variation and association studies of human disease. Nat Genet 39**:** S37-42.

Metzker, M. L., 2010 Sequencing technologies - the next generation. Nat Rev Genet 11**:** 31-46.

Nebert, D. W., and J. S. Felton, 1976 Importance of genetic factors influencing the metabolism of foreign compounds. Fed Proc 35**:** 1133-1141.

Newell, D. W., A. Barth, T. N. Ricciardi and A. T. Malouf, 1997 Glycine causes increased excitability and neurotoxicity by activation of NMDA receptors in the hippocampus. Exp Neurol 145**:** 235-244.

Obenrader, S., 2007 The Sanger Method, pp., edited by S. Obenrader.

Pacifici, G. M., A. Viani, M. Franchi, S. Santerini, A. Temellini *et al.*, 1990 Conjugation pathways in liver disease. Br J Clin Pharmacol 30**:** 427-435.

Pritchard, J. K., 2001 Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69**:** 124-137.

Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris *et al.*, 2012 A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13**:** 341.

Ronaghi, M., M. Uhlen and P. Nyren, 1998 A sequencing method based on real-time pyrophosphate. Science 281**:** 363, 365.

Rothberg, J. M., W. Hinz, T. M. Rearick, J. Schultz, W. Mileski *et al.*, 2011 An integrated semiconductor device enabling non-optical genome sequencing. Nature 475**:** 348-352.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409**:** 928-933.

Sanger, F., S. Nicklen and A. R. Coulson, 1977 DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74**:** 5463-5467.

Schuster, S. C., 2008 Next-generation sequencing transforms today's biology. Nat Methods 5**:** 16-18.

Sweetman, L., and J. C. Williams, Chapter 93: Branched Chain Organic Acidurias, pp. The McGraw-Hill Companies, Inc.

van der Sluis, R., C. P. Badenhorst, F. H. van der Westhuizen and A. A. van Dijk, 2013 Characterisation of the influence of genetic variations on the enzyme activity of a recombinant human glycine N-acyltransferase. Gene 515**:** 447-453.

van der Westhuizen, F. H., P. J. Pretorius and E. Erasmus, 2000 The utilization of alanine, glutamic acid, and serine as amino acid substrates for glycine N-acyltransferase. J Biochem Mol Toxicol 14**:** 102-109.

Vetting, M. W., S. d. C. LP, M. Yu, S. S. Hegde, S. Magnet *et al.*, 2005 Structure and functions of the GNAT superfamily of acetyltransferases. Arch Biochem Biophys 433**:** 212-226.

Wain, L. V., J. A. Armour and M. D. Tobin, 2009 Genomic copy number variation, human health, and disease. Lancet 374**:** 340-350.

Watts, P. J., M. C. Davies and C. D. Melia, 1990 Microencapsulation using emulsification/solvent evaporation: an overview of techniques and applications. Crit Rev Ther Drug Carrier Syst 7**:** 235-259.

Williams, R. T., 1978 Species variations in the pathways of drug metabolism. Environ Health Perspect 22**:** 133-138.

Wilson, K., and J. Walker, 2007 *Principles an Techniques of Biochemistry an Molecular Biology*. Cambridge University Press.

Wood, S. G., M. R. Al-Ani and A. Lawson, 1978 Hippuric acid excretion after benzylamine ingestion in man. Br J Ind Med 35**:** 230-231.

Xu, C., C. Y. Li and A. N. Kong, 2005 Induction of phase I, II and III drug metabolism/transport by xenobiotics. Arch Pharm Res 28**:** 249-268.